

THE 23RD EUROPEAN MODELING & SIMULATION SYMPOSIUM

SEPTEMBER 12-14 2011

ROME, ITALY



EDITED BY

AGOSTINO G. BRUZZONE

MIQUEL A. PIERA

FRANCESCO LONGO

PRISCILLA ELFREY

MICHAEL AFFENZELLER

OSMAN BALCI

PRINTED IN RENDE (CS), ITALY, SEPTEMBER 2011

ISBN 978-88-903724-4-5

© 2011 DIPTTEM UNIVERSITÀ DI GENOVA

RESPONSIBILITY FOR THE ACCURACY OF ALL STATEMENTS IN EACH PAPER RESTS SOLELY WITH THE AUTHOR(S). STATEMENTS ARE NOT NECESSARILY REPRESENTATIVE OF NOR ENDORSED BY THE DIPTTEM, UNIVERSITY OF GENOVA. PERMISSION IS GRANTED TO PHOTOCOPY PORTIONS OF THE PUBLICATION FOR PERSONAL USE AND FOR THE USE OF STUDENTS PROVIDING CREDIT IS GIVEN TO THE CONFERENCES AND PUBLICATION. PERMISSION DOES NOT EXTEND TO OTHER TYPES OF REPRODUCTION NOR TO COPYING FOR INCORPORATION INTO COMMERCIAL ADVERTISING NOR FOR ANY OTHER PROFIT - MAKING PURPOSE. OTHER PUBLICATIONS ARE ENCOURAGED TO INCLUDE 300 TO 500 WORD ABSTRACTS OR EXCERPTS FROM ANY PAPER CONTAINED IN THIS BOOK, PROVIDED CREDITS ARE GIVEN TO THE AUTHOR(S) AND THE WORKSHOP.

FOR PERMISSION TO PUBLISH A COMPLETE PAPER WRITE TO: DIPTTEM UNIVERSITY OF GENOVA, DIRECTOR, VIA OPERA PIA 15, 16145 GENOVA, ITALY. ADDITIONAL COPIES OF THE PROCEEDINGS OF THE *EMSS* ARE AVAILABLE FROM DIPTTEM UNIVERSITY OF GENOVA, DIRECTOR, VIA OPERA PIA 15, 16145 GENOVA, ITALY.

THE 23RD EUROPEAN MODELING & SIMULATION SYMPOSIUM
SEPTEMBER 12-14 2011
ROME, ITALY

ORGANIZED BY



DIPTTEM - UNIVERSITY OF GENOA



LIOPHANT SIMULATION



SIMULATION TEAM



IMCS - INTERNATIONAL MEDITERRANEAN & LATIN AMERICAN COUNCIL OF SIMULATION



MECHANICAL DEPARTMENT, UNIVERSITY OF CALABRIA



MSC-LES, MODELING & SIMULATION CENTER, LABORATORY OF ENTERPRISE SOLUTIONS



MODELING AND SIMULATION CENTER OF EXCELLENCE (MSCOE)



MISS LATVIAN CENTER - RIGA TECHNICAL UNIVERSITY



LOGISIM



LSIS - LABORATOIRE DES SCIENCES DE L'INFORMATION ET DES SYSTEMES



MISS - UNIVERSITY OF PERUGIA



MISS - BRASILIAN CENTER, LAMCE-COPPE-UFRJ



MISS - MCLEOD INSTITUTE OF SIMULATION SCIENCES



M&SNET - MCLEOD MODELING AND SIMULATION NETWORK



LATVIAN SIMULATION SOCIETY



ECOLE SUPERIEURE D'INGENIERIE EN SCIENCES APPLIQUEES



FACULTAD DE CIENCIAS EXACTAS. INGENIERIA Y AGRIMENSURA



Universidad de La Laguna

UNIVERSITY OF LA LAGUNA



CIFASIS: CONICET-UNR-UPCAM



INSTICC - INSTITUTE FOR SYSTEMS AND TECHNOLOGIES OF INFORMATION, CONTROL AND COMMUNICATION

I3M 2011 INDUSTRIAL SPONSORS



PRESAGIS



CAE



CAL-TEK



MAST



AEGIS TECHNOLOGIES

I3M 2011 MEDIA PARTNERS



MILITARY SIMULATION & TRAINING MAGAZINE

MILITARY SIMULATION & TRAINING MAGAZINE



EURO MERCI

EDITORS

AGOSTINO BRUZZONE

MISS-DIPTM, UNIVERSITY OF GENOA, ITALY

agostino@itim.unige.it

MIQUEL A. PIERA

AUTONOMOUS UNIVERSITY OF BARCELONA, SPAIN

MiquelAngel.Piera@uab.es

FRANCESCO LONGO

MSC-LES, UNIVERSITY OF CALABRIA, ITALY

f.longo@unical.it

PRISCILLA ELFREY

NASA-KSC, FL, USA

priscilla.r.elfrey@nasa.gov

MICHAEL AFFENZELLER

UPPER AUSTRIAN UNIVERSITY OF APPLIED SCIENCES, AUSTRIA

Michael.Affenzeller@fh-hagenberg.at

OSMAN BALCI

VIRGINIA TECH, USA

balci@vt.edu

**THE INTERNATIONAL MEDITERRANEAN AND LATIN AMERICAN MODELING
MULTICONFERENCE, I3M 2011**

GENERAL CO-CHAIRS

AGOSTINO BRUZZONE, *MISS DIPTM, UNIVERSITY OF GENOA, ITALY*
MIQUEL ANGEL PIERA, *AUTONOMOUS UNIVERSITY OF BARCELONA, SPAIN*

PROGRAM CHAIR

FRANCESCO LONGO, *MSC-LES, UNIVERSITY OF CALABRIA, ITALY*

THE 23RD EUROPEAN MODELING & SIMULATION SYMPOSIUM, EMSS 2011

GENERAL CO-CHAIRS

FRANCESCO LONGO, *MSC-LES, UNIVERSITY OF CALABRIA, ITALY*
PRISCILLA ELFREY, *NASA-KSC, USA*

PROGRAM CO-CHAIRS

OSMAN BALCI, *VIRGINIA TECH, USA*
MICHEAL AFFENZELLER, *UPPER AUSTRIAN UNIVERSITY OF APPLIED SCIENCES, AUSTRIA*

EMSS 2011 INTERNATIONAL PROGRAM COMMITTEE

MICHAEL AFFENZELLER, *UPPER AUSTRIAN UNIV. OF AS, AUSTRIA*
WERNER BACKFRIEDER, *UPPER AUSTRIAN UNIV. OF AS, AUSTRIA*
OSMAN BALCI, *VIRGINIA TECH, USA*
STJEPAN BOGDAN, *UNIVERSITY OF ZAGREB, CROATIA*
ENRICO BOCCA, *SIMULATION TEAM, ITALY*
FELIX BREITENECKER, *TECHNICAL UNIVERSITY OF WIEN, AUSTRIA*
AGOSTINO BRUZZONE, *UNIVERSITY OF GENOA, ITALY*
SORIN DAN COTOFANA, *DELFT UNIV. OF TECHNOLOGY, GERMANY*
GIANLUCA DE LEO, *VMASC-ODU, USA*
STEPHAN DREISEITL, *UPPER AUSTRIAN UNIV. OF AS, AUSTRIA*
PRISCILLA ELFREY, *NASA-KSC, USA*
WENHUI FAN, *TSINGHUA UNIVERSITY, CHINA*
MARIA PIA FANTI, *POLYTECHNIC UNIVERSITY OF BARI, ITALY*
IDALIA FLORES, *UNIVERSITY OF MEXICO, MEXICO*
CLAUDIA FRYDMAN, *LSIS, FRANCE*
MURAT M. GÜNAL, *TURKISH NAVAL ACADEMY, TURKE*
GRAHAM HORTON, *MAGDEBURG UNIVERSITY, GERMANY*
AMIR HUSSAIN, *UNIVERSITY OF STIRLING, SCOTLAND, UK*
WITOLD JACAK, *UPPER AUSTRIAN UNIV. OF AS, AUSTRIA*
EMILIO JIMÉNEZ, *UNIVERSITY OF LA RIOJA, SPAIN*
BERTHOLD KERSCHBAUMER, *UPPER AUSTRIAN UNIV. OF AS, AUSTRIA*
DANKO KEZIC, *UNIVERSITY OF SPLIT, CROATIA*
PETER KULCZYCKI, *UPPER AUSTRIAN UNIV. OF AS, AUSTRIA*
JUAN IGNACIO LATORRE BIEL, *UNIV. PÚBLICA DE NAVARRA, SPAIN*
FRANCESCO LONGO, *MSC-LES, UNIVERSITY OF CALABRIA, ITALY*
MARINA MASSEI, *LIOPHANT SIMULATION, ITALY*
YURI MERKURYEV, *RIGA TECHNICAL UNIVERSITY, LATVIA*
TANYA MORENO CORONADO, *UNIV. NAC. AUT. DE MEXICO, MEXICO*
MIGUEL MÚJICA MOTA, *UAB, SPAIN*
GASPER MUSIC, *UNIVERSITY OF LJUBLJANA, SLOVENIA*
GABY NEUMANN, *TECH. UNIV. APPL. SCIENCES WILDAU, GERMANY*
MUAZ NIAZI, *COMSATS INSTITUTE OF IT, PAKISTAN*
TUDOR NICULIU, *UNIVERSITY OF BUCHAREST, ROMANIA*
VERA NOVAK, *HARVARD MEDICAL SCHOOL, USA*
TUNCER ÖREN, *M&SNET, UNIVERSITY OF OTTAWA, CANADA*
FEDERICA PASCUCCI, *UNIVERSITY OF ROMA 3, ITALY*
SEGURA PEREZ, *UNIV. NAC. AUT. DE MEXICO, MEXICO*
MERCEDES PEREZ DE LA PARTE, *UNIVERSIDAD DE LA RIOJA, SPAIN*
MIQUEL ANGEL PIERA, *UAB, SPAIN*
CESAR DE PRADA, *UNIVERSIDAD DE VALLADOLID, SPAIN*
ROCCO RONGO, *UNIVERSITY OF CALABRIA, ITALY*
STEFANO SAETTA, *UNIVERSITY OF PERUGIA, ITALY*
ROBERTO SETOLA, *UNIVERSITY OF ROMA 3, ITALY*
ROGER SMITH,
WILLIAM SPATARO, *UNIVERSITY OF CALABRIA, ITALY*
JERZY W. ROZENBLIT, *UNIVERSITY OF ARIZONA, USA*
ALBERTO TREMORI, *SIMULATION TEAM, ITALY*
LEVENT YILMAZ, *AUBURN UNIVERSITY, USA*
STEPHAN WINKLER, *UPPER AUSTRIAN UNIV. OF AS, AUSTRIA*
LIN ZHANG, *BEIHANG UNIVERSITY, CHINA*
XUESONG ZHANG, *JILIN UNIVERSITY, CHINA*
GERALD ZWETTLER, *UPPER AUSTRIAN UNIV. OF AS, AUSTRIA*

TRACKS AND WORKSHOP CHAIRS

WORKSHOP ON MODELING & SIMULATION IN HEALTHCARE

CHAIRS: WITOLD JACAK, *UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA (AUSTRIA)*; WERNER BACKFRIEDER, *UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA (AUSTRIA)*; MURAT M. GÜNAL, *TURKISH NAVAL ACADEMY (TURKEY)*; JERZY W. ROZENBLIT, *UNIVERSITY OF ARIZONA*

AGENT DIRECTED SIMULATION

CHAIRS: TUNCER ÖREN, *UNIVERSITY OF OTTAWA (CANADA)*; LEVENT YILMAZ, *AUBURN UNIVERSITY (USA)*

CLOUD SIMULATION AND HIGH PERFORMANCE COMPUTING

CHAIRS: PROF. LIN ZHANG, *BEIHANG UNIVERSITY, (BEIJING, CHINA)* PROF. WENHUI FAN, *TSINGHUA UNIVERSITY (CHINA)*

DISCRETE AND COMBINED SIMULATION

CHAIR: GASPER MUSIC, *UNIVERSITY OF LJUBLJANA, (SLOVENIA)*

HUMAN-CENTRED AND HUMAN-FOCUSED MODELLING AND SIMULATION

CHAIRS: GABY NEUMANN, *TECHNICAL UNIVERSITY OF APPLIED SCIENCES WILDAU (GERMANY)*; AGOSTINO BRUZZONE, *MISS-DIPTÉM, UNIVERSITY OF GENOA, (ITALY)*

INDUSTRIAL PROCESSES MODELING & SIMULATION

CHAIR: CESAR DE PRADA, *UNIVERSIDAD DE VALLADOLID, (SPAIN)*

INDUSTRIAL ENGINEERING

CHAIR: ENRICO BOCCA, *SIMULATION TEAM, (ITALY)*

PETRI NETS BASED MODELLING & SIMULATION

CHAIRS: EMILIO JIMÉNEZ, *UNIVERSITY OF LA RIOJA (SPAIN)*; JUAN IGNACIO LATORRE, *PUBLIC UNIVERSITY OF NAVARRE (SPAIN)*

SIMULATION AND ARTIFICIAL INTELLIGENCE

CHAIR: TUDOR NICULIU, *UNIVERSITY "POLITEHNICA" OF BUCHAREST (ROMANIA)*

SIMULATION OPTIMIZATION APPROACHES IN INDUSTRY, SERVICES AND LOGISTICS PROCESSES

CHAIRS: IDALIA FLORES, *UNIVERSITY OF MEXICO*; MIGUEL MÚJICA MOTA, *UNIVERSITAT AUTONOMA DE BARCELONA (SPAIN)*.

GENERAL CO-CHAIRS' MESSAGE

WELCOME TO EMSS 2011

Building on the long success of 22 editions, the 23th European Modeling & Simulation Symposium (also known since 1996 as "Simulation in Industry") is an important forum to discuss theories, practices and experiences on M&S (Modeling & Simulation).

EMSS 2011 brings together people from Academia, Agencies and Industries from all over the world, despite the Symposium name referring just to Europe; in fact EMSS 2011 represents a unique opportunity within I3M2011 framework to share experiences and ideas and to generate a new archival source for innovative papers on M&S-related topics.

The Symposium is also meant to provide information, identify directions for further research and to be an ongoing framework for knowledge sharing; the structure of the conferences, strongly based on Tracks, allows to create synergies among different groups keeping pretty sharp each framework; the quality of the papers is the stronghold of this event and even this year it was possible to apply severe selection on the submissions and to guarantee top level papers, therefore the conference in Rome is one of the largest EMSS organized during last years.

In fact, another strong feature for the 2011 edition of EMSS is the site: the event is located in Rome, a city that for his history and cultural background was called Caput Mundi (Head of the world).

The EMSS 2011 Program, Presentations, People and Place make it a professionally worthwhile and a personally enjoyable experience: so welcome to EMSS 2011.



*Francesco Longo
MSC-LES
University of Calabria*



*Priscilla Elfrey
Kennedy Space Center
NASA*

ACKNOWLEDGEMENTS

The EMSS 2011 International Program Committee (IPC) has selected the papers for the Conference among many submissions; therefore, based on this effort, a very successful event is expected. The EMSS 2011 IPC would like to thank all the authors as well as the reviewers for their invaluable work.

A special thank goes to all the organizations, institutions and societies that have supported and technically sponsored the event.

LOCAL ORGANIZATION COMMITTEE

AGOSTINO G. BRUZZONE, *MISS-DIPTM, UNIVERSITY OF GENOA, ITALY*

ENRICO BOCCA, *SIMULATION TEAM, ITALY*

FRANCESCO LONGO, *MSC-LES, UNIVERSITY OF CALABRIA, ITALY*

FRANCESCA MADEO, *UNIVERSITY OF GENOA, ITALY*

MARINA MASSEI, *LIOPHANT SIMULATION, ITALY*

LETIZIA NICOLETTI, *CAL-TEK SRL*

FEDERICO TARONE, *SIMULATION TEAM, ITALY*

ALBERTO TREMORI, *SIMULATION TEAM, ITALY*




This International Conference is part of the I3M Multiconference: the Congress leading *Simulation around the World and Along the Years*

I3M Simulation around the World & along the Years
I3M2010, Fes, Morocco



www.liophant.org/i3m

I3M Simulation around the World & along the Years
I3M2011, Rome, Italy



www.liophant.org/i3m

I3M Simulation around the World & along the Years
I3M2012, Wien, Austria



www.liophant.org/i3m

I3M Simulation around the World & along the Years
I3M2007, Bergeggi, Italy



www.liophant.org/i3m

I3M Simulation around the World & along the Years
I3M2008, Calabria, Italy




www.liophant.org/i3m

I3M Simulation around the World & along the Years
I3M2009, Tenerife, Spain



www.liophant.org/i3m

I3M Simulation around the World & along the Years
I3M2004, Liguria, Italy



www.liophant.org/i3m

I3M Simulation around the World & along the Years
I3M2005, Marseille, France



www.liophant.org/i3m

I3M Simulation around the World & along the Years
I3M2006, Barcelona, Spain



www.liophant.org/i3m

I3M Simulation around the World & along the Years
HMS2002, Marseille, France



www.liophant.org/i3m

I3M Simulation around the World & along the Years
HMS2003, Riga, Latvia



www.liophant.org/i3m

I3M Simulation around the World & along the Years
HMS2004, Rio de Janeiro, Brazil



www.liophant.org/i3m

EMSS Simulation around the World & along the Years
ESS1993, Delft, The Netherlands




www.liophant.org/i3m

EMSS Simulation around the World & along the Years
ESS1994, Istanbul, Turkey



www.liophant.org/i3m

EMSS Simulation around the World & along the Years
ESS1996, Genoa, Italy



www.liophant.org/i3m

Index

Controllable Equivalent Resistance CMOS Active Resistor with Improved Accuracy and Increased Frequency Response <i>Cosmin Popa</i>	1
A Local Search Genetic Algorithm For The Job Shop Scheduling Problem <i>Mebarek Kebabla, Hayet Mouss, Nadia Mouss</i>	5
Temporal Neuro-Fuzzy Systems in Fault Diagnosis and Prognosis <i>Rafik Mahdaoui, Hayet Mouss, Djamel Mouss, Ouahiba Chouhal</i>	11
Optimization of automated guided vehicle rules for a multiple-load AGV system using simulation and SAW, VICOR and TOPSIS methods in a FMS environment <i>Parham Azimi, Masuomeh Gardeshi</i>	17
A New Method for the Validation and Optimisation of Unstable Discrete Event Models <i>Hans-Peter Barbey</i>	25
Simulation Helps Assess and Increase Airplane Manufacturing Capacity <i>Marcelo Zottolo, Edward Williams, Onur Ulgen</i>	30
Calibration of process algebra models of discretely observed stochastic biochemical system <i>Paola Lecca</i>	36
Denervated Muscle Undergoing Electrical Stimulation: Development Of Monitoring Techniques Based On Medical Image Modelling <i>Paolo Gargiulo, Thomas Mandl, Egill A. Friðgeirsson, Ilaria Bagnaro, Thordur Helgason, Páll Ingvarsson, Marcello Bracale, Winfried Mayr, Ugo Carraro, Helmut Kern</i>	44
3d Segmented Model Of Head For Modelling Electrical Activity Of Brain <i>Egill A. Friðgeirsson, Paolo Gargiulo, Ceon Ramon, Jens Hauelsen</i>	50
An Agent-Based Information Foraging Model of Scientific Knowledge Creation and Spillover in Open Science Communities <i>Ozgur Ozmen, Levent Yilmaz</i>	55
Simulation Highway - Direct Access Intelligent Cloud Simulator <i>Egils Ginters, Inita Sakne, Ieva Lauberte, Artis Aizstrauts, Girts Dreijja, Rosa Maria Aquilar China, Yuri Merkurjev, Leonid Novitsky, Janis Grundspenkis</i>	62
Simulation of Gesmey Generator Manoeuvres <i>Amable López, José A. Somolinos, Luis R. Núñez, Alfonso M. Carneros</i>	72
Using Semantic Web Technologies to Compose Live Virtual Constructive (LVC) Systems <i>Warren Bizub, Julia Brandt, Meggan Schoenberg</i>	78
Side Differences in Mri-Scans In Facial Palsy: 3-D Modelling, Segmentation And Voxel Gradient Changes <i>Paolo Gargiulo, Carsten Michael Klingner, Egill A. Friðgeirsson, Hartmut Peter Burmeister, Gerd Fabian Volk, Orlando Guntinas-Lichius</i>	87
Exploiting Variance Behavior in Simulation-based Optimization <i>Pasquale Legato, Rina Mary Mazza</i>	93

Accelerated fully 3D iterative reconstruction in SPECT <i>Werner Backfrieder, Gerald Zwettler</i>	100
Study on the servilization of simulation capability <i>Y L Luo, L Zhang, F Tao, Y Bao, L Ren</i>	105
Gyrus And Sulcus Modelling Utilizing a Generic Topography Analysis Strategy for Processing Arbitrarily Oriented 3d Surfaces <i>Gerald Zwettler, Werner Backfrieder</i>	111
Fast Marching Method Based Path Planning for Wheeled Mobile Robots <i>Gregor Klanar, Gasper Music</i>	118
Modeling and Simulation Architecture For Cloud Computing and Internet of Things (IoT) Based Distributed Cyber-Physical Systems (DCPS) <i>Xie Lulu, Wang Zhongjie</i>	127
Transport Network Optimization: Self-Organization by Genetic Programming <i>Johannes Göbel, Anthony E. Krzesinski, Bernd Page</i>	137
3D Physics Based Modeling and Simulation of Intrinsic Stress in SiGe for Nano PMOSFETs <i>Abderrazzak EL boukili</i>	144
Simulation model for the calcination process of cement <i>Idalia Flores, Guillermo Perea</i>	150
Job Satisfaction Modelling in Agent-Based Simulations <i>Alexander Tarvid</i>	158
Simultaneous Scheduling o Machines ad Operators in a Multi-Resource Coinstrained Job-Shop Scenario <i>Lorenzo Tiacchi, Stefano Saetta</i>	166
Effect of reject option on classifier performance <i>Stephan Dreiseitl, Melanie Osl</i>	176
New discrete Topology Optimization method for industrial tasks <i>Sierk Fiebig</i>	181
3G Mobile Network Planning Based On A Traffic Simulation Model And A Cost-Benefit Model To Service Los Cabos International Airport <i>Aida Huerta Barrientos, Mayra Elizondo Cortes</i>	187
Providing Semantic Interoperability for Integrated Healthcare Using a Model Transformation Approach <i>Barbara Franz, Herwig Mayr</i>	195
Management of supply networks using PDES <i>Carmine De Nicola, Rosanna Manzo, Luigi Rarità</i>	201
Sugar Factory Benchmark <i>Rogelio Mazaeda, Alexander Rodriguez, Alejandro Merino, Cesar De Prada, Luis Felipe Acebes,</i>	211
Designing and Implementing a Model to Examine R&D section ´s capabilities with emphasis on reversed engineering in chemical Factory <i>Neda Khadem Geraili, Mona Benhari</i>	220
On the Use of Minimum-Bias Computer Experimental Designs <i>Husam Hamad</i>	229
A Practical Guide for the Initialisation of Multi-Agent Systems with Random Number Sequences from Aggregated Correlation Data <i>Volker Nissen, Danilo Saft</i>	235

Multi-agent multi-level Modeling - A methodology to simulate complex systems <i>Jean-Baptiste Soyeux, Gildas Morvan, Rochdi Merzouki, Daniel Dupont</i>	241
Indoor Pedestrian Navigation Simulation Via A Network Framework <i>John Usher, Eric Kolstad</i>	247
Reconfigurable Human-System Cosimulation <i>Tudor-Razvan Niculiu, Maria Niculiu</i>	254
An Asynchronous Parallel Hybrid Optimization Approach To Simulation-Based Mixed-Integer Nonlinear Problems <i>Kathleen Fowler, Timothy Kopp, Jacob Orsini, Josh Griffin, Genetha Gray</i>	264
Reconstruction of clinical workflows based on the IHE integration profile "Cross-Enterprise Document Workflow" <i>Melanie Strasser, Franz Pfeifer, Emmanuel Helm, Andreas Schuler, Josef Altmann</i>	272
A Methodology For Developing Des Models: Event Graphs And Sharpsim <i>Arda Ceylan, Murat Gunal</i>	278
Revisitation o The Simulation Methodologies And Applications In Manufacturing <i>T. Radha Ramanan, Ihsan Sabuncuoglu</i>	283
Insights into the Practice of Expert Simulation Modellers <i>Rizwan Ahmed, Mahmood Shah</i>	289
Modelling Resilience in Cloud-Scale Data Centres <i>John Cartlidge, Ilango Sriram</i>	299
Advanced Container Transportation Equipment using Transfer Robot and Alignment System <i>Young Jin Lee, Dong Seop Han, Dae Woo Kang, Duk Kyun Lee, Geun Choi, Kwon Soon Lee,</i>	308
Global Context Influences Local Decision <i>Terry Bossomaier, Michael Harrè</i>	314
Modeling And Simulation Of Petri Nets For Complex Scheduling Rules Of Automated Manufacturing Systems <i>Chulhan Kim, Tae-Eog Lee</i>	319
A Business Process Modeling Approach To Support Production Systems Analysis And Simulation <i>Claudia Battista, Giulia Dello Stritto, Francesco Giordano, Raffaele Iannone, Massimiliano M. Schiraldi</i>	325
Online Collaborative Simulation Conceptual Model Development <i>Bhakti S. S. Onggo, Suchismita Hoare</i>	333
dSPACE Based direct-driven permanent magnet synchronous wind power system modeling and simulation <i>Yan-xia Shen, Xiang-xia Liu, Zhi-cheng Ji, Ting-long Pan</i>	340
Missing Data Estimation for Cancer Diagnosis Support <i>Witold Jacak, Karin Proell</i>	345
A New Devs-Based Generic Artificial Neural Network Modeling Approach <i>Samuel Toma, Laurent Capocchi, Dominique Federici</i>	351
An Improved Time-Line Search Algorithm To Optimize Industrial Systems <i>Miguel Mujica, Miquel Angel Piera</i>	357

Simulation and Model Calibration With Sensitivity Analysis For Threat Detection in Brain <i>Keegan Lowenstein, Brian Leventhal, Kylie Drouin, Robert Dowman, Katie Fowler, Sumona Mondal</i>	363
Simulation And Modelling Of The Flat-Band Voltage For Below 200nm SOI Devices <i>Cristian Ravariu, Florin Babarada</i>	371
Simulation of Human Behavior in Situation of Emergency <i>Samira Benkhedda, Fatima Bendella, Karima Belmabrouk</i>	375
Simulation, Optimisation and Design a Platform for in-vivo Electrophysiological Signals Processing <i>Florin Babarada, Cristian Ravariu, Janel Arhip</i>	380
A Simulation-Based Framework for Industrial Automated Wet-Etch Station Scheduling Problem In The Semiconductor Industry <i>Adrian Aguirre, Vanina Cafaro, Carlos Mendez, Pedro Castro</i>	384
Data Stream Management in Income Tax Microsimulation Models <i>Istvan Molnar, Gyorgy Lipovszki</i>	394
Experimental manufacturing system for research and training on human-centred simulation <i>Diego Crespo Pereira, David del Rio Vilas, Rosa Rios Prado, Nadia Rego Monteil, Adolfo Lamas Rodriguez</i>	400
Modelling and Simulation of a Wireless Body Area Network Prototype for Health Monitoring <i>Yeray Callero, Rosa María Aguilar</i>	410
An Extreme Learning Machine Algorithm for Higher Order Neural Network Models <i>Shuxiang Xu</i>	418
Increasing Availability of Production Flow Lines Through Optimal Buffer Sizing: a Simulative Study <i>Vittorio Cesarotti, Alessio Giuiusa, Vito Introna</i>	423
Using Query Extension And User Feedback To Improve Pubmed Search <i>Viktoria Dorfer, Sophie A. Blank, Stephan M. Winkler, Thomas Kern, Gerald Petz, Patrizia Faschang</i>	433
Simulation of The Vessel Traffic Schedule In The Strait Of Istanbul <i>Şirin Özlem, Ihan Or, Birnur</i>	439
New Genetic Programming Hypothesis Search Strategies for Improving the Interpretability in Medical Data Mining Applications <i>Michael Affenzeller, Christian Fischer, Gabriel Kronberger, Stephan M. Winkler, Stefan Wagner</i>	448
On the Use of Estimated Tumor Marker Classifications in Tumor Diagnosis Prediction - A Case Study for Breast Cancer <i>Stephan Winkler, Michael Affenzeller, Gabriel Kronberger, Michael Kommenda, Stefan Wagner, Witold Jacak, Herbert Stekel</i>	454
Automatic Selection of Relevant Data During Ultrasonic Inspection <i>Thouraya Merazi Meksen, Malika Boudraa, Bachir Boudraa</i>	460
CT Based Models for Monitoring Bone Changes in Paraplegic Patients Undergoing Functional Electrical Stimulation <i>Páll Jens Reynisson, Benedikt Helgason, Stephen J. Ferguson, Thordur</i>	465

<i>Helgason, Rúnar Unnpórsson, Páll Ingvarsson, Helmut Kern, Winfried Mayr, Ugo Carraro, Paolo Gargiulo</i>	
A High Resolution Distributed Agent-Based Simulation Environment for Large-Scale Emergency Response <i>Glenn Hawe, Graham Coates, Duncan Wilson, Roger Crouch</i>	470
Eigenfrequency Based Sensitivity Analysis Of Vehicle Drivetrain Oscillations <i>Oliver Manuel Krieg, Jens Neumann, Bernhard Hoess, Heinz Ulbrich</i>	478
Improving Job Scheduling on a Heterogeneous Cluster by Predicting Job Execution Times Using Heuristics <i>Hannes Brandstätter-Müller, Bahram Parsapour, Andreas Hölzlwimmer, Gerald Lirk, Peter Kulczycki</i>	488
Agent-Based Simulation Of Electronic Marketplaces With Ontology-Services <i>Maria João Viamonte, Nuno Silva, Paulo Maio</i>	496
Key Issues In Cloud Simulation Platform Based On Cloud Computing <i>Lei Ren, Lin Zhang, Yabin Zhang, Yongliang Luo, Qian Li</i>	502
Research On Key Technologies Of Resource Management In Cloud Simulation Platform <i>Ting Yu Lin, Xu Dong Chai, Bo Hu Li</i>	508
A Framework For Enhanced Project Schedule Design To Aid Project Manager's Decision Making Processes <i>Sanja Lazarova-Molnar, Rabeb Mizouni, Nader Kesserwan</i>	516
A Novel Approach To Realistic Modeling And Simulation Of State-Varying Failure Rates <i>Sanja Lazarova-Molnar</i>	526
Modeling and Simulation of Order Sortation Systems <i>Fahrettin Eldemir, Elif Karakaya</i>	535
Development Of The Surface To Air Missile Simulator Through The Process Of Component Composition And Dynamic Reconfiguration Of Weapon System <i>Jeebeom Suk, Jaeoh Lee, Yoonho Seo</i>	541
Modeling and simulating a benchmark on dynamic reliability as a Stochastic Activity Network <i>Daniele Codetta-Raiteri</i>	545
A Stochastic Approach To Risk Modeling For Solvency Ii <i>Vojo Bubevski</i>	555
Design and Implementation of a Fuzzy Cognitive Maps Expert System for Oil Price Estimation <i>Mohamad Ali Azadeh, Zeinab sadat Ghaemmohamadi</i>	562
How to benefit more from intuitive power and experience of the human simulation knowledge stakeholder <i>Gaby Neumann</i>	568
Object-oriented modelling and verification aided by model simplification techniques <i>Anton Sodja, Borut Zupancic</i>	574
The exclusive entities in the formalization of a decision problem based on a discrete event system by means of Petri nets <i>Juan Ignacio Latorre-Biel, Emilio Jimenez-Macias</i>	580
Matrix-based operations and equivalent classes in alternative Petri nets <i>Juan Ignacio Latorre, Emilio Jimenez</i>	587

Synthesis of Feedback Controller for Stabilization of Chaotic Henon Map Oscillations by Means of Analytic Programming <i>Roman Senkerik, Zuzana Oplatkova, Ivan Zelinka, Donald Davendra, Roman Jasek</i>	593
A Simulation Study of The Interdependence Of Scalability And Cannibalization In The Software Industry <i>Francesco Novelli</i>	599
Security in sending and storage of Petri nets by signing and encryption <i>Iñigo León Samaniego, Mercedes Pérez de la Parte, Eduardo Martínez Camara, Juan Carlos Sáenz-Díez Muro</i>	605
Petri net transformation for decision making: compound Petri nets to alternatives aggregation Petri nets <i>Juan Ignacio Latorre, Emilio Jimenez</i>	613
Improvements in the optimization of Flexible Manufacturing Cells modelled with Discrete Event Dynamics Systems: Application to a real factory problem <i>Diego Rodriguez, Mercedes Perez, Juan Manuel Blanco</i>	619
A Simulation-Based Capacity Planning Model: a Case Study In a Contract Furnishing Sme <i>Nadia Rego Monteil, David del Rio Vilas, Diego Crespo Pereira, Rosa Rios Prado, Arturo Nieto de Almeida</i>	626
PN as a tool for innovation in industry: a review <i>Jesús Fernandez de Miguel, Julio Blanco Fernandez, Mercedes Perez</i>	635
Stochastic Optimization of Industrial Maintenance Strategies <i>Francisco Castellanos, Ann Wellens</i>	642
Design And Development of Data Analysis Modules For The Aermod And Calpuff Simulation Models <i>Ann Wellens, Gamar García</i>	648
Development Of A Simulation Tool For Consequence Analysis In Industrial Instalations <i>Victor Pérez, Gamar García, Maria Guadalupe Ávila, Francisco Castellanos, Ann Wellens</i>	654
Using Markov Chain and Graph Theory Concepts to Analyze Behavior in Complex Distributed Systems <i>Christopher Dabrowski, Fern Hunt</i>	659
Formal Framework For The Devs-Driven Modeling Language <i>Ufuoma Ighoroje, Oumar Maïga, Mamadou Traoré</i>	669
A Methodology for the DEVS Simulation Graph Construction <i>Adedoyin Adegoke, Ibrahima Hamadou, Hamidu Togo, Mamadou Traoré,</i>	675
Efficient exploration of Coloured Petri net based scheduling problem solutions <i>Gasper Music</i>	681
Plant Capacity Analysis in a Dairy Company, Applying Montecarlo Simulation <i>Joselito Medina-Marin, Gilberto Perez-Lechuga, Juan Carlos Seck-Tuoh-Mora, Norberto Hernandez-Romero, Isaias Simon-Marmolejo</i>	690
GPGPU Programming and Cellular Automata: Implementation of The Sciara Lava Flow Simulation Code <i>Giuseppe Filippone, William Spataro, Giuseppe Spingola, Donato D'Ambrosio, Rocco Rongo, Giovanni Perna, Salvatore Di Gregorio</i>	696

Neighborhood Concept for Modeling an Adaptive Routing in Wireless Sensors Network	703
<i>Jan Nikodem, Maciej Nikodem, Ryszard Klempous, Marek Woda, Zenon Chaczko</i>	
A DEVS-based Simplified Business Process Modelling Library	709
<i>Igor Rust, Deniz Cetinkaya, Mamadou Seck, Ivo Wenzler</i>	
Research on co-simulation task scheduling in Cloud Simulation Platform	715
<i>Chen Yang, Bo Hu Li, Xudong Chai</i>	
An Optimal Non-Blocking Dispatching in Free-Choice Manufacturing Flowlines by Using Machine-Job Incidence Matrix	722
<i>Ivica Sindicic, Stjepan Bogdan, Tamara Petrovic</i>	
Simulation of Vascular Volume Pulsation of Radial Index Artery	728
<i>Pichitra Uangpairoj, Masahiro Shibata</i>	
Coding Tcqn Models Into The Simio Simulation Environment	734
<i>Miguel Mujica, Miquel Angel Piera</i>	
Developing a Simulation Training Tool from a Medical Protocol	740
<i>Catherine M. Banks, John A. Sokolowski</i>	
Model Synthesis Using a Multi-Agent Learning Strategy	747
<i>Sebastian Bohlmann, Arne Klauke, Volkhard Klinger, Helena Szczerbicka</i>	
Service Optimization For System-Of-Systems Based On Pool Scheduling And Inventory Management Driven By Smart Simulation Solutions	755
<i>Agostino Bruzzone, Marina Massei, Enrico Bocca</i>	
Modeling Of Obesity Epidemics By Intelligent Agents	768
<i>Agostino Bruzzone, Vera Novak, Francesca Madeo, Cecilia Cereda</i>	
Maritime Security: Emerging Technologies for Asymmetric Threats	775
<i>Agostino Bruzzone, Marina Massei, Alberto Tremori, Francesco Madeo, Federico Tarone, Francesco Longo</i>	
On The Short Period Production Planning in Industrial Plants: A Real Case Study	782
<i>Agostino Bruzzone, Francesco Longo, Letizia Nicoletti, Rafael Diaz</i>	
Authors' Index	792

CONTROLLABLE EQUIVALENT RESISTANCE CMOS ACTIVE RESISTOR WITH IMPROVED ACCURACY AND INCREASED FREQUENCY RESPONSE

Cosmin Popa

University Politehnica of Bucharest, Faculty of Electronics, Telecommunications and Information Technology, Romania

cosmin_popa@yahoo.com

ABSTRACT

A new active resistor circuit will be further presented. The main advantages of the original proposed implementation are the improved linearity, the small area consumption and the improved frequency response. An original technique for linearizing the $I(V)$ characteristic of the active resistor will be proposed, based on the simulation of the Ohm law using two linearized differential amplifiers, a multiplier and a current-pass circuit. The controllability of the active resistor circuit is excellent, existing the possibility of modifying the value of the equivalent resistance by changing the ratio between a control voltage and a control current. Additionally, the value of the simulated resistance is not function on technological parameters, with the result of improved circuit accuracy. The errors introduced by the second-order effects will be also strongly reduced, while the area consumption of the active resistor will be minimized by replacing the classical MOS transistor with FGMOS (Floating Gate MOS) devices.

Keywords: active resistor, differential amplifiers, linearity error, second-order effects

I. INTRODUCTION

CMOS active resistors are very important blocks in VLSI analog designs, mainly used for replacing the large value passive resistors, with the great advantage of a much smaller area occupied on silicon. Their utilization domains includes amplitude control in low distortion oscillators, voltage controlled amplifiers and active RC filters. These important applications for programmable floating resistors have motivated a significant research effort for linearising their current-voltage characteristic.

The first generation of MOS active resistors [1], [2] used MOS transistors working in the linear region. The main disadvantage is that the realised active resistor is inherently nonlinear and the distortion components were complex functions on MOS technological parameters.

A better design of CMOS active resistors is based on MOS transistors working in saturation [3], [4], [5]. Because of the quadratic characteristic of the MOS transistor, some linearisation techniques were developed in order to minimize the nonlinear terms from the current-voltage characteristic of the active resistor. Usually, the resulting linearisation of the $I-V$ characteristic is obtained by a first-order analysis. However, the second-order effects which affect the MOS transistor operation

(mobility degradation, bulk effect and short-channel effect) limits the circuit linearity introducing odd and even-order distortions, as shown in [4]. For this reason, an improved linearisation technique has to be design to compensate the nonlinearities introduced by the second-order effects.

II. THEORETICAL ANALYSIS

The original idea for implementing a linear current-voltage characteristic of the active resistor, similar to the characteristic of a classical passive resistor is to simulate the Ohm law using two linearized differential amplifiers and a multiplier circuit. Because of the requirements for a good frequency response, only MOS transistors working in saturation could be used.

2.1. The block diagram of the active resistor

The structure of the proposed active resistor is based on four important blocks: two differential amplifiers ADI and $AD2$ with linear transfer function, a multiplier circuit $MULT$ and a current-pass circuit I , the block diagram being presented in Figure 1.

The I_{XY} current, which is passing through the I block, is generated by the multiplier circuit, $I_{XY} = I_O I_2 / I_1$, while I_1 and I_2 currents are obtained from the differential amplifiers ADI and $AD2$, $I_1 = g_{m1} V_O$ and $I_2 = g_{m2} (V_X - V_Y)$. It results:

$$I_{XY} = I_O \frac{g_{m2} (V_X - V_Y)}{g_{m1} V_O} \quad (1)$$

The equivalent resistance of the circuit having the block diagram presented in Figure 1 will be:

$$R_{ech.} = \frac{V_X - V_Y}{I_{XY}} = \frac{V_O}{I_O} \frac{g_{m1}}{g_{m2}} = \frac{V_O}{I_O} \sqrt{\frac{(W/L)_1 I_{O1}}{(W/L)_2 I_{O2}}} \quad (2)$$

The great advantage of the proposed implementation of the active resistor is the very good controllability of the equivalent resistance by the ratio of a control voltage V_O and a control current I_O . As it is shown in (2), the value of the resistance does not depend on technological parameters, $(W/L)_1$ and $(W/L)_2$ representing aspect ratios of the transistors composing the differential

amplifiers $AD1$ and $AD2$, respectively, while I_{O1} and I_{O2} being biasing currents of these amplifiers.

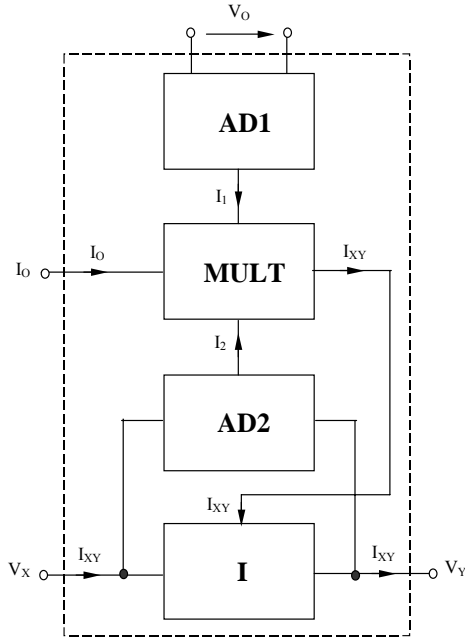


Figure 1: The block diagram of the active resistor

2.2. Classical CMOS differential amplifier

The most common approach of a differential amplifier in CMOS technology is based on strong-inverted MOS transistors (usually working in the saturation region), having the most important advantage of a much better frequency response with respect to the weak-inverted MOS differential amplifiers. As a result of the quadratic characteristic of a MOS transistor operating in saturation, the transfer characteristic of the classical CMOS differential amplifier will be strongly nonlinear, its linearity being in reasonable limits only for a very limited range of the differential input voltage. The drain currents of the classical CMOS differential amplifier will have the following nonlinear dependence on the differential input voltage, v_{id} :

$$I_{d1,2} = \frac{I_O}{2} \pm \frac{I_O}{2} \left(\frac{Kv_{id}^2}{I_O} - \frac{K^2v_{id}^4}{4I_O^2} \right)^{1/2}, \quad (3)$$

having a Taylor expansion around $v_{id} = 0$, fifth-order limited expressed by:

$$I_{d1,2}(v_{id}) \cong \frac{I_O}{2} \pm \frac{K^{1/2}I_O^{1/2}}{2} v_{id} \mp \frac{K^{3/2}}{16I_O^{1/2}} v_{id}^3 \mp \dots, \quad (4)$$

where I_O is the biasing current for the differential amplifier. In order to improve the circuit linearity, especially for large values of the differential input voltage (THD has relatively large values for v_{id} of about hundreds of mV), a linearization technique has to be implemented.

2.3. The original linearized differential amplifier

The original proposed differential structure from Figure 2 is based on a symmetrical structure that assures, in a first-order analysis, the linearization of the transfer characteristic, equivalent to a constant circuit transconductance.

Supposing a saturation operation of all MOS active devices from the previous circuit, it is possible to write that the output current expression is $I_2 = I_X - I_Y$:

$$I_2 = \frac{K}{2} (V_{GS_X} - V_{GS_Y}) (V_{GS_X} + V_{GS_Y} - 2V_T). \quad (5)$$

Because:

$$V_{GS_X} - V_{GS_Y} = 2(V_X - V_Y) \quad (6)$$

and:

$$V_{GS_X} + V_{GS_Y} = 2V_{GS_O}, \quad (7)$$

it results a linear dependence of the output current on the differential input voltage:

$$I_2 = \sqrt{8KI_O} (V_X - V_Y), \quad (8)$$

equivalent to a constant transconductance of the circuit:

$$g_m = \sqrt{8KI_O}. \quad (9)$$

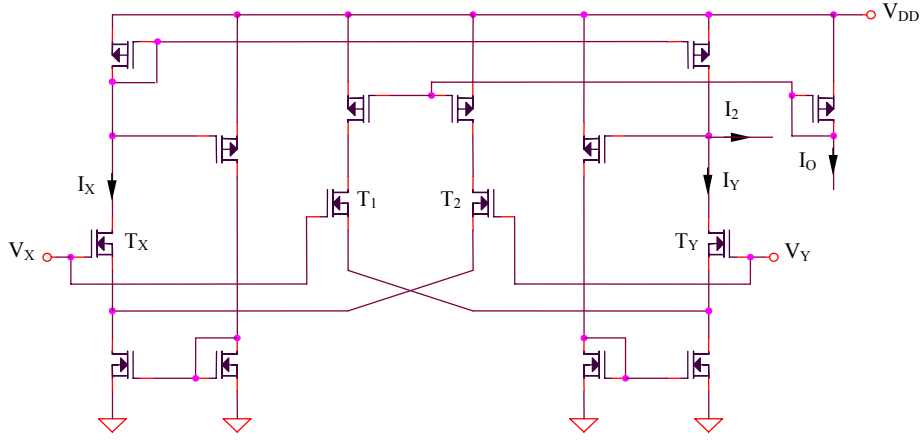


Figure 2: The linearized differential structure

The important advantages of the previous circuit is the improved linearity that could be achieved in a first-order analysis and the possibility of controlling the value of the transconductance by modifying a continuous current (I_O).

2.4. The second-order effects

The linearity (8) of the transfer characteristic of the differential amplifier from Figure 2 is slightly affected by the second-order effects that affect the MOS transistor operation, modeled by the following relations: channel-length modulation (10) and mobility degradation (11).

$$I_D = \frac{K}{2} (V_{GS} - V_T)^2 (1 + \lambda V_{DS}) \quad (10)$$

$$K = \frac{K_0}{[1 + \theta_G (V_{GS} - V_T)][1 + \theta_D V_{DS}]} \quad (11)$$

Considering that the design condition $\lambda = \theta_D$ is fulfilled, the gate-source voltage of a MOS transistor working in saturation at a drain current I_D will be:

$$V_{GS} = V_T + \sqrt{\frac{2I_D}{K}} + \theta_G \frac{I_D}{K} \quad (12)$$

The last term represents the error which affects the quadratic characteristic of the MOS transistor biased in saturation, caused by the previous presented second-order effects. The result will be a small accuracy degradation of the entire circuit linearity, quantitative evaluated by the superior-order terms in the transfer characteristic of the differential amplifier:

$$I_{XY} = \sum_{k=1}^{\infty} a_k (V_X - V_Y)^k \quad (13)$$

Because of the circuit symmetry, the odd-order terms from the previous relation are usually cancel out, so the main circuit nonlinearity caused by the second-order effects will be represented by the third-order error term from the previous relation, having much smaller value than the linear term.

2.5. The current-pass circuit

The necessity of designing this circuit is derived from the requirement that the same current to pass between the two output pins, X and Y. The implementation in CMOS technology of this circuit is very simple, consisting in a simple and a multiple current mirrors (Figure 3).

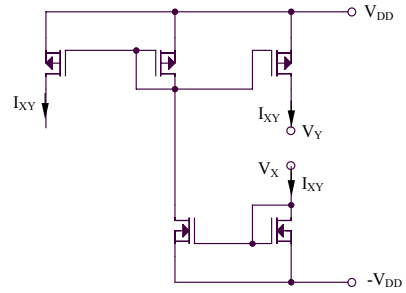


Figure 3: The current-pass circuit

2.6. The multiplier circuit

The original idea for obtaining the multiplying function is to use two identical square-root circuits, implementing the following functions:

$$I_{OUT1} = 2\sqrt{I_O I_2} \quad (14)$$

and:

$$I_{OUT2} = 2\sqrt{I_{XY} I_1} \quad (15)$$

I_{OUT1} and I_{OUT2} being the output currents of these square-root circuits. Using a classical current mirror, it is possible to impose that $I_{OUT1} = I_{OUT2}$, resulting the necessary multiplying function:

$$I_{XY} = I_O \frac{I_2}{I_1} \quad (16)$$

The important advantages of this implementation are represented by the increased frequency response that could be obtained as a result of the current-mode operation of the multiplier circuit and of the biasing in saturation of all the MOS active devices and, additionally, by the reduced

silicon occupied area achieved by using exclusively MOS transistors.

2.7. Active resistor with negative equivalent resistance

Active resistors with controllable negative equivalent resistance cover a specific area of VLSI designs, finding very large domains of applications such as the canceling of an operational amplifier load or the design of integrators with improved performances. In order to obtain a negative equivalent active resistance circuit, the block diagram from Figure 1 must be modified by inverting the sense of the I_{XY} current passing through the I block, resulting an equivalent resistance expressed by:

$$R_{ech.}' = -\frac{V_0}{I_0} \sqrt{\frac{(W/L)_1 I_{O1}}{(W/L)_2 I_{O2}}}. \quad (17)$$

III. CONCLUSIONS

A new active resistor circuit has been presented. The main advantages of the original proposed implementation are the improved linearity, the small area consumption and the improved frequency response. An original technique for linearizing the $I(V)$ characteristic of the active resistor has been proposed, based on the simulation of the Ohm law using two linearized differential amplifiers, a multiplier and a current-pass circuit. The controllability of the active resistor circuit is excellent, existing the possibility of modifying the value of the equivalent resistance by changing the ratio between a control voltage and a control current. Additionally, the value of the simulated resistance is not function on technological parameters, with the result of an improved circuit accuracy. The errors introduced by the second-order effects have been also strongly reduced, while the area consumption of the active resistor has been minimized by replacing the classical MOS transistor with FG MOS devices. As a result of the proposed linearization technique designed for the differential amplifier from Figure 2, the linearity (9) of its transfer characteristic is referring both to small and large signal operation, being limited only by the second-order effects that affect the MOS transistors' operation. The consequence will be a relatively large range of the input voltage that could be applied across the input pins (V_X and V_Y from Figure 1), respecting the important restriction of maintaining the estimated circuit linearity.

ACKNOWLEDGMENTS

The work has been co-funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement, POSDRU/89/1.5/S/62557.

REFERENCES

1. Z. Wang, "Current-controlled Linear MOS Earthed and Floating Resistors and Application", IEEE Proceedings on Circuits, Devices and Systems, 1990, pp. 479-481.
2. L. Sellami, "Linear Bilateral CMOS Resistor for Neural-type Circuits", Proceedings of the 40th Midwest Symposium on Circuits and Systems, 1997, pp. 1330-1333.
3. E. Ozalevli, P. Hasler, „Design of a CMOS Floating-Gate Resistor for Highly Linear Amplifier and Multiplier Applications”, Custom Integrated Circuits Conference, 2005. Proceedings of the IEEE 2005, 18-21 Sept. 2005, pp. 735-738.
4. K. Kaewdang, W. Surakampontom, N. Fujii, „A Design of CMOS Tunable Current Amplifiers”, Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium on, 26-29 Oct. 2004, pp. 519-522.
5. F. Bahmani, E. Sanchez-Sinencio, „A Highly Linear Pseudo-Differential Transconductance, Solid-State Circuits Conference, 2004. ESSCIRC 2004. Proceeding of the 30th European, 21-23 Sept. 2004, pp. 111-114.

AUTHORS BIOGRAPHY

Cosmin Popa is with University Politehnica of Bucharest, Faculty of Electronics, Telecommunications and Information Technology, Romania. His area of interest includes analog integrated and mixed-signal VLSI designs. He is author of more than 140 scientific papers and of 5 research books.

A LOCAL SEARCH GENETIC ALGORITHM FOR THE JOB SHOP SCHEDULING PROBLEM

Kebabla Mebarek, Mouss Leila Hayat and Mouss Nadia

Laboratoire d'automatique et productique, Université Hadj Lakhdar -Batna
kebabla@yahoo.fr, hayet_mouss@yahoo.fr, kinzmouss@sahoo.fr

Abstract--Scheduling of job-shop is very important in the fields of production management and combinatorial optimization. This paper proposes a method for solving general job-shop scheduling problems based on hybridized algorithm that combines a genetic algorithm with a taboo search in two distinct phases research. In the first phase an operations-coded genetic algorithm is used to find an elite population. The set of elite solutions obtained from the first phase acts as the initial population of the second phase, in which a taboo search algorithm is applied to each one of them to intensify the research. The effectiveness of this algorithm is confirmed by applying it to a set of benchmarks with the makespan as the objective function. The results obtained show that local search applied at the final population can improve greatly the research.

Index Terms-- Job-Shop Scheduling, Hybrid Meta-Heuristic, Genetic Algorithm, Local Search, Taboo Search.

INTRODUCTION

The job-shop scheduling problem (JSSP) plays an important role in the scheduling theory and finds many practical applications. It deals with the sequencing of a set of jobs on a set of machines, in order to minimize an objective function [9]. For years, job-shop scheduling has attracted the attention of many researchers in the fields of both production management and combinatorial optimization. Efficient methods for solving the JSSP have significant effects on performance of production system. It has been demonstrated that this problem is usually an NP-complete (nondeterministic polynomial time complete) problem [5]. An indication for this is that one 10×10 problem formulated by Muth & Thompson in 1963 [10] remained unsolved for twenty years [12]. For this hardness, exact methods become quickly inapplicable in practice. Instead, it is often preferred to use approximation algorithms such as heuristics and meta-heuristics e.g. simulated annealing, genetic algorithms, and taboo search.

In recent years, much attention has been devoted to four general heuristics: simulated annealing (SA), taboo search (TS), genetic algorithm (GA), and neural network (NN) [13]. These methods are capable of providing high-quality solutions with reasonable computational effort. However, the problem is hard that cannot be solved efficiently by applying any single technique and a great deal of research have focused on hybrid methods [14]. Several authors pointed out that the performance of genetic algorithm on some combinatorial optimization problems was a bit inferior to that of neighborhood search algorithms (e.g., local search, simulated annealing and taboo search). Hybrid methods of genetic algorithms and those neighborhood search algorithms were proposed, and their high performance was reported [6].

In this paper, we propose a genetic local search algorithm to improve the search process, the proposed algorithm acts in two phases. In the first one, a genetic algorithm is developed to find a set of best solutions. In the second, a local search algorithm with a memory has to intensify the research around each solution to improve it. The performance of the algorithm will be assessed through an experimental analysis with a set of benchmark problems. The remainder of this paper is organized as follows: in the next section we describe the problem of general job-shop scheduling to be solved. In section 3 we describe the proposed algorithm and its process. A numerical experiment is presented in the section 4.

PROBLEM DESCRIPTION

The problem studied in this paper is a deterministic and static n -job, m -machine job-shop scheduling problem (JSSP). The aim behind is to optimally allocate different operations for each job across a set of machines respecting temporal and resource constraints. This problem is formulated as follows:

There are n jobs J_1, \dots, J_n to be scheduled on m machines M_1, \dots, M_m . Each job j consists of a sequence of n_j operations

O_{ij} ($i = 1, \dots, n_j$) that must be processed on the m machines in a given order $O_{1j} - O_{2j} - \dots - O_{n_j j}$ (in our case $n_j = m$). Each operation is characterized by specifying both the required machine $M \in \{M_1, \dots, M_m\}$, and the fixed processing time $p_{ij} > 0$ [4].

Furthermore, several constraints considered on jobs and machines, are listed as follows:

- (i) Each job must pass through each machine once and only once.
- (ii) Each job should be processed through the machines in a particular order.
- (iii) Each operation executed must be uninterrupted on a given machine (no pre-emption is allowed).
- (iv) Each machine can only process one operation at a time.

The objective is to determine a feasible schedule with minimal *makespan* (i.e. minimizing the completion time of the last job) which is the most common goal for these problems:

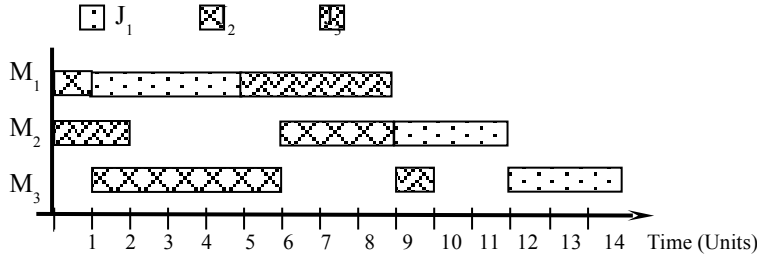


Fig. 1. A Gantt-Chart representation of a solution for the instance in TABLE I.

I. THE PROPOSED GENETIC LOCAL SEARCH ALGORITHM

For solving this problem, we introduced a fast and easily implemented hybrid algorithm. The proposed algorithm combines a genetic algorithm with a local search one. The former has to find a set of best solutions, and the local search procedure is applied to each solution generated by genetic operations to "dig" around for improving it. The

$$C_{max} = \max_{j=1, \dots, n} \{C_{j}\},$$

where C_j is the completion time of job J_j .

Table 1 presents an example of a job-shop problem formed of 3 jobs (J_1, J_2, J_3) which processed on 3 machines (M_1, M_2, M_3). For this problem, a solution schedule is presented by Gantt chart in Fig.1.

TABLE I

An example of a job-shop scheduling problem

J_1 :	$M_1:4$	$M_2:3$	$M_3:3$
J_2 :	$M_1:1$	$M_3:5$	$M_2:3$
J_3 :	$M_2:2$	$M_1:4$	$M_3:1$

global procedure of this algorithm is described briefly as follows:

A. First phase

In the first phase, we apply a genetic algorithm which begin with an initial population and attempt to improve it through successive generations. The process of this algorithm is presented in Algorithm 1.

Algorithm 1: The genetic algorithm

Input A scheduling problem instance P ;

Output A set of best schedules for instance P ;

1. Generate the initial population;
2. Evaluate the population;

while No termination criterion is satisfied **do**

3. Select chromosomes from the current population;
4. Apply the recombination operator to the chromosomes selected at step 3 to generate new ones;
5. Evaluate the chromosomes generated at step 4;
6. Apply the mutation operator to a randomly selected chromosomes;
7. Apply the selection criteria to replace new chromosomes;

return A set of best schedules evaluated so far;

1 Chromosome Representation: To represent the chromosome, we base on an *Operation-Based* representation that uses an unpartitioned permutation with m -repetitions of job numbers [3]. For an n -job m -machine problem, a chromosome contains $n \times m$ genes. Each job appears in the chromosome exactly m times, and each repeating gene does not indicate a concrete operation of a job but refers to an operation which is context-dependent.

For example, considering the problem above (TABLE I), one of the chromosomes may be [231213123], which should be interpreted to a schedule as shown in Fig. 2. Each job number is repeated three times because each job has three operations. The first job number represents the first operation of the job, and the second represents the second operation. The order of genes in the chromosome represents the order in which the operations of jobs are scheduled.

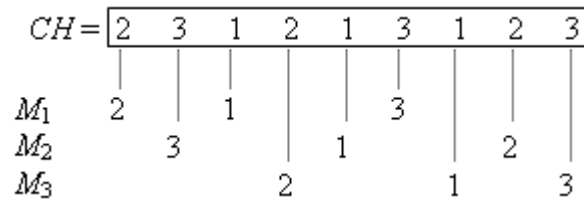


Fig. 2. A schedule building from a chromosome of the problem showed in TABLE I.

2 Initialization: In this algorithm the initial population consists of randomly generated chromosomes.

3 Gentic operators: The genetic evolution is done using three main genetic operators: crossover, mutation, and selection.

Crossover: In this study we adopt the *Generalized Order-Crossover* (GOX) scheme that generates only feasible solutions [3]. In GOX, one chromosome (donor) contributes a substring of length in the range of one third to half of his length. This substring is inserted in the receiver chromosome in the same position of the substring first gene after deleting all genes from the receiver with the same position as the genes in the substring according to their order in the jobs.

Mutation: The mutation operator has to bring a change to the chromosome. In our algorithm, we use a special mutation operator, that we called *Job Based Mutation* (JBM). In this mutation two jobs are randomly chosen. After that, all genes of the considered chromosome corresponding to one job are changed to the other. For example, chromosome [2 3 1 2 1 3 1 2 3] become [1 3 2 1 2 3 2 1 3] if considering jobs 1 and 2 to be swapped.

Selection: The selection mechanism for reproduction in this paper is based on the fitness ranking of the chromosomes. Two chromosomes are chosen with a probability proportional to their fitness for crossover among best individuals (some rate of worst solutions is excluded from being reproduced). Then deleting the worst member of the population.

4 Fitness Function: Solutions in both phases are evaluated according to their fitness. In this study, we use the makespan value of schedules as the fitness function.

B. Second phase

In this phase a local search procedure is applied to the best solutions among final offspring solutions generated by genetic operations. Essentially, local search consists in moving from a solution to another one in its neighborhood. So, we implant a simple taboo search method in order to avoid recycling. For this, two elements are necessary to define: the neighborhood structure and the memory (taboo list). The basic role of the taboo list is to prevent the search process from turning back to solutions visited in previous steps. The taboo list stores the arcs that have recently been reversed rather than the whole solutions. The length of the taboo list is usually of critical importance. Thus, we have used a dynamic taboo list that varying between two values [min, max] as it is proposed in [11]. Its length is decreased by one unit when the current solution is better than the previous one; otherwise it is increased in same amount.

For the neighborhood structure, in our taboo search, we adopt the technique used by Nowocki and Smutnicki [11]. In this neighborhood, a critical path composed of b blocks is generated. A critical block of operations is defined as a set with the maximum successive operations that belong to the critical path and that are processed on a same machine. If $1 < l < b$ (l : block order), then swap the first two operations on the last block and the last two operations on the first block. However, if $l = 1$ swap only the last two operations in this block, on the other hand if $l = b$ swap the first two operations [11].

Algorithm 2 shows the taboo search algorithm we have considered here. This algorithm is a simple one, in the first step, the initial solution is taken from the set given by the first phase. Then, it iterates over a number of steps. In each iteration, the neighborhood of the current solution is built and one of the neighbors is selected for the next iteration.

The tabu search finishes after a fixed number of iterations

that is not so much high. Then it passes to the following solution of the elite set to begin with it again.

Algorithm 2: The local search algorithm

Input A set of best schedules C and a problem instance P ;

Output A schedule for instance P ;

for all set of best solutions i **do**

1. Evaluate schedule $C(i)$;
2. Generate the NS neighborhood of the current solution;

If there is a better solution **then**

3. replace the current solution with and return to 1;
4. update the tabu list;

else

5. replace the current solution with best solution among the examined;
6. update the tabu list;

If iterations threshold is not reached **then** go to 2;

return the best schedule evaluated so far;

II. NUMERICAL EXPERIMENTS

The presented above genetic local search algorithm was programmed in Pascal Object (Delphi environment) and was run on a PC computer with a processor Intel P Dual 2.2

GHz for all the experiments. Fig. 3 shows a screenshot of this program solving the famous FT10 proposed in Muth & Thompson [10].

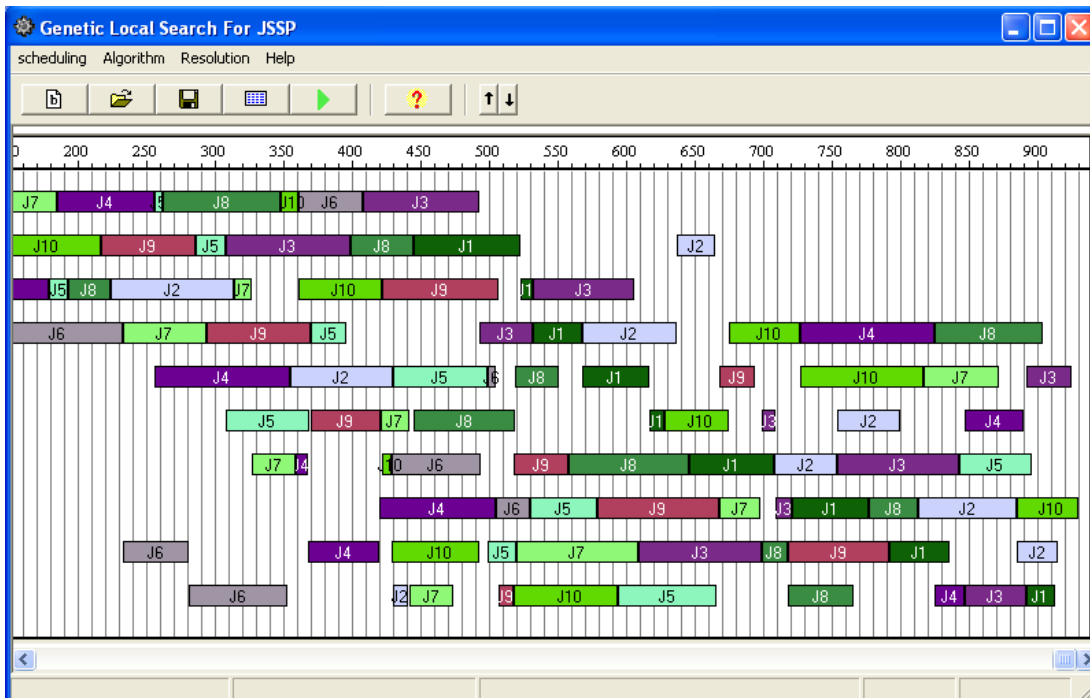


Fig.3. A screenshot of the program showing optimal solution of FT10.

The performance of the algorithm is analyzed on a set of benchmarks on the job-shop scheduling problem instances from literature. The size of the benchmark instances varies from 10 to 20 jobs and from 5 to 20 machines. We consider (FT10, FT20) proposed by Fisher and Thompson [10];

three problems (ABZ5-7) generated by Adams, Balas & Zawack [1]; ten 10×10 problems (ORB01-10) generated by Applegate and Cook [2] and 15 problems of different sizes (LA01-15) generated by Lawrence [8].

Table II shows the makespan performance statistics of the proposed GLS algorithm for the selected benchmark problems comparatively with a simple genetic algorithm that constitute the first phase of the whole GLS algorithm. It lists problem name, problem size (number of jobs, number

of operations), the best-known solution, the best solution obtained by simple genetic algorithm in five attempts (i.e after first phase processing only), and the best solution obtained by our GLS algorithm in five attempts.

TABLE II
Computational results obtained by the proposed algorithm on benchmark problems

Instance	Size (n,m)	Best known solution	Best generated solution with GA in 5 attempts	Best generated solution with GLS in 5 attempts
FT10	(10,10)	930	938	930
FT20	(20,5)	1165	1173	1165
ABZ5	(10,10)	1234	1238	1234
ABZ6	(10,10)	943	944	943
ABZ7	(15,20)	656	680	667
ORB01	(10,10)	1059	1068	1059
ORB02	(10,10)	888	897	889
ORB03	(10,10)	1005	1011	1008
ORB04	(10,10)	1005	1032	1012
ORB05	(10,10)	887	890	887
ORB06	(10,10)	1010	1010	1010
ORB07	(10,10)	397	397	397
ORB08	(10,10)	899	899	899
ORB09	(10,10)	934	934	934
ORB10	(10,10)	944	944	944
LA01	(10,5)	666	666	666
LA02	(10,5)	655	665	665
LA03	(10,5)	597	597	597
LA04	(10,5)	590	590	590
LA05	(10,5)	593	593	593
LA06	(15,5)	926	926	926
LA07	(15,5)	890	890	890
LA08	(15,5)	863	863	863
LA09	(15,5)	951	951	951
LA10	(15,5)	958	958	958
LA11	(20,5)	1222	1222	1222
LA12	(20,5)	1039	1039	1039
LA13	(20,5)	1150	1150	1150
LA14	(20,5)	1292	1292	1292
LA15	(20,5)	1207	1207	1207

From the computational results of the table 1, it could be concluded that the proposed algorithm produced good solutions on all instances tested. In most of cases, it returns optimal solutions. In the few other cases the solutions are very near of the optimal. It could be concluded also, that the GLS algorithm returns best solutions or at least equal solutions to those returned by the simple genetic algorithm that constitute the first phase of the proposed algorithm. This means that the second phase improves really the research process.

CONCLUSION

Due to the stubborn nature of job-shop scheduling, much effort shown in the literature has focused on hybrid methods, since most single techniques cannot solve this problem efficiently [7]. In this paper, we have considered a general job-shop problem, and we have proposed a genetic local search algorithm. The proposed algorithm acts in two steps. Firstly, a genetic algorithm with operation-based encoding, GOX crossover, JBM mutation and elitist strategy selection is applied to get a set of best solutions. In

a second step, a tabu search procedure is applied to each solution of the elite set generated by genetic operations hopefully to improve some of them.

An empirical study is carried out to test the proposed strategies on a set of standard JSP instances taken from the literature. The results show that the proposed algorithm is an efficient mean in solving the problem considered. This algorithm gives better results than the two algorithms separately. It's concluded that combination of local search and genetic algorithms is a promising approach for improving resolution of job-shop scheduling problems.

REFERENCES

- [1] Adams, J., E. Balas, and D. Zawack. "The Shifting Bottleneck Procedure for Job Shop Scheduling", *Management Science*, vol.34 (1988), pp. 391–401.
- [2] Applegate, D. & Cook, W., "A Computational Study of Job-shop Scheduling". *ORSA J. Computing*. vol. 2 (1991), pp. 149-156.
- [3] Bierwirth C., "A generalized permutation approach to job shop scheduling with genetic algorithms", *OR Spektrum*, vol. 17 (1995), pp. 87-92.
- [4] French, S., *Sequencing and Scheduling: An introduction to the Mathematics of the Job-Shop*, John Wiley & Sons, Inc., New York (1982).
- [5] Garey M.R., Johnson D.S., Sethi R., "The Complexity of Flow-Shop and Job-Shop Scheduling", *Mathematics of Operations Research*, vol. 1 (1976), pp. 117-129.
- [6] Ishibuchi H. & Tadahiko M., "Multi-Objective Genetic Local Search Algorithm", *Proceedings of IEEE International Conference on Evolutionary Computation*, 1996, pp. 119-124.
- [7] Jain A.S. & Meeran S., "Deterministic job shop scheduling: Past, present, Future", *European Journal of Operation Research*, vol. 113, pp. 390-434, 1999.
- [8] Lawrence S., "Resource constrained project scheduling: an experimental investigation of heuristic scheduling techniques", *Technical Report, Graduate School of Industrial Administration*, Pittsburgh, Carnegie Mellon University, 1984.
- [9] Mati Y. & Xie X., "The complexity of two-job shop problems with multi-purpose unrelated machines", *European Journal of Operational Research*, vol. 152 (2004), pp. 159–169.
- [10] Muth J. F. & Thompson G., *Industrial scheduling*, Prentice Hall, Englewood Cliffs, NJ, 1963.
- [11] Nowicki E. and Smutnicki C., "A fast taboo search algorithm for the job shop scheduling problem" *Management Science*, vol. 42, pp. 797-813, 1996.
- [12] Ombuki B. M. & Ventresca M., "Local Search Genetic Algorithms for the Job Shop Scheduling Problem", *Applied Intelligence*, vol. 21, pp. 99-109, 2004.
- [13] Ponnambalam S. G., Aravindan P. and Rajesh S. V., "A Tabu Search Algorithm for Job Shop Scheduling", *Int. J. Adv. Manuf. Technol.*, vol. 16, pp. 765-771, 2000.
- [14] Vazquez M. & Whitley D., "A Comparison of Genetic Algorithms for the Static Job Shop Scheduling Problem", *Proceedings of the 6th International Conference on Parallel Problem Solving from Nature*, pp.303-312, Sept.2000.
- [15] Waiman C. and Hong Z., "Using Genetic Algorithms and Heuristics for Job Shop, Scheduling with Sequence-Dependent Setup Times" *Annals of Operations Research*, vol. 107, pp. 65-81, 2001.

TEMPORAL NEURO-FUZZY SYSTEMS IN FAULT DIAGNOSIS AND PROGNOSIS

Mahdaoui Rafik , Mouss Hayet Leyla, Mouss Djamel , Chouhal Ouahiba

Laboratoire d'Automatique et Productique (LAP) Université de Batna,
1, Rue Chahid Boukhrouf 05000 Batna, Algérie
Centre universitaire Khenchela Algérie,
Route de Batna BP:1252, El Houria, 40004 Khenchela Algérie

mehdaoui.rafik@yahoo.fr, chouhal_wahiba@yahoo.fr, hayet_mouss@yahoo.fr, moussdj@yahoo.fr

ABSTRACT

Fault diagnosis and failure prognosis are essential techniques in improving the safety of many manufacturing systems. Therefore, on-line fault detection and isolation is one of the most important tasks in safety-critical and intelligent control systems.

Computational intelligence techniques are being investigated as extension of the traditional fault diagnosis methods. This paper discusses the properties of the TSK/Mamdani approaches and neuro-fuzzy (NF) fault diagnosis within an application study of an manufacturing systems. The key issues of finding a suitable structure for detecting and isolating ten realistic actuator faults are described.

Within this framework, data-processing interactive software of simulation baptized NEFDIAG (NEuro Fuzzy DIAGnosis) version 1.0 is developed. This software devoted primarily to creation, training and test of a classification Neuro-Fuzzy system of industrial process failures. NEFDIAG can be represented like a special type of fuzzy perceptron, with three layers used to classify patterns and failures. The system selected is the workshop of SCIMAT clinker , cement factory of Ain Touta " Batna, Algeria ".

Keywords: Diagnosis; artificial neuronal networks; fuzzy logic; Neuro-fuzzy systems; pattern recognition; AMDEC.

1. INTRODUCTION

The function of diagnosis is a very complex task and can be only one part solved by the technique of pattern recognition(PR) , the diagnosis by PR can be presented as an alternative solution at the model approach since the operating modes are modeled, not analytical manner, but by using only one whole of measurements of this modes [8]. Therefore the human expert in his mission of diagnosing the cause of a failure of a whole system uses quantitative or qualitative information. On another side, in spite of the results largely surprising obtained by the ANN in monitoring and precisely in diagnosis they remain even enough far from equalizing the sensory capacities and of reasoning

human being. Fuzzy logic makes another very effective axis in industrial diagnosis.

Also, can we replace the human expert for automating the task of diagnosis to 100%, by using the neuro-fuzzy approach? And How made the human expert to gather all information allowing him to learn its decision? Our objective consists to make an association (adaptation) techniques of fuzzy logic with the neuronal techniques (a system Neuro-fuzzy), to choose the types of networks of neuron, to determine the fuzzy rules and finally the structure of the system Neuro-Fuzzy to automate the maximum of the task diagnosis.

In order to achieve this goal we organize this article thus. The first part presents principal architectures and principles Neuro-Fuzzy systems operation and their applications. The second part is dedicated to the workshop of clinker on the level of cement factory. Lastly, in the third part we propose a Neuro-Fuzzy system for system of production diagnosis.

Neuro-Fuzzy systems

The Neuro-fuzzy model combines, in a single framework, both numerical and symbolic knowledge about the process. Automatic linguistic rule extraction is a useful aspect of NF especially when little or no prior knowledge about the process is available (Brown and Harris, 1994; Jang, 1995). For example, a NF model of a non-linear dynamical system can be identified from the empirical data.

This model can give us some insight about the nonlinearity and dynamical properties of the system.

The most common NF systems are based on two types of fuzzy models TSK (Takagi and Sugeno, 1985; Sugeno and Kang, 1988) and Mamdani (1995, 1996) combined with NN learning algorithms. TSK models use local linear models in the consequents, which are easier to interpret and can be used for control and fault diagnosis (Füssel, et al 1997; Ballé et al 1997). Mamdani models use fuzzy sets as consequents and therefore give a more qualitative description. Many Neuro-fuzzy structures have been successfully applied to a wide range of applications from industrial processes to financial systems, because of the ease of rule base design, linguistic modeling, application to complex and

uncertain systems, inherent non-linear nature, learning abilities, parallel processing and fault-tolerance abilities. However, successful implementation depends heavily on prior knowledge of the system and the empirical data (Ayoubi, 1995).

Neuro-fuzzy networks by intrinsic nature can handle limited number of inputs. When the system to be identified is complex and has large number of inputs, the fuzzy rule base becomes large.

NF models usually identified from empirical data are not very transparent. Transparency accounts a more meaningful description of the process i.e less rules with appropriate membership functions. In ANFIS (Jang, 1993, 1995) a fixed structure with grid partition is used. Antecedent and consequent parameters are identified by a combination of least squares estimate and gradient based method, called hybrid learning rule. This method is fast and easy to implement for low dimension input spaces. It is more prone to lose the transparency and the local model accuracy because of the use of error back propagation that is a global and not locally nonlinear optimization procedure. One possible method to overcome this problem can be to find the antecedents & rules separately e.g. clustering and constrain the antecedents, and then apply optimization.

Hierarchical NF networks can be used to overcome the dimensionality problem by decomposing the system into a series of MISO and/or SISO systems called hierarchical systems (Tachibana and Furuhashi, 1994). The local rules use subsets of input spaces and are activated by higher level rules[12].

The criteria on which to build a NF model are based on the requirements for faults diagnosis and the system characteristics. The function of the NF model in the FDI scheme is also important i.e. Preprocessing data, Identification (Residual generation) or classification (Decision Making/Fault Isolation).

For example a NF model with high approximation capability and disturbance rejection is needed for identification so that the residuals are more accurate.

Whereas in the classification stage, a NF network with more transparency is required.

The following characteristics of NF models are important:

- Approximation/Generalisation capabilities
- transparency: Reasoning/use of prior knowledge /rules
- Training Speed/ Processing speed
- Complexity
- Transformability: To be able to convert in other forms of NF models in order to provide different levels of transparency and approximation power.
- Adaptive learning

Two most important characteristics are the generalising and reasoning capabilities. Depending on the application requirement, usually a compromise is made between the above two.

In order to implement this type of fuzzy perception and exploited to diagnose of dedicated

production system we have to propose data-processing software NEFDIAG.

2. NEFDIAG PRESENTATION

NEFDIAG is a data processing program of interactive simulation, carried out within LAP (university of Batna) written under DELPHI, dedicated primarily to creation, the training and the test of a Neuro-Fuzzy system of classification of the breakdowns of a dedicated industrial process. NEFDIAG models a fuzzy classifier Fr with a whole of classes $C = \{c1, c2, \dots, cm\}$.

Structure and training NEFDIAG can be represented like a special type of fuzzy perception, with three layers to use to classify failures.

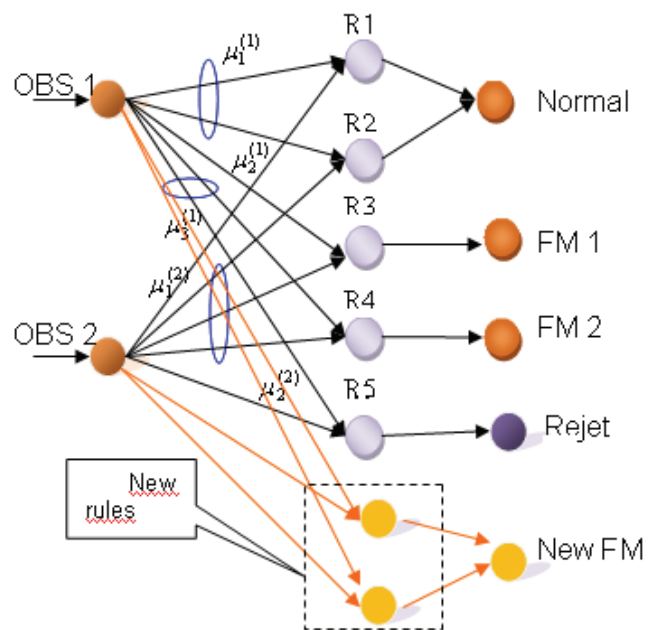


Fig. 1. Neuro-Fuzzy Architecture

NEFDIAG makes it's training by a set of forms, such as each form will be affected (classified) towards one of the preset classes. then NEFDIAG generates the fuzzy rules by a course of the data optimizes the rules by training the parameters of the subsets fuzzy which are used for partitioned the data " characteristic " of the forms with classified and the parameters of the data. NEFDIAG can be used to classify a new observation; the system can be represented in the form of fuzzy rules:

If symptom1 is A1 Symptom2 is A2
 Symptom3 is A3 Symptom N is An
 Then the form (x1, x2, x3..., xn) is belonged to class
 « Failure mode 1».

Such as A1 A2 A3 An are linguistic terms represented by fuzzy sets. This characteristic will make it possible to know the analyses on our data, and to use this knowledge to classify them. The training of the

networks of artificial Neuro-Fuzzy is a phase which makes it possible to determine or modify the parameters of the network, in order to adopt a desired behavior. The stage of training is based on the descent of gradient of average quadratic error made by network RNF.

System NEFDIAG can start with a base of knowledge partial of the forms, and can then refine it during the training, or it can start with an empty base of knowledge. The definite user the initial number of the functions of membership for partitioning fields of the data input. And it is also necessary to specify the number K, a number maximum of the neurons of the rules which will be created in the hidden layer. The principal steps of algorithm of training are thus presented.

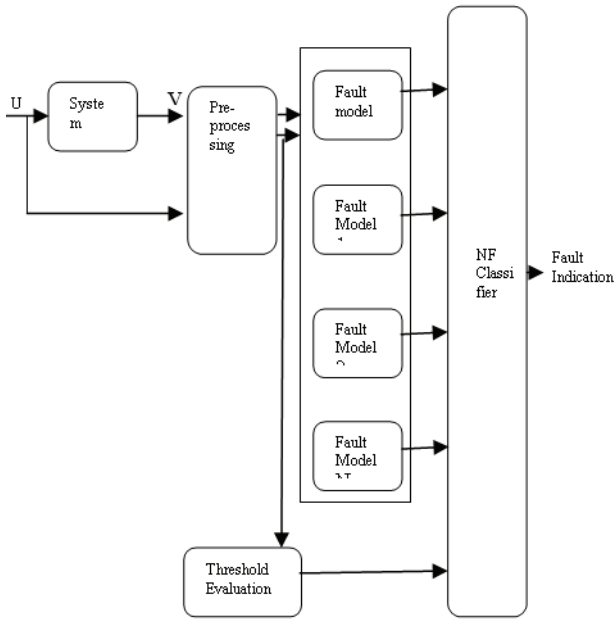


Fig. 2. Neuro-Fuzzy Fault detection.

Initialization: for each data resulting by sensors there is an input unit, for each mode of failure there is an output unit. For each input unit an initial fuzzy partition is specified « exp: A number of the of triangular membership functions.

2.1. Training of the rules

System NEFDIAG can start with a base of knowledge partial of the forms, and can then refine it during the training «Fig. 3», the rule will be created by research (for a given form F) the combination of the functions of membership such as each produced entry the greatest function of membership «fig.3 ». If this combination is not identical for the rules exist in the rules base, and numbers of rules is not maximum, then rules will be created and added to the rules base « Fig. 1»,

$$\varepsilon_r = \tau_r (1 - \tau_r) \frac{1}{m} \sum_{j=1}^m (2v_r^{(j)}(t_j) - 1) |E_j| \quad (1)$$

The algorithm of training detects (calculates) all the antecedents of the rules and then creates the list of the antecedents. In the first time this list is empty, or contains antecedents of rules of knowledge a priori. The algorithm selects then consequent to seek antecedent A and to create the basic list of rule candidates. The best rules will be selected base of the rules candidates, in base of measurement of performance [7].

In this case some classes (mode of failure) would not be represented in the base of rules, if the rules for this mode of failure to a value of very small performance.

Training of the Functions of Membership

For the training of the membership functions, a simple retro propagation is used. It depends on the error of output for each unit of rules. Each rule changes its membership functions by the change as of the their supports « fig. 7 ».

It is necessary that the error of each rule is calculated [5].

τ_r is fulfillment of a rule r.

After the appearance of another new mode of failure in the phase of training our system is the degree Neuro-fuzzy will make an adaptation or a reorganization of the system to be adapted has the new situation «fig. 2».

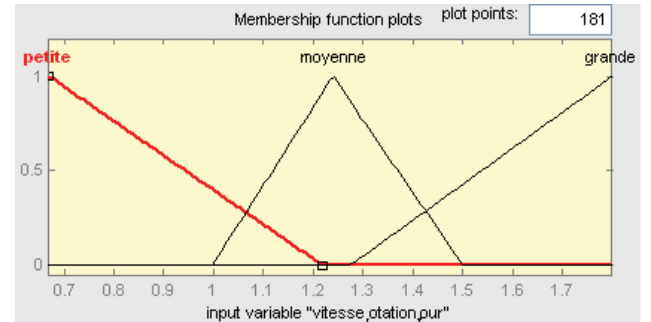


Fig. 3. V_R_F before training

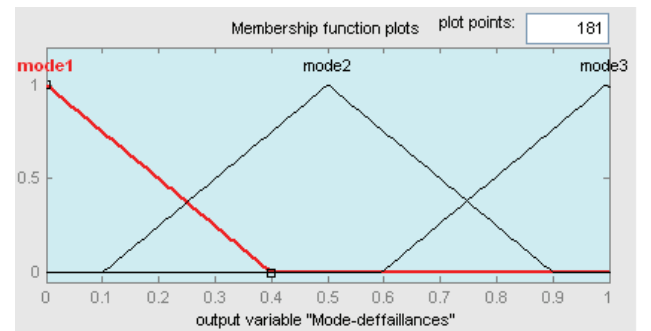


Fig. 4. V_R_F after training

Initially layers of rules (or rules bases) add all rules of the mode of failure detected. Then in the layer of the modes of failure, another node will be connected to the Neuro-Fuzzy network «fig.7 »

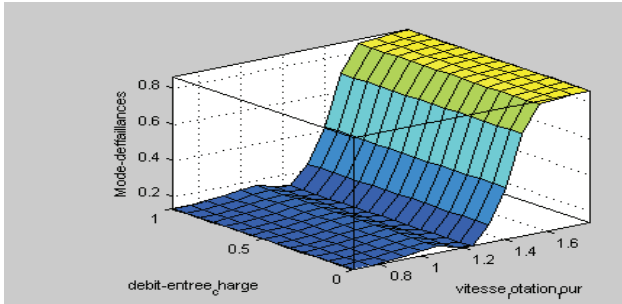


Fig. 5. The training of membership function

3. THE WORKSHOP OF CLINKER

Our application is illustrated on an industrial process of manufacture of cement. This installation belongs to cement factory of Ain-Touta (SCIMAT) ALGERIA. This cement factory have a capacity of 2.500.000 t/an " Two furnaces " is made up of several units which determine the various phases of the manufacturing process of cement. The workshop of cooking gathers two furnaces whose flow clinker is of 1560 t/h. The cement crushing includes two crushers of 100t/h each one. Forwarding of cement is carried out starting from two stations, for the trucks and another for the coaches «Fig.6, ".

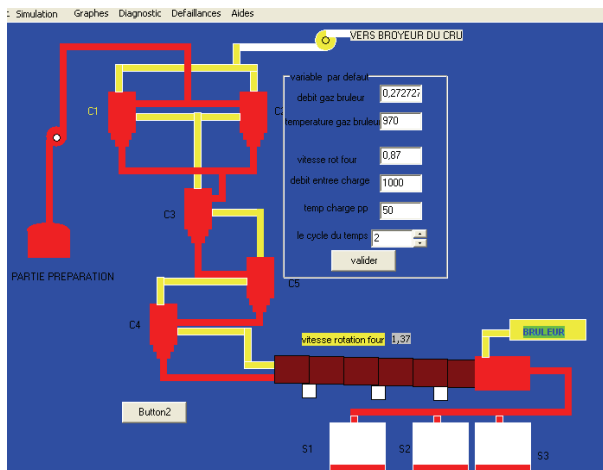


Fig. 6. Alarme message

4. NEURO FUZZY DIAGNOSIS

4.1 Dysfunctions analyse

This step has an objective of identification the dysfunctions which can influence the mission of the system. This analysis is largely facilitated by the recognition of the models structural and functional of

the installation. For the analysis of the dysfunctions we adopted the method of analysis of the modes of failure, their effects and their criticality (AMDEC). While basing itself on the study carried out by [6], on the workshop of cooking, we worked out an AMDEC by considering only the most critical modes of the failures (criticality >10) and this for reasons of simplicity [6]. Therefore we have a system Neuro-fuzzy of 27 inputs and four outputs which were created to make a diagnosis of our system. The rules which are created with the system are knowledge a priori, a priori the base of rule. Each variable having an initial partition will be modified with the length of the phase of training (a number of sets fuzzy for each variable). The reasoning for the diagnosis is described in the form of fuzzy rules inside our Neuro-fuzzy system.

The principal advantage of the use of the base of fuzzy rules lies in its modularity and its facility of extension (suppression or addition of other rules). The initial rules base to establish the diagnosis of the failures is built by exploiting the model worked out in phase's dysfunction of our system (AMDEC). Indeed, this analysis makes it possible to establish the bonds of causes for purposes between the components failing and the symptoms observed. These bonds will be represented in the forms of fuzzy rules building the knowledge base which will be training later and then tested, to carry out the fuzzy reasoning necessary and to lead to the results expressing the function of diagnosis.

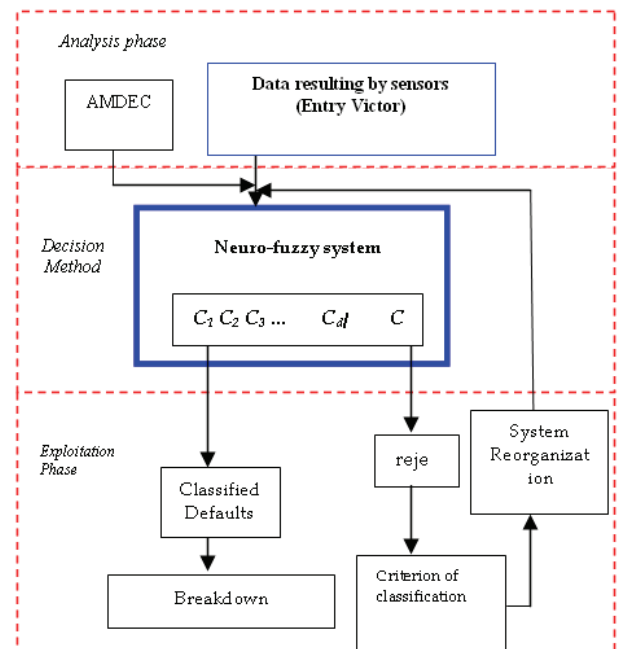


Fig. 7. The diagnosis by NEFDIAG.

Then the detection of the anomalies is represented in the form of alarm message intended to announce to the operator (user) the appearance of an anomaly (or anomalies) and makes it possible to identify the component responsible using a base for data which

stock all the information provided by the AMDEC (mode of failure, possible causes, equipment, effects on the system).

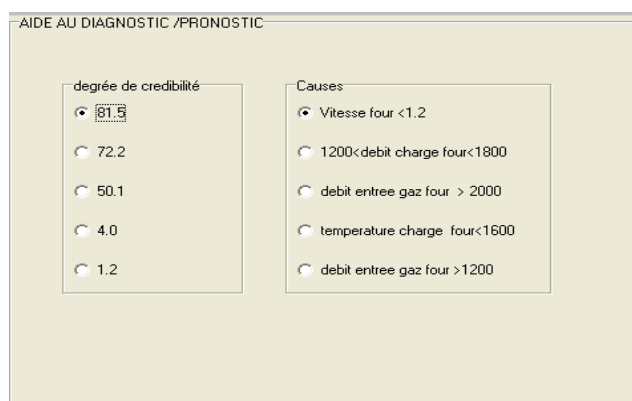


Fig. 8. diagnosis mode1

After appearance of an anomaly, a message of alarm makes it possible to the operator to detect the dysfunction and also to locate the responsible component. The " fig. 2, " illustrates the system of cooking with the presence of a dysfunction. Let us note that in this study, the anomalies or dysfunctions indicate functional anomalies.

After the posting of the message, the operator can consult this last for more information or to remove it " fig. 6, ". Then NEFDIAG makes interventions to control the variables which are the origin of the current failure "Fig. 8, ".

5. Conclusions

In this article we presented a new tool for diagnosis by Neuro-Fuzzy systems following approaches AMDEC, we detailed the implementation of an example of industrial application by the development tool NEFDIAG. We illustrated use of our tool of assistance to the prediction diagnosis in the form of a prototype NEFDIAG to install on a PC. We approached the various stages to be followed for the development of assistance system to the diagnosis starting from the methods of classification and fuzzy recognitions of the forms. NEFDIAG is represented like a special type of fuzzy perceptron, with three layers used to classify failures, by using the neuro-fuzzy system of the type 3. NEFDIAG makes training with two phases. A training of rules, and generates the fuzzy rules by the course of data and optimizes the rules by training of the parameters of the fuzzy sets which are used for partitioning the data of the forms to classify and the parameters of the data.

In spite of great importance of fuzzy neural networks for solving wide range of real-world problems, unfortunately, little progress has been made in their development.

we have discussed recurrent neural networks with fuzzy weights and biases as adjustable parameters and internal

feedback loops, which allows capturing dynamic response of a system without using external feedback through delays. In this case all the nodes are able to process linguistic information.

As the main problem regarding fuzzy and recurrent fuzzy neural networks that limits their application range is the difficulty of proper adjustment of fuzzy weights and biases, we put an emphasize on the RFNN training algorithm.

REFERENCES

- [1] H.K. Kwan and Y.Cai," A fuzzy neural network and its application to pattern recognition" IEEE Transactions on Fuzzy Systems, 3 pp. 185-193. 1994
- [2] J.-S. Roger Jang "ANFIS: Adaptive-network-based fuzzy inference system " ,IEEE Trans. Syst., Man, and Cybernetics, 23(1993) 665-685.
- [3] J.M. Keller and H.Tahani "Implementation of conjunctive and disjunctive fuzzy logic rules with neural networks" International Journal of Approximate Reasoning,6(1992) 221-240.
- [4] G. Zweingelstein " Diagnostic des défaillances, théories et pratique pour les systèmes industriels " Col. Traité des nouvelles technologies, séries diagnostic et maintenance, Hermès ,1995.
- [5] L.-X wang and J.M Mendel "Generation fuzzy rules by learning from examples" IEEE trans. Syst., Man, and Cybernetics, (1992) 22(6):1414-1427.
- [6] D. Mouss "Diagnostic et conduite des systèmes de production par approche a base de connaissances " Thèse de doctorat Université de Batna, 2006.
- [7] D. Nauck "Neuro-fuzzy systems: review and prospects"European congress on intelligent technique and sift computing (EUFIT'97), Aachen, sep.8.11, (1997),
- [8] D. Racoceanu "contribution à la surveillance des systèmes de production en utilisant les techniques de l'intelligence artificielle "
- [9] M. Orchard and G. Vachtsevanos, "A particle filtering-based frame-work for real-time falut diagnosis and faillure prognosis in a tur-bine engine," Mediterranean Conference on Control and AutomationMED'07, Athens, Greece, July 2007.
- [10] A. Al-Ghamd and D. Mba, "A comparative experimental study onthe use of acoustic emission and vibration analysis for bearing defectidentification and estimation of defect size," Mechanical Systems andSignal Processing, vol. 20, pp. 1537–1571, 2006.
- [11] B. Li, M.-Y. Chow, Y. Tipsuwan and J. Hung, "Neural-network-based motor rolling bearing fault diagnosis," IEEE Trans. IndustrialElectronics, vol. 47, no. 5, pp. 1060–1069, Oct. 2000.
- [12] Orchard, M., "A Particle Filtering-Based Framework for On-Line Fault Diagnosis and

Failure Prognosis,” Ph.D. Thesis. Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, 2007.

- [13] Orchard, M., Wu, B. and Vachtsevanos, G., “A Particle Filter Framework for Failure Prognosis,” Proceedings of WTC2005, World Tribology Congress III, Washington D.C., USA, 2005.
- [14] Rafik Mahdaoui, Mouss” Industrial dynamics monitoring by Temporals Neuro-Fuzzy systems: Application to manufacturing system” GVIP 2009.

OPTIMIZATION THE AUTOMATED GUIDED VEHICLES RULES FOR A MULTIPLE-LOAD AGV SYSTEM USING SIMULATION AND SAW, VICOR AND TOPSIS METHODS IN A FMS ENVIRONMENT

Parham Azimi^(a), Masuomeh Gardeshi^(b)

^(a) Islamic Azad University (Qazvin Branch), Department of Mechanical and Industrial Engineering

^(b) Allameh Tabatabaee University, Department of Management and Accounting

^(a) p.azimi@yahoo.com, ^(b) Gardeshi_maryam@yahoo.com

ABSTRACT

In this paper, several pickup/delivery and pickup/dispatching rules have been examined in an automated guided vehicle system which is the biggest set of strategies in the literature. The best control strategy has been determined considering some important criteria such as System Throughput(ST), Mean Flow Time of Parts (MFTP), Mean Tardiness of Parts (MFTP), AGV Idle Time (AGVIT), AGV Travel Full (AGVTF), AGV Travel Empty (AGVTE), AGV Load Time (AGVLT), AGV Unload Time (AGVUT), Mean Queue Length (MQL) and Mean Queue Waiting (MQW). All strategies have been ranked using SAW, VICOR and TOPSIS methods. For this reason, several simulation experiments were conducted to obtain the best solution. As the experimental results show the approach is effective enough to be used in real world environments.

Keywords: AGV, Control strategy, Ranking Methods

1. INTRODUCTION

According to Tompkins and White (1984), about 20%-50% of total operational costs can be attributed to material handling system. As a result, researchers have been looking for methods to minimize their material handling cost. One solution is to automate material handling operations. Because of the rapid advances of automation, computer and control technologies, many automated material handling systems are available to us today [5]. Because of their routing flexibility, automated guided vehicles (AGVs) have been used in many manufacturing systems such as parts manufacturing systems with diverse and complex processing routes, warehouses, dispatching systems, local and international transportation systems, ports and etc [2]. In recent years, there have been many studies on AGV-related problems. Automated guided vehicles (AGVs) are known for their routing flexibility advantage. AGVs are driverless transportation systems that are being used in horizontal movements. The concept was introduced by [1] in 1995. Since that time, several applications have been developed. In designing an AGV system, many tactical (e.g., system design like

pickup/delivery or P/D points, the fleet size, flow path layout, etc.) and operational (e.g., routing or dispatching strategies) problems have been addressed. For example, the older ones were addressed by Co and Tanchoco [15], King and Wilson [16], Ganesharajah and Sriskandarajah [17], Johnson and Brandeau [18], Manda and Palekar [19], and Hoff and Sarker [20]. Co and Tanchoco discussed the operational issues of dispatching, routing, and scheduling of AGVs. According to [3], several key points must be considered in designing an AGV system:

- Flow path layout,
- Traffic management for preventing any deadlocks and collisions,
- Position and number of P/D points,
- Fleet size
- Dispatching rules
- Routing rules
- Locating the idle AGVs
- Breakdown management

Despite lots of advantages, AGVs has a famous disadvantage which is being more difficult to be controlled. Many issues need to be resolved in AGV controlling system such as pickup-dispatching problem [4]. The AGV control problem involves determining a place that the AGV should visit in order to perform its pickup or delivery task [5]. One important characteristic of this problem is that a vehicle load in any given route is a mix of pickup and delivery loads [6]. The pickup and delivery problem (PDP), with or without time windows, has been widely studied in the literature by many researchers, from various formulations to several solution methods, have been proposed to deal with different versions of the PDP. Most exact and heuristic methods have been developed to solve real instances of static and dynamic problems under either stochastic or deterministic demand. In most dynamic versions of the PDP (with demand that appears in real-time), it is assumed that the dispatcher manages reliable advanced information with regard to service requests. Over the last few years, the interest in studying the dynamic and stochastic versions of the PDP (associated with dial-a-ride systems) has been grown rapidly, mainly due to the

access to communication and information technologies, as well as the current interest in real-time dispatching and routing environments [7]. A multiple load AGV that can carry several loads for pickup or delivery has four major problems. The first problem is the task-determination problem that determines whether the next task is a pickup task or a delivery task. The second problem is referred as the delivery-dispatching problem in which the best delivery point is determined if its next task is a delivery task. The third problem is referred to as the pickup-dispatching problem. In this problem, the best pickup point is selected if its next task is a pickup task. Finally, the fourth problem is the load-selection problem, in which the best load is selected to be picked up from the output queue of a pickup point [5].

Now, the four major problems are defined in more details as follows

1.1. Task-determination problem

In order to select a task between pickup and delivery tasks for a semi loaded AGV, [3] proposed three strategies as follows:

- Delivery-task first (DTF): according to this rule, the AGV must deliver its load first then can pick up another load even in a coincident case when it receives both pickup and delivery requests.
- Pickup-task first (PTF): this rule is the opposite of DTF.

Load ratios (LR): LR can be formulated as follows:
 $LR = \frac{\text{number of loads in AGV}}{\text{AGV capacity}}$

LR strategy could have several forms in application. An example can be found in Table 1 by [5].

Table 1: An example for LR strategy

Criteria for LR	Probability (%)	Next Task
$0 < LR << 35\%$	D	30
	P	70
$35 < LR << 65\%$	D	00
	P	00
$65 < LR < 100\%$	D	70
	P	30

According to [5] and [8], it was shown that DTF has the best performance, so this strategy was used in the simulation model. One may define control strategies as follows.

1.2. Delivery-dispatching problem

If the next task is delivery and there are several P/D points in the selected route, in order to determine the best pickup point, one can define some strategies such as: Longest time in system (LTS), Longest Waiting Time At Pick up point (LWTAP), Longest Average Waiting Time At Pick up point (LAWTAP), Shortest Distance (SD), Greatest QUEUE Length (GQL), Earliest Due Time (EDT), Earliest Average Due Time (EADT), Smallest Remaining Processing Time

(SRPT) and Smallest Slack Time (SST) according to [5].

1.3. Delivery problem

An AGV faces to this case when it has several loads and it must be determined a P/D point in its route to deliver its loads. According to [5], the same strategies as previous section can be developed.

1.4. Selection problem

If there are several loads in the queue of a P/D point, an AGV must select the best load to be picked up. According [9], because the best strategy is First-In-Queue-First-Out (FIQFO), this strategy was used in the simulation model.

2. SIMULATION MODEL

In the simulation model, some specific assumptions were considered. All vehicles are multiple-load AGVs and the fleet size in the system is 3 units. The flow path layout and all model information are the same as the one which was adopted by [8] and [9] for the best comparison. The flow path layout is shown in Figures 2, 3, and 4 where all paths are unidirectional with the capacity of one unit to prevent any conflicts. In order to unload the loads before picking up more loads from a machine by an AGV, the delivery point and the pickup point of every machine has been arranged. Every machine has a buffer area, at which idle AGVs can stay and wait for pickup requests. All AGVs have the same loading capacity and same speed (1.8 m/s). Parts are placed in the pallets and in each pallet, there is only one type of product and for each part, the production sequence and the Mix-Ratio are known (Table 3). The load-carry capacity of these AGVs is four loads. There are 12 machines in the manufacturing system as mentioned in Figure 2. Workstations 1 and 12 are the entry and sink stations, respectively. The workstations 2–11 are processing machines. The number of part types made in the system is six. Table 4 shows the distribution functions of each machine processing time. It is assumed that parts will go through the same operations on the same workstations. It is also assumed that the setup times are included in the related processing times. Furthermore, in the simulations, a part is assigned with a due time when it arrives at the system randomly. The due time is generated by adding the arrival time with a random number. According to the levels which were shown in Table 2, there are 20 different strategies which will be used in the simulation model as control strategies. We used a coding system for referring any kind of strategies using the capital letters shown in the columns of Table 2. For example, a strategy (or problem) TIP1D1L1 refers to a strategy where the task rule is DTF, the pickup-dispatching rule is LTIS, the delivery dispatching rule is SQL, and the load-selection rule is FIQFO [9]. Meanwhile, in the simulation model, we used NV as a workstation-initiated approach for assigning the AGVs to the next

task. In order to evaluate the control strategies, the following criteria were used in the model and the number of each criterion was used as a reference in the simulation experiments:

1. System throughput (ST),
2. mean flow time of parts (MFTP),
3. The mean tardiness of parts (MTP),
4. Percentage of vehicles idle time (AGVI),
5. Percentage of time moving vehicles with full capacity (AGVTF),
6. percent time on moving vehicles with empty capacity (AGVTE),
7. Percentage of load time (AGVL),
8. Percentage of unload time (AGVUL),
9. The average queue length in pickup and delivery points (MQL),
10. The average waiting time in pickup and delivery points (MQW) [9].

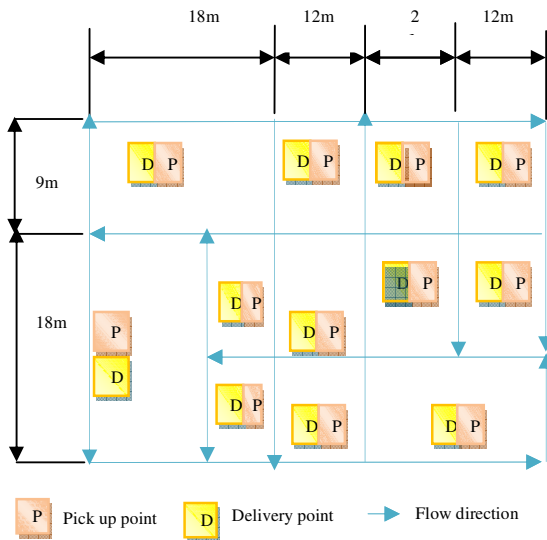


Figure 1: The flow path layout

Load-Selection	Delivery-Dispatching	Pickup-Dispatching	Tasks	Levels
FIQFO	SQL	LTIS	DFT	1
	EDT	GOL		2
	SD	EDT		3
	LIQFO	SRPT		4
	FIFO	LWTAP		5
		SD		6

Table 2: The levels of controlling strategies

Table 3: The mix-ratio and process sequence of each part

Part Type	Mix-Ratio	Sequence
1	0.16	1-3-5-7-9-11-12
2	0.17	1-2-4-6-8-10-12
3	0.18	1-4-5-7-9-10-12
4	0.15	1-4-5-7-9-10-12

5	0.14	1-3-4-5-9-11-12
6	0.20	1-2-3-6-8-9-12

All simulation experiments were run by Enterprise Dynamics V8.1 software. The number of replications for each calculation was set at 30 by independent sunburns. The simulation period for each replication was 170,000 seconds. For determining the warm-up period, the throughput criterion was used in 30 runs. The results were shown in Figure 3.

As the figure shows, when the total production reaches 750 units (480,000 seconds), the system reaches a stable state. Therefore, for simulation replications, at first a warm-up period of 480,000 seconds ran then 30 replications were executed afterward for each calculation (Fig. 3). Due to several criteria, we have used the mean of three criteria decision making methods such as VICOR, SAW and TPSIS to evaluate and rank the results for the control strategies (Table 7). Other data has been taken from [9] in Fig. 2 and Table 5.

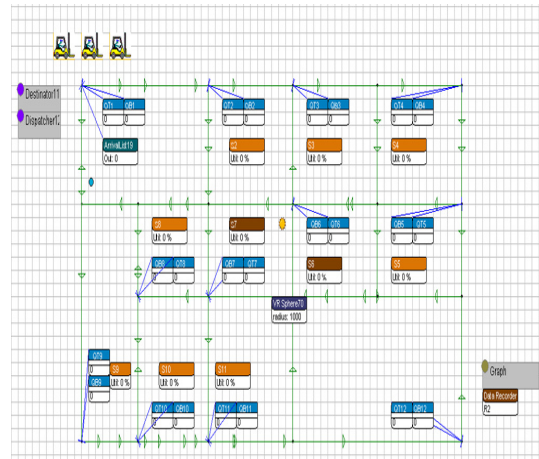


Figure 2: Two-dimensional view

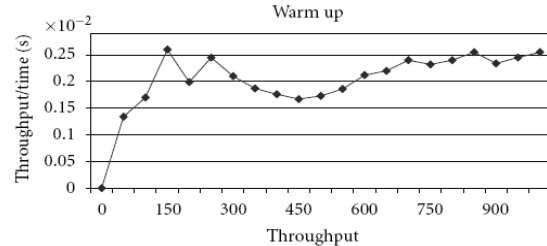


Figure 3: The warm-up diagram

3. COMPUTATIONAL RESULTS

For calculating the weight of each criterion, the experts' views which had been based on a field study were taken. At first, we had some interviews with 10 special experts. All experts were the production managers and the financial managers of 5 local auto manufacturers which are using multiple-load AGVs in their production sites and the weights are listed in Table 4.

Table 4: The weight of strategies

ST +	MFTP -	MTP -	MQL -	MQW -
16	14	11	15	16
AGVI -	AGVTF +	AGVTE -	AGVL -	AGVUL -
9	4	3	6	6

As the results show, the greatest weights belong to ST and MQW and the lowest ones belong to AGVTE criterion.

Table 5: The processing-time distribution and the production sequence of each product type

Work station	Processing time distribution (min)
2	N (1,0.1)
3	N (1.5,0.15)
4	N (2,0.2)
5	N (1,0.1)
6	N (2,0.2)
7	N (2,0.2)
8	N (1.5,0.15)
9	N (1.5,0.15)
10	N (2,0.2)
11	N (1,0.1)

The results of each ranking methods show in next tables.

Table 7 :Ranking result by SAW method

Rank	Strategy SAW	Grade	Rank	Strategy SAW	Grade
1	21	0.8075	16	14	0.4754
2	23	0.7963	17	13	0.4678
3	25	0.7702	18	16	0.4624
4	22	0.7379	19	18	0.4614
5	24	0.734	20	6	0.4347
6	2	0.681	21	8	0.4269
7	5	0.6799	22	9	0.4186
8	1	0.6787	23	15	0.4104
9	4	0.6781	24	26	0.4081
10	3	0.6721	25	20	0.4077
11	7	0.5113	26	17	0.4072
12	10	0.5041	27	19	0.4028
13	12	0.4825	28	27	0.4026
14	29	0.4781	29	28	0.395
15	11	0.476	30	30	0.3843

Table 8 :Ranking result by TOPSIS method

Rank	Strategy TOPSIS	Grade	Rank	Strategy TOPSIS	Grade
1	21	0.883	16	13	0.5312
2	23	0.8399	17	7	0.5275
3	22	0.8388	18	29	0.525
4	25	0.8381	19	9	0.5008
5	24	0.8334	20	30	0.4865
6	15	0.5855	21	6	0.4837
7	1	0.5789	22	8	0.4807
8	3	0.5781	23	16	0.4776
9	14	0.5758	24	18	0.4736
10	4	0.5738	25	26	0.4626
11	5	0.5707	26	27	0.4626
12	11	0.5706	27	28	0.4584
13	2	0.5706	28	20	0.4571
14	12	0.5628	29	27	0.4534
15	10	0.5567	30	19	0.4201

Table 9-Ranking result by VICOR method

Rank	Strategy VICOR V=0.5s	Grade	Rank	Strategy VICOR V=0.5s	Grade
1	15	0.2807	16	10	0.673
2	22	0.2834	17	29	0.7449
3	21	0.3577	18	7	0.7905
4	23	0.4143	19	6	0.7975
5	24	0.4263	20	8	0.8022
6	25	0.4423	21	30	0.8061
7	12	0.4821	22	16	0.8297
8	14	0.5168	23	9	0.8406
9	11	0.586	24	28	0.8458
10	1	0.5972	25	18	0.8483
11	3	0.6036	26	26	0.8583
12	4	0.6095	27	27	0.8941
13	5	0.6187	28	20	0.9001
14	2	0.6271	29	17	0.9692
15	13	0.6306	30	19	0.9721

According to the results in Table 6, the best strategy regarding ST criterion is T1P2D3L1 and

The worst one is T1P1D2L1, because it uses GQL as pickup-dispatching rule and SD as Delivery-dispatching rule, so these strategies have the greatest influences on the system throughput. Regarding MQW as the second important criterion, the best strategy is T1P5D5L1 and the worst one is T1P4D2L1. Maximum queue length happens when it uses T1P4D2L1 and the shortest length belongs to the strategy T1P5D5L1.

As Tables 7, 8 and 9 shows the best and worst strategy of each method is different because of this reason we have used BORDA method, for combining the result of these three methods.

The main contribution of this paper is using the mean of TOPSIS, SAW and VICOR weights for selecting the

best control strategies by mixing them with mentioned weights. The related results were shown in Table 10.

Because, ED 8.1 has several graphical tools, the verification of the simulation model was easy task and regarding the model validation, we compared our model to the one developed by [9]. Because both models have the same parameters, the system throughput must be the same at 95% as the significant level. Because our models results had not Normal distribution and the fact that both models have not been paired with different variances, we used the Smith-Satterwaithe test as the validity criterion. The number of samples was 20 and the p-value of the test was 0.0004 so the H_0 hypothesis was accepted, so both models have the same results.

Table 6- Simulation results

Criterion Strategy	ST	MTP	MFTP	AGVI	MQW	MQL	AGVUL	AGVL	AGVTE	AGVTF
TIP1D1L1	695.33	358.1344	229.2505	0.00	2097.33	10.63	0.06	0.06	0.53	0.35
TIP1D2L1	688.50	371.2092	231.4515	0.00	2103.86	10.16	0.06	0.06	0.52	0.35
TIP1D3L1	694.30	357.8818	225.8998	0.00	2128.03	11.38	0.06	0.06	0.53	0.35
TIP1D4L1	692.40	365.5618	232.0686	0.00	2083.89	10.65	0.06	0.06	0.52	0.35
TIP1D5L1	690.43	370.8603	226.8865	0.00	2037.88	10.95	0.06	0.06	0.52	0.35
TIP2D1L1	1034.07	16.2724	702.8447	0.02	10455.54	50.42	0.10	0.10	0.47	0.31
TIP2D2L1	1013.33	19.2254	774.6444	0.00	10959.81	52.91	0.10	0.10	0.43	0.37
TIP2D3L1	1036.40	19.02247	704.0995	0.02	10547.30	50.61	0.10	0.10	0.46	0.32
TIP2D4L1	1023.23	24.2128	744.1661	0.01	11060.93	51.79	0.10	0.10	0.44	0.35
TIP2D5L1	1022.73	24.5994	749.2228	0.00	10332.17	42.97	0.10	0.10	0.44	0.35
TIP3D1L1	954.10	86.1634	822.0297	0.00	7664.92	33.76	0.09	0.09	0.49	0.32
TIP3D2L1	1013.90	95.717	782.3533	0.00	8,020.67815	35.9226	0.10	0.10	0.44	0.37
TIP3D3L1	966.6	110.8199	827.92455	0.00	8,546.09154	39.04	0.09	0.09	0.49	0.32
TIP3D4L1	933.70	69.5078	780.081	0.00	7801.11	37.52	0.09	0.09	0.48	0.33
TIP3D5L1	955.27	51.2697	649.5937	0.01	7152.91	38.71	0.09	0.09	0.48	0.33
TIP4D1L1	1,018	18.21646	283.6282	0.03	10794.58	60.67	0.10	0.10	0.47	0.31
TIP4D2L1	1010.07	23.18528	855.147	0.01	11315.18	65.03884	0.10	0.10	0.46	0.33
TIP4D3L1	1,017	19.02247	276.5005	0.03	10902.08	62.83	0.10	0.10	0.46	0.31
TIP4D4L1	1012.47	25.53951	690.31	0.03	11108.33	63.33	0.10	0.10	0.46	0.32
TIP4D5L1	1012.20	25.2164	634.67	0.02	10956.82	62.94	0.10	0.10	0.46	0.32
TIP5D1L1	725.97	29.1373	145.1971	0.00	2087.37	6.60	0.07	0.07	0.52	0.34
TIP5D2L1	750.13	10.2596	210.8495	0.01	1517.19	11.43	0.07	0.07	0.50	0.35
TIP5D3L1	717.80	9.7899	150.8138	0.01	1976.14	6.77	0.07	0.07	0.51	0.34
TIP5D4L1	719.33	11.5195	228.7747	0.01	2027.73	6.63	0.07	0.07	0.51	0.34
TIP5D5L1	711.43	13.7399	148.4764	0.01	2001.99	6.50	0.07	0.07	0.51	0.34
TIP6D1L1	975.80	23.9437	663.6724	0.02	10628.23	59.85	0.09	0.09	0.50	0.30
TIP6D2L1	962.83	26.03228	670.0885	0.02	10814.46	62.67	0.09	0.09	0.49	0.31
TIP6D3L1	973.03	34.0848	652.5999	0.02	10532.46	59.74	0.09	0.09	0.50	0.30
TIP6D4L1	972.93	40.5459	648.8595	0.00	10381.81	59.98	0.09	0.09	0.50	0.32
TIP6D5L1	978.03	53.8118	624.6368	0.01	10363.30	60.69	0.10	0.10	0.48	0.31

According to the final results, the best strategy is TIP5D1L1 and the worst one is TIP3D4L1, that is, when we use LWTAP for pickup dispatching and SQL for delivery-dispatching activities.

Table 10 :Final Ranking result by BORDA method

Rank	Strategies	Grade	Rank	Strategies	Grade
1	21	29	16	13	14
2	23	28	17	7	13
3	22	27	18	29	12
4	25	26	19	6	11
5	24	25	20	8	10
6	15	24	21	30	8
7	6	23	22	16	8
8	3	22	23	9	8
9	14	21	24	18	6
10	4	20	25	26	5
11	5	19	26	28	3
12	12	18	27	17	3
13	11	17	28	20	3
14	2	17	29	27	1
15	10	15	30	19	0

The results show the importance of due time for selecting the best control Strategies. The role of due time in the current consuming market conditions where the market is full of different brands with suitable quality and services is a key factor to keep the customers satisfied by agreed delivery times. For comparison purposes, we selected the study done by [8]. They used 3 criteria and 18 different strategies but here we used 10 criteria and 30 different strategies. They used ANOVA analysis for ranking the strategies, but here, we used TOPSIS together with SAW and VICOR method for ranking them. The approach used here is more close to the real applications where the top managers can change the strategies based on the production, market situation, and financial issues. For another comparison, the best strategy reported by [9] was TIP2D3L1 and the worst one was TIP1D3L1. We have developed more control strategies than [9] and it helped us to find better solutions for some criteria like MQW and MQL. Another important finding is that it is not reasonable to just focus on one criterion. As mentioned in the literature, most previous researches were focused on optimizing the system throughput. According to Table 11 and 6, taking TIP2D3L1 has the greatest value of system throughput but stands in the 20st rank.

Table 11: Strategies ranking by BORDA method

Strategy	rank	Strategy	Rank
TIP1D1L1	7	TIP4D1L1	22
TIP1D2L1	14	TIP4D2L1	27
TIP1D3L1	8	TIP4D3L1	24
TIP1D4L1	10	TIP4D4L1	30
TIP1D5L1	11	TIP4D5L1	28
TIP2D1L1	19	TIP5D1L1	1
TIP2D2L1	17	TIP5D2L1	3
TIP2D3L1	20	TIP5D3L1	2
TIP2D4L1	23	TIP5D4L1	5
TIP2D5L1	15	TIP5D5L1	4
TIP3D1L1	12	TIP6D1L1	25
TIP3D2L1	20	TIP6D2L1	29
TIP3D3L1	16	TIP6D3L1	26
TIP3D4L1	9	TIP6D4L1	18
TIP3D5L1	6	TIP6D5L1	21

4. CONCLUSIONS

In this paper, pickup-dispatching problem together with delivery-dispatching problem of a multiple-load automated guided vehicle (AGV) system has been studied. Several different rules of these problems were used to create the best control strategies. For selecting the best strategy several important criteria were considered, such as System, Throughput (ST), Mean Flow Time.

of Parts (MFTP), Mean Tardiness of Parts (MFTP), AGV Idle Time (AGVIT), AGV Travel Full (AGVTF), AGV Travel Empty (AGVTE), AGV Load Time (AGVLT), AGV Unload Time (AGVUT), Mean Queue Length (MQL), and Mean Queue Waiting (MQW), in a part manufacturing system where each part has a due date. The criteria were evaluated by a filed study with 10 experts who work in 5 local auto manufacturing companies to make the results as applicable as possible. For evaluating each criterion, we used three MADM methods: TOPSIS, VICOR and SAW methods. Finally, for ranking and selecting the best control strategy, BORDA method and importance weights were applied.

Here we defined the distances between workstations and calculated the warm-up period in order to make the simulation more practical while the total strategies examined were 30 strategies with 10 criteria which had been the biggest sets tested so far.

The first contribution of the paper is using several criteria for selecting the best control strategy. Most previous researches just focused one or two criteria. The second contribution of the current research is using a large number of control strategies in comparison to latest studies like [8] and [9] that helped us to obtain better results. The results show that the proposed algorithm is efficient and robust enough to be used in applications. Regarding the research limitations, we have not considered the optimization process together

with the FMS which can be carried out in future researches.

REFERENCES

- [1] T. Muller *Automated Guided Vehicles*, IFS (Publications)/Springer, Berlin, Germany, 1983.
- [2] I. F. A. Vis, "Survey of research in the design and control of automated guided vehicle systems," *European Journal of Operational Research*, vol. 170, no. 3, pp. 677–709, 2006.
- [3] Malmborg, C.J., 1990. A model for the design of zone control automated guided vehicle systems. *International Journal of Production Research* 28 (10), 1741–1758.
- [4] Gilbert Laporte a, Reza Zanjirani Farahani b,* , Elnaz Miandoabchi b, c. Designing an efficient method for tandem AGV network design problem using tabu search/Elsevier,2006.
- [5] Y.C. Hoa and S.H. CHIEN. A simulation study on the performance of task-determination rules and delivery-dispatching rules for multiple-load AGVs, *International Journal of Production Research*, Vol. 44, No. 20, 15 October 2006, pp 4193–4222.
- [6] Ferimn Alfredo Tang Montané, Roberto Diéguez Galvao, A tabu search algorithm for the vehicle routing problem with simultaneous pick-up and delivery service /Elsevier, 2006,pp 595–619.
- [7] Doris Saeza, Cristian E. Cortésb, Alfredo Nõõeza, Hybrid adaptive predictive control for the multi-vehicle dynamic pick-up and delivery problem based on genetic algorithms and fuzzy clustering/ Elsevier, 2008, pp 3412 – 3438.
- [8] Y.-C. Ho and H.-C. Liu, "A simulation study on the performance of pickup-dispatching rules for multiple-load AGVs, *Computers and Industrial Engineering*, vol. 51, no. 3, pp. 445–463, 2006.
- [9] Parham Azimi, Hasan Haleh, and Mehran Alidoost, The Selection of the Best Control Rule for a Multiple-Load AGV System Using Simulation and Fuzzy MADM in a Flexible Manufacturing System/ Hindawi Publishing Corporation Modeling and Simulation in Engineering, Volume 2010, Article ID 821701, 11 pages doi:10.1155/2010/821701.
- [10] Incontrol simulation solution B.V.
- [11] J. A. Tompkins and J. A. White, *Facility Planning*, Wiley, New York, NY, USA, 1984.
- [15] C. G. Co and J. M. A. Tanchoco, "A review of research on AGVS vehicle management," *Engineering Costs and Production Economics*, vol. 21, no. 1, pp. 35–42, 1991.
- [16] R. E. King and C. Wilson, "A review of automated-guided vehicle systems design and scheduling," *Production Planning and Control*, vol. 2, no. 1, pp. 44–51, 1991.
- [17] T. Ganesharajah and C. Sriskandarajah, "Survey of scheduling research in AGV-served manufacturing systems," in *Proceedings of the Instrumentation Systems Automation Technical Conference (IAS '95)*, vol. 50, pp. 87–94, Toronto, Canada, April 1995.
- [18] M. E. Johnson and M. L. Brandeau, "Stochastic modeling for automated material handling system design and control," *Transportation Science*, vol. 30, no. 4, pp. 330–350, 1996.
- [19] B. S. Manda and U. S. Palekar, "Recent advances in the design and analysis of material handling Systems," *Journal of Manufacturing Science and Engineering*, vol. 119, no. 4, pp. 841–848, 1997.
- [20] E. B. Hoff and B. R. Sarker, "An overview of path design and dispatching methods for automated guided vehicles," *Integrated Manufacturing Systems*, vol. 9, no. 5, pp. 296–307, 1998.

A NEW METHOD FOR THE VALIDATION AND OPTIMISATION OF UNSTABLE DISCRETE EVENT MODELS

Hans-Peter Barbey

University of Applied Sciences Bielefeld

hans-peter.barbey@fh-bielefeld.de

ABSTRACT

Logistic systems can be designed as push-systems or pull-systems. In a pull-system, one parameter, e.g. the stock items have to be closed-loop controlled. It will be shown that the closed-loop controlled model can work in an unstable manner like a “logistic oscillating circuit”. In comparison with a closed-loop controlled electronic system, the elements of a logistic system have relatively long dead times. In discrete event simulation, there is no method to optimally calculate the parameters of the system also taking account of the dead times. Furthermore, there is no validation technique to date which can determine the unstable behaviour of a system. In many different areas of engineering, the FFT analysis is a frequently used method. It will be shown that the FFT analysis is a suitable method to determine the unstable behaviour of discrete event simulation models. However, FFT analysis is only a method to determine the unstable behaviour; the elimination can only be done by trial and error.

Keywords: Simulation, closed-loop control, discrete event models, validation, FFT analysis

1. INTRODUCTION

Discrete event simulation is a frequently used method for designing new, complex production systems. The advantage of this method is that all elements of the system can be described with only a few parameters. Objectives of such a simulation study are to obtain knowledge of, for example, cycle times or utilisation of the machinery. There are numerous commercially available programs for this.

The execution of a complete simulation study is described in VDI 3633 (2000). An important point within a simulation study is the validation of the model and the results. There is no generally applicable guideline for this. Depending on the particular project, different methods can be applied. These methods are described in numerous articles. A good summary of all the literature and methods of validation is in Rabe et al (2008) (table 1).

2. SIMULATION OF LOGISTIC SYSTEMS

2.1. Characteristics of a closed-loop controlled system

Push-systems are often used for the management of a production system. These systems are producing goods

Table 1: Validation methods (according to Rabe et al.)

	Target description	Task-specification	Concept model	Formal model	Executable model	Simulation results	Basic data	Processed data
Animation					X	X		
Review	X	X	X	X	X	X	X	X
Dimensional Consistency test				X	X	X	X	X
Event validity test					X			
Fixed value test				X	X	X		
Extreme condition test				X	X	X		
Monitoring					X	X		X
Desk checking	X	X	X	X	X	X	X	X
Sensitivity Analysis					X	X		X
Statistical Techniques					X	X	X	X
Structural walkthrough	X	X	X	X	X	X	X	X
Internal validity test					X	X		
Sub model testing			X	X	X			
Trace analysis					X			
Turing test					X			
Cause-effect graph			X	X	X			
Face validity	X	X	X	X	X	X	X	X
Predictive validation					X			
Comparison to other models					X	X		
Historical data validation					X			

without a customer order. Therefore, the stock in such a system is relatively high. To reduce stock, a pull-system

is the better solution. Here, the production only starts on demand. The higher the demand, the faster is the production rate. These systems are closed-loop controlled.

These closed-loop controlled systems can of course also be designed using discrete event simulation. As there is usually no special module for this, the controller has to be designed using a special program code.

First, it is necessary to point out the differences between a logistics closed-loop controlled system and closed-loop controlled systems in for example automation technology. In the automation technology, the systems are designed using electronic modules. The signals in these systems are therefore electric currents or voltages. Therefore the speed of the signals is practically infinite. In contrast, the speed of the signal in a logistics system is very low. The speed of a conveyor is low or the processing time on a production machine is high. Thus, the speed of the signal which is connected with the flow of objects through the production system is very low. Described in terms of control technology, the time difference between the entry of an object or signal to a machine and the exit is the dead time. It is generally known that dead times in a closed-loop controlled system can result in oscillations and thus unstable systems. There are methods in control technology to design the parameters of a system so that these oscillations can be prevented. These methods could not be applied in the discrete event simulation due to missing mathematical relationship between the system parameters.

2.2. Design of an unstable closed-loop controlled system

First, it is necessary to demonstrate, that unstable behaviour can also occur in logistics systems. A very simple model will be designed for this (Figure 1): The production system delivers the goods to a stock 1. This stock has a constant storage time, e.g. for cooling. Then the items are stored in a stock 2, e.g. for distribution. The stock output rate is constant. The difference between the current stock and the target stock is the parameter which controls the production cycle time. The higher the difference, the faster is the production (Figure 1).

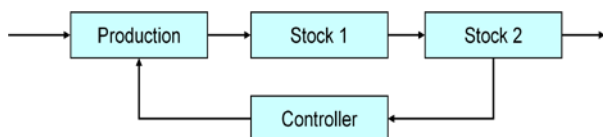


Figure 1: Simple closed-loop controlled production system

The system is designed with the following parameters:

- The cycle time of the production is closed-loop controlled
- The production has an infinite capacity
- The storage time of stock 1 is 1 day

- The target stock is 15 units for stock 2
- Both stocks have an infinite capacity
- The stock output rate of stock 2 is constant at 1 unit/4.8 h

This system is closed-loop controlled by a controller with the following characteristics:

- The current stock is recorded once per day
- The cycle time of the production is calculated with a delay of one half day
- The difference of the stock is fed into the production system on the next day

A discrete event model is created using these values (Figure.2). A similar model is presented in Barbey (2008).

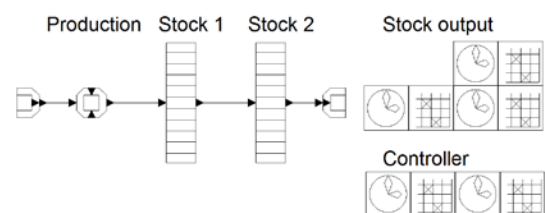


Figure 2: Discrete event model designed with DOSIMIS

2.3. Validation and results

In the following diagram the results for the stock are recorded over a period of 5 weeks

At the very beginning, the controller is switched off. The stock has a constant figure of 30 units here. After approx. two weeks, the controller is switched on. Then the stock has on average the target value. However, it is overlaid with an oscillation with an amplitude of 7 units (Figure 3).

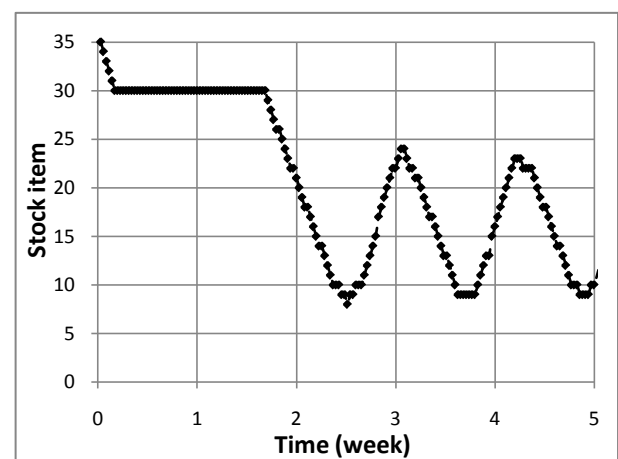


Figure 3: Stock of the first model

This model demonstrates that unstable behaviour in logistics systems can occur if an unsuitable set of parameters is applied. All input parameters are constant, but the stock is only constant on average. The model is acting like a “logistic oscillating circuit”.

If unstable behaviour can be produced in one model, then it is also possible to produce unstable behaviour in any other model. Therefore, validation of the model is an important issue.

According to Table 1, there are many validation methods which can be applied. However, only the reviewing of the results is suitable for this model. However, this model has been designed as simple as possible with the only objective of indicating an error: in this case unstable behaviour. It is also obvious that this error will be found with the selected validation methods. However, the models are generally much more complex and the unstable behaviour is thus possibly not immediately visible.

All the methods in Table 1 can find almost any kind of error in discrete simulation models, but they are not suitable for detecting unstable behaviour. Therefore, a new validation method must be found. In many areas of engineering, the Fourier analysis is a frequently used method to analyse dynamic processes. This method states that each periodic signal can be compiled from a sum of sine and cosine-functions. The mathematical theory for this method is described in Brigham (1997).

The Fourier analysis is initially applied to the described model. The results of the Fourier analysis which have been produced here using Excel are shown in Figure 4. The frequency range was analysed up to a frequency of 2 1/d.

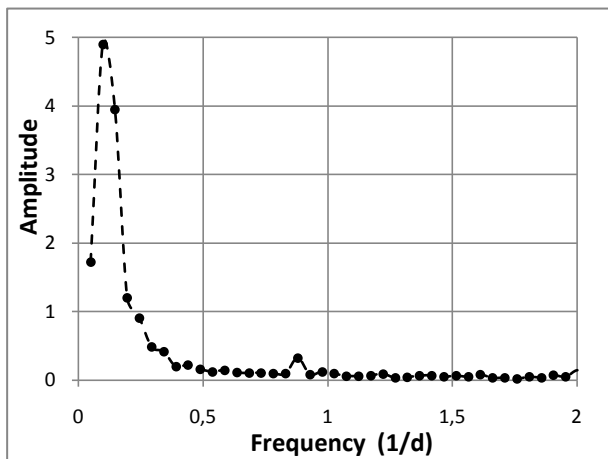


Figure 4: FFT analysis with Excel for the first model

There is only one peak at a frequency of approx. 0.12 1/d. The same result can be obtained from the consideration of Figure 3. This frequency is the characteristic frequency of the unstable system. The application of the Fourier analysis to this simple model demonstrates that this method is suitable for the validation of discrete event models.

3. VALIDATION AND OPTIMISATION OF COMPLEX MODELS

Chapter 2 illustrates the application of a new validation method to a very simple model. Normally, models are significantly more complex. Therefore, this simple model will be somewhat modified in the following.

3.1. Model with an additional output

Two parameters are changed in the next model:

- The current stock is recorded once per day and the cycle time of the production is calculated immediately
- There is an additional stock output each 1.4d. The additional output is a random number between 4 and 8.

The results are shown in Figure 5. The result still seems to be periodic. There must be a periodic part in, because the additional stock output is periodic. But there is no method according to Table 1 in this case which detects the unstable behaviour of the model.

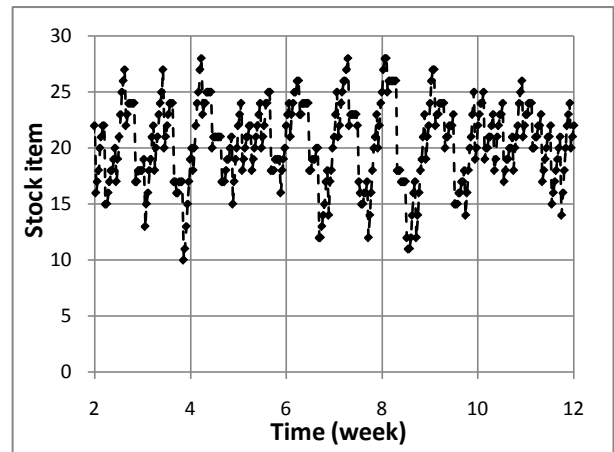


Figure 5: Stock with an additional output

The unstable behaviour is also not visible in Figure 5. Therefore, the Fourier analysis is now applied to the results. The diagram in Figure 6 shows some typical characteristics of the model.

There is one peak at a frequency of 0.7 1/d. This is related to the additional output which has exactly this frequency. A second peak is at a frequency of 1.4 1/d which is double the frequency of the stock output. An extension of the frequency range would show additional peaks at all frequency multiples of the output.

However, the largest peak is at a frequency of 0.15 1/d. This frequency cannot be explained by the additional output; it is related to the unstable behaviour of the model itself.

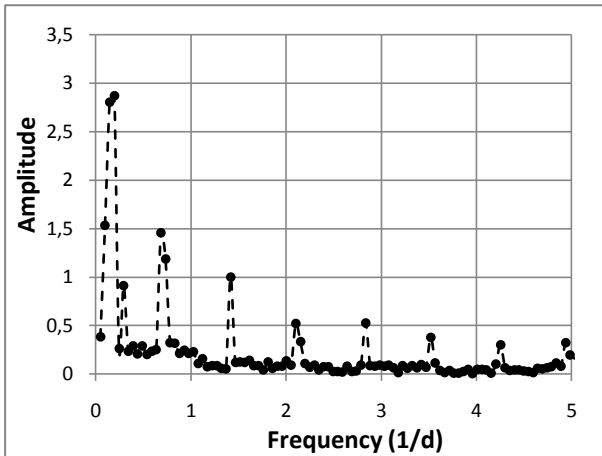


Figure 6: FFT analysis with an additional output

Using the frequency analysis, it is possible to determine the unstable behaviour of a model. For this, each characteristic frequency has to be compared with the parameters of the model. In this case, the additional output is a parameter which produces a periodic signal. Therefore, this frequency and its multiples can be excluded. All other frequencies which could not be explained by the parameters of the model show an unstable behaviour of the model.

However, the unstable behaviour cannot be eliminated using the Fourier analysis as there is no functional relationship between the model parameters. The elimination of the unstable behaviour can only be realised by trial and error. A variation of the model parameters and a following Fourier analysis shows whether the results are satisfactory.

3.2. Optimisation

Optimisation is always project-specific. Each model has a specific set of parameters which can be changed individually. The model is now only used to demonstrate how unstable behaviour can be eliminated by using the Fourier analysis.

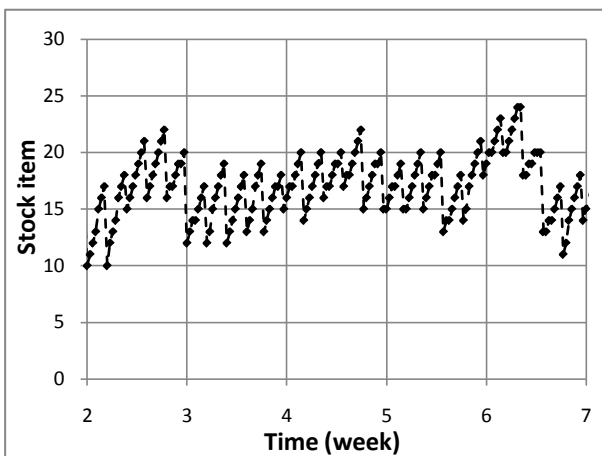


Figure 7: Stock with the modified controller

In the original model, the controller was designed so that the difference in the stock is supplied by the transport system to the production system on the next day. Now the controller will be modified so that only one third of the difference will be supplied to the production system on the next day. Figure 7 shows the stock with the modified controller. Compared with Figure 5, the fluctuation of the stock is reduced. Figure 8 shows the Fourier analysis for this.

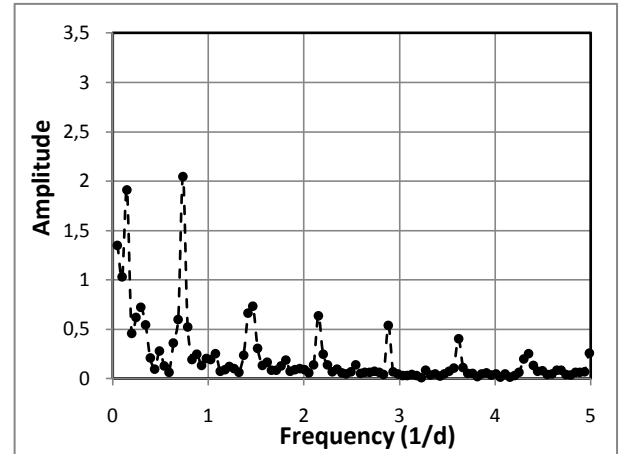


Figure 8: FFT analysis of the modified model

The peak at a frequency of 0.15 1/d is clearly reduced but still exists. All other peaks caused by the additional stock output are still there. An additional simulation with this model without the additional output shows the improvement actually achieved (Figure 9).

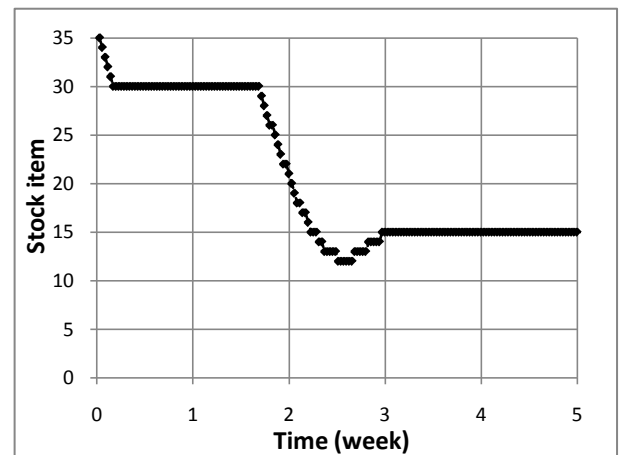


Figure 9: Stock without the additional output

Figure 3 shows a permanent oscillation after the controller is switched on. In contrast, Figure 9 only shows one oscillation. The signal then has a constant value of 15. This demonstrates an improvement, but the optimum has not yet been reached. Therefore, further variation of the parameters has to be done.

This procedure shows that the Fourier analysis is suitable for finding the instability of a model and optimising it afterwards.

REFERENCES

- Barbey, H.-P. 2008. "Simulation des Stabilitätsverhaltens von Produktionssystemen am Beispiel einer lagerbestandsgeregelten Produktion." *Proceedings of the 2008 ASIM-conference*, Berlin.
- Brigham, E.O. 1997. *FFT-Anwendungen*. R.Oldenburger Verlag, München.
- Rabe M., S. Spiekermann, S. Wenzel. 2008. *Verifikation und Validierung in Produktion und Logistik*. Springer Verlag, Berlin. Heidelberg.
- VDI 3633. *Simulation von Logistik-, Materialfluss- und Produktionssystemen. 2000*, Beuth Verlag, Berlin.

AUTHORS BIOGRAPHY

HANS-PETER BARBEY was born in Kiel, Germany, and went to the University of Hannover where he studied mechanical engineering and obtained his degree in 1981. At the same university, he obtained his doctorate in 1987. He worked for 10 years for different companies for plastic machinery and plastic processing before moving in 1997 to the University of Applied Sciences Bielefeld. There, he is teaching transportation technology, plant planning and discrete simulation. His research is focussed on the simulation of production processes.

His e-mail address is:

hans-peter.barbey@fh-bielefeld.de

And his web page can be found at

<http://www.fh-bielefeld.de/fb3/barbey>

SIMULATION HELPS ASSESS AND INCREASE AIRPLANE MANUFACTURING CAPACITY

Marcelo Zottolo
Edward J. Williams
Onur M. Ülgen
PMC
15726 Michigan Avenue
Dearborn, Michigan 48126 USA
(mzottolo | ewilliams | Ulgen)@pmcorp.com

ABSTRACT

Simulation has long been used in the manufacturing industry to help determine, and suggest ways of increasing, production capacity under a variety of scenarios. Indeed, historically, this economic sector was the first to make extensive use of simulation. Over the last several decades, and continuing today, the most numerous applications of simulation to manufacturing operations involve mass production facilities such as those fabricating motor vehicles or home appliances. Less frequently, but very usefully, simulation has been applied to customized manufacturing or fabrication applications, such as the building of ships to individualized specifications. In the case study described in this paper, simulation was successfully applied, in synergy with other techniques of industrial engineering, to assess and increase the throughput capacity of a manufacturer of custom-built personal jet airplanes with a four-to-six passenger (plus moderate amounts of luggage) carrying capacity.

Keywords: Manufacturing, “job shop,” customization, capacity planning, discrete-event simulation, bottleneck analysis

1. INTRODUCTION

Very likely, the most long-standing user of simulation, as distinguished by economic sector, is the manufacturing sector (Miller and Pegden 2000). Within this sector, simulation analysis helps production and industrial engineers (and their managers) assess and improve production capacity, identify and ameliorate bottlenecks, improve deployment of resources (whether labor, equipment, or both), and hence strengthen a company’s economic performance (Harrell and Tumay 1995). Frequently, these applications of simulation analyze a mass-production process, such as those producing automobiles or home appliances. Such processes are typically high-volume, have largely linear flow, and have a relatively low ratio of workers to machines. Somewhat less frequently, simulation analysis is applied to “job-shop” manufacturing, which typically involves a lower volume of production, with markedly higher cost and price per unit, directed toward often customized requests. Such manufacturing systems typically have more, and more highly skilled, workers relative to machines and equipment (El Wakil

2002). In view of the lower number of units produced and their higher prices and cost, each unit is “high stakes,” meriting careful attention to work flow, buffer capacities, and buffer placements to streamline workflow and minimize total process time (Heragu 2008). Various examples of “job shop” simulation appear in the literature. Implementation of an application to model the custom production of trains (general, fast, freight, etc.) is discussed in (Lian and Van Landeghem 2002); significantly, this analysis combines value stream mapping with simulation. The application of simulation to design-build construction projects is discussed in (Orsoni and Karadimas 2006). The expansion plan of a marine container terminal, incorporating production of custom equipments to be installed therein, is discussed in (Ambrosino and Tānfani 2009).

In the study described here, discrete-event process simulation was successfully applied to the paint-shop processes involved in the manufacture of custom-built jet airplanes for personal and corporate use. Such airplanes are a publicly inconspicuous but economically and logistically important part of the overall aviation infrastructure (McCartney 2011). The manufacturing company aspired, in view of trends indicating increasing order volume, to produce two or even three airplanes per day, yet initially was unable to produce 1½ airplanes on average per day. Since the painting operation was already known to be a painfully obvious bottleneck, simulation analysis was concentrated on it, and coupled synergistically with other industrial engineering techniques such as value stream mapping, layout analysis, and lean manufacturing.

2. OVERVIEW OF PAINTING PROCESSES

The airplane manufacturing facility comprises three large buildings, and the painting processes occupy all of the intermediate (in the process flow sense) building. This building, in turn, is divided into four major “positions.” Position 1 handles preparatory work: body work, washing, chemical coating, and thermal baking (hardening) of the chemical coating. Position 2 handles the vast majority of the actual painting: wrapping, spraying the primer coat, two consecutive sprayings of the top coat (to achieve durability and opacity), and thermal baking of these three coats. Position 3 handles the painting of custom-ordered markings, such as

signature stripes and corporate logos, on the airplane. The work done in this position is labor-intensive due to the necessity of frequently applying and then removing masking tape. Each of these first three positions involves work done in either of two parallel floor spaces within this building. Position 4 handles final detailing, cleaning and varnishing, and painting the airplane door and its frame. This basic work flow is shown in Figure 1, Appendix.

3. OBJECTIVES DEFINITION AND MODEL DEVELOPMENT

3.1 Setting Objectives and Scope

The project charter specified that the consultants (1) examine the overall process flow to determine the maximum number of planes per day (two? three?) given the current painting facility “footprint” (overall square meters and building cross-section) as a binding constraint, and (2) use simulation and allied techniques to suggest revisions to the painting process to achieve that maximum. Value stream mapping and time studies, conducted before the simulation model-building effort began, soon convinced both the consultants and the client managers that “two planes per day” would plausibly be achievable but “three planes per day” would not be. Given this firm and well-defined foundation for the simulation portion of the study, the consultants undertook the design and construction of the base-case simulation model. Much of the input data needed for this model, such as cycle times, worker requirements, buffer capacities, and transfer times for airplanes between workstations, had just been collected during the value stream mapping and time studies. Indeed, the “double use” of these data is one of many strong justifications for using simulation synergistically with other industrial engineering analysis methods (Chung 2004). All additional data needed was collected during a two-month period whose final two weeks coincided with the base model development described in the next section. As the construction of the base case model began, client and consultant engineers brainstormed promising modifications of the current system.

3.2 Choice of Software

The clients and the consultants concurred on the use of the WITNESS® simulation software for model development. This software provides convenient high-quality animation, logical support for both “pull” and “push” operational logic, the ability to build reusable sub-models, and a powerful “labor” construct capable of modeling operationally complex rules for the deployment and transit of both laborers and portable pieces of equipment (Mehta and Rawles 1999). A small, vivid, and typical example of WITNESS® flexibility appears in the following “output rule” (a rule specifying whether, to where, and when a machine

sends an entity [here, an airplane] which has just finished processing at that machine:

```
IF vPaint_02_Done = 0
  PUSH to PAINT_02_1
ELSE
  Wait
ENDIF
```

This output rule relies on the current value of the variable vPaint_02 to decide whether to send the airplane downstream (in this case, to machine PAINT_02_1) or to hold the airplane at its current location until the variable becomes equal to zero.

WITNESS® also provides automatic collection and graphical display of system metrics such as minimum, average, and maximum queue lengths, number of cycles undertaken by each machine, utilization of each labor resource, and total entities throughput.

The animation layout constructed within the WITNESS® simulation model is shown in Figure 2 in the Appendix.

3.3 Choice of Stochastic Distributions

Arrival of WITNESS® “parts” (planes) to the model was based on historical records of planes leaving the upstream operation. Historical time-to-fail (or number-of-cycles-to-fail) and time-to-repair data were entered into a distribution fitter, ExpertFit® (Law and McComas 2003) to determine suitable closed-form distributions (if indeed, such existed) using techniques such as Kolmogorov-Smirnov and Anderson-Darling goodness of fit tests for maximum-likelihood estimators (Leemis 2004). As examples of these data, the paint booths routinely require a filter change every thirty airplanes on average, with out-of-service time averaging eight hours. Similarly, the preparation booths routinely require a filter change every twenty airplanes on average, with out-of-service time averaging four hours. Routine preventative maintenance lasting four hours on average is done at the paint booths weekly. Major equipment breakdowns, lasting an average of three days, occur on average every three months at paint booths and once a year at preparation booths. With few exceptions, times-to-fail were modeled with exponential or Weibull distributions, and times-to-repair were modeled with gamma (of which the Erlang is a special case), Weibull, or log-normal distributions.

3.4 Model Development Timing

Setting of the objectives and construction of the base case (reflective of the current system) model required two calendar weeks and three person-weeks. During those two weeks one simulation analyst worked on the model full time and another contributed additional work on the model part time.

4. MODEL VERIFICATION AND VALIDATION

4.1 Documentation of Assumptions

As data collection efforts drew to a close, the clients and the analysts agreed upon and documented the following assumptions pertinent to building the model of the base case system:

1. Planes are always available from upstream to be painted (consistent with the long-standing recognition that the paint shop was the bottleneck *blocking* upstream processes).
2. No downstream blocking occurs relative to planes leaving the painting operations (consistent with the long-standing recognition that the paint shop was the bottleneck *starving* downstream processes).
3. Labor resources are not the constraint (consistent with anecdotal evidence, and also with the observation that – contrary to many manufacturing contexts – in this context, capital equipment is more expensive and harder to obtain than the relatively unskilled labor needed for various operations [e.g., the application and removal of masking tape mentioned above]).
4. Equipment preventive maintenance and unscheduled downtime data are still valid as provided from historical data.

4.2 Verification, Validation, and Credibility

Early in the project, even the most casual observations of the painting process convinced both clients and consultants that the system was conceptually steady-state (indeed, some queues were *never* observed empty). As initial settings for verification and validation of the base model, warm-up time was set to one month and run time (with gathering of performance statistics) to one year. Typical techniques were then used for model verification and validation. As a fundamental basis for initial high-level verification and validation, the “observed” versus “estimated” collective cycle times for each of the four positions (shown in Figure 1, Appendix) were examined for reasonably close agreement. These methods included running the model with all variability eliminated for easy checking against spreadsheet computations, running one entity through the model. Structured walkthroughs held by the two modelers and their technical leader, careful examination of the animation, extreme condition tests, and discussion of plausibility of preliminary results with the client’s process engineers (including Turing tests) all proved useful to the tasks of verification and validation (Sargent 2004). After routine errors (e.g., mismatched variable names) were found and corrected, the analysts graphed performance metrics of the base model against simulated time. These graphs demonstrated that accurate determination of performance metrics, with sufficiently narrow 95% confidence intervals required increasing the warm-up length to two months and the run length to two years of

simulated time, with 10 replications for each situation to be examined. Note that even with a two-year statistics-gathering run length, on average only two major equipment breakdowns will occur at preparation booths. The usual analytical recommendation is that the most unusual event in a system be expected to occur five or six times during each replication (Law 2004). As a countermeasure, with replication length already two years, the analysts checked that several different but representative numbers of these breakdowns occurred among the replications – an approach conceptually akin to stratified sampling. Next, the model achieved credibility with the client engineers and managers by predicting currently observed performance metrics within 4%.

5. RESULTS AND IMPROVEMENTS

In agreement with current observation, the base case model indicated average production of 1.45 planes per day – and also correctly indicated severe blocking (19% of time each paint booth blocked) just downstream from both paint booths (the key operations in Position 2) and hence just upstream from the detailing operation of Position 3. Meanwhile, results of layout analysis had suggested workflow enhancements, *not* involving capital expenditure, having the potential to create buffer space for at least one plane, maybe two, between the pair of paint booths and the detailing operation. Accordingly, the first two alternative scenarios modeled introduced an as yet hypothetical buffer at this point. Introducing this buffer into the model required less than ½ day of modeling time. Setting the buffer capacity to 1 yielded average production of 1.64 planes per day, with paint booth blocked time reduced from 19% to 9%. Increasing the buffer capacity to 2 yielded average production of 1.75 planes per day, with paint booth blocked time further reduced to 4%.

Next, the collaborating engineers (consultants and clients) turned their attention to the possibility of adding a second paint-detailing station within Position 3. This modification to the model was similarly added, verified, and validated for reasonableness of results within one day. However, its results proved disappointing, especially considering that a second detailing station represented significant capital and operating expense. Indeed, this addition did reduce blocked time at both upstream booths to less than 2%. However, the key performance metric “average planes per day” increased to only 1.81 from 1.75.

6. CONCLUSIONS

The client’s engineers promptly implemented the workflow enhancements suggested by the layout analysis, and concurrently created and used a buffer of capacity 2 between the paint booths and detailing operations. The key performance metric “average planes per day” promptly increased from 1.45 to 1.74, an increase of 20% with no capital investment required. Blocked time at the paint booths also decreased as

predicted by the simulation study. Although most welcome, this throughput increase fell short of the “two planes per day” aspirations. Therefore, client and consultant engineers agreed upon follow-up studies, now in progress. These studies are investigating these throughput improvement opportunities:

Standardization of various operations to minimize variability of time required.

Workplace organization and visual controls, partly to manage inventories of paint and partly to minimize wasted time (“muda”) searching for tools.

Development of templates for setup of striping operations (part of detailing) to minimize detailing time; this suggestion came from a client engineer familiar with the practice of “SMED” [Single Minute Exchange of Die] as practiced in many manufacturing industries and pioneered by the Japanese engineer Shigeo Shingo (Collier and Evans 2007).

ACKNOWLEDGMENT

The authors gratefully acknowledge the leadership and help provided by colleague Ravi Lote on this industrial engineering project.

Additionally, constructive criticisms from two anonymous referees have provided the authors significant help in improving the paper.

REFERENCES

- Ambrosino, Daniela, and Elena Tànfani. 2009. A Discrete Event Simulation Model for the Analysis of Critical Factors in the Expansion Plan of a Marine Container Terminal. In *Proceedings of the 23rd European Conference on Modelling and Simulation*, eds. Javier Otamendi, Andrzej Bargiela, José Luis Montes, and Luis Miguel Doncel Pedrera, 288-294.
- Chung, Christopher A. 2004. *Simulation Modeling Handbook: A Practical Approach*. Boca Raton, Florida: CRC Press.
- Collier, David A. and James R. Evans. 2007. *Operations Management: Goods, Services and Value Chains*. Mason, Ohio: Thomson South-Western.
- El Wakil, Sherif D. 2002. *Processes and Design for Manufacturing*, 2nd edition. Long Grove, Illinois: Waveland Press, Incorporated.
- Harrell, Charles, and Kerim Tumay. 1995. *Simulation Made Easy: A Manager's Guide*. Norcross, Georgia: Engineering & Management Press.
- Heragu, Sunderesh S. 2008. *Facilities Design*, 3rd edition. Boca Raton, Florida: CRC Press.
- Law, Averill, and Michael G. McComas. 2003. How the ExpertFit Distribution-Fitting Software Can Make Your Simulation Models More Valid. In *Proceedings of the 2003 Winter Simulation Conference*, Volume 1, eds. Stephen E. Chick, Paul J. Sánchez, David Ferrin, and Douglas J. Morrice, 169-174.
- Law, Averill M. 2004. Statistical Analysis of Simulation Output Data: The Practical State of the Art. In *Proceedings of the 2004 Winter Simulation Conference*, Volume 1, eds. Ricki G. Ingalls, Manuel D. Rossetti, Jeffrey S. Smith, and Brett A. Peters, 67-72.

- Leemis, Lawrence M. 2004. Building Credible Input Models. In *Proceedings of the 2004 Winter Simulation Conference*, Volume 1, eds. Ricki G. Ingalls, Manuel D. Rossetti, Jeffrey S. Smith, and Brett A. Peters, 29-40.
- Lian, Yang-Hua, and Hendrik Van Landeghem. 2002. An Application of Simulation and Value Stream Mapping in Lean Manufacturing. In *Proceedings of the 14th European Simulation Symposium*, eds. Alexander Verbraeck and Wilfried Krug, 300-307.
- McCartney, Scott. 2011. Lagging Private-Jet Industry Resumes Takeoff. *Wall Street Journal* CCLVII:33(D5).
- Mehta, Arvind, and Ian Rawles. 1999. Business Solutions Using WITNESS. In *Proceedings of the 1999 Winter Simulation Conference*, Volume 1, eds. Phillip A. Farrington, Harriet Black Nembhard, David T. Sturrock, and Gerald W. Evans, 230-233.
- Miller, Scott, and Dennis Pegden. 2000. Introduction to Manufacturing Simulation. In *Proceedings of the 2000 Winter Simulation Conference*, Volume 1, eds. Jeffrey A. Joines, Russell R. Barton, Keebom Kang, and Paul A. Fishwick, 63-66.
- Orsoni, Alessandra, and Nikolaos V. Karadimas. 2006. The Role of Modelling and Simulation in Design-Build Projects. In *Proceedings of the 20th European Conference on Modelling and Simulation*, eds. Wolfgang Borutzky, Alessandra Orsoni, and Richard Zobel, 315-320.
- Sargent, Robert G. 2004. Validation and Verification of Simulation Models. In *Proceedings of the 2004 Winter Simulation Conference*, Volume 1, eds. Ricki G. Ingalls, Manuel D. Rossetti, Jeffrey S. Smith, and Brett A. Peters, 17-28.

AUTHOR BIOGRAPHIES

MARCELO ZOTTOLO, born in Buenos Aires, Argentina, came to the United States to finish his college studies. He was graduated from the University of Michigan - Dearborn as an Industrial and Systems Engineer in December 2000, and subsequently earned his master's degree in the same field in June 2004. He was awarded the Class Honors distinction and his Senior Design Project was nominated for the Senior Design Competition 2001. This project studied the improvement of manufacturing processes for the fabrication of automotive wire harnesses, ultimately proposing an automation tool leading to improvements in future designs. Additionally, he was co-author of a paper on simulation in a distribution system which earned a “best paper” award at the Harbour, Maritime, and Simulation Logistics conference held in Marseille, France, in 2001. He is now a Consulting Project Manager at PMC with a solid background in data-driven process improvement methodologies to optimize the performance of different systems. He is experienced in the concurrent application of lean thinking, theory of constraints, workflow measurement, and simulation modeling across multiple economic sectors including retail, service, healthcare, insurance, and manufacturing. His responsibilities include leading teams in building, verifying, validating, and analyzing simulation models in Enterprise Dynamics®,

WITNESS®, ProModel®, and SIMUL8® for large corporate clients; he also presents in-house training seminars. His email address is mzottolo@pmcorp.com

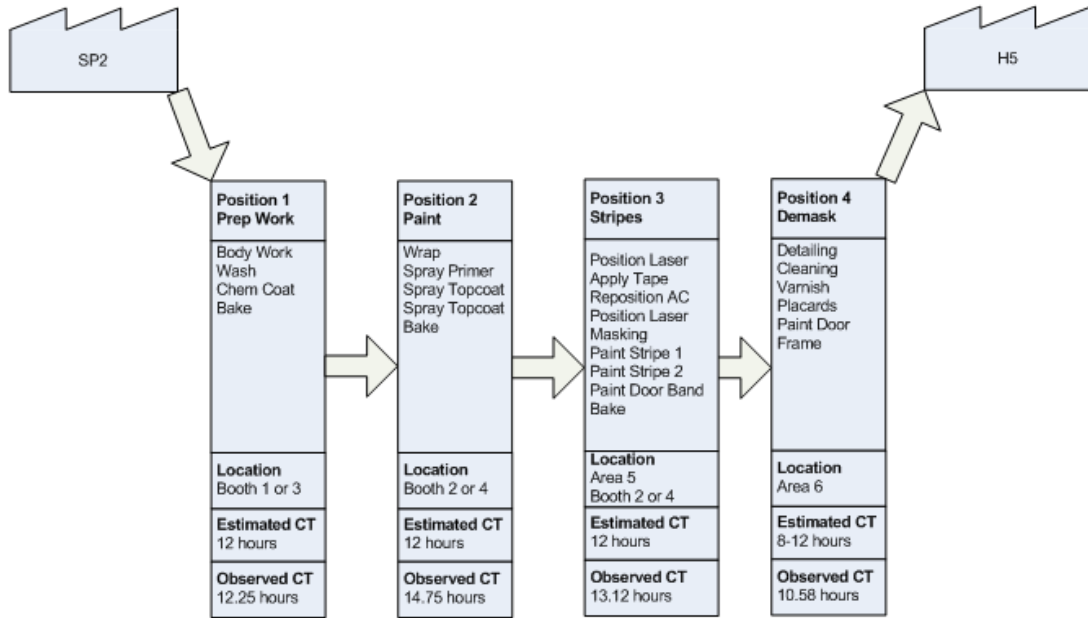
EDWARD J. WILLIAMS holds bachelor's and master's degrees in mathematics (Michigan State University, 1967; University of Wisconsin, 1968). From 1969 to 1971, he did statistical programming and analysis of biomedical data at Walter Reed Army Hospital, Washington, D.C. He joined Ford Motor Company in 1972, where he worked until retirement in December 2001 as a computer software analyst supporting statistical and simulation software. After retirement from Ford, he joined PMC, Dearborn, Michigan, as a senior simulation analyst. Also, since 1980, he has taught classes at the University of Michigan, including both undergraduate and graduate simulation classes using GPSS/H™, SLAM II™, SIMAN™, ProModel®, SIMUL8®, or Arena®. He is a member of the Institute of Industrial Engineers [IIE], the Society for Computer Simulation International [SCS], and the Michigan Simulation Users Group [MSUG]. He serves on the editorial board of the International Journal of Industrial Engineering – Applications and Practice. During the last several years, he has given invited plenary addresses on simulation and statistics at conferences in Monterrey, México; İstanbul, Turkey; Genova, Italy; Rīga, Latvia; and Jyväskylä, Finland. He served as a co-editor of Proceedings of the International Workshop on Harbour, Maritime and Multimodal Logistics Modelling & Simulation 2003, a conference held in Rīga, Latvia. Likewise, he served the Summer Computer Simulation Conferences of 2004, 2005, and 2006 as Proceedings co-editor. He is the Simulation Applications track co-ordinator for the 2011 Winter Simulation Conference. His email address is ewilliams@pmcorp.com.

ONUR M. ÜLGEN is the president and founder of Production Modeling Corporation (PMC), a Dearborn, Michigan, based industrial engineering and software services company as well as a Professor of Industrial and Manufacturing Systems Engineering at the University of Michigan-Dearborn. He received his Ph.D. degree in Industrial Engineering from Texas Tech University in 1979. His present consulting and research interests include simulation and scheduling applications, applications of lean techniques in manufacturing and service industries, supply chain optimization, and product portfolio management. He has published or presented more than 100 papers in his consulting and research areas.

Under his leadership PMC has grown to be the largest independent productivity services company in North America in the use of industrial and operations engineering tools in an integrated fashion. PMC has successfully completed more than 3000 productivity improvement projects for different size companies including General Motors, Ford, DaimlerChrysler, Sara

Lee, Johnson Controls, and Whirlpool. The scientific and professional societies of which he is a member include American Production and Inventory Control Society (APICS) and Institute of Industrial Engineers (IIE). He is also a founding member of the MSUG (Michigan Simulation User Group).

APPENDIX



TOTAL LEAD TIME = 50.7 HOURS

Figure 1: Overview of Work Flow

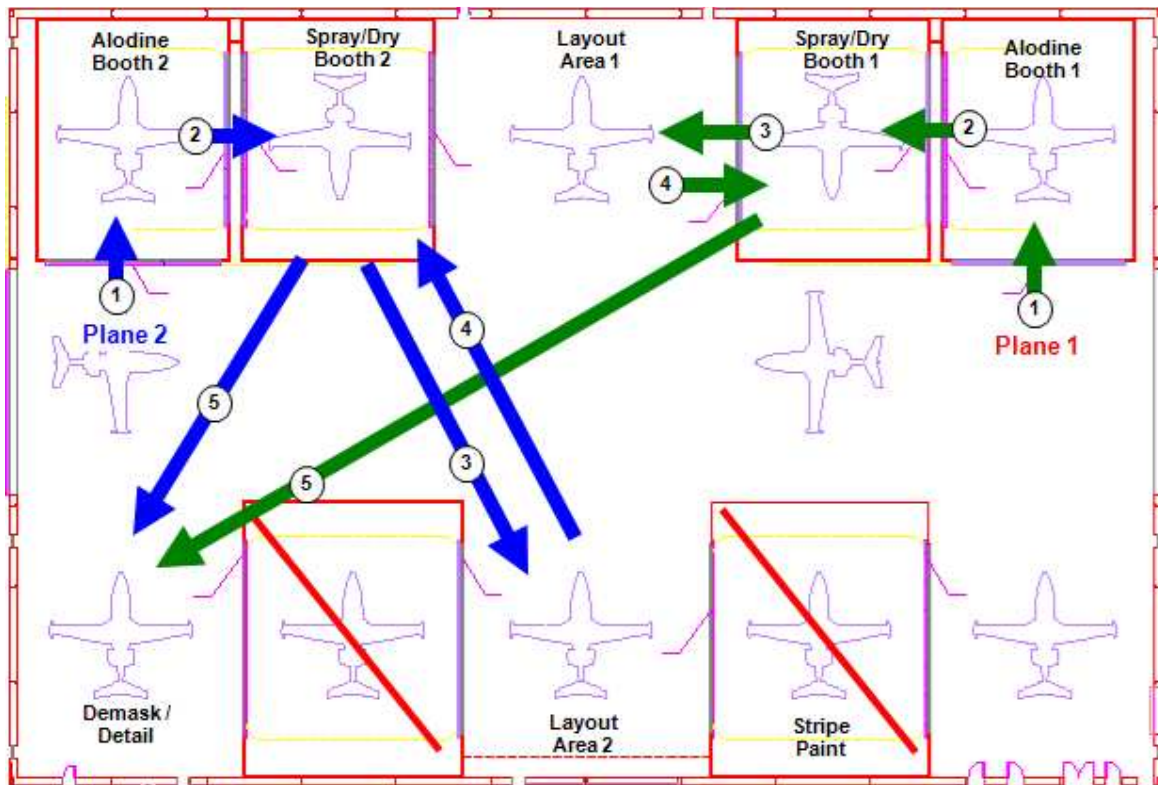


Figure 2: Abstraction of Work Flow Used in WITNESS® Simulation/Animation

CALIBRATION OF PROCESS ALGEBRA MODELS OF DISCRETELY OBSERVED STOCHASTIC BIOCHEMICAL SYSTEMS

Paola Lecca

The Microsoft Research – University of Trento
Centre for Computational and Systems Biology
Piazza Mancini 17, 38123 Povo (Trento), Italy

lecca@cosbi.eu

ABSTRACT

We present a maximum likelihood method for inferring kinetics of stochastic systems of chemical reactions, given discrete time-course observations of the abundance of either some or all of the molecular species and a BlenX model of the system. BlenX is a process calculus providing a tool and algebraic laws for a high-level description of interactions, communications, and synchronizations between processes representing the biomolecules. BlenX offers an efficient alternative to differential equations, but it poses different challenges to the model calibration. The main difficulty is the sampling of the reaction pathways between two observed states. We define a maximum likelihood function in terms of reaction propensities and we estimate it by sampling the intermediate pathways from the transition system of a BlenX. The method of sampling the transition system is inspired to the elementary mode analysis. Our method is illustrated with the example of a BlenX model of chaperone-assisted protein folding.

Keywords: BlenX, parameter estimation, maximum likelihood estimation.

1. INTRODUCTION

Modelling the time evolution of biological systems requires the specification of the interactions among the biochemical species and the kinetic parameters of these interactions. The choice of a language able to express the main features of biological systems and the methods for estimating the model parameters (model calibration) are two interdependent. In this article, we introduce the BlenX language (Dematté, et al., 2008 (a) and (b)) for modeling biological processes and a way to calibrate a BlenX-specified model using discrete time-course observations of either some or all of the molecular species.

The majority of the models of inter- and intra-cellular dynamics are specified in ordinary differential equations. These equations are usually employed by physicist to formalize the natural laws and to describe the dynamics of the inert matter. However, the recent achievement of the systems biology paradigm

highlights the need of a new systematic approach to modeling systems belonging to living matter. This need can be translated into the necessity to develop new mathematical methods and tools to model living systems, considering that these systems require mathematical and computational approaches substantially different from those used to model inert matter. Multifunctionality of biochemical complexes, parallelism and concurrency of their interactions and modular structure of the network of interactions are the main features characterizing a biological system. Process algebras (or process calculi) are currently proposing as suitable formalisms for the specification of a biological process (some examples are: stochastic π -calculus (Priami, 1995), BioAmbients (Regev, et al., 2004), Brane Calculi (Cardelli, 2005), CCS-R (Danos & Krivine, 2004), k-calculus (Danos & Laneve, 2004), PEPA (Gilmore & Hillstone, 1994)). Usually, these algebras are applied to the study of concurrent processes. The tools of the process algebras are algebraic languages for the specification of processes and the formulation of statements about them, together with calculi for the verification of these statements. Process calculi provide a tool for the high-level description of interactions, communications, and synchronizations between a collection of independent agents or processes. The use of these calculi in modelling biological system is based on a new abstraction of the physical concepts of interacting molecules, interactions and change of state. Interacting molecules are represented by processes. Interactions between molecules are represented by synchronized communication. The change of state consequent to an interaction is described by the modification that processes undergo after the realization of their communication.

In this article we focus on a new member of the family of process algebra: the BlenX language. It has been developed by our lab (CoSBI, 2011) to extend the expression capabilities of the stochastic π -calculus. As in stochastic π -calculus, also in BlenX models consist of agents (processes) which stochastically engage actions. However, with BlenX we can describe more easily spatial structures like membranes, compartments,

interaction domains, and the formation of biological complexes driven by chemical/physical affinities. Currently, BlenX can be used to specify continuous time stochastic systems. In fact, the stochastic simulation algorithm of the BlenX simulator is an efficient variant of the Gillespie algorithm (Gillespie, 1977). In Gillespie-like approaches every reaction is explicitly simulated. When simulated, a Gillespie realization represents a random walk that exactly provides the distribution of the Chemical Master Equation.

The communications representing the physico-chemical interactions between the biological entities are associated to a rate constant that quantifies the specific speed of the communication and reflects the kinetic rate constant (and/or the affinity) of the reaction. This rate constant is the parameter of an exponential probability distribution of the waiting time of reactions. Thus, the waiting time of a reaction is a realization of a random variable exponentially distributed with parameter equal to the rate constant of the reaction. The stochastic simulation algorithm generates random numbers to determine the next reaction to occur as well as the time at which the reaction occurs. The time evolution of the system proceeds by jumps from one state to another. The state of the system at time t is given by the number of molecules (or more generally of biological - biochemical entities) of each species included in the system at that time. Therefore, first of all, the calibration of a BlenX model is the calibration of a stochastic model.

The existent method for parameter inference in stochastic system belongs to one of the following two categories: maximum-likelihood based approach and Bayesian inference approach. A comprehensive review of these methods can be found in the introduction of recent work of Y. Wang et al. (Wang, Christley, Mjolsness, & Xie, 2010), R. J. Boys et al. (Boys, Wilkinson, & Kirkwood, 2008) and in P. Lecca et al. (Lecca, Palmisano, Ihekwaba, & Priami, 2010). Here we focus on maximum likelihood (ML) approaches, following our previous studies on ML inference models (Lecca, Palmisano, Ihekwaba, & Priami, 2010).

Most proposed methods for parameter inference in stochastic biochemical models consider how to calculate the maximum likelihood for the rate parameter values given a stochastic model and discrete experimental data of the amount of molecules of all or only of some species. Since for biological systems of realistic size and complexity, the likelihood function is computationally intractable, these methods either perform exact inference on an approximated model where the likelihood computation is tractable, or they approximate the likelihood with a more tractable function, or some combination of the two. In this paper we refer to a method of parameter estimation presented in (Wang, Christley, Mjolsness, & Xie, 2010) and we show how it can be adapted to estimate the kinetic parameters of a BlenX model. In particular, we show how the transition system of a BlenX model can be

sampled to calculate the maximum likelihood function without any need of simulating the model. The paper proceeds as follows: in Section 2 we introduce the reader to the BlenX language and present the model on which we show the performance of the inference method; in Section 3 we present the method and the results, and finally in Section 4 we give some conclusions.

2. THE BLENX LANGUAGE: AN OVERVIEW

BlenX is a process algebra-based stochastic programming language that shares features with stochastic π -calculus (Priami 1995) and Beta-binders (PQP). BlenX, as these other members of the family of process algebra-based languages, has a strong focus on the interactions of entities. BlenX is explicitly designed to model the interactions of biological entities such as proteins and other biochemical species. It is a stochastic language in the sense that the probability and speed of the interactions and actions governing the time evolution of the system are specified in the body of the programs written in this language.

In BlenX, each species is given with an abstract entity that we call a *box*. Each box has a number of connectivity interfaces called *binders*, and it is equipped with an internal program. The sites of interaction are represented as binders on the box surface. For example in Figure 1, each box has only one binder. Binders are identified by their names, e.g., x and their types, e.g., A .

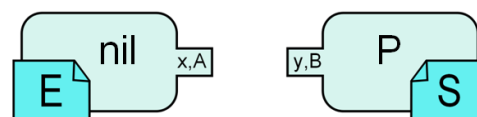


Figure 1: an enzyme E and its substrate S can be represented by boxes equipped with interaction sites on the interfaces, i. e. the binders (x, A) and (y, B), and an internal processes, e.g. the deadlock process *nil* and a process *P*, respectively.

A box can stochastically interact with another box, and change state as a result of this interaction with respect to the actions specified in its internal program. Alternatively, a box can autonomously change state by stochastically performing an action that is given in its internal program. For instance, the complexation of an enzyme E and a substrate S can be described in a BlenX model with the boxes depicted in Figure 1, where these boxes interact and bind with their binders. Then the interaction rate, specified in the BlenX code, determines the rate of the association. The internal program, which can be *nil* as it is the case for E here, determines the actions the box can undertake after this interaction.

The *nil* process does nothing (it is the deadlock process). Other stochastic actions that a BlenX box can perform are summarized as follows: a box can

- communicate with another box that is bound to it (or with itself) by performing

- an input action, e.g., `x?(message)` that is complementary to the output action, e.g., `x!(message)`, of the other box, or vice versa; and this way send or receive a message;
- perform a stochastic `delay` action;
- change (`ch`) the type of one of its interfaces;
- eliminate itself by performing a `die` action;
- expose a new binder;
- hide one of its binders;
- unhide a binder which is hidden.

In addition to these actions, there are also other programming constructs available such as if-then statements and state-checks. For example, let us consider the box *S* in Figure 1. We can program this box by defining program *P* so that it will change its type from *B* to *C* if it is bound:

```
if (y,B) and (y,bound) then ch(y,C) endif
```

In BlenX, following the process algebra tradition, we can compose actions by using algebraic composition operators to define increasingly complex behaviors. We can sequentially compose actions by resorting to the prefix-operator, which is written as an infix dot. For instance,

```
ch(y,C).hide(y).nil
```

denotes a program that first performs change action and then hides the changed binder.

Programs can be composed in parallel. Parallel composition (denoted by the infix operator “|”, for instance $P|Q$), allows the description of programs, which may run independently in parallel and also synchronize on *complementary actions* (i.e., *input* and *output* over the same channel).

The `rep` operator replicates copies of the process passed as argument. Only guarded replication is used, i. e. the process argument of this operator must be prefixed by an action that forbids any other action of the process until the first action has been executed.

Programs can also be composed by *stochastic choice*, denoted with the summation operator “+”. The sum of processes *P* and *Q*, $P + Q$ behaves either as *P* or as *Q*, determined by specific speed (i. e. rate constants) defined for *P* and *Q*. The selection of one discards the other forever.

In BlenX, we use `events`, which are programming constructs for expressing actions that are enabled by global conditions. For example, in the model presented here, we use the `new` construct to introduce new molecules of a species if their amount reaches a minimum threshold. For instance in:

```
when (protein : |protein| < 10000 : r)
new(500);
```

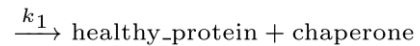
the amount of species `protein` is increased by 500 units when it becomes less 10,000 units. The increasing rate is `r`.

2.1. The case study: chaperone-assisted protein folding

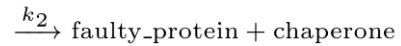
In this section we briefly describe the main mechanisms of chaperone-assisted protein folding and present the code of a BlenX model of this process. A more detailed descriptions of the biological processes and of the BlenX specification can be found in (Lecca P., 2011).

The ability of the cell to handle misfolded proteins is expressed by some complexes of macromolecules, called *chaperones*. Molecular chaperones interact with unfolded or partially folded protein subunits, e.g. nascent chains emerging from the ribosome, or extended chains being translocated across sub-cellular membranes. They prevent inappropriate association or aggregation of exposed hydrophobic surfaces and direct their substrates into productive folding, transport or degradation pathways. In the healthy cells, if a protein does not assume the correct 3D shape, or a cellular stress induces a right-folded protein to assume a wrong folding, the chaperones act to re-shape it correctly.

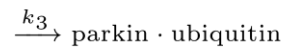
R_1 : protein + chaperone



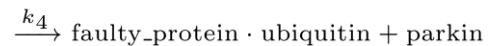
R_2 : protein + chaperone



R_3 : parkin + ubiquitin



R_4 : faulty_protein + parkin · ubiquitin



R_5 : faulty_protein · ubiquitin + proteasome

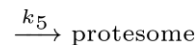


Figure 2: interactions between protein and chaperone, and between faulty protein and ubiquitin-proteasome system.

In the case in which the protein is still not correctly refolded, the cellular ubiquitin-proteasome targets and degrades it before the faulty protein can cause damages. The protein parkin mediates the targeting of misfolded proteins for degradation by moving the molecules of ubiquitin on these proteins. The proteasome machinery recognizes the ubiquitinated proteins and degrades

them. The set of reactions driving the dynamics of the model is reported in Figure 2.

In the following, we report the BlenX code describing the main interaction among the component of the systems: the nascent protein, the chaperone, the ubiquitin, the parkin and the proteasome, all represented by boxes. The parameters of the model are specified in the body of the code and are bold-faced and grey highlighted. The rate parameters used in this model and the initial amounts of molecular species are expressed in arbitrary units, and qualitatively reproduce the real dynamic observed dynamics.

The Figure 3 depicts the boxes representing each entity involved in the systems and the interactions on dedicated binders.

```

1 [time=200] // absolute simulation time
2
3 <<BASERATE: 100>> // basal rate
4
5 // A nascent protein interacts with a
6 // molecular chaperone
7
8 let protein : bproc =
9   #(y:100, P), #(ubi:100, U), 10
10  #(prot:100000,PTSP)
11  [
12    [
13      y?().ch(1000,y,DR).hide(1,ubi).nil
14      + y?().ch(1000,y,DW).ubi?().
15      prot?().die(1).nil
16    ]

```

```

17 // Definition of chaperone
18
19 let chaperone : bproc = #(x:100, C)
20 [rep x!().nil];
21
22 // Definition of parkin bioprocess
23
24 let parkin : bproc =
25   #(to_ubiquitin:0.5, T_UB)
26   [to_ubiquitin!().nil];
27
28 let ubiquitin : bproc =
29   #h(u:1, UB), #(actp:1, UB2),
30   #(from_parkin:0.5, F_PARK)
31   [
32     from_parkin?().unhide(0.5,u).u!().actp!()
33   ];
34
35 let proteasome : bproc =
36   #h(pt:50000, PTS), #(actv:500000, ACT),
37   [
38     delay(0.1).(actv?().unhide(600000,pt).
39     pt!().nil)
40   ];
41
42 // Production of ubiquitin, parkin and nascent
43 // proteins
44
45 when (ubiquitin: |ubiquitin| = 0 : inf) new(10000);
46 when (parkin: |parkin| = 0 : inf) new(10000);
47 when (protein :: protein_prod) new(1);
48
49 // Initial amount of components
50 run 5000 protein || 10000 parkin ||
51 10000 chaperone || 10000 ubiquitin ||
52 10000 proteasome

```

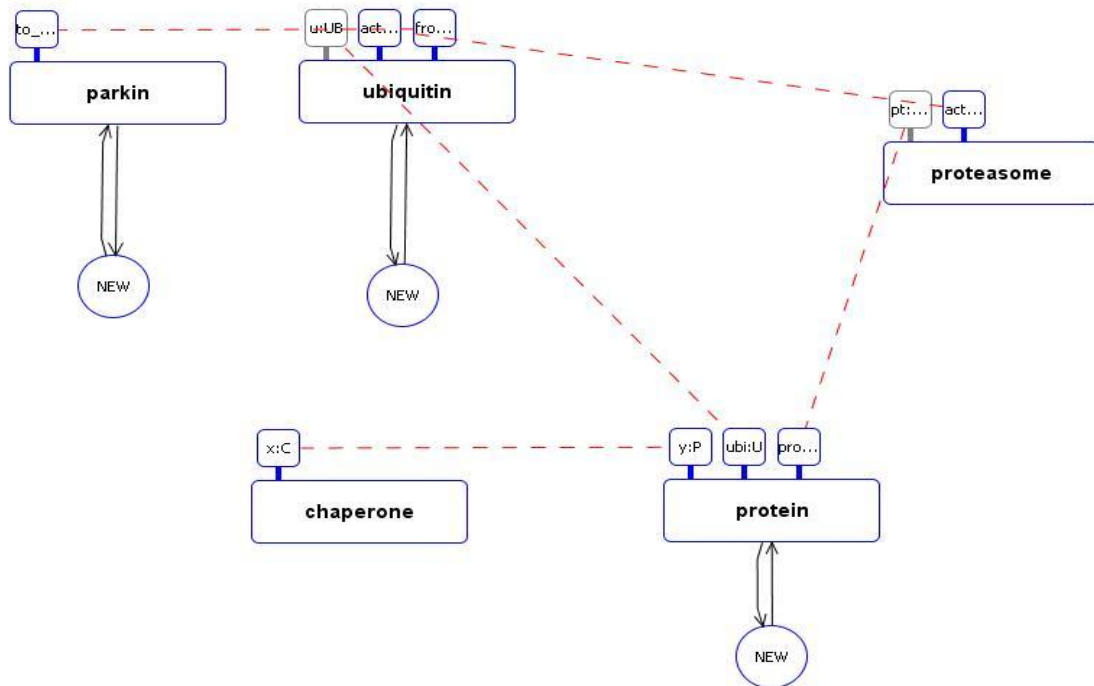


Figure 3: picture representing the boxes of protein, chaperone, parkin, ubiquitin and proteasome. A dotted red line connect the interacting boxes through the communication binders. The model also consider the production of protein, parkin and ubiquiting molecules represented with the circle “NEW”.

The structured operational (interleaving) semantics of the language is used to generate a labelled transition system. A state transition system is an abstract machine consisting of a set of states and transitions between states. The state transition system of the model presented in this paper is shown in Figure 4.

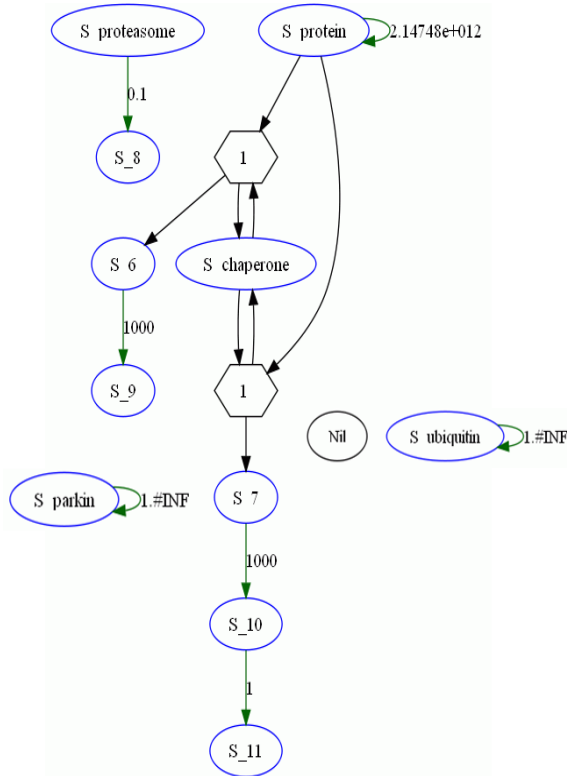


Figure 4: graph of the state transition system of the BlenX model of chaperone-assisted protein folding and protein ubiquitination. “S” stays for species. All the intermediate species are showed and numbered. The labels on the arrows indicate the rate constant values. Where no label is specified, it is assumed to be equal to the basal rate constant defined in the BlenX code.

3. ML-BASED PARAMETER INFERENCE

In this section we first review the key-points of method developed by Y. Wang et al. for the estimation of parameters of stochastic systems. Then, we show how these ideas can be adapted and used for the estimation of the rate constant of a BlenX model from experimental measurements of the amount of molecules at discrete time points.

3.1. The likelihood function

Our goal is to estimate the rate parameters of a stochastic Markov process algebra based on the observations at a set of discrete time points. In this

section we review the methods of Y. Wang et al. (Wang, Christley, Mjolsness, & Xie, 2010).

Suppose we have the vector of the observations

$$\mathbf{X}(t) \equiv \{\hat{X}_\Gamma(t_1), \hat{X}_\Gamma(t_2), \dots, \hat{X}_\Gamma(t_m)\} \quad (3.1.1)$$

of the system at m discrete time points $\{t_1, t_2, \dots, t_m\}$ for a subset of species $\Gamma \subseteq \{1, \dots, N\}$. Denoting the likelihood of the observations for a given set of rate parameters by

$$L(\hat{X}_\Gamma(t_1), \hat{X}_\Gamma(t_2), \dots, \hat{X}_\Gamma(t_m); \Theta),$$

we estimate the rate parameters by maximizing the likelihood function with respect to the parameters.

Suppose that the reaction system involves M reactions, R_1, \dots, R_M . Denote with $X = (x_1, \dots, x_n)$ the state vector of the system. Each reaction has an associated propensity, also called *hazard function*, $h_i(X, \Theta)$. Θ is the set of rate parameters associated with the reactions. The hazard function determines the rate of the transition probability out of state X due to the reaction of type i . For convenience, as in (Wang, Christley, Mjolsness, & Xie, 2010), we adopt the following compact representation of a reaction system:



where $\mathcal{U} = [u_{ij}]$ are $M \times N$ stoichiometry matrices. It is useful to introduce also the *net effect reaction matrix*

$$\mathcal{A} = \mathcal{U} - \mathcal{V} \quad (3.1.2)$$

which reports the net change of species numbers associated with a reaction.

Denote $P(X, t)$ the probability of the system in state X at time t . For a time increment Δt , $P(X, t + \Delta t)$ can be written as the sum of probabilities of the number of ways in which the system can reach or leave the current state:

$$P(X, t + \Delta t) = \sum_{i=1}^M \left[h_i(X - A_i, \Theta) P(X - A_i, t) \Delta t + \left(1 - \sum_{i=1}^M h_i(X, \Theta) \Delta t \right) \right] P(X, t) \quad (3.1.3)$$

where A_i is the i -th row of the matrix \mathcal{A} . In the limit of $\Delta t \rightarrow 0$, Eq. (3.1.3) becomes

$$\frac{d}{dt}P(X, t) = \sum_{X'} \sum_{i=1}^M [h_i(X', \Theta) \delta_{X', X-A_i} - h_i(X, \Theta) \delta_{X', X}] \quad (3.1.4)$$

where $\delta_{X', X}$ is the Kronecker delta function. For our convenience we introduce $H_{X', X}$ as follows:

$$H_{X', X} = \sum_{i=1}^M [h_i(X', \Theta) \delta_{X', X-A_i} - h_i(X, \Theta) \delta_{X', X}] \quad (3.1.5)$$

For simplicity, consider a single time interval $[t_k, t_{k+1}]$ between two measurements of the abundance of the species j ($j = 1, \dots, N$), $\hat{X}_j(t_k)$ and $\hat{X}_j(t_{k+1})$. We discretize this time interval in K subintervals and denote the system state by $\{X^\nu | \nu = 1, 2, \dots, K\}$. Therefore $\hat{X}^0 \equiv \hat{X}(t_k)$ and $\hat{X}^K \equiv \hat{X}(t_{k+1})$ are the full observations available at the start and at the end of this time interval. All X^ν are the intermediate states not directly observable. Using the Markov property of the stochastic process, the likelihood of observing \hat{X}_j^0 and \hat{X}_j^K under a model with parameters Θ is

$$L(\hat{X}_j^0, \hat{X}_j^K; \Theta) = \sum_{\hat{X}_j^1, \dots, \hat{X}_j^{K-1}} P(\hat{X}_j^0) \prod_{\nu=0}^{K-1} P(\hat{X}_j^{\nu+1} | \hat{X}_j^\nu; \Theta) \quad (3.1.6)$$

If $K \gg 1$, then

$$P(\hat{X}_j^{\nu+1} | \hat{X}_j^\nu; \Theta) \approx \delta_{\hat{X}_j^\nu, \hat{X}_j^{\nu+1}} + \frac{1}{K} H_{\hat{X}_j^\nu, \hat{X}_j^{\nu+1}}(t_{s+1} - t_s) \quad (3.1.7)$$

If we choose K equal to the number of reactions occurring in the time interval $[t_s, t_{s+1}]$, $H_{\hat{X}_j^\nu, \hat{X}_j^{\nu+1}}$ can be expressed as follows

$$H_{X^\nu, X^{\nu+1}} = h_{R_\nu}(X^\nu, \Theta) \delta_{X^\nu, X^\nu - A_\nu} - h_{R_\nu}(X^{\nu+1}, \Theta) \delta_{X^\nu, X^{\nu+1}} \quad (3.1.8)$$

where R_ν is the reaction transforming X^ν into $X^{\nu+1}$ ($\nu = 0, \dots, K-1$). For a biological system of realistic size, the number of reactions occurring between two measured state is usually much greater than 1, so that the condition $K \gg 1$ is usually satisfied.

3.2. Sampling the BlenX state transition system

Since the system is stochastic K is not constant, but it can change simulation by simulation. $\{R_\nu\}$ with

$\nu = 0, \dots, K-1$ can be considered a latent reaction pathway. To calculate the likelihood we have to find an efficient way to sample this latent reaction pathway conditioned to the observations. Namely, we have to sample the latent reaction processes that match the initial and the end state in the time interval. The parameter estimation is then formulated as maximization of the likelihood function. The parameter estimate Θ^* is calculated as

$$\Theta^* = \arg \max_{\Theta} L(\hat{X}_j^0, \hat{X}_j^K; \Theta) \quad (3.2.1)$$

and the likelihood function over the entire duration of the observation is the product of the likelihood of each subinterval.

$$L(\mathbf{X}(t); \Theta) = \prod_{j \in \Gamma} \prod_{s=1}^m L(X_j(t_s), X_j(t_{s+1}); \Theta) \quad (3.2.2)$$

To sample latent path that are consistent with the observations means to generate a Markov chain that match the initial and the end state of the system in the considered time interval. One commonly used sampling method is the stochastic simulation algorithm (SSA). However, SSA is computationally inefficient when the total number of possible state is high. Y. Wang et al (Wang, Christley, Mjolsness, & Xie, 2010) suggested a Markov chain sampler working as follows:

1. generate an initial path
2. generate a set of reaction, by adding or removing reactions front he initial set
3. estimating the acceptance probability of a new set
4. accept or reject a pathway

Note that both the initial path and the processed path have to match the observations at the start and the end of the interval, implying that only a subset of the reactions can be used for either initialization or addition/deletion. In this work, the first path is randomly generated.

After an initial path is generated we can use the *elementary mode analysis* to generate a new sample. In this study we select randomly the first path. An elementary mode of a biochemical network is a set of reactions that does not change the observed number of molecular species. Therefore, an elementary mode is a column vector \mathbf{q}_k of non-negative integers that satisfy the following condition

$$A_\Gamma^T \mathbf{q}_k = 0 \quad (3.2.3)$$

where A_Γ is the net effect matrix of the system Γ . The set of all independent elementary mode $\{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_J\}$ is called *null set* of the biochemical reaction system. Provided with a reaction path and a null set, after randomly choosing an elementary mode \mathbf{q}_k from the null set, we can proceed as follows:

- with probability $p_u = 0.25$ (Wang, Christley, Mjolsness, & Xie, 2010), add the set of reactions in \mathbf{q}_k with random waiting time of reaction from a uniform distribution within the considered time interval;
- with probability $p_r = 0.25$ (Wang, Christley, Mjolsness, & Xie, 2010), remove one set of randomly selected reactions in \mathbf{q}_k from the current path within the considered time interval;

A new sample have to undergo two additional constraints:

- the number of any reaction type must be positive after the move
- the population numbers for all species have to remain positive throughout the whole process.

If either of the two conditions is violated we set the probability of the new sample path to be zero and reject the new path.

Each time a new pathway is sampled, we determine the acceptance probability of the proposed pathway according to the formula

$$p = \frac{p_{\text{current}}}{p_{\text{previous}}}$$

where p_{previous} is the probability of the previous pathway, and p_{current} is the probability of the current proposed pathway. The probability of a pathway is calculated as in the following formula.

$$p_{\text{pathways}} = \prod_{\{j|q_{k,j} \neq 0\}} \frac{k_j n \tau}{\alpha_j \pi}$$

where n is the total number of components in the reaction system, τ is the time length of the subinterval, k_j is the rate constant of the reaction of type j , $\alpha_j = 1$ if the reaction is monomolecular, and $\alpha_j = 2$ if the reaction is bimolecular. $\pi \in (0, 1)$ is a uniform deviate, as in Colvin et al. (Colvin, Monine, Faeder, Hlavacek, Von Hoff, & Posner, 2009). A pathways is accepted if $p \geq 0.25$.

4. RESULTS

We generated the time series of the components of the system in Figure 3 synthetically by running the BlenX code with the values of parameters reported in the code in previous pages, and then we applied the procedure of parameter inference described in the previous section. In Figure 5, we show the time series taken as an input for the model calibration procedure. In Table 2 we report the results of the inference for the main reactions (i. e. the rate-limiting step reactions) listed in Figure 2.

Good agreement between inferred and expected values has been obtained within the parameter variance estimated around 1.

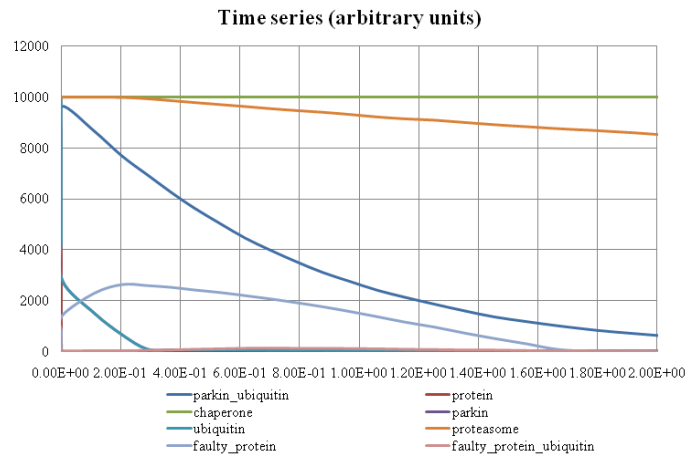


Figure 5: time-series synthetically generated from the model with given parameters are used as an input to the parameter inference procedure.

Table 1: comparison between inferred and expected values of the rate –limiting step reaction in the system of chaperon-assisted protein folding.

Parameter	Inferred	Expected
k_1	0.03203	0.01
k_2	0.03249	0.01
k_3	0.42681	0.5
k_4	0.03793	0.5
k_5	0.00032	0.0002

5. CONCLUSIONS

We presented a maximum-likelihood method for inferring rate parameters of reactions of a stochastic biochemical systems from discrete time observations. The core of the method has been proposed in 2010 by Wang et al. (Wang, Christley, Mjolsness, & Xie, 2010). In this work we illustrated how it can be adapted to calibrate a process-algebra model of a biochemical system. We showed that the mathematical approach of the method is suitable to the identification of parameters in language- reaction-based model. In this paper we reported a simple example to give to the reader the flavor both of the BlenX process algebra language and of the capabilities of the inference method. From the results obtained from synthetic and real case data (not described in this paper) we conclude that this procedure is trustable.

REFERENCES

- CoSbi. (2011). Retrieved from <http://www.cosbi.eu>.
- Boys, R. J., Wilkinson, D. J., & Kirkwood, T. L. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Stat. Comput.*, 18, 125-135.
- Cardelli, L. (2005). Brane Calculi - Interactions of Biological Membranes. *Workshop on Computational Methods in sSystems Biology (CMSB'04) Lecture Notes in Computer Science*. 3082, pp. 257-278. Springer.
- Colvin, J., Monine, M. I., Faeder, J. R., Hlavacek, W. S., Von Hoff, D. D., & Posner, R. G. (2009). Simulation of large-scale rule-based models. *Bioinformatics*, 25(7), 910-917.
- Danos, V., & Krivine, J. (2004). Reservable communicating systems. *CONCUR 2004. 3170 of LNCS*, pp. 292-307. Springer-Verlag.
- Danos, V., & Laneve, C. (2004). Formal molecular biology. *TCS*.

Dematté, L., Priami, C., & Romanel, A. (2008). The BetaWorkbench: a computational tool to study the dynamics of biological systems. 9(5), 437-448.

Dematté, L., Priami, C., & Romanel, A. (2008). *The BlenX language: a tutorial*. Trento, Italy: The Microsoft Research - University of Trento centre for Computational and Systems Biology.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Chemical Physics*, 81, 2340-2361.

Gilmore, S., & Hillstone, J. (1994). The PEPA Workbench: A Tool to Support a Process Algebra-based Approach to Performance Modelling. *Lecture Notes in Computer Science*, 794, 353-368.

Lecca, P. (2011). BlenX models of alpha-synuclein and parkin kinetics in neuropathology of Parkinson's disease. *Journal of Biological Systems*, 19(2).

Lecca, P., Palmisano, A., Ihekwa, A. E., & Priami, C. (2010). Calibration of dynamic models of biological systems with KInfer. *European Journal of Biophysics*, 39(6), 1019.

Priami, C. (1995). Stochastic p-calculus. *The Computer Journal*, 38, 578-589.

Regev, A., Panina, E. M., Silverman, L., Cardelli, L., & Shapiro, E. (2004). Bioambients: an abstraction for biological compartments. *Theor. Comput. Sci.*, 325(1), 141-167.

Wang, Y., Christley, S., Mjolsness, E., & Xie, X. (2010). Parameter inference for discretely observed stochastic kinetic models using gradient descent. *BMC Systems Biology*, 4(99).

AUTHORS BIOGRAPHY

Dr. **Paola Lecca** received a Master Degree in Theoretical Physics from the University of Trento (Italy) and a PhD in Computer Science from the International Doctorate School in Information and Communication Technologies at the University of Trento (Italy). Currently Paola Lecca is the Principal Investigator of the Inference and Data manipulation research group at The Microsoft Research – University of Trento Centre for Computational and Systems Biology (Trento, Italy). Dr. Paola Lecca's research interests include stochastic biochemical kinetic, biological networks inference, optimal experimental design in biochemistry, and computational cell biology. She designed prototypes for biological model calibration and for the simulation of diffusion pathways in cells and tissues. She has published articles in leading medical, biological and bioinformatics Journals and Conferences (<http://www.cosbi.eu/>).

DENERVATED MUSCLE UNDERGOING ELECTRICAL STIMULATION: DEVELOPMENT OF MONITORING TECHNIQUES BASED ON MEDICAL IMAGE MODELLING

Paolo Gargiulo^{(a) (b)}, Thomas Mandl^(c), Egill A. Friðgeirsson^{(a) (b)}, Ilaria Bagnaro^(d), Thordur Helgason^{(a) (b)}, Páll Ingvarsson^(e), Marcello Bracale^(d), Winfried Mayr^(f), Ugo Carraro^(g), Helmut Kern^(h)

^(a) Department of Development and Consultancy HUT, University Hospital Landspítali

^(b) Department of Biomedical Engineering, University of Reykjavík

^(c) Department of Computer Science, University of Applied Sciences Technikum Wien, Vienna, Austria

^(d) Department of Biomedical Engineering, University Federico II Napoli

^(e) Department of Rehabilitation Medicine, Landspítali-University Hospital, Reykjavík, Iceland

^(f) Medical University of Vienna, Center of Biomedical Engineering and Physics

^(g) Laboratory of Translational Myology of the University of Padova, Department of Biomedical Sciences, Padova, Italy

^(h) Ludwig Boltzmann Institute of Electrostimulation and Physical Rehabilitation, Department of Physical Medicine, Wilhelminenspital. Vienna, Austria

^{(a),(b)} paologar@landspitali.is, ^(b) thomas.mandl@technikum-wien.at, ^{(a),(b)} egillf05@ru.is, ^(d) i.bagnaro@gmail.com,
^{(a),(b)} thordur@landspitali.is, ^(e) palling@landspitali.is, ^(d) bracale@unina.it, ^(f) winfried.mayr@meduniwien.ac.at,
^(g) ugo.carraro@unipd.it, ^(h) helmut.kern@wienkav.at

ABSTRACT

Muscle tissue composition accounting for the relative content of muscle fibres and intramuscular adipose and loose fibrous tissues can be efficiently analyzed and quantified using images from spiral computed tomography (S-CT) technology and the associated distribution of Hounsfield unit (HU) values. Muscle density distribution, especially when including the whole muscle volume, provides remarkable information on the muscle condition.

We analyse the content of fat, connective tissue, normal muscle and dense fibrous connective tissue in spinal cord injured patients undergoing electrical stimulation treatment using 3-D modelling and segmentation tools.

The results show in a novel way and quantitatively the muscle restoration and growth induced by electrical stimulation; the amount of normal muscle fibres increases from 45% to 60% of the whole volume while connective tissue and fat reduce respectively of 30% and 50%.

Moreover the effectiveness of the FES treatment using surface electrodes is evaluated calculating the density distribution along rectus femoris cross sectional areas. The results show that muscles undergoing FES restore in certain areas and decline in others depending on patient anatomy and surface electrodes positioning.

Keywords: Functional electrical stimulation, Segmentation, Numerical methods, spinal cord injury.

1. INTRODUCTION

Loss of muscle mass occurs with many pathological conditions and is linked to increased patient disability, morbidity and mortality (Janssen, 2002; Fisher 2004) thus, it is important to discover how to deter muscle

degeneration. Empirical clinical observations (Kern 1999) revealed that lower motor neuron (LMN) denervated degenerated muscle can recover by a specific variation of home based daily functional electrical stimulation (FES) therapy. This is in contradiction to earlier data which suggested that FES was effective only when started immediately after LMN lesion. These observations led to the founding of the European funded project RISE in November 2001. The project's aim was to establish the biological basis for a clinical rehabilitation treatment for patients who have permanent muscle LMN denervation in the lower extremities. To this end, it funded research designed to reverse muscle degeneration induced by the permanent lack of innervation in spinal cord injured (SCI) patients using muscle FES. Some of the funding has been used to fund research in rehabilitative centres in Vienna (Austria), Heidelberg (Germany), Hamburg (Germany), Tübingen (Germany), Reykjavik (Iceland) and Vicenza (Italy). The RISE project has achieved its goal (Kern, 2010) using a multidisciplinary approach to optimize technology to stimulate LMN denervated muscle with custom-designed electrodes and stimulators (see figure 1) developed in Vienna, Austria (Mayr, 2001).



Figure 1: A RISE patient during FES treatment.

The project encompassed a clinical trial with over 25 voluntary patients and additional animal experiments to research the muscle restoration process by combining physiological, histological, immunohistochemical, and biochemical analyses with anthropometric techniques (Kern, 2009; Mödlin 2005).

The results of the EU RISE Project, and of the related animal research, provide different perspectives. Twenty out of 25 patients completed a 2 years h-b FES program (Kern 2010a; Kern 2010b), which resulted in: 1. Significant increase of thigh muscle size and of the muscle fibers, with striking improvements of the ultra-structural organization of contractile material; 2. Significant increase in muscle force output during electrical stimulation (knee extension torque); 3. The recovery of quadriceps m. force was sufficient to allow compliant subjects to perform FES-assisted stand-up and stepping-in-place exercises; 4. Ultra structural analyses demonstrated that the shorter was the time elapsed from SCI to the beginning of h-b FES, the larger were the number and the size of recovered fibers. The study demonstrates that h-b FES of permanent LMN denervated muscle is an effective home therapy that results in rescue of muscle mass and tetanic contractility. Important benefits for the patients are the improved cosmetic appearance of lower extremities, the enhanced cushioning effect for seating (Kern 2004; Kern 2010a; Kern 2010b; Boncompagni 2007) and the early result of impressive reduction of the leg edema (Bizzarrini 2007). The last observation is supported by changes of the capillary networks observed in the muscle biopsies harvested from subjects suffering with long-lasting LMN denervation before and after h-b FES (Scelsi, 2006) and thoracic level SCI (Lotta 1991; Lotta 2001; Scelsi 1991; Scelsi 1995; Scelsi 2001; Scelsi 2005).

Many of the tissue analyses employed to study structural changes occurring in LMN denervated muscle (both after long term LMN denervation and during electrical stimulation) were performed with biopsies which meant that only a few milligrams of muscle could be analyzed (Kern 2010b). Complementary imaging techniques, such as X-ray computed tomography (CT), were also employed in order to assess and validate histological information. The value of the imaging methods demonstrates that the development and use of non-invasive anthropometric techniques is critical to this area of research.

2. MATERIAL AND METHODS

As a consequence of long term denervation the muscle degenerates dramatically. The muscles become very thin and the single bellies are no longer recognizable in their shape. Only rectus femoris (RF) remains recognizable among the quadriceps muscles though it is severely degenerated compared to the normal situation. Therefore 3-D modelling and segmentation techniques are used to isolate RF from other bellies and to monitor changes occurring during the degeneration and restoration process.

Besides, accurately measuring changes in RF during electrical stimulation treatment is important to evaluate treatment effectiveness. Surface electrodes are placed on top of the quadriceps and since RF occupies the middle of the thigh it is especially exposed to the current distribution (Mandl 2008). RF is thus the optimal target for monitoring therapeutical effects and morphological changes with a 3D approach (Gargiulo 2008; Gargiulo 2010).

2.1 Data set source

X-ray computed tomography is an imaging method that uses X-rays to produce images of structures ‘inside’ the body. Patients are scanned slice-per-slice and each slice is scanned several times from different angles. Each imaged volume element (voxel) is traversed, during the scan, by numerous X-ray photons and the intensity of the transmitted radiation is measured by detectors. The measured intensity profile contains information on the densities the beam encountered on its path through the body (i.e. the denser the regions the weaker the signal). With suitable mathematical methods the measured profiles of each slice are transformed into an image of the structures inside – the image is reconstructed, the grey value corresponding to the linear attenuation coefficient.

This way, 3-Dimensional data are gathered scanning the patient’s lower limbs with spiral CT. The scan starts above the head of the femur and continues down to the knee joint, both legs being covered in one scan. Slice increment is set to 0.625 mm resulting in a total of about 750–900 CT slices, depending on the patient’s size.

Every acquired CT slice is subdivided into a matrix of different size, from a minimum of 128×128 up to 1024×1024 volume elements. The average linear attenuation coefficient μ of the tissue contained in each voxel is represented by floating numbers in the computer which range from 0.0 up to values equal to 1.0.

Once the image is calculated from the 3D data set it is converted into a matrix of picture elements (pixels) with each pixel assigned the attenuation value of the corresponding voxel. Linear attenuation coefficients are rescaled to an integer range that encompasses 4096 values, between -1000 and 3095. From these intensity readings, the density or attenuation value of the tissue at each point in the slice can be calculated. This scale is called CT number or Hounsfield unit (HU) and it is expressed by the following formula:

$$\text{CT Number} = 1000 \times \frac{\mu_{\text{pixel}} - \mu_{\text{water}}}{\mu_{\text{water}}} \quad (1)$$

With this scaling, if the linear attenuation coefficient of a given pixel (μ_{pixel}) is equal to that of water, the CT number will be 0. If μ_{pixel} is less than μ_{water} the CT number will be negative which is typical for air spaces, lung tissues and fatty tissues. Values of μ_{pixel} greater than μ_{water} will result in positive CT numbers. Very dense tissue such as bone has large

positive numbers. Table 1 displays main organic tissues and respective HU intervals.

Table 1: The table displays the main human tissues and their empirical HU intervals. (Gargiulo 2011)

Anatomical Tissues	Hounsfield intervals	
	Min	Max
Bone	250	3071
Compact bone	601	1988
Spongy bone	250	600
Normal Muscle	41	80
Dense fibrous connective (Tendons-dense muscle)	81	200
Loose connective (low dense muscle)	-5	40
Fat	-200	-6
Skin	-30	60
Tooth	1200	3071

2.2 Segmentation of RF

The threshold interval chosen to segment RF is: [-5, 200] HU. A wide interval is chosen because it must display muscle tissue and allow monitoring of changes, particularly the restoration-degeneration process. Within the selected interval the displayed pixels are representing both normal and degenerated muscles, additionally connective tissue and water but subcutaneous fat is excluded. In this way the surrounding fat in RF is automatically excluded from the segmentation mask (Fig.2 A). After thresholding, the next step for the segmentation is to isolate RF from the other muscles. For this purpose the following procedure is used. The process starts from a cross section where the muscle boundaries are well visible (usually in the middle, along the length of the muscle). A contour is manually drawn around the muscle and projected to the next cross sections in both directions. If the contour fits the new cross sectional area well then it is projected unchanged forward to the next slice, otherwise it is adapted and then projected ahead. We assume that shapes change little from one slice to the next the adaptation necessary is small. It is done with active contours that ‘snap’ to boundaries.

The process continues until all cross sections containing RF cross sections are edited (Fig.2 B). The contour areas are then erased creating a gap between segmentation target and surrounding. Finally, a new segmentation mask representing RF is created by applying a region growing procedure which creates a new mask separating the edited structure that is no longer connected to the surrounding (Fig.2 C). The result of the segmentation process and a 3D rendering thereof is shown in Fig.2 D.

All Segmentation was performed with MIMICS 10.1. (www.materialise.com)

2.3 Muscle tissue analysis

In order to discriminate the biological tissues in the data set, different thresholds are established using the HU scale. Specific attenuation values are assigned to each

individual voxel. The degree of attenuation depends on the energy spectrum of the x-rays as well as on the average atomic number of the mass density of the patient tissue. Most computer display hardware requires integer numbers and therefore the linear attenuation coefficients are rescaled to an integer range that encompasses 4096 values, between -1000 and 3095. Dense tissue such as bone has large positive CT number while negative CT numbers are typical for air spaces, lung tissues and fatty tissues. Muscle tissues are normally displayed with HU values between 50 and 100 HU though within a normal muscle belly there are also other tissue elements such as connective and fat which are coded with much lower HU values. Anyway the specific HU value depends also on the pixel size. Indeed every element can express its absolute HU value if it occupies completely the specific pixel volume otherwise this value will be an average between the different parts contained in it. This fact explains the wide range of values present inside a data set and suggests the definition of various intervals to study muscle structural changes. Therefore to monitor and quantify tissue composition in the stimulated muscle volume the HU values present within the segmented volume are divided in four HU intervals [-200, -6], [-5, -40], [41, 80] and [81, 200] representing respectively fat, loose connective (low dense muscle), normal muscle and Dense fibrous connective (Tendons-dense muscle).

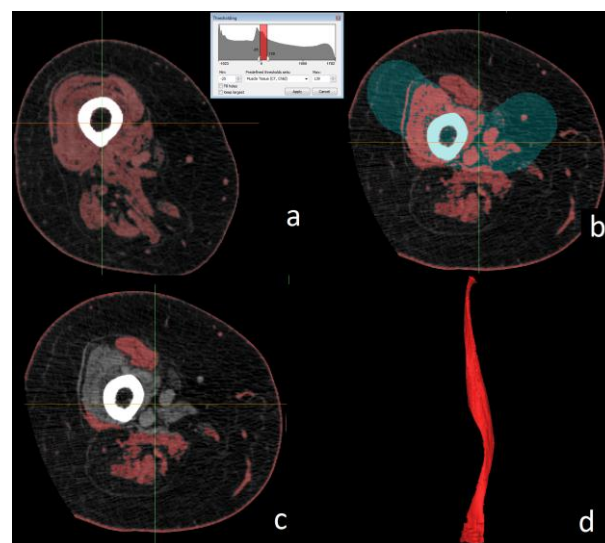


Figure 2: Main steps for the Segmentation of rectus femurs (A) Thresholding (B) A contour is manually drawn around the muscle and projected to the next cross sections in both directions (C) The contour is erased and rectus femoris is isolated from the surrounding muscles (D) 3D rendering of rectus femoris.

2.4 Cross sectional density analysis

In order to evaluate the effectiveness of the FES treatment using surface electrodes we develop a Matlab (Matworks Inc) subroutine to analyse the density distribution along rectus femoris cross sectional areas. Each cross section provides a mean HU value. The

measured rectus femoris lengths are between 400 and 500 mm depending from patient anatomy (starting from the pelvis attachment and ending at patellar tendons) the slice increment is 0.625 mm therefore the number of mean values is between 640 and 800.

Figure 3 shows the computation results on a healthy subject. It can be noticed how the mean values on the cross sectional areas are rather uniform and displayed in the interval [50-70] HU. Density values are higher at the muscle extremities where the muscle attaches to the tendons.

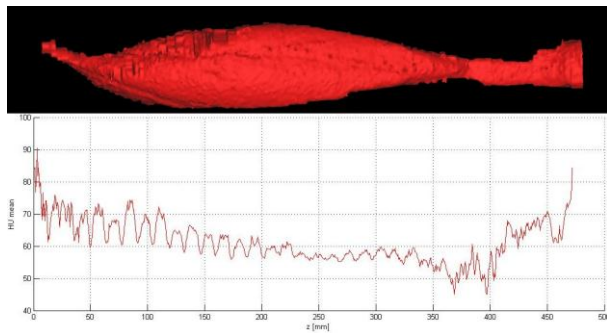


Figure 3: Density distribution along the cross sections in healthy subject.

3. RESULTS

Muscle restoration due to electrical stimulation can be seen quantitatively using the muscle tissue analysis. Figure 4 shows muscle tissue changes after 4 years of electrical stimulation treatment. Due to stimulation treatment the fraction of normal muscle fibres increases from 45% to 60% of the whole volume while connective tissue and fat reduce to 30% and 50% respectively.

Figure 6 shows the results from cross sectional density analysis from two RISE patients:

- Patient 1 started the FES treatment 1 year after the paralysis (fig.5a). He was compliant to the stimulation protocol for approximately 1.5 years and successively interrupting the treatment.
- Patient 2 started the FES treatment 4 year after the paralysis (fig.5b). He was compliant over all the monitored period.

Muscle density increase: Patient 1 and 2 are stimulated in the same way but the stimulation is more effective on patient 1 because muscles at beginning are in better conditions and adipose tissue is thinner. The area indicated with *a* in figure 6 (Patient 1) displays a localised increase of density - from 45 to 50 HU - starting at 120 mm and ending approximately at 400 mm from the pelvis attachment

The areas indicated with *d* and *f* (patient 2) displays a localised increase of density - from 40 to 55 HU - starting at 170 mm and ending approximately at 350 mm from the pelvis attachment.

Muscle density decline in non-compliant patient: comparing the areas indicated with *a* and *b* (patient 1) can be seen a general muscle density decrease of 10 HU

over the all length after 2 years of interruption from FES treatment.

Muscle density decline in compliant patient: the areas *c*, *e*, and *g* (patient 2) display zones on rectus femoris between 50 mm and 150 mm from the pelvis attachment where muscle density continues to decrease - from 30 down to 5 HU- during the FES treatment. Here the electrical stimulus does not reach efficiently the muscle fibers because the adipose tissue between skin and muscle is too thick (fig. 5A).

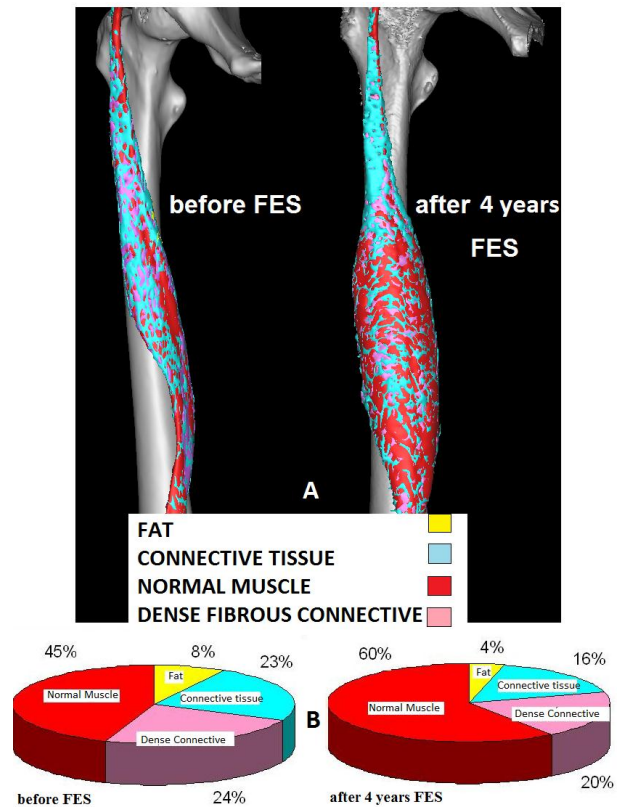


Figure 4: 3D models of rectus femoris before and after 4 years of electrical stimulation treatment (A). Chart pies showing the muscle composition at beginning and after 4 years FES according to the Hounsfield intervals referred in table 1 (B).

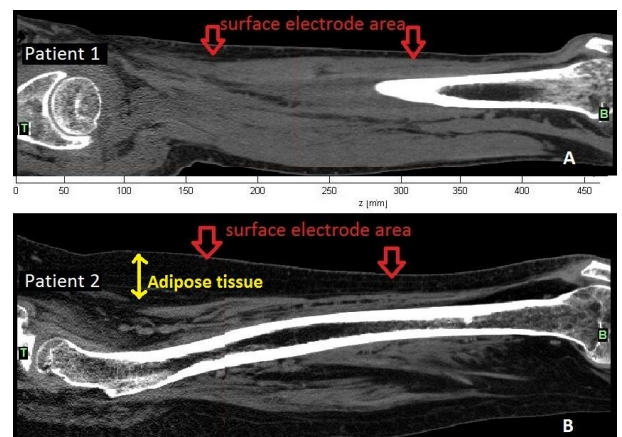


Figure 5: Spiral CT pictures sagittal view at beginning of the FES treatment: patient 1 (A), patient 2 (B).

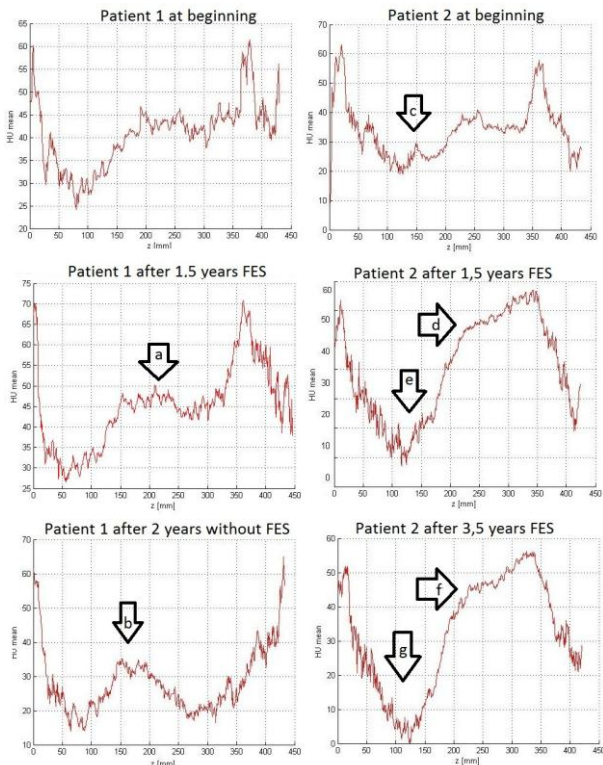


Figure 6: Cross sectional density analysis performed on two patients at beginning of the FES treatment, after 1,5 years and after 3,5 years. Patient 1 interrupt the FES treatment after 1,5 years.

4. CONCLUSION

The 3D approach combined with muscle tissue analysis, gradient distributions and cross sectional density analysis provides information on the whole muscle and on its structural changes during electrical stimulation treatment otherwise not accessible with other monitoring techniques. Muscle growth and tissue restoration can be efficiently monitored; beside the density analysis along the segmented muscle can also drive to an improved surface electrode design and positioning.

The computational method developed in this work is associated to thresholding criteria and HU values which are used to define the different tissues within the muscle. Various physical factors can influence the CT number representation during a scan session. The parameter that mostly affects the accuracy and the spatial distribution of HU values is the applied voltage across an X-ray tube; this amplitude is measured in kilo volt (kV) and determines the highest X-ray quantum energy and therefore the attenuation coefficient. CT number distribution is also influenced by phantom (or patient) orientation and position in scan aperture. Therefore it is necessary to know and account these variability's when CT numbers are used for tissue characterization and comparison.

ACKNOWLEDGMENTS

This work has been supported by the University Hospital Landspitali research fund

REFERENCES

- Bizzarini E, 2009. "Epidemiology and clinical management of Conus-Cauda Syndrome and flaccid paraplegia in Friuli Venezia Giulia: Data of the Spinal Unit of Udine". *Basic Appl Myol* 2009; 19: 163-167.
- Boncompagni S, 2007. "Structural differentiation of skeletal muscle fibers in the absence of innervation in humans". *Proc Natl Acad Sci USA* 2007; 104: 19339-19344.
- Fisher AL, 2004. "Of worms and women: sarcopenia and its role in disability and mortality". *J Am Geriatr Soc.*;52:1185-90
- Gargiulo P, 2008. "Restoration of Muscle Volume and Shape Induced by Electrical Stimulation of Denervated Degenerated Muscles: Qualitative and Quantitative Measurement of Changes in Rectus Femoris Using Computer Tomography and Image Segmentation" *Artif Organs.* 32: 609-613.
- Gargiulo P, 2010. "Quantitative color 3-dimensional computer tomography imaging of human long-term denervated Muscle" *Neurological Research.* 32:13-20.
- Gargiulo P, 2011. "Monitoring of Muscle and Bone Recovery in Spinal Cord Injury Patients Treated With Electrical Stimulation Using Three-Dimensional Imaging and Segmentation Techniques: Methodological Assessment". *Artificial Organs Volume 35, Issue 3, pages 275-281.*
- Janssen I, 2002. "Low relative skeletal muscle mass (sarcopenia) in older persons is associated with functional impairment and physical disability". *J Am Geriatr Soc.*;50:889-896.
- Mandl T, 2008. "Functional Electrical Stimulation of Long-term Denervated, Degenerated Human Skeletal Muscle: Estimating Activation Using T2-Parameter Magnetic Resonance Imaging Methods" *Artificial Organs* 2008; 32(8): 604-608.
- Mayr W, 2001. "Basic design and construction of the Vienna FES implants: existing solutions and prospects for new generations of implants". *Med Eng Phys* 2001; 23: 53-60.
- Mödlin M, 2005. "Electrical stimulation of denervated muscles: first results of a clinical study". *Artif Organs*; 29: 203-206.
- Kern H, 1999. "Standing up with denervated muscles in humans using functionalelectrical stimulation". *Artif Organs*; 23: 447-452.
- Kern H, 2009. "European Project RISE: Partners, protocols, demography". *Basic Appl Myol/European Journal of Translational Myology* 2009; 19:211-216.
- Kern H, 2004. Long-term denervation in humans causes degeneration of both contractile and excitation-

contraction coupling apparatus that can be reversed by functional electrical stimulation (FES). A role for myofiber regeneration? *J Neuropathol Exp Neurol* 2004;63: 919-931.

- Kern H, 2010a. "Home-based Functional Electrical Stimulation (h-bFES) recovers permanently denervated muscles in paraplegic patients with complete lower motor neuron lesion". *Neurorehab Neur Rep*. vol. 24(8):709-721. Epub 2010 May 11.
- Kern H, 2010b. "One year of home-based Functional Electrical Stimulation (FES) in complete lower motor neuron paraplegia: Recovery of tetanic contractility drives the structural improvements of denervated muscle". *Neurol Res* 2010; 32: 5-12
- Lotta S, 1991. "Morphometric and neurophysiological analysis of skeletal muscle in paraplegic patients with traumatic cord lesion". *Paraplegia* 1991; 29: 247-252.
- Lotta S, 2001. "Microvascular changes in the lower extremities of paraplegics with heterotopic ossification". *Spinal Cord* 2001;39: 595-598.
- Scelsi R, 1991. "Morphological properties of skeletal muscle in spastic paraplegia". *Basic Appl Myol* 1991; 1: 317-326.
- Scelsi R, 1995. "Morphological changes in the skin microlymphatics in recently injured paraplegics with ileo-femoral venous thrombosis". *Paraplegia* 1995; 33: 472-475.
- Scelsi R. 2001 "Skeletal muscle pathology after Spinal Cord Injury: Our 20 year experience and results on skeletal muscle changes in paraplegics, related to functional rehabilitation". *Basic Appl Myol* 2001; 11: 75-85.
- Scelsi R, 2005. "Morphological alterations of microvasculature and neoangiogenesis in the pressure ulcers repair in paraplegics". *Basic Appl Myol* 2005; 15: 203-208.
- Scelsi R, 2006. "Flaccid paraplegia: Improvement of the muscle capillary supply after early-started daily functional electric stimulation (FES) in human permanent lower motoneuron denervation". *Basic Appl Myol* 2006; 16: 105-107.

AUTHORS BIOGRAPHY

Paolo Gargiulo is assistant professor at Reykjavik University and works as biomedical engineer and researcher at the University Hospital of Iceland. He studied electrical engineering at University Federico II in Naples, and obtained a PhD at Technical University of Vienna, Austria. His research interests are in the field of electrical stimulation medical modeling and rapid prototyping for clinical applications.

Thomas Mandl received his "Diplomingenieur" (MSc level) in Technical Physics from the Technical University of Vienna late 2005. He has since been a phd student in Medical Physics at the Medical University of Vienna in Prof. Moser's MR group. His (research) interests include computer simulation and visualization as well as muscle relevant MR methods.

Egill Axfjord Fridgeirsson is a master student at Reykjavik University Iceland in Biomedical Engineering. His research interests are Medical modeling, Biomagnetism and clinical applications of medical models. He currently works in the Clinical Engineering department of Landspítali University hospital.

Ilaria Bagnaro is a master student at University Federico II in Naples in Biomedical Engineering. She is on her second year of graduate studies and this work represents part of her master thesis.

Thordur Helgason is associate professor at Reykjavik University and works as biomedical engineer and researcher at the University Hospital of Iceland. He studied electrical engineering at University of Iceland, and obtained a PhD at Technical University of Karlsruhe, Germany. His research interests are in the field of electrical stimulation biomedical technologies.

Páll Ingvarsson is neurologist specialised at Goteborg University, Sweden. He works at Department of Rehabilitation Medicine, Landspítali-University Hospital, Reykjavik, Iceland

Marcello Bracale is professor in biomedical engineering at University Federico II, Naples. He received his degree in Electrical Engineering from the University of Naples in 1965 and his post-graduate specialization in Biomedical Technologies from the University of Bologna. His main fields of scientific and professional interest are: electronic and biological instrumentation; bio signal and data analysis; health care system and management; health telematics and telemedicine.

Winfried Mayr is Associate Professor of Biomedical Engineering and Rehabilitation Technology. He works at Medical University of Vienna Center for Medical Physics and Biomedical Engineering. His research interests are in the field of Functional electrical stimulation (mobilization of paraplegics, phrenic pacing, EMG-controlled stimulation, pelvic floor applications, denervated muscles, application of FES in microgravity and bed-rest) Implant technology (Microelectronic and electromechanical implants).

Ugo Carraro is Associate Professor of General Pathology at the Faculty of Medicine of the University of Padova, since 1983. Editor-in-Chief of Basic and Applied Myology/The European Journal of Translational Myology since 1991, he founded and chair since 2005 the University of Padova Interdepartmental Research Center of Myology.

Helmut Kern is head of the "Department of Physical Medicine and Rehabilitation of the Wilhelminenspital" (Vienna, Austria) since 1984 and director of the research institute "Ludwig Boltzmann Institute of Electrical Stimulation and Rehabilitation" (Vienna, Austria) since 1988.

3D SEGMENTED MODEL OF HEAD FOR MODELLING ELECTRICAL ACTIVITY OF BRAIN

Egill A. Friðgeirsson^{(a),(b)}, Paolo Gargiulo^{(a),(b)}, Ceon Ramon^{(b),(d)}, Jens Haueisen,^(c)

^(a) Department of Development and Consultancy UTS

^(b) Department of Biomedical Engineering, University of Reykjavik

^(c) Institute of Biomedical Engineering and Informatics, Technical University Ilmenau, Germany

^(d) Department of Electrical Engineering, University of Washington, United States

^{(a),(b)} egillf05@ru.is, ^{(a),(b)} paologar@landspitali.is, ^{(b),(d)} ceonramon@yahoo.com, ^(c) jens.haueisen@tu-ilmenau.de

ABSTRACT

Computer simulation and modelling of the human body and its behaviour are very useful tools in situations where it is either too risky to perform an invasive procedure or too costly for in vivo experiments or simply impossible for ethical reasons. In this paper we describe a method to model the electrical behaviour of human brain from segmented MR images. The aim of the work is to use these models to predict the electrical activity of the human brain under normal and pathological conditions. The image processing software package MIMICS is used for 3D volume segmentation of MR images. These models have detailed 3D representation of major tissue surfaces within the head, with over 12 different tissues segmented. In addition, computational tools in Matlab were developed for calculating normal vectors on the brain surface and for associating this information to the equivalent electrical dipole sources as an input into the model.

Keywords: Segmentation, EEG modelling; Finite element; Realistic head model.

1. INTRODUCTION

The relationship between neuronal sources and the recorded scalp electroencephalograms (EEG's) has for a long time been of interest (Abraham and Marsan 1958). This relationship is characterized by the spread of electrical currents in inhomogeneous tissues, which govern the scalp electrical potentials.

Forward EEG modelling is a discipline which uses numerical techniques such as finite element method (FEM) modelling to study the relationship between electric sources in the brain and the resulting electrical potentials at the scalp (Hallez and Vanrumste 2007). Sources in the form of current dipoles are placed in the brain and then the FEM equations are solved for the resulting potential at the scalp.

For accurate forward EEG modelling detailed segmentation of tissues is needed, especially between the electrical source and scalp. In former studies the model complexity, or the number of tissue types, has been shown to affect the results (Ramon 2006). These

have emphasized the role of accurate segmentation of the cerebrospinal fluid (CSF) and bone (Ramon 2004).

The kind of 3D segmentation used here with the software platform MIMICS (Materialize Inc, USA) has previously been applied to monitor quadriceps femoris in paraplegic patients undergoing electrical stimulation as described in (Gargiulo 2010; Gargiulo 2011).

In this work we develop high (1.0 mm) resolution human head models from segmented MR images. These models have a detailed 3D representation of major tissue surfaces with over 12 tissue types defined.

Brain geometry is then used to locate the position and orientation of the cortical neurons to use as sources in the models. The cortical pyramidal neurons are thought to make up most of the EEG signal (Linás and Nicholson 1974; Okada 1982).

We will use these FEM models to simulate electrical activity of brain under normal and pathological conditions.

2. MATERIAL AND METHODS

2.1 MR data

A whole head T1 weighted volumetric MR scan of an adult female subject was used for this work. Each slice has 256×256 pixels, and each pixel has a gray value within the 4096 gray-scale values, meaning that it is represented with a 12-bit value. The contiguous slice thickness was 1.0 mm and pixel resolution was 1.0 mm. A total data set from a single scan is therefore $(256 \times 256 \times 192 \times 12)/8 = 1.9 \times 10^7$ bytes. This data set gives a complete 3D description of the tissue within the head.

2.2 Segmentation of the head

We develop detailed 3D representation of major tissues within the head including white and gray matter, cerebellum, CSF, cortical and trabecular bone, dura layer, skin, eyes and crystalline lens and so on. Some of these tissues have a distinguished threshold while other are displayed within the same gray values interval. In these cases special segmentation techniques and manual editing are employed to isolate single tissues from the surroundings.

The process starts from a cross section where the selected tissue boundaries are well visible. A contour is manually drawn around the region of interest (ROI) and projected to the next cross sections in both directions. If the contour fits well the new cross sectional area then it is projected unchanged forward to the next slice, otherwise it is adapted using manual editing and then projected ahead to the next slice. If the cross sectional area doesn't change much between slices it is enough to identify the contour in only a few slices and interpolate between them. This process continues until all cross sections containing the selected ROI are covered. The contour areas are then erased creating a gap between the ROI and the surroundings. Finally, a new segmentation mask representing ROI is created applying a region growing procedure which creates a new mask separating the edited structure that is no longer connected to the surrounding tissues.

Some tissues like the dura layer and the skin were segmented using other tools in mimics. The dura layer was segmented using wrapping functions on the CSF tissue mask. Similarly the skin was segmented using a wrapping function on the soft tissue mask.

The result of the segmentation process and a 3D rendering thereof is shown in Fig.1.

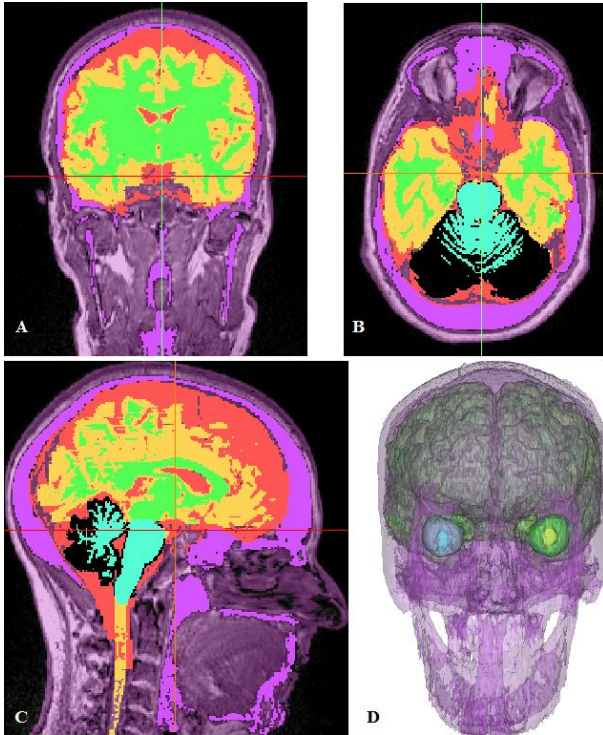


Figure 1: Segmentation results showing different tissue surfaces. Coronal view (A), axial view (B), sagittal view (C) and 3-D reconstruction of the segmented tissues (D).

2.3 Normal vectors

After the head has been adequately segmented the next step is to locate the surface boundary between the white and gray matter. This was done by using isosurface

algorithms in Matlab version 7.10 (Mathworks Inc, USA) with a binary image matrix with one as the white matter and zero elsewhere.

After the surface nodes were extracted, normal vectors were computed using the central finite difference approximation to the numerical gradient at those nodes in the binary matrix. So for each surface node $f(x, y, z)$ the three gradient components are found by (1):

$$\begin{aligned}\frac{\partial f}{\partial x} &= f\left(x + \frac{h_x}{2}\right) - f\left(x - \frac{h_x}{2}\right) \\ \frac{\partial f}{\partial y} &= f\left(y + \frac{h_y}{2}\right) - f\left(y - \frac{h_y}{2}\right) \\ \frac{\partial f}{\partial z} &= f\left(z + \frac{h_z}{2}\right) - f\left(z - \frac{h_z}{2}\right)\end{aligned}\quad (1)$$

Where h_x , h_y and h_z are the separation between the adjacent points in x , y and z directions, respectively.

Since the surface nodes represent a level set the gradient components are perpendicular to the surface and are therefore the normal vector components. This can be seen in Fig 2.

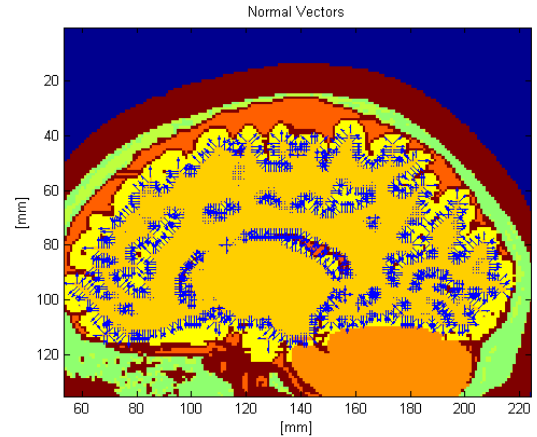


Figure 2: The normal vectors within a segmented slice of the brain.

2.4 Electrical Activity Modelling

An example of modelling the spontaneous electrical activity of the normal brain is presented here using the normal vectors to represent location and orientation of dipoles. A FEM model was constructed out of 192 segmented axial slices extending from top of the head to the bottom of the neck. The voxel resolution was $1 \times 1 \times 1$ mm. The electrical conductivities of various tissues were obtained from the literature and are summarized in our previous work (Ramon 2006). The conductivity of the dura matter is not well established and it was found to have a large range from 0.02 to 0.1 S/m (Oozeer 2005). For our work, we used a midrange value of 0.06 S/m.

The electrical activity in the top portion of the brain, above the eye level, was simulated with 125 dipoles randomly located in different parts of the brain. The dipole intensity distribution was in the range of 0.0 to 0.4 mA meter with a uniform random distribution.

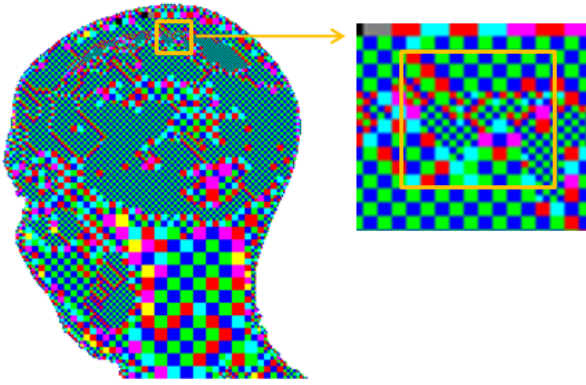


Figure 3: The FEM mesh of the adaptive solver.

An adaptive FEM solver, developed by us (Schimpf et al. 1998; Ramon et al. 2006), was used to compute flux and potential distribution in the whole head model. Fig. 3 shows an adaptive FEM grid and the details of the grid in the vicinity of one of the dipolar source. It shows an adaptive FEM grid and the details of the grid in the vicinity of one of the dipolar source. The FEM software automatically adjusts the grid resolution in each pass to achieve the desired L2 norm while keeping the computational errors to a minimum level.

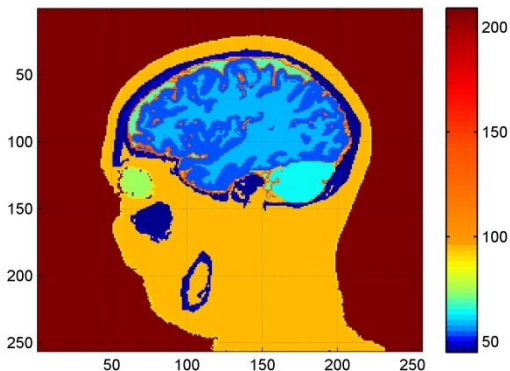


Figure 4: The segmented anatomical slice.

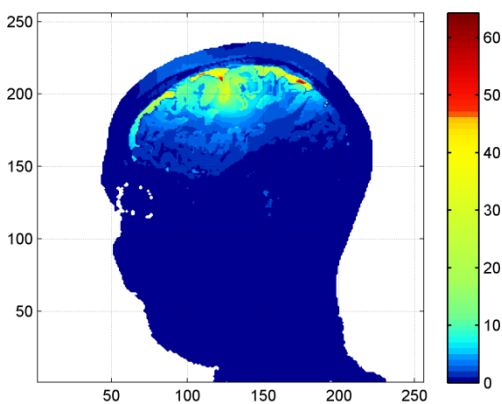


Figure 5: Current density distribution within the segmented slice. The intensity scale is in $\mu\text{A}/\text{cm}^3$. The nose is on the top.

An example of the current distribution is given in Figs. 4 and 5. The Fig. 4 shows the segmented anatomical slice and Fig. 5 shows the volume current density distribution. The current flow pattern follows the anatomical tissue boundaries very accurately. This is very pronounced at the CSF, gray and white matter boundaries. This shows that our segmentation and FEM modelling works well for computing flux and potential distributions in human head models.

2.5 Differential Contributions of Dipolar Activity to Scalp Potentials

The above described tools were used to analyze how neuronal activity at different depths contribute towards scalp EEGs. For this purpose we computed the scalp potentials due to dipoles in two layers at two different depths. The first layer, called Layer 1, was from top of the brain surface to the depth of 0.3 cm and the second layer, called Layer 2, was from the depth of 0.3 to 0.6 cm from top of the brain surface. There were 820 dipoles in the first layer and 3780 dipoles in the second layer. The total number of dipoles was 4,600. The dipole intensity distribution was in the range of 0.0 to 0.4 mA meter with a uniform random distribution.

The adaptive FEM solver described above was used to compute flux and potential distribution in the whole head model. Two models were studied. For one model the dipoles in the first layer were used and in the other model the dipoles in the second layer were used. The scalp potentials were extracted from the node potentials in the finite element models of the head. All computations were performed on an Intel quad core, second generation, 2.4 GHz workstation with 8 GB memory. Each run took about 10 minutes. Post-processing and visualizations were done using Matlab software, version 7.10.

3 RESULTS

Scalp potentials due to the first and second layers are given in Fig. 6 and a combination of both layers are given in Fig. 7. The scalp potentials of Layer 1 and Layer 2 both have an equivalent dipolar activity patterns. For the Layer 1, the positive contours are on the left side and the negative contours are on the right side and zero-crossing line of yellow colour is almost vertical in the middle of the positive and negative contours. An equivalent dipolar source can visualized as extending from the centre of the negative contours to the centre of the positive contours. For the Layer 2, almost circular negative contours are visible in the plot. This will suggest that the equivalent dipolar activity is pointing from the top to the bottom of the head, i.e., from superior to inferior position. The combined activity of the dipolar sources in Layer 1 and Layer 2 is given in Fig. 7 and it shows that it is dominated by the scalp potentials due to dipoles in Layer 2. This is feasible because there are 3780 dipoles in Layer 2 as compared with only 820 dipoles in Layer 1.

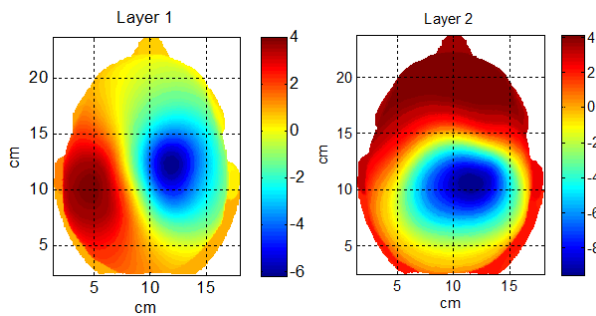


Figure 6: Scalp potentials due to dipoles in layers 1 and 2. The magnitude scale of the colour bar is in micro volts (μV).

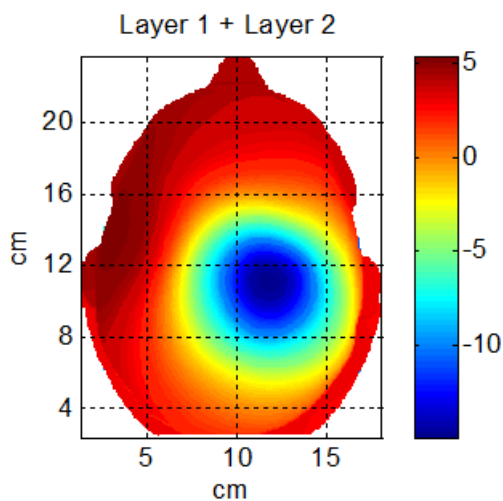


Figure 7: Scalp potentials due to combined dipolar activity from layer 1 and 2. The nose is on the top. The magnitude scale of the colour bar is in micro volts (μV).

4 CONCLUSION

These are our preliminary results to show our capabilities to segment the tissue boundaries with the use of the MIMICS software, to build detailed accurate anatomical models and to accurately compute current and potential distributions in the model. An application of our tools for differential contributions of the dipolar activity to the scalp potentials is also presented. In future, we plan to use these tools for modelling of the electrical activity of the brain under normal and pathological conditions and to research the influence of different brain structures on the EEG signal.

ACKNOWLEDGMENTS

This work has been supported by the University Hospital Landspítali research fund.

REFERENCES

Abraham K. and Marsan C.A., 1958. Patterns of cortical discharges and their relation to routine scalp electroencephalography, *Electroencephalography and Clinical Neurophysiology* 10(3):447-461.

Gargiulo, P., Kern, H., Carraro, U., Ingvarsson, P., Knútsdóttir, S., Guðmundsdóttir, V., Yngvason,

S., Vatnsdal, B. and Helgason, T., 2010. Quantitative color 3-dimensional computer tomography imaging of human long-term denervated Muscle. *Neurological Research* 32:13-20

Gargiulo, P., Vatnsdal, B., Ingvarsson, P., Knútsdóttir, S., Guðmundsdóttir, V., Yngvason, S. and Helgason, T., 2008. Restoration of Muscle Volume and Shape Induced by Electrical Stimulation of Denervated Degenerated Muscles: Qualitative and Quantitative Measurement of Changes in Rectus Femoris Using Computer Tomography and Image Segmentation. *Artificial Organs* 32: 609–613.

Hallez, H., Vanrumste, B., Grech, R., Muscat, J., De Clercq, W., Verguult, A., D'Asseler, Y., Camilleri, K.P., Fabri, S.G., Van Huffel, S. and Lemahieu, I. 2007. Review on solving the forward problem in EEG source analysis. *Journal of Neuroengineering and Rehabilitation* 4: 46.

Linás, R.R., Nicholson, C., 1974. Analysis of field potentials in the central nervous system. *Handbook Electroencephalography Clinical Neurophysiology* 2B:61-83.

Okada, Y.C., 1982. Neurogenesis of evoked magnetic fields. In: Williamson, S.J., Romani, G.L., Kaufman, L. & Modena, L., eds. *Biomagnetism; an Interdisciplinary Approach*. New York:Plenum Press, 399-408,

Oozeer M., Veraart, C., Legat, V., Delbeke, J., 2005. Simulation of intra-orbital optic nerve electrical stimulation. *Medical and Biological Engineering Computing* 43(5):608-17.

Ramon, C., Schimpf, P.H., Haueisen, J., Holmes, M, and Ishimaru, A., 2004, Role of soft bone, CSF and gray matter in EEG simulations. *Biomedical Engineering Online* 16(4):245-8

Ramon, C., Schimpf, P.H and Haueisen, J., 2006. Influence of head models on EEG simulations and inverse source localizations. *Biomedical Engineering Online* 5:10

Schimpf PH, Haueisen J, Ramon C, and Nowak H., 1998. Realistic computer modeling of electric and magnetic fields of human head and torso. *Parallel Computing* 24:1433-1460.

AUTHORS BIOGRAPHY

Egill Axfjord Fridgeirsson is a 24 year old master student at Reykjavik University Iceland in Biomedical Engineering. He is on his second year of graduate studies and this work represents part of his master thesis. His research interests are Medical modeling, Biomagnetism and clinical applications of medical models. He currently works in the Clinical Engineering department of Landspítali University hospital.

Paolo Gargiulo is assistant professor at Reykjavik University and works as biomedical engineer and researcher at the University Hospital of Iceland. He studied electrical engineering at University Federico II

in Naples, and obtained a PhD at Technical University of Vienna, Austria. His research interests are in the field of electrical stimulation medical modeling and rapid prototyping for clinical applications.

Ceon Ramon obtained his Ph.D. in Electrical Engineering from University of Utah in 1973 specializations in lasers and quantum optics. For the past 20 years, his research efforts have been largely involved with neuroscience and the developmental genesis of human EEG under normal and epileptic conditions. For the past 40 years, Ceon has held teaching and research appointments at the University of Utah, State University of New York, Stony Brook and at the University of Washington. Presently, he is an Affiliate Professor of Electrical Engineering at the University of Washington and a Professor of Biomedical Engineering at Reykjavik University, Iceland.

Jens Haueisen received a M.S. and a Ph.D. in electrical engineering from the Technical University Ilmenau, Germany, in 1992 and 1996, respectively. From 1996 to 1998 he worked as a Post-Doc and from 1998 to 2005 as the head of the Biomagnetic Center, Friedrich-Schiller-University, Jena, Germany. Since 2005 he is Professor of Biomedical Engineering and directs the Institute of Biomedical Engineering and Informatics at the Technical University Ilmenau, Germany. His research interests are in the numerical computation of bioelectric and biomagnetic fields and biological signal analysis.

An Agent-Based Information Foraging Model of Scientific Knowledge Creation and Spillover in Open Science Communities

Özgür Özmen^(a), Levent Yilmaz^(b)

M&SNet: Auburn Modeling and Simulation Lab

Samuel Ginn College of Engineering

Auburn University

^(a) ozo0002@auburn.edu, ^(b) yi.lmaz@auburn.edu

ABSTRACT

Motivation and problem-domain preferences of scientists can affect aggregate level emergence and growth of problem domains in science. In this study, we introduce an agent-based model that is based on information foraging and expectancy theory to examine the impact of rationality and openness on the growth and evolution of scientific domains. We simulate a virtual socio-technical system, in which scientists with different preferences search for problem domains to contribute knowledge, while considering their motivational gains. Problem domains become mature and knowledge spills occur over time to facilitate creation of new problem domains. We conduct experiments to examine emergence and growth of clusters of domains based on local interactions and preferences of scientists and present preliminary qualitative observations.

Keywords: information foraging, agent based modeling, science of science, open innovation, knowledge spillover

1. INTRODUCTION

Knowledge spillovers are studied widely in literature and linked to innovation measures and outputs (Jaffe, 2008; Feldman and Audretsch, 1996). Spillovers are defined as the migration of knowledge beyond the domain borders (Fallah and Ibrahim, 2004). In our study we express spillovers not only in terms of knowledge transfer but also mobility of scientists. Hence, spillovers result in formation of new domains as a consequence of knowledge and skill transfer.

A research question of interest is to find out connection between individual rationality and aggregate efficiency. Axtell and Epstein (1999) discuss empirical data, which demonstrate that all individuals should not necessarily be rational to produce efficiency in macro level outcomes of a system. Given that individual rationality is bounded, they explore how much rationality should exist in a system to generate macro-level patterns. In this work, we do not propose to discern minimum level of rationality, but rather aim to address how rationality affects the spread of knowledge spillovers as well as growth and development of domains.

Scientists join or leave a problem domain on the basis of problems to be solved and tasks to be accomplished, and their position in the scientific landscape depends upon their knowledge, levels of interest, personal learning objectives, resources, and commitments (Hollingshead et al., 2002). Moti-

vation of scientists as a personal interest is one of the main indicators for willingness of contribution. We take individual motivation as a main force driving the commitment of a scientist to a problem domain.

Metaphorically, scientists can be viewed as predators. Predators are expected to abandon their current patch (e.g., domain) when local capture rate (e.g., problem solving success) is lower than estimated capture rate in the overall environment (Bernstein et al., 1988). Information foraging theory, which is derived from this evolutionary phenomenon developed by Pirolli and Card (1999) assumes that people, if they have an opportunity, would adjust their strategies or the topology of their environment to maximize their rate of information gain. In our study, scientists join or abandon problem domains based on perceived cues about their performance in attaining the desired outcome. The cues are represented by the “instrumentality” component of an individual’s motivation, which is described in detail in section 3. Also, in our model, motivated and successful scientists recruit new scientists just as genetically more adapted predators are more likely to have a offspring in their natural environment.

Interaction surface between scientists and information repositories in real life determines the time costs, resource costs, and opportunity costs of different information foraging strategies (Pirolli and Card, 1999). It is also suggested that people are selfish and lazy in applying their cost-benefit analyses (Nielsen, 2003). In accordance with these observations, we define three different characteristic of the scientists in relation to their cost- benefit decisions.

David (1998) defines the force of Open Science’s universalist pattern as providing entry into scientific artifacts and open discussion by all participants, while promoting “openness” in regard to new findings. Carayol and Dalle (2007) explain open-science phenomenon as significant freedom of scientists to choose what they want to do and how they want to do.

In light of these observations, we present an agent based model, called “KnowledgeSpill” to create a virtual environment where scientists have limited omniscience. In the model, opportunities in a particular problem domain deplete over time. We visualized the impacts of individual rationality and openness on the growth of scientific domains. In section 2, we present the conceptual basis for our model. In sections 3 and 4, we describe the model structure and mechanisms in detail along with the initial parameter values and description of the visualizations. Section 5 discusses preliminary results

and the qualitative observations derived from them. In section 6 we conclude by summarizing our findings in relation to reviewed literature and suggest potential avenues of future work.

2. BACKGROUND AND RELATED WORK

It is demonstrated that user innovation communities are self-organizing complex adaptive systems (Yilmaz, 2008). However, not all complex systems are self-organizing (Monge and Contractor, 2003). A system is self organizing when the network is self-generative (e.g., spawning agents), there is mutual causality between parameters, imports energy into system (e.g., creating new problem-domains and opportunities) and is not in an equilibrium state.

Agent based modeling is widely used to study such complex systems, and it is especially helpful to understand, explore, and parameterize those systems (Bonabeau, 2002). Monge and Contractor (2003) describe the main elements of complex systems in terms of the network of agents, their attributes or traits, the rules of interaction, and the structures that emerge from these micro-level interactions. Typical classes of agent traits are location, capabilities, and memory.

Simulation is used to study scientific domains by different scholars. For instance, Nigel (1997) introduces a model to determine whether it is possible to reproduce observed regularities in science using a small number of simple assumptions. His model generates knowledge structures consistent with observed Zipf distributions involving scientific articles and their authorship. Naveh and Sun (2006) explore the effects of cognitive preferences on the aggregate number of scientific articles produced and argue that simulations with credible cognition mechanisms may lead to creativity in academia.

In scientific knowledge generation, resources may include knowledge, people with skills and abilities, financial support and/or access to tools and raw materials (Mohrman and Wagner, 2008; Powell et al., 1996). The following assertion is stated by Carayol and Dalle (2007): “When the discipline grows, the relative rewarding of problems located in already developed fields increases: because their audience becomes larger, contributions to such domains are more likely to be cited.” Another point expressed in Carayol and Dalle (2007) is that more specialized disciplines are more likely to get more specialized through time, and this phenomenon is more striking when the scientists are more focused on rewarding areas. We describe this phenomenon as *imitation behavior*. Furthermore, scientists make their decisions according to their perceptions and anticipations of their own performance. It is suggested that, intrinsic task motivation plays a critical role in creativity and innovations (Amabile, 1996; Kanter, 1988). Expectancy theory (Vroom, 1964) is a widely known popular model of motivation based on intrinsic motivation. The *rationality* perspective explored in this study can be defined as the behavior towards the maximum motivational outcome based on this theory.

3. CONCEPTUAL MODEL

The location of a scientist in our simulation context is a metaphor for disciplines. Scientists can not migrate between domains or become actively involved within the problem-domains in different disciplines without proper orientation and enculturation. So, the location can be perceived as a discipline and scientists can move in a limited environment at a given time during simulation. There are two basic agent types in the model:

- *Scientists* quest for knowledge in different problem domains. Scientists are members of different disciplines and are more likely to be aware of the problems within their area of expertise. Transferring to other disciplines or areas is difficult because of entrance threshold and the need for enculturation prior to making acceptable contributions. In our model, enculturation is interpreted as the process of searching environment to find a problem-domain to contribute. Our simplifying assumption is that scientists initially move randomly and have an omniscience of 1 cell around, known as Von Neumann neighborhood. A change in the area of awareness and shift to a more open environment with a lower entry threshold facilitates browsing a wider scope in the knowledge space.
- *Problem Domains* are different areas in disciplines (e.g., database management in computer science). Their maturity reflects the knowledge level. As maturity grows, the receptivity of the domain decreases due to higher levels of inertia in mature domains. Increased inertia results in knowledge spillover. Scientists migrate to new problem-domains formed as a result of these spillovers. (e.g., data-mining emerges via transfer of scientists from database management and machine learning fields)

3.1. Maturity and Receptivity

Maturity of a problem-domain is the knowledge level of that particular domain and increases with each contribution. This is represented in the model initially by assigning to a domain a randomly selected maturity value between (0, 0.5]. At every time tick, maturity level of domain i at time $t + 1$ is defined as follows:

$$M_{i,t+1} = \text{Min}(M_{i,t} + \sum_{j=0}^{C_i} \gamma_j, 1) \quad (1)$$

where C_i is the number of accepted contributions at time t in domain i and γ_j is the incremental increase on knowledge repository in the domain caused by each contribution which is drawn between 0 and maximum increment value. Receptivity R at time $t + 1$ of domain i is defined as:

$$R_{i,t+1} = 1 - M_{i,t} \quad (2)$$

3.2. Traits of Scientists

The area of perception (openness) and memory of past successes can be thought of as elements of scientist’s traits. The

choice of the new domain in the area of awareness (scope) of a scientist is based on their preference. We assume that a scientist belongs to a character group with certain probability. There are three different groups:

- *Rationals* look for the least mature problem domain to contribute in their scope. Our interpretation is that less mature domains are more likely to accept contributions, so scientists are more likely to gather experience, early reputation, and motivation by working in those domains.
- *Imitators* have the characteristic of being influenced by the trend or crowd. They look around for a domain to contribute, but are more likely to choose crowded ones. Preferential attachment (Barabasi and Albert, 1999) is used as the guiding principle.
- *Random scientists* randomly choose their problem domains.

Scientists can be in two basic states. They can be “free” not working on a domain or can be “active” by contributing to a domain. Free scientists are searching their scope and when they find one or more domains, they make the decision based on the aforementioned traits above. Scientists, who are not free, contribute to the particular domain they reside on. At every time interval, they have fixed 10% chance of contribution, and once they contribute, the domain decides to accept it or not based on its receptivity level. Scientists also stop practicing when they reach to their maximum age.

As the simulation unfolds, experienced and highly motivated scientists spawn new scientists who are in an initial state. A highly motivated scientist inspires a new scientist with a fixed probability of 0.1. Also, each scientist has a susceptibility level. The larger the tenure of the scientist, the lower its susceptibility level. Susceptibility starts from a higher level and decreases exponentially over time but not below a certain threshold, which is different for each scientist. Susceptibility is defined as the following:

$$S_{i,t+1} = \beta_0 + \frac{1}{2} e^{-\beta_1 \times A_i} \quad (3)$$

where $S_{i,t+1}$ is the susceptibility level of scientist i at time $t + 1$, β_0 is the lowest susceptibility level of a scientist, β_1 is the function parameter and A_i is the current age of the scientist i . Susceptibility is used to determine experience increment after every successful contribution. According to this definition experience of scientist i at time $t + 1$ is:

$$E_{i,t+1} = E_{i,t} + S_{i,t} \times \lambda_s \quad (4)$$

where λ_s is the success increment determined initially and fixed. Success increment can be described as the gain of experience after a successful contribution.

3.3. Motivation Theory and Memory

In expectancy theory, motivation is defined as a product of three factors: how much one desires a *reward* (*valence*), one’s

estimate of the probability that effort will result in successful *performance* (*expectancy*), and one’s estimate that performance will result in receiving the *reward* (*instrumentality*). It is given with the formula below:

$$\text{Motivation} = \text{Expectancy} \times \text{Instrumentality} \times \text{Valence} \quad (5)$$

We assume valence as a fixed value and is different for each scientist, while expectancy can be perceived as the experience of an individual. The experience is described as the attained level of skill through successful contributions. That is, the more experienced the scientists are, the more likely to be motivated they are. Instrumentality is a dynamic parameter which is updated over time, as it denotes the estimate of the award for successful performance. Scientists have a time window of θ which is used to calculate success rate below:

$$Sr_{i,t+1} = \frac{N_{i,t+1}}{\theta} \quad (6)$$

while N is the number of successful contribution of scientist i during the time-window and θ is the pre-defined time-window length.

Also there is a memory factor α , which is different for each scientist. It denotes the impact of success rate on instrumentality and indicates the significance and weight of the current observation with respect to prior experience. A small value of α results in a conservative behavior by avoiding overriding of the past experience. The relative weight of past and present capture rate which is controlled by a parameter denominated the ‘memory factor’ is seen as a common approach in psychological models of simple learning (Bernstein et al., 1988). Following is the formula for instrumentality of scientist i at time $t + 1$

$$I_{i,t+1} = \text{Min}(1, (\alpha_i \times Sr_{i,t+1} + (1 - \alpha_i) \times I_{i,t})) \quad (7)$$

4. IMPLEMENTATION AND VISUALIZATION

Our model is coded in the RePast (Recursive Porous Agent Simulation Toolkit) environment, which is a free and open-source agent based modeling toolkit. Illustration of the aforementioned concepts are analyzed by simulating the spillovers over time. In the figures below, we illustrate that a mature domain spills over to the location on its north-east creating a new problem domain while also transferring 10% of its scientists. The brightness of the color represents maturity. The brighter the color of a domain is, the more is the maturity of that particular domain.

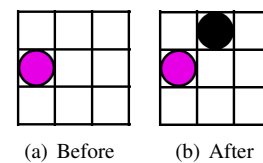


Figure 1. Creation of a new domain

Table 1. Initial settings of the model

Parameter Name	Initial Value
Rational rate	0.75
θ (Time window)	1
Imitator rate	0.2
Maximum Age	3000
Random rate	0.1
Minimum Age	1500
λ_s (Success)	0.01
Run time	3000
Contribution rate	0.1
Initial number of Scientists	500
Maturity threshold	0.8
Initial number of Domains	10
Spillover rate	0.1
World Height	50
World Width	50
Transfer rate	0.1
Motivation threshold(for spawning)	0.95
Initial Age	0
Spawning rate	0.1
Initial Experience	(0,0.5]
Motivation threshold(for leaving)	0.05
Initial Valence	(0.5,1]
Initial Maturity	(0,0.1]
β_0	(0,0.3]
β_1	-0.001
Initial Instrumentality	(0,1]
α (Memory)	(0,1]
Increment of Maturity	(0,0.002]

In order to understand the illustration better, snapshots of the model over time are shown in Figure 2. Also, Table 1 lists the initial values of the parameters of the model.

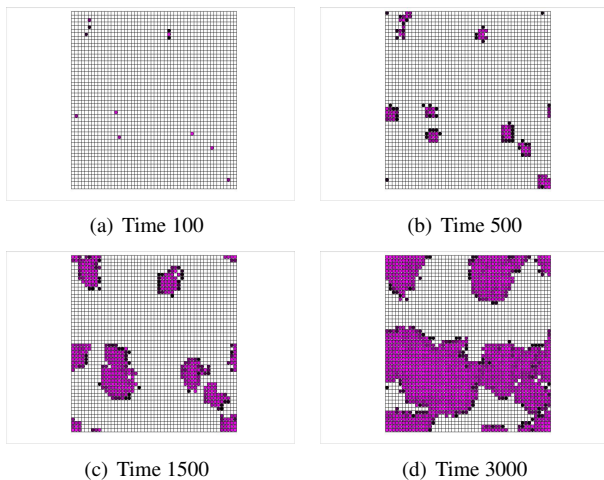


Figure 2. Emergent patterns over time

5. PRELIMINARY EXPERIMENTS AND OBSERVATIONS

We simulate the creation of knowledge (spillovers) under various scenarios observing the spread of domain structures, proportion of mature domains, proportion of the population with regard to motivation levels and the spillovers occurred over time. We examine the implications of the model under 9 basic scenarios, which are created by 3 dimensions of rationality and 3 dimensions of the openness as shown in Table 2.

Table 2. Main Scenarios

Rationality	Low	Medium	High
Rational	25%	50%	75%
Imitator	65%	40%	15%
Random	10%	10%	10%
Scope	1 cell	5 cells	10 cells

Emergent patterns represented in figures 3, 4 and 5 are recorded at time tick 3000. 3000 time ticks can be interpreted as a scientist’s maximum work-life, which we approximate as 60 years. Each time tick denotes 1 week. The first interesting observation is that the growth direction of the clusters changes with the ratio of the rationals. Another expected result is that new domains (less mature problem venues) are grouped at the edge of the main clusters.

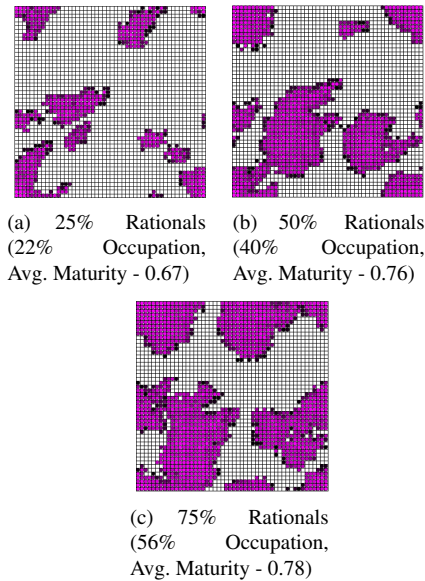


Figure 3. The visualization of the domain clusters for each rationality level after 3000 time ticks. Occupation rate indicates how many percent of the grid is occupied by domains and Avg. maturity is the average maturity level of all domains. This Scenario is with Scope of 1 cell.

When the proportion of rationals increases from 50% to 75%, the increase of the average maturity slows down. More interestingly, when agents can browse wider scope for opportunities (e.g., open-science cases) the occupation rate increases, but 5-cell scope results in higher rates of active do-

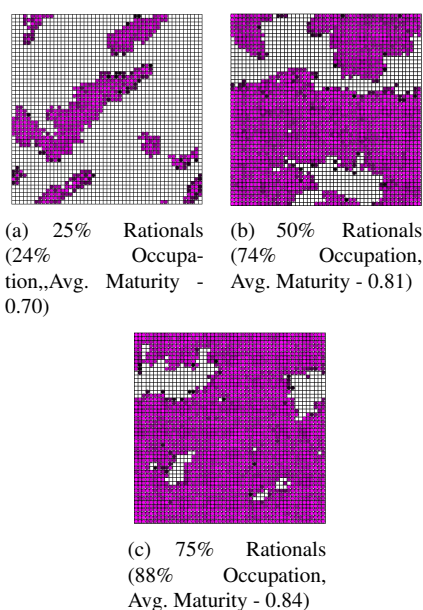


Figure 4. Scenario with Scope of 5 cells.

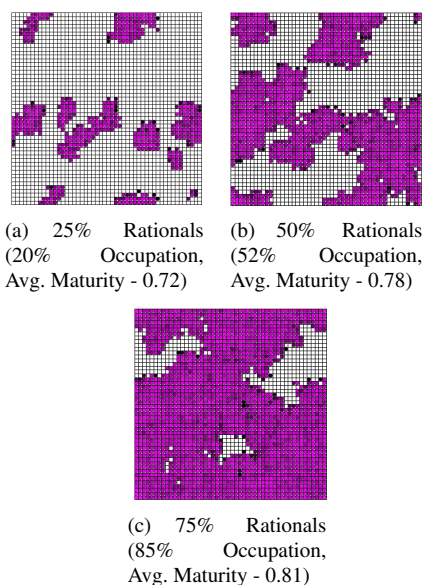


Figure 5. Scenario with Scope of 10 cells.

main occupation in comparison to the 10-cell scope. This phenomenon needs further examination. The degree of openness matters and there seems to be diminishing return between openness and the development of the domains. When the scope increases to 10-cells, increased levels of rational scientists does not yield higher levels of domain growth as compared to 5-cell scope.

In Figure 6, we examine the distribution of maturity levels across domains. The plots for different population characteristics suggests that there are more mature domains in the case of populations with higher rationals, while there are almost same amount of less mature domains in all scenarios. The graph shifts up with more omniscience scientists. The 5-cell

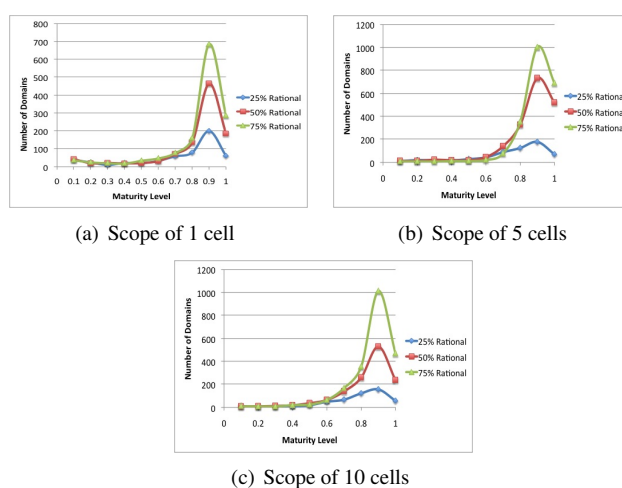


Figure 6. Number of domains vs. the range of the maturity levels for each rationality level

(e.g., moderate openness) case outperforms all other cases regarding knowledge creation.

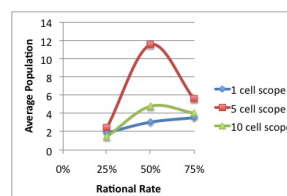


Figure 7. Average population per domain with different rational rates and different scopes

Although our expectation is to observe highly populated domains when the rationality is lower, increased levels of rational agents resulted in more spillovers, and spawning, and hence more number of scientists start practicing in the context. However, as shown in Figure 7, there is a level of diminishing returns. At moderate levels of openness, more scientists are practicing due to increasing spawning rates when the rationality is set at 50%. Beyond this point, average population per domain starts decreasing. To better understand this observation, we examine distribution of motivational levels among the members of the scientific community.

It is expected that increased rationality within a community yields higher levels of motivation throughout the population. In Figure 8, we can observe that the 5-cell case has the most motivated scientist population at 50% rational rate. Our interpretation is that when the scope gets larger, it diminishes the effect of rationality and at around level of 50% rationals, moderate level of openness results in higher levels of motivation. Furthermore, when we examine the distribution of scientists across domains, we observe that 20% of the domains host around 80% of the scientists under each scenario. Finally, as shown in Figure 9, to determine whether power law distribution over spillovers exists, we generate log-log plot of spillovers and their frequency. The results are indicative of existence of a power-law; that is, the frequency of spillovers

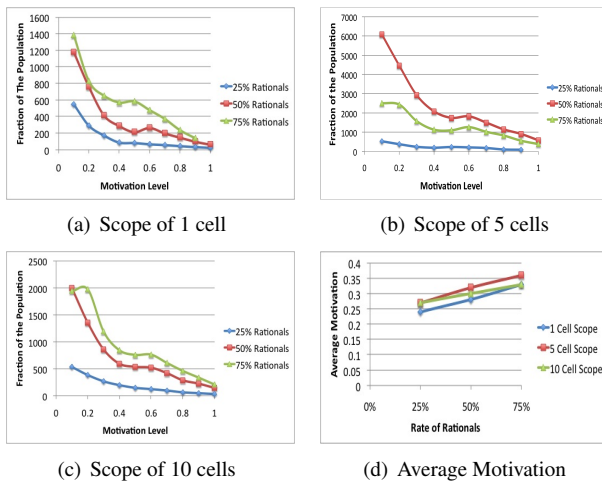


Figure 8. (a), (b) and (c) illustrate the fraction of the population in different range of motivation levels for each rationality level (d) Average motivation of a scientist for each rationality level and scope

is inversely proportional to its size.

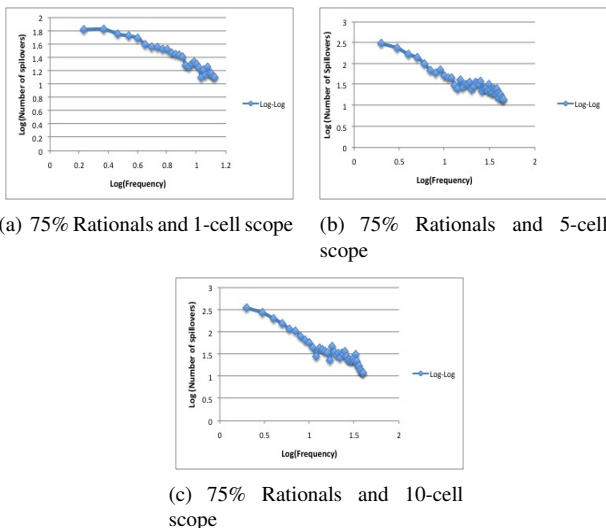


Figure 9. Logarithm of the number of spillovers vs. logarithm of the frequency of that number

6. CONCLUSION

In this study, our main objective is to adopt and analyze the implications of computational mechanisms of well known theories such as information foraging and of behaviors such as motivation, susceptibility, and maturity on the growth and development of scientific domains. Preliminary observations indicate that with more rational populations, the allocation of efforts are distributed more efficiently, resulting in faster growth of domains and a community climate indicative of high motivation. These implications are consistent with our definition of rationality and expectations from information

foraging theory. But when we increased the degree of openness, the growth was not significantly fostered with higher rational population size. Considering population dynamics, with less rational population (e.g., more imitators), our expectation is to observe dense domains; however, as a result of increasing motivation and spawning, increased rates of activity occurred rather in high rationality populations. The future work will explore if there is a diminishing return of openness in terms of motivation and number of mature domains. Potential extensions of the model include ecological aspects that relate to allocation of funds and resources across emergent domains. As feedback mechanisms, funding policies and their effects on the growth of clusters would be an interesting avenue of future research.

REFERENCES

- Amabile, T. (1996). *Creativity in context: Update to the social psychology of creativity* (boulder, co. Westview Press.
- Baer, J. (1998). *The case for domain specificity of creativity. Creativity Research Journal 11*, 173–178.
- Axtell, R. and J. Epstein (1999). Coordination in transient social networks: an agent-based computational model of the timing of retirement. *Behavioral dimensions of retirement economics*, 161–83.
- Barabasi, A. and R. Albert (1999). Emergence of scaling in random networks. *Science 286*(5439), 509.
- Bernstein, C., A. Kacelnik, and J. Krebs (1988). Individual decisions and the distribution of predators in a patchy environment. *The Journal of Animal Ecology 57*(3), 1007–1026.
- Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America 99*(Suppl 3), 7280.
- Carayol, N. and J. Dalle (2007). Sequential problem choice and the reward system in open science. *Structural Change and Economic Dynamics 18*(2), 167–191.
- David, P. (1998). Common agency contracting and the emergence of “open science” institutions. *The American Economic Review 88*(2), 15–21.
- Fallah, M. and S. Ibrahim (2004). Knowledge spillover and innovation in technological clusters. *Proceedings, IAMOT 2004 Conference, Washington, DC*.
- Feldman, M. and D. Audretsch (1996). Location, location, location: the geography of innovation and knowledge spillovers. *Discussion Papers FS IV 96 28*.
- Hollingshead, A., J. Fulk, and P. Monge (2002). Fostering intranet knowledge sharing: An integration of transactive memory and public goods approaches. *Distributed work*, 335–355.

- Jaffe, A. (2008). The “science of science policy”: reflections on the important questions and the challenges they present. *The Journal of Technology Transfer* 33(2), 131–139.
- Kanter, R. M. (1988). When a thousand flowers bloom: Structural, collective and social conditions for innovation in organization. *Research in Organizational Behavior*, eds. Staw BM and Cummings LL 10.
- Mohrman, S. and C. Wagner (2008). The dynamics of knowledge creation: Phase one assessment of the role and contribution of the department of energy’s nanoscale science research centers. Technical report.
- Monge, P. and N. Contractor (2003). *Theories of communication networks*.
- Naveh, I. and R. Sun (2006). A cognitively based simulation of academic science. *Computational and Mathematical Organization Theory* 12(4), 313–337.
- Nielsen, J. (2003). Information foraging: Why google makes people leave your site faster.
- Nigel, G. (1997, Jan). A simulation of the structure of academic science. *Sociological Research Online*.
- Pirolli, P. and S. K. Card (1999, Jan). Information foraging. *Psychological Review*.
- Powell, W., K. Koput, and L. Smith-Doerr (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative science quarterly* 41(1), 116–145.
- Vroom, V. (1964). *Work and motivation*.
- Yilmaz, L. (2008). Innovation systems are self-organizing complex adaptive systems. *Association for the Advancement of Artificial Intelligence*.

SIMULATION HIGHWAY – DIRECT ACCESS INTELLIGENT CLOUD SIMULATOR

Egils Ginters^(a), Inita Sakne^(a), Ieva Lauberte^(a), Artis Aizstrauts^(a), Girts Dreija^(a), Rosa Maria Aguilar Chinae^(b), Yuri Merkurjev^(c), Leonid Novitsky^(c), Janis Grundspenkis^(c)

^(a)Sociotechnical Systems Engineering Institute
Vidzeme University of Applied Sciences
Cesu Street 4-2A, Valmiera LV-4200, Latvia

^(b)Universidad de La Laguna
Pabellón de Gobierno, C/ Molinos de Agua s/n., 38200 La Laguna, Spain

^(c)Faculty of Computer Science and Information Technology
Riga Technical University
Kalku Street 1, Riga LV-1050

^(a)egils.ginters@va.lv, ^(b)rosi@isaatc.ull.es, ^(c)merkur@itl.rtu.lv

ABSTRACT

Simulation nowadays plays an increasing role in the assessment of possible solutions and situation analysis. However, the tools used and previously created models are often incompatible and territorially distributed. Domain specialists lack the expertise and specific programming skills to use them effectively. The development of the Future Internet creates new challenges for simulation engineering by offering extended access. However, primarily a conception is needed, which would create unified access to the simulation architecture on the Future Internet. The paper discusses development of a new conceptual approach to simulation engineering, aiming to support domain experts in performing simulation of complicated socio-technical systems on the Cloud.

Keywords: simulation engineering, simulation highway, Future Internet, Cloud simulation

1. INTRODUCTION

The requirements for situations forecasting and transparency of the scenarios in manufacturing and politics before decision making are topical due to high losses of possible faults and impact on the society and environment.

One of the most effective methods to verify possible solutions while saving financial resources and minimizing security risks is simulation. Simulation is a critically vital component of decision making.

The world is varied, as are the methods of specification that correspond to core systems: continuous, discrete, determined, stochastic etc. Formal languages and analytical methods are used to describe these systems. However, the mentioned specifications can be done by an information technology (IT) professional or systems analyst rather than by a domain expert. All of these applied systems can be simulated by using a predefined set of instruments: discrete-event tools (DEVS), system

dynamics (SD), an agent-based approach (ABM), multilevel models, micro analytical models etc. Naturally, all of these instruments have corresponding popular and concrete simulation tools which have their own modeling languages, formats and rules. Real socio-technical systems are not homogenous, therefore, to make a decision, multiple distinctive simulation models have to be combined in a unified environment. There are exist distributed communication environments like HLA that provide such functionality, however, that is not possible without significant financial investments, even more each case is a custom engineering solution. The development of Future Internet, Internet of Things, Service oriented Architecture and Cloud computing creates a new challenge and possibilities for simulation engineering. The aim of this article is to discuss the project findings related to developing the new conceptual approach to simulation engineering on the Future Internet giving various domain experts an immediate possibility to specify different simulation cases as well as translate, distribute and implement these specifications on the Cloud through the Simulation Highway.

2. SIMULATION AND TOOLS

A system is a set of entities, real or abstract, comprising a whole where each component interacts with or is related to at least one other component and they all serve a common objective. In a system we can always find different types of organization in it, and such organization can be described by concepts and principles which are independent from the specific domain at which we are looking. Socio-technical systems are open to, and interact with, their environments, and that they can acquire qualitatively new properties through emergence, resulting in continual evolution (Von Bertalanffy 1976). A system would be defined as group of objects that are joined together in some regular interaction or interdependence

towards the accomplishment of some purpose (Banks 1996). Otherwise, a system can be defined as a collection of interacting components that receives input and provides output for some purpose (Chang 2004).

The socio-technical systems are tended to self-organization, cognition and continual evolution. The systems can be classified as physical and conceptual or abstract systems, open or closed, continuous or discrete systems, static or dynamic systems, linear or non-linear and deterministic or stochastic systems.

A model of the goal systems is an external and explicit representation of part of reality as seen by the people who wish to use that model to understand, to change, to manage and to control that part of reality (Pidd 1996). Model can be defined as a representation of a system for the purpose of studying the system (Banks 1996).

Simulation is the imitation of the operation of a real-world process or system over time (Banks 1998). Simulation modeling and analysis is the process of creating and experimenting with a computerized mathematical model of a physical system (Chang 2004). Simulation is the way how to research the model.

Nowadays different simulation technologies and mostly non-compatible set of simulation software tools are used. For example, some groups can be mentioned (Gilbert and Troitzsch 2006; A.Bruzzone, A.Verbraeck, E.Ginters et. al. 2002): System dynamics and world (large systems) (DYNAMO, VenSim, PowerSim, STELLA, iThink etc.); Queuing models (discrete-event systems) simulation software (DELSI 1.1., OMNET++, QSIM, QPR Process, EXTEND, SIMPLE++, ARENA, SIMUL8, AutoMod, WITNESS, AnyLogic etc.); Micro analytical simulation solutions (MICSIM, UMDBS, STINMOD, DYNAMOD, CORSIM etc.); Multilevel simulation tools (HLM, MIMOSE, MLwiN, aML etc.); Cellular automata CAFUN, LCAU, CASim, Trilife, JCSim, CellLab, CAGE, SITSIM etc.); Agent-based tools (AgentSheet, NetLogo, SWARM, RePast etc.);

Learning and evolutionary models simulation tools (Artificial Neural Nets, Evolutionary models, Reinforcement Learning (Research simulators - Stuttgart Neural Network Simulator (SNNS), PDP++, JavaNNS, XNBC and the BNN Toolbox for MATLAB etc.; Data analysis simulators - Alyuda NeuroIntelligence, BrainMaker, EasyNN-plus, MATLAB Neural Network Toolbox, NeuralTools, Netlab, Palisade etc.; Component based development environments - JOONE, Peltarion Synapse and NeuroDimension, NeuroSolutions etc.). The simulation software tools (HLA, DIS etc.) ensure elaboration of distributed models. Sometimes also adjacent approaches like CORBA are applied and new distributed simulation environments are elaborated (Aizstrauts et.al 2010).

Unfortunately approaches are very different and simulation platforms and alphabets do not compatible. Up today selection of the simulation software tools are intuitive. Most part of the simulators cannot be used by domain decision makers in real-time and immediate way due to complexity, heavy architecture and special knowledge on programming and mathematics.

3. COMMUNICATION IN SIMULATION

The development of simulation enables the researchers to explore more sophisticated problems at a more detailed level. At the same time this implies a necessity to use several models, or even different modelling tools. Different modelling approaches (multi agent systems, discrete events simulation, system dynamics, etc.) usually envisage the use of different modelling tools (software). This makes the issue of their mutual communication a central concern to the researcher. In Figure 1 development of simulation tools architecture is shown spreading from stand-alone simulators to Web solutions and homogeneous or heterogeneous distributed simulators.

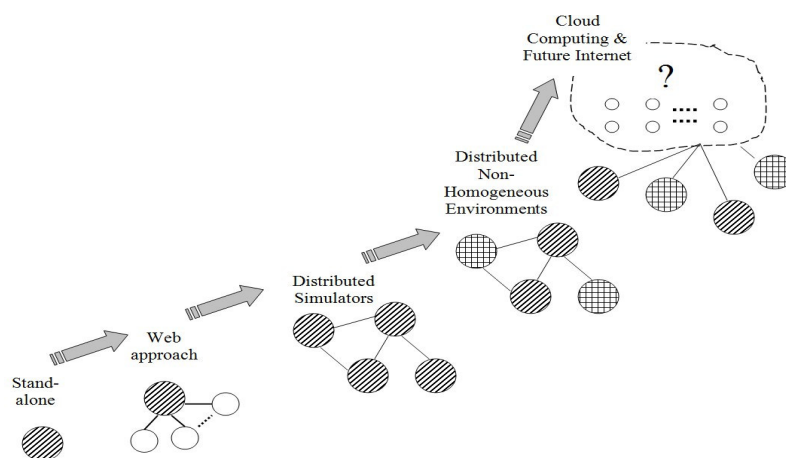


Figure 1: Development of simulation tools architecture

At present, it is very difficult to combine different models made by different simulators because there are no unified descriptions (specifications) for simulation models and a common and well suitable approach for joining simulation tools has not been developed. Therefore, designers cannot co-operate easily when using different simulator models.

Of course, there are several solutions, but they are complicated and not suitable for people without specific knowledge.

This is an important constraint that impedes the development of distributed simulation especially within social and behavioural sciences. These sciences often deal with complicated socio-technical systems that cannot be explored at the sufficient level of quality with one modelling tool.

High Level Architecture (HLA) (Carley 2002) is a concept of the architecture for distributed simulation systems. HLA ensures interoperability and reuses among simulations. It consists of rules that separate parts of distributed simulation model (federates) must follow to achieve proper interaction during a federation execution (see Figure 2); Object Model Template that defines the format for specifying the set of common objects used by a federation (federation object model), their attributes, and relationships among them; Interface Specification, which provides interface to the Run-Time Infrastructure, which can be distributed and ties together federates during model execution.

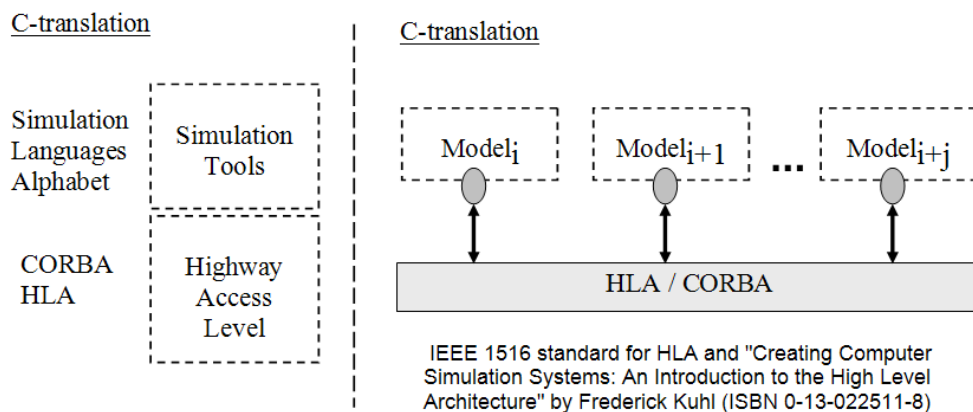


Figure 2: Communication in distributed simulation model

The distributed time management can be done, because all federates' nodes directly undertake synchronization roles. Therefore, the total simulation takes less time and the system is safer, unfortunately, implementation is more complex and laborious.

HLA was created for certain purposes and it was developed according to the needs of industry, where HLA was used. Accordingly innovations of HLA were encapsulated and less attention was paid to the developments within the field of simulations. HLA is a powerful tool for complex simulation systems. But as the technologies develop, the scope and functionality of simulations becomes wider. Inevitably some spheres emerge where HLA does not meet the needs anymore, it is not always convenient. Among one of shortcomings of HLA is its complexity. The wide variety of HLA functionality is rarely needed, besides expenses of HLA and its implementation are rather high (Aizstraus et.al 2010). Some different protocols and methods (CORBA, HLA, FIPA, ALSP, DIS and other) (Strassburger 2006; Verbraeck 2004; Bolton 2001) exist for elaboration the communication environments, unfortunately any of it has their own disadvantages and right selection is still problematic.

One of the most important problems of the existing communication tools (environments) is compatibility with the architectural solutions of the Future Internet, because environments are non-flexible, and with low adaptability to the requirements of the SoA and the Future Internet of services.

4. THE FUTURE INTERNET AND THE CLOUD

Over the last decade, most Europeans and more than two billion people worldwide have become Internet users. More than 67.6% in the European Union and almost 26% people worldwide use the Internet. In 2010 the number of Internet users in China reached 400 million, which is more than the population of the entire United States (Blackman et.al 2010). With over 1 trillion pages and billions of users the Web is one of the most successful engineering phenomena ever created. At the end of 2009 there were 234 million websites of which 47 million were added in the year. The Web is now a rich media repository, the current upload to Flickr is equivalent to 30 billion new photos per year and YouTube now serves over 1 billion videos per day (Pedrinaci 2010).

The rapid development of mobile technology, providing Internet access to individual remote devices (Internet of Things), has accelerated the pace of development; as a result it is expected that the number of users in 2020 will reach seven billion. The Internet moved from the technical to the social category, where development is more driven by the efforts of interested Internet users than the pressure of technological achievements (Pedrinaci 2010). Some of these users see it as a business environment, others as a provider for social networking opportunities. An inevitable fusion of technology and the social environment can be observed, where the Future Internet is actually a representation of the public, a complex, but integral part of the social system. Future research must conceptualize the Internet as the global social machine. Research shows that (Blackman et.al 2010) technology development is not the key driver of Internet development. The vital part is social factors, which determine that no revolutionary technical changes are possible in the Internet development at the moment. That is why the implementation of new methodologies, algorithms and services gains a special status. They change and improve the Internet step by step by improving the quality of service a user receives i.e. performance, security and intelligence. Simulation is not an exception as it can become one of such services in the Future Internet environment.

Mobile technology has long been offering personal simulation tools such as AgentSheets and others. As a result, it is expected that simulation will become one of the services in the future network Internet of Services, which, in turn, will be promoted by the Internet of Things, which will provide access to remote and mobile equipment.

The Semantic Web is an extension of the current human – readable Web, adding formal knowledge representation so that intelligent software can reason with the information in an automatic and flexible way (Pedrinaci 2010). Semantic Web research has therefore largely focussed on defining languages and tools for representing knowledge in a way that can be shared, reused, combined, and processed over the Web. This research has led to a plethora of standards such as RDF(S), OWL, as well as corresponding tools such as ontology editors, RDF(S) storage and querying systems. The semantic approach can be one of the basic elements in the development of a unified approach for the specification of simulation models and the translation of further constructions to execute them in the Future Internet environment.

It is noted that the human-machine interface plays a special role in the development of the Internet where the emphasis on browser-type access and search-engines gradually moves to Facebook, Twitter and national social network analogues. However, this is just the beginning as 3D immersion and virtual and augmented reality (VR/AR) applications development and introduction (Future Internet 3D) is expected. The

development of this direction could have a direct impact on the progress of simulation engineering by providing the visualization of simulation results, which could gradually replace built-in visualization tools, contributing to the unification of a simulation approach.

There are different types and practices of standardization: with ISO-styled standardization, the process may be heavy-handed; or IETF types – global and motivated by desire to keep the internet running effectively, a place where some consensus will be found, and based between ISO and IEEE; or IEEE types of standards – in some ways the opposite of ISO in process and being purely technical; finally we have various Web consortia (e.g. OASIS) becoming even more important since they are high level, including the various open source standards for interfaces and whole applications, which may or may not be normalized (Blackman et.al 2010).

Currently, the most popular methods to describe data and semantic information are considered XML, RDF or OWL, SOAP or REST notations are used for protocol description, but BPEL or BPMN are used to specify orchestration mechanisms. Efforts are being made to make these languages the standard tools for describing Future Internet services, although their possibilities to describe the functioning of a socio-technical system raise serious doubts. At least stochastic process specification could lead to the solving of a series of difficult problems. Opposed to the orchestration approach, the use of choreographies CHOREOS (<http://www.choreos.eu>) seems more actual and promising, because there is no single monitoring and synchronization service, but all members of the service network work independently and within the extent of their competence to reach the determined goal. Such architecture is more viable, as there are no central administration resources, the disruption of which equals the doom of the service.

To describe the essence of Future Internet, the following concept set is offered: Internet of Contents (IoC), which is provided by the Internet of Services (IoS), while the remote access to specific devices is provided by the Internet of Things (IoT), but it is possible that the user network is part of a larger network, and then it is considered the Internet of Networks (IoN). There is a need for both researchers and practitioners to develop platforms made up of adaptive Future Internet applications. In this sense, the emergence and consolidation of Service-Oriented Architectures (SOA), Cloud Computing and Wireless Sensor Networks (WSN) rise benefits, such as flexibility, scalability, security, interoperability, and adaptability, for building applications.

As one of important parts of the Internet of Services the Cloud computing could be mentioned. Cloud computing ensure scalable storage, computation facilities,

application hosting, or even the provisioning of entire applications accessed remotely through the Internet. Some well-known commercial Cloud solutions are for instance Google's AppEngine, Gmail, and Docs, Amazon's Elastic Computing, or Salesforce (Pedinaci 2010). The European Commission supports approximately 140 Future Internet projects to a greater or lesser extent and some research of them are related with Cloud (<http://www.future-internet.eu/activities/fp7-projects.html>), for instance, Cloud4SOA (<http://www.cloud4soa.eu/>) focuses on the semantic interoperability and on introducing a user-centric approach for applications which are built upon and deployed using Cloud resources; Cloud-TM (<http://www.cloudtm.eu/>) working on the development and administration of Cloud applications; CONTRAIL (<http://contrail-project.eu/>) designs an open source system for integration of heterogeneous resources into a single homogeneous Federated Cloud; CumuloNimbo (<http://www.cumulonimbo.eu/>) provides consistency, availability, and simpler programming abstractions, such as transactions; Morfeo 4CaaS (<http://4caast.morfeo-project.org/>) - platform for elastic hosting of Internet-scale multi-tier applications; mOSAIC (<http://www.mosaic-cloud.eu/>) develops a platform allowing to the users to tune Cloud services; OPTIMIS (<http://www.optimis-project.eu/>) establishes an open Cloud Service Ecosystem for adaptable and reliable IT resources support; VISION Cloud (<http://www.visioncloud.eu/>) introduces an infrastructure for reliable delivery of data-intensive storage services and many other projects launched during previous calls.

Cloud-based services because are expected to become one of the main IT market niches and as a consequence many large companies are working on the creation of their own solution to retain a competitive position (The Economist 2009). Due to the reason mentioned above it would be rational to provide Cloud-based simulation service as integral component of the Internet of Services.

5. SIMULATION LAYOUT DESIGN AND VISUALIZATION OF THE RESULTS

One of the essential actualities is the display of simulation results in a form that is understandable to domain specialists and as close as possible to the specifics of the business sector.

Currently, built-in tools with limited functionality are used to display simulation results. Furthermore, it is not clear how to visualize the results of a non-homogenous and distributed simulation.

There are several different definitions for Virtual Reality (VR), but one of the formulations determines that it is the simulation of the goal system using computer graphics and providing the user with the ability to interact with the researchable system by using three and more levels of freedom (Burdea 2003). It all depends on the research object. If reality is the object of research, the virtual constructions complement the

visible object. In the present case, there is talk of an Augmented Reality (AR) solution. If the research object is a virtual object, where a real life existing installation or component is added, then it can be considered an Augmented Virtuality (AV) situation. Currently, virtual reality questions relate to a scientific sub-sector, which combines various fields such as computers, robotics, graphics, engineering and cognition. VR worlds are 3D environments, created by computer graphics techniques, where one or more users are immersed totally or partially to interact with virtual elements. Mainly, special devices stimulate the sight, hearing and touch. Higher immersion level can be achieved with output devices, which are mainly directed to humans visual sense, for example, head mounted displays (HMD), stereoscopic monitors, special glasses, projection walls, CAVE systems, etc. Multiple sound sources positioned provides 3D sound, and touch can be simulated by the use of haptic devices (Moraos and Machado 2009).

The basic element of a VR system is the authoring platform which provides the import of models from other 3D graphics tools (AutoCAD, Maya, 3D Max etc.), the generating and rendering of scenes, and the building and operation of scenarios. Although the construction of VR/AR systems is still expensive and time consuming, which is caused by the incompatibility of hardware and authoring platforms, gradual development is taking place (Ginters et.al 2007). At least VRLM supports most authoring platforms. In recent years VR platform providers (Bluemel 2011) have been working on the creation of tools for simulation layout planning. Initially, VR/AR might be a good supplement to any of the simulation environments to improve the clearness of simulation, but in the near future the agreement between simulation and VR professionals could lead to the development of a unified VR-simulation interface concept. Another useful VR application could be the visualization of simulation results by adapting them to the perception and industry specifics of domains experts. In any case, it is clear that the fusion of VR with simulation environments and tools is a matter of the nearest future.

There are more than twenty EC FP7 projects, which currently do research in the field of virtual reality, however, there are few with a connection to the development of Future Internet, for instance, IRMOS (<http://www.irmosproject.eu>) will design, develop, integrate and validate a Service Oriented Infrastructure that enables a broad range of interactive real-time applications. It will support the development and deployment of real time applications in a distributed way. The infrastructure will be demonstrated by focusing on virtual and augmented reality; VirtualLife (<http://www.ict-virtuallife.eu>) aims to combine a high quality immersive 3D virtual experience with the trustworthiness of a secure communication infrastructure, focusing on the creation of secure and ruled places within the virtual world where important transactions can occur; 2020 3D Media

(www.20203dmedia.eu) is aimed to the development of new technologies to support the acquisition, coding, editing, networked distribution, and display of stereoscopic and immersive audiovisual media, capable of providing novel and more compelling forms of entertainment both for home and for public grounds. The users of the resulting technologies will be both media industry professionals across the current film, TV and 'new media' sectors producing programme material as well as the general public.

For the time being, virtual reality experts are busy with their internal problems and, it seems, are not ready to deploy and adapt their systems for the Future Internet environment. A close cooperation between Future Internet architects and VR/AR ideologists has not been formed because both sides are not ready for serious negotiations, although opinions can be heard, that Web

3.0 will be in 3D. However, it is clear that sooner or later it will happen, and simulation experts should participate in the development of this unified concept.

6. SIMULATION HIGHWAY – THE CONCEPT

Simulation Highway - common approach and rules to deploy, access, join and exploit the different and heterogeneous simulation models in distributed environment on the Future Internet and Cloud.

The Simulation Highway (Ginters and Vorslovs 2008; Ginters and Aquilar 2008) ensures translation and distribution the simulation requests in the Cloud. These simulation requests address a set of simulation cells organizing implementation highway during the simulation session of defined task.

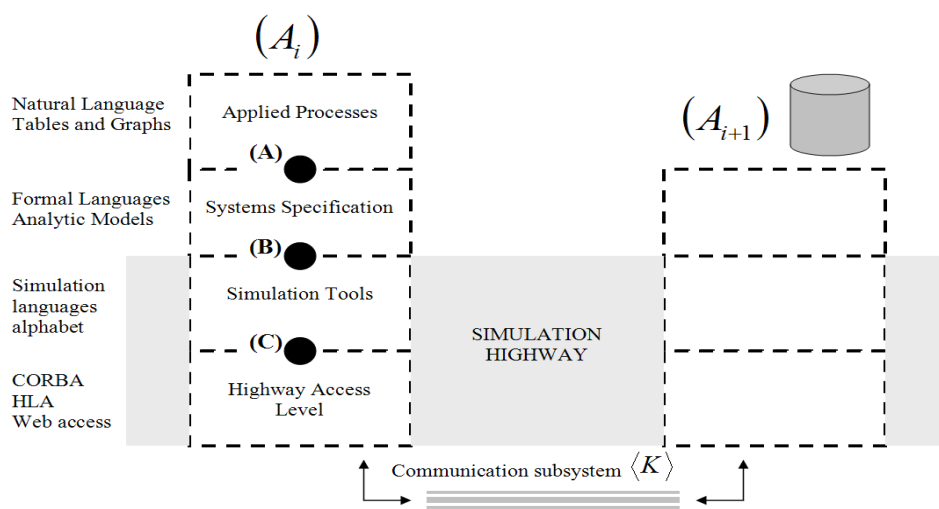


Figure 3: Multilevel Requests Translation and Distribution on Future Internet (Ginters and Aquilar 2008)

Each simulation cell serves as a server and simultaneously as a switch so that various and previously existing simulation models that are registered to the cell and participates in the decision

making task in real-time can be connected to the simulation highway of the task (see Figure 4). Cognition is integral attribute of each simulation cell.

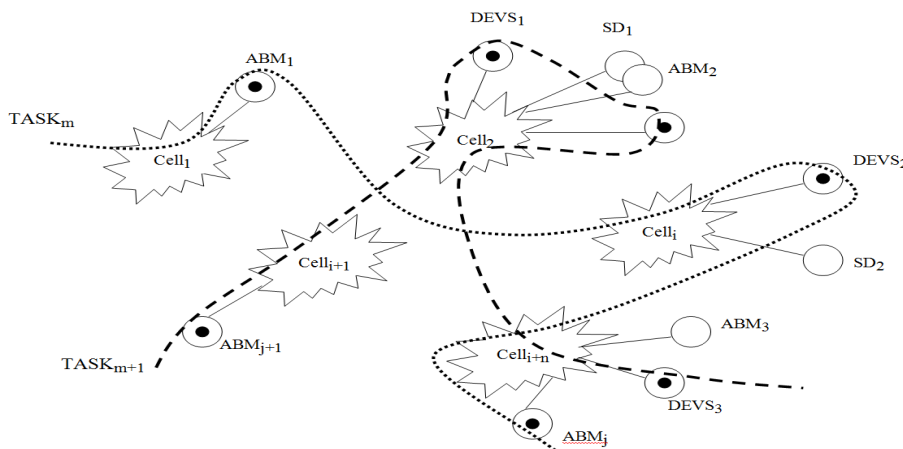


Figure 4: Simulation Highway – Distributed Multiple Access Simulation Environment

To each cell different models: discrete-event (DEVS), system dynamics (SD), agent-based (ABM) can be registered. Each task, formulated by the domain expert, generates a heterogeneous chain of interacting models, or highway. The results are visualized in a manner that is understandable and demonstrative to the client using VR facilities provided by the Future Internet 3D.

The important problem for distributed simulation is placity of the Cloud platforms. For example, Amazon's EC2 supports the Message Passing Interface (MPI), the message-based communications protocol used by parallel programs that run on clusters (Fujimoto et.al 2010). This provides theoretically a variant to making parallel and distributed simulations.

Nevertheless this approach is too general to be successfully used in such a specific field as simulation, because Cloud platforms are better oriented at providing high bandwidth communications among applications for longer sessions than to interchange of many small messages requiring quick delivery. Theoretically it is possible that simulation societies would arrange common Cloud platform suitable for implementation of

the simulation tasks, but such approach would be the step away from the aim, because the benefit is use of the common Cloud solutions by all the tasks. Therefore, selection of the right Cloud platform would be the challenging task.

7. ONTOLOGY BASED ADAPTABLE UNIVERSAL SIMULATION SPECIFICATION LANGUAGE

One of the major problems is a different process specification and the performance of further transmissions, ensuring cooperation with the Simulation Highway.

Existing simulation languages are different and problem-oriented; the abstraction level is low enough. On the other hand, the languages used by software engineers are not suitable to describe real and complex processes. There have been several attempts to achieve universal solutions (see Figure 5).

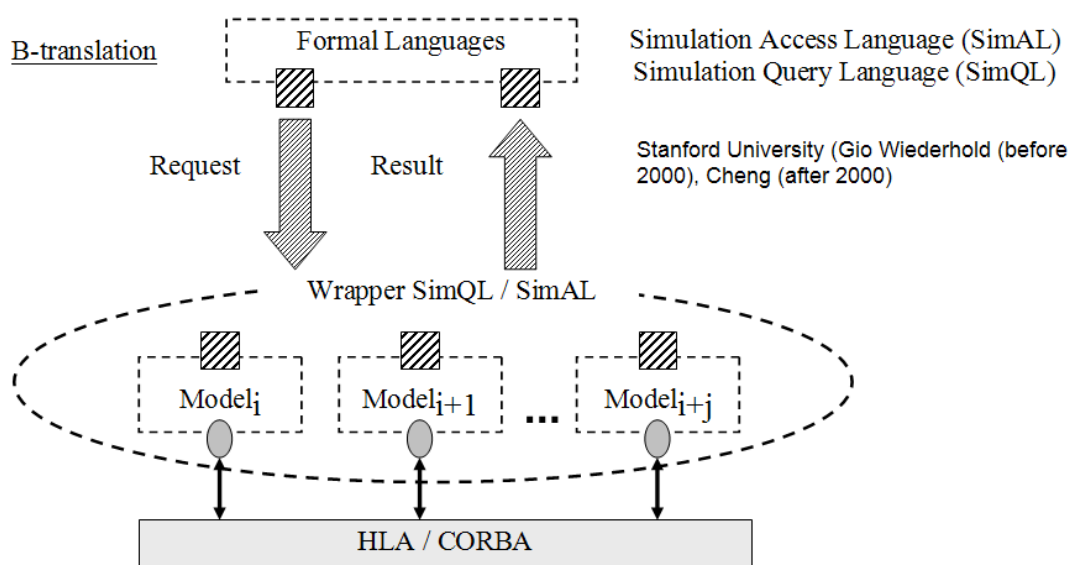


Figure 5: Client-Server Access Model SimAL/SimQL

In the beginning of the 21st century client-server access was offered for simulation models by integrating SQL query analogues, but the proposed solution did not gain wide acceptance, because it was not enough, i.e. access to the simulation environment did not improved for domain experts, because the query still had to be made in what a person without programming skills would consider an unfriendly language.

Ontologies provide formal methods for describing the concepts, categories, and relationships within a domain.

Domain ontologies may be particularly helpful to simulation modellers since they can be used to communicate domain information to simulation and modelling tools with limited human intervention.

Ontology driven simulation (ODS) takes advantage of this feature by using software tools to align knowledge resident in domain ontologies with knowledge resident in a modelling ontology in order to facilitate the creation of simulation models. In ODS, a tool is used to map concepts from domain ontologies to concepts in a

modelling ontology and then create instances of modelling ontology classes to represent a model.

Ontology driven simulation uses ontologies to drive the creation of simulation models and in doing so makes use of an agreed upon set of terms and relationships that are shared by domain experts, modelers, and model development tools. These terms and relationships provide a semantic grounding and structure for the executable model. The domain ontology exists for a specific application area, and its classes and instances determine the types of components that will make up the model.

The modeling ontology is developed independently of any specific domain ontology but may rely on general upper ontologies. Modellers receive the benefit of having a stable set of terms for the domain available through the design tools that they are using.

Domain experts and others who make use of the simulation models benefit by having simulations use a common set of terms with which they are familiar.

Under the framework of the Simulation Highway project is intended to create Ontology Based Adaptable Universal Simulation Specification Language based on a knowledgeable, ontological and semantic approach that would give various domain experts an immediate possibility to specify different simulation cases as well as translate, distribute and implement these specifications in the Cloud through the Simulation Highway.

The language would be based on UML and its followers BPMN and BPEL, and supplemented with the

possibilities to describe stochastic and heterogeneous processes, as well as additions that would provide domain experts with more friendly access. The use of BPMN and BPEL would facilitate the introduction of the new simulation specification language in the Future Internet environment.

8. SIMULATION HIGHWAY AND VISUALIZATION OF THE RESULTS

It is clear that the presentation form of the simulation results should be demonstrative and close to the perceptual characteristics of domain experts. The higher is the level of immersion, the better the quality of the gained knowledge.

The important reason for VR/AR use is the requirements for the quality and the performance of simulation layout design. Of course, in industrial tasks it does not a critical factor like military applications (Smith 2010), but in any case intelligent application of VR facilities would substantially reduce the potential errors and time for the layout design.

The development of virtual and augmented reality (VR/AR) solutions and evolution of simulation environments creates a convergence of both sides where VR/AR will become an integral part of simulation tools. However, the next step is the demand for universalism and sufficiently open access, which could be provided by the Future Internet of Services and Cloud facilities.

This means, that requirements have to be defined and developed for an interface between the Simulation Highway and the set of virtualization tools on the Future Internet 3D (see Figure 6).

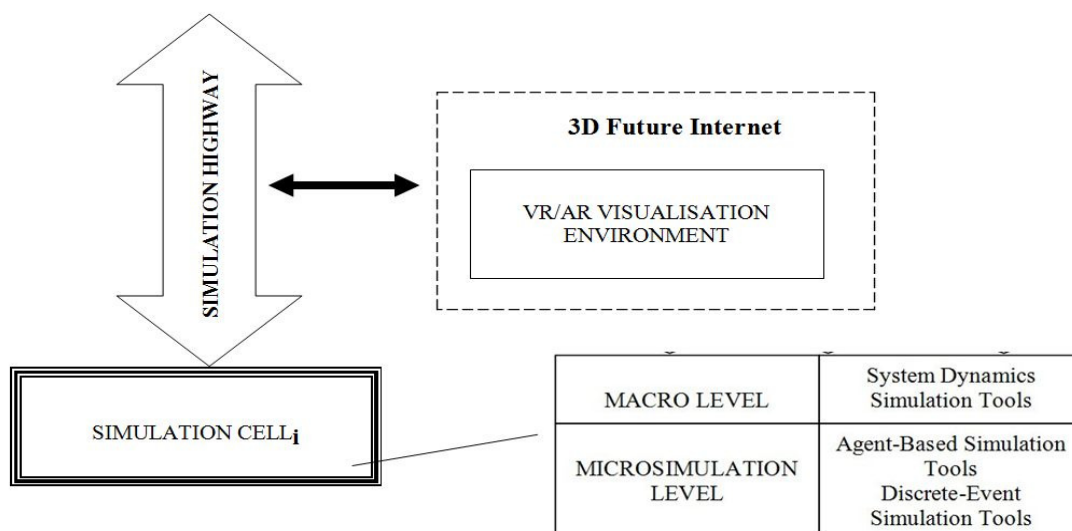


Figure 6: Simulation Highway - part of the Internet of Networks

The Future Internet of Devices (Things) ensuring access to the net for the mobile users dramatically increases the number of potential modellers. However, in the case of

simulation, it must be remembered that the resource capacity of mobile devices (screen, memory, battery) is limited and cannot be compared to stationary users.

9. CONCLUSIONS

Simulation, in its half a century long existence, has achieved structurally diverse and heterogeneous solutions. New possibilities have emerged, which are analogous to software design, such as prototyping, automated model generation and others, and that is why simulation can be defined as simulation engineering.

The diversity and time consuming creation of models claim for requirements by unification and standardization. This can slow down the development of simulation technologies while expanding its scope of application, which could create the possibility of convenient decision making for domain experts.

Unification is doubtful without the creation of a unified approach to the distributed simulation and Ontology Based Adaptable Universal Simulation Specification Language based on a knowledgeable, ontological and semantic approach, which should preferably be based on habitual, well supported and with the Future Internet key persons understandings compatible constructions such as UML, BPEL, BPMN, etc.

The development of the Future Internet solutions is one of the priorities of European research. Therefore, service providers, including simulation professionals, should appropriately focus on solution development for the Future Internet of Services on the Cloud. A major advantage of simulation in the Cloud is that it is scalable. As the number of modellers increases, simulation servers can be added to increase the computational and storage capacity. Serious challenge is selection of the right platform for the Cloud due to real-time requirements of simulation.

The visualization of simulation results has to be done in an understandable and acceptable form, which VR/AR solutions can provide. These have to be integrated into the Future Internet 3D, but the Simulation Highway, as a part of Internet of Networks, has to ensure an interface between VR/AR and simulation models.

The proposed Simulation Highway solution is an attempt to unify the access to heterogeneous simulation models, to provide domain experts with the immediate specification possibilities and to ensure the economy of computer resources by deploying the simulation on the Future Internet of Services. The solution promotes the development tendencies of the Internet of Things, because it intends to deploy simulation tools to mobile communication devices.

Project partners will realize a demonstration tasks which will relate to a decision making simulation on port logistics and policy modeling. Expected results would allow for the consolidation of varied and different previously developed simulation models into a

unified environment and ensure their direct usage by miscellaneous industry domain experts and specialists.

ACKNOWLEDGMENTS

The current article was prepared in the framework of ERAF project No.2DP/2.1.1.2.0/10/APIA/VIAA/001 "Support for preparation of IST FP7 STREP project "Simulation Highway".

REFERENCES

- Von Bertalanffy, Ludwig, 1976. General System Theory: Foundations, Development, Applications, Rev. ed., ISBN-978-0807604533.
- Banks, J., Carson, S. II, and Nelson, B.L., 1996. Discrete-Event System Simulation, 2nd ed., Prentice Hall, Upper Saddle River, N.J.
- Chang, Chris, 2004. Simulation modeling handbook: a practical approach, CRC Press LLC, ISBN 0-8493-1241-8.
- Pidd, Michael, 1996. Tools for Thinking: Modelling in Management Science, ISBN 978-0471964551, 1996, 360p.
- Banks, Jerry, 1998. Handbook of simulation, John Wiley & Sons, ISBN 0-471-13403-1.
- Gilbert, Nigel, and Troitzsch, Klaus G., 2006. Simulation for the Social Scientist. Second Edition. ISBN 139780335221493, Open University Press.
- Bruzzone, A., Verbraeck, A., Ginters, E., et. al., 2002. Logistics Information Systems, Part 1-2, Jumi, ISBN 9984-30-021-8, Riga, 2002, 682 p.
- Carley, K.M., 2002. Computational organizational science and organizational engineering, Simulation Modelling Practice and Theory 10 (2002) 253–269.
- Aizstrauts, A., Ginters, E., Aizstrauta, D., 2010. Step by Step to Easy Communication Environment for Distributed Simulation. //In: *Annual Proceedings of Vidzeme University College „ICTE in Regional Development”*, ISBN ISBN 978-9984-633-20-6, Valmiera: Vidzeme University of Applied Sciences, Sociotechnical systems engineering institute, 2010, pp.11-13.
- Silins, A., Ginters, E., Aizstrauta, D., 2010. Easy Communication Environment for Distributed Simulation. //In: *World Scientific Proceedings Series on Computer Engineering and Information Science 3 “Computational Intelligence in Business and Economics”*/ Proceedings of the MS'10 International Conference, Barcelona, Spain, July 15-17, 2010, ISBN 978-981-4324-43-4, pp.91-98
- Strassburger, S., 2006. The road to COTS-interoperability: from generic HLA-interfaces towards plug-and-play capabilities, in *Proceedings of the 37th conference on Winter simulation*, December 03-06, 2006, Monterey, California.
- Verbraeck, A., 2004. Component-based Distributed Simulations. The Way Forward?, in *Proceedings 18th Workshop on Parallel and Distributed Computer Simulation*. (Kufstein, Austria, 16-19

- May 2004), IEEE Computer Society Press, Los Alamitos, CA, 2004. pp. 141-148.
- Bolton, F., 2001. *Pure CORBA*, Sams, ISBN 978-0672318122, 2001, P. 944.
- Blackman, C., Brown, I., Cave, J., Forge, S., Guevara, K., Srivastava, L., Tsuchiya, M., Popper, R., 2010. Towards a Future Internet. Interrelation between Technological, Social and Economic Trends. Final Report for DG Information Society and Media, European Commission DG INFSO, Project SMART, 2008/0049 November 2010
- Pedrinaci, C., and Domingue, J., 2010. Toward the Next Wave of Services: Linked Services for the Web of Data, *Journal of Universal Computer Science*.
- The Economist. 2009. Clash of the clouds. *The Economist*. 392, 80-82, 2009
- Burdea, G., Coiffet, P., 2003. *Virtual Reality Technology*. 2nd ed., Wiley Interscience, 2003.
- Marcos de Morales, R., Dos Santos Machado, L., 2009. Online Training Evaluation in Virtual Reality Simulators Using Possibilistic Networks.//In: Proceedings SHEWC 2009 Safety, Health and Environmental World Congress, July 26 - 29, 2009, Mongaguá, Brazil.
- Ginters, E., Vorslovs, I., 2008. Simulation Highway for Food Quality Management. *Proceedings of the International Scientific Conference „Applied Information and Communication Technologies”*, Latvia University of Agriculture, Jelgava, Latvia, April 10-12, ISBN 978-9984-784-68-7, 2008, pp.86-95.
- Ginters, E., Aguilar Chinaea, M. R., 2008. Simulation Highway for Applied Systems Management. *Proceedings of the 7th WSEAS Int. Conf. on CIRCUITS, SYSTEMS, ELECTRONICS, CONTROL & SIGNAL PROCESSING (CSECS '08)*, Puerto de La Cruz, Tenerife, Spain, December 15-17, 2008, ISBN 978-960-474-035-2, 2008, pp.180-186
- Towards a Future Internet, 2010. Interrelation between Technological, Social and Economic Trends. *Final Report for DG Information Society and Media. European Commission DG INFSO Project SMART 2008/0049*, November 2010, Available from: <http://www.internetfutures.eu/wp-content/uploads/2010/11/TAFI-Final-Report.pdf>, [accessed 01.04.2011]
- Fujimoto, R.M, Malik, A.W., Park, A.J., 2010. Parallel and Distributed Simulation in the Cloud, *SCS M&S Magazine – 2010 / n3*
- Smith, R., 2010. Simulation in the Cloud, ITEC 2010, London, 18-20 May.

AUTHORS BIOGRAPHY

Egils GINTERS is director of Socio-technical Systems Engineering Institute. He is full time Professor of Information Technologies in the Systems Modelling Department at the Vidzeme University of Applied Sciences. He is a member of the Institute of Electrical and Electronics Engineers (IEEE), European Social

Simulation Association (ESSA) and Latvian Simulation Society. He participated and/or coordinated some of EC funded research and academic projects: FP7 FUPOL project No. 287119 (2011-2014), FP7-ICT-2009-5 CHOREOS project No. 257178 (2010-2013), e-LOGMAR-M No.511285 (2004-2006), SocSimNet LV/B/F/PP-172.000 (2004-2006), LOGIS MOBILE LV/B/F/PP-172.001 (2004-2006), IST BALTPORTS-IT (2000-2003), LOGIS LV-PP-138.003 (2000-2002), European INCO Copernicus DAMAC-HPPL976012 (1998-2000), INCO Copernicus Project AMCAI 0312 (1994-1997). His main field of interests involves: systems simulation, logistics information systems and technology acceptance and sustainability assessment.

Yuri MERKURYEV is Professor and Head of the Department of Modelling and Simulation at Riga Technical University in Riga, Latvia. His professional interests include methodology of discrete-event simulation, supply chain simulation and management, as well as education in the areas of simulation and logistics management. Prof. Merkurjev is a corresponding member of the Latvian Academy of Sciences, President of Latvian Simulation Society, Board Member of the Federation of European Simulation Societies (EUROSIM), SCS Senior Member and Director of the Latvian Center of the McLeod Institute of Simulation Sciences, and Chartered IT Professional Fellow of the British Computer Society. He authors about 300 scientific publications, including 6 books, and is a co-editor (with Galina Merkurjeva, Miquel Angel Piera and Antoni Guasch) of a recently published by Springer-Verlag book “Simulation-Based Case Studies in Logistics: Education and Applied Research”. He is editorial board member of several journals, including “Simulation: Transactions of the Society for Modeling and Simulation International,” and “International Journal of Simulation and Process Modelling”. Prof. Merkurjev regularly participates in organising international conferences in the area of modelling and simulation. In particular, he has served as General Chair of the International Conference "European Conference on Modelling and Simulation", ECMS'2005. He is permanently involved into organising of the HMS (The International Conference on Harbour, Maritime & Multimodal Logistics Modelling and Simulation) series of conferences within the annual International Mediterranean and Latin American Modelling Multiconference, I3M.

Rosa Maria AQUILAR CHINEA is Professor and Vice-Rector at Universidad de la Laguna in Santa Cruz de Tenerife, Spain. His professional interests include ontologies of discrete-event simulation environments. Prof. Aquilar Chinaea regularly participates in organising international conferences in the area of modelling and simulation. In particular, he has served as Chair of the International Mediterranean Modelling Multiconferences EMSS 2006 and other.

SIMULATION OF GESMEY GENERATOR MANOEUVERS

Amable López^(a), José A. Somolinos^(b), Luis R. Núñez^(c), Alfonso M. Carneros^(d)

^{(a) (b) (c)} Universidad Politécnica de Madrid, GIT-ERM R&D Group
^(b) F.C.T. SOERMAR

^(a)amable.lopez@upm.es, ^(b)joseandres.somolinos@upm.es, ^(c)luisramon.nunez@upm.es, ^(d)alfonso.carneros@soermar.es

ABSTRACT

GESMEY generator is a Tidal Energy Converter -TEC- device designed to operate in medium and high depth locations with an original design based on a moored system and a main structure of star aspect. After a brief introduction where the interest of this type of generators for the marine currents energy harnessing is presented, the objectives of the GESMEY Project, the design procedure and some results and the automation process scheduled for moving it from the submerged operation state to the floating maintenance situation are presented too. The needs of new models and tools for the study of the dynamical response of these kind of devices under these operation states together with the design solutions that have been taken in this original design are described. Finally, simulation results of some of the GESMEY manoeuvres obtained with a tool we have developed and with some commercial simulation tools are presented. The comparative study of both simulation responses validates the simplest dynamical model used for controlling the generator.

Keywords: marine current converter, moored device, manoeuvres modelling and simulation.

1. INTRODUCTION

Marine currents are one of the most promising marine renewable energy -MRE- sources that mainly is derived from tides movement. The main advantages for harnessing tidal energy are:

- Specific locations in the oceans with high energy density, mainly near shore.
- Reliable long-term prediction of speed and power.
- Better relationship between mean and nominal power than other MREs.
- Very low environmental impact.
- High reliability compared to other devices like wave converters.

The energy that could be extracted from ocean currents is estimated around 800 TWh/year -about 4% of global electricity consumption- (IEC 2011), but currently it is not possible to exploit the most important part of this huge energy potential since most of this

energy -about 80%- is concentrated inside areas with depths over 40 meters. Then, it is necessary a second generation of converters capable of extracting this energy from these high depth sites.

At this moment the development of TEC devices for the stream exploitation, is focusing on the first generation devices (King 2009; Myers 2011) that work supported on the sea bottom, and then suitable only for sites with depths below 40 m. Figure 1 shows some prototypes of this kind of devices from hundreds of kW to 1 MW in testing period -Atlantis, Open-Hydro- and one -the Sea-Gen device, of 1.2 MW- working under commercial test stage from July 2009.



Figure 1: Sea-Gen, Atlantis, Morild & Open-Hydro TECs

2. THE GESMEY PROJECT

2.1. Generator Design

The initial goal of GESMEY Project -Spanish acronym from Submarine Electrical Generator with Y shape Framework- was to develop a device specially designed to harnessing the currents of the Strait of Gibraltar. This strait has a very irregular bathymetric profile, with zones between 90 m and 960 m depth in the channel axis (figure 2). The energetic resource that the Strait

offers is made up by a double current, a superficial one from the Atlantic to the Mediterranean and the other one that goes at a lower level and reverse direction.

There are several places with a “mean spring tide” speed up 2 m/s in the Strait, but normally they are in deep sites, usually over 80 to 100 meters depth locations (G^a-Lafuente 2010).

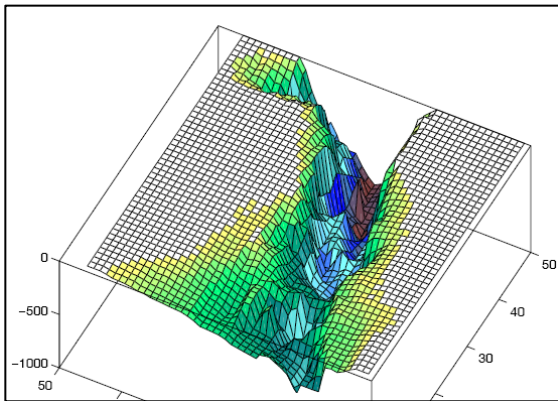


Figure 2. Bathymetric Profile of the Gibraltar Strait

Thus, the main objective of the GESMEY Project was to develop a second generation device with a low life cycle cost, designed for the Strait of Gibraltar and others world sites with water depths over 40 m where the present devices cannot operate. The main goals that the GESMEY design (López 2009) can be resumed as:

- Simplified deployment
- Minimum environmental impact
- No surface elements on operation
- Robust and simple construction
- Easily scalable (depth, stream speed, nominal power)
- Use of commercial off-the-shelf (COTS) technologies.

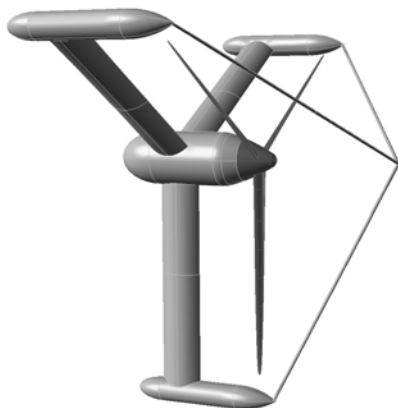


Figure 3. GESMEY Device in Operation State

As a result of the study of alternatives, we choose the design drawn in figure 3. It can be seen that the GESMEY TEC comprises the following elements:

- Rotor: With fixed pitch blades to improve efficiency and reliability.
- Central POD: Power Take-Off -PTO- components and ancillary systems,.
- Columns: Main structural parts and ancillary ballast tanks.
- End Torpedoes: Main ballast tanks.

An important portion of the inner volume of the columns and torpedoes is used as water ballast tanks. The changes on their ballast volume lets handle its floatability and then the position and/or the orientation of the device are controlled. More details of the design, distribution of elements, location of components, and dimensions are described in Núñez (2010).

2.2. Main States of Operation

Under operation, as is shown in figure 3, the device is maintained on position by a mooring systems adapted to the site environmental conditions. By controlling the ballast water level on torpedoes -the uppers with net buoyancy and the lowers with net weight- an adequate stability is achieved to keep the device vertical with reduced heel and trim angles on despite the torque and force produced by the rotor.

For maintenance -when it is necessary to extract the device form water- the procedure is very simple (figure 4). First, removing some water ballast, the device goes up to surface smoothly. When it reaches the sea surface, a new change on ballast tanks produces a self rotation. And finally the device floats on sea surface with the rotor outside water (figure 5). The device is self supported for transport.

For the device commissioning or recovering the operation state after a maintenance procedure, it can be used the reverse sequence. The whole procedure will be fully automatized with only a remote supervisory control from the tidal farm control station.

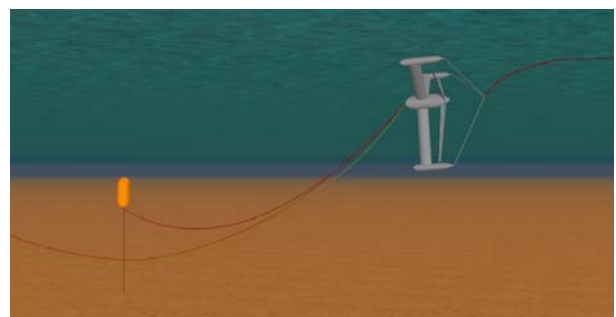


Figure 4. Image of the Emerging Manoeuvre

2.3. Project Development

The “five stages protocol” showed on table 1 (Southampton 2008) for MER converters development has been adopted during the GESMEY Project, and the executed stages are summarized below.

The starting point of the Project was the Universidad Politécnica de Madrid -UPM- patent (López 2007). The Project Stage 1 -named Functional

Definition Phase- was developed between 2008 and 2009 by the UPM GIT-ERM R&D Group on Marine Renewable Energy and the Fundación SOERMAR that is the R&D centre of the private Spanish Shipyards.



Figure 5. Scale Model of GESMEY Device in Maintenance State.

Table 1: MER Devices Development Stages

Stage	Development Level	Main Tasks
1	Conceptual Design	Explain main concepts and components. Identify R&D needs.
2	Construction Design	Detailed machinery and structure design. Physical tests at middle scale and/or CFD simulation
3	Operation Design	Physical tests at large scale, with integration between subsystems
4	Technical Demonstration Prototype	Full-scale prototype testing at sea
5	Commercial Demonstration Prototype	Full-scale commercial demonstrator testing

As results of this stage various designs adapted to various power and current speed profiles were carried out. The chosen design for 1 MW unidirectional currents -named GSY-U1M- is showed on figure 2. Their main values are summarized on table 2.

During 2010 and 2011 we are developing Stages 2 and 3 of the MRE protocol by a consortium of the UPM GIT-ERM Group, SOERMAR and Balenciaga Shipyard. The main delivery from these Stages will be a 10 kW prototype that will be intensively tested on sea at the end of 2011.

Next prototypes of 100 kW and 1 MW (Stages 4 and 5) are under technical design in order to start their test at 2013. The geographical testing areas have already been selected around the Spanish Coast, where marine currents offer good values and there are high

depths where first-generation devices cannot operate nowadays.

Table 2: GESMEY U1M Main Specifications

Nominal power	1.0 MW
Nominal current speed	1.8 m/s
Rotor blades diameter	32 m
Site depth	80 m
Device maximum diameter	38 m
Force over mooring	1.0 MN
Structure weight	80 t
Device weight	140 t
Buoy volume (x3)	200 m ³

3. GESMEY MODELLING

3.1. GESMEY General Modelling

During the GESMEY Project Stage 1, different possibilities for making a computerized tool which facilitates the calculations on operation state were considered. The final developed tool HACERIC - Spanish acronym from: tool for the analysis of radial bodies inside currents flow- let the user enter data - sizes, weights...- corresponding to the device under analysis and adjust different ballast tanks levels, obtaining as results the most significant forces, torques, and orientation angles of the device.

A special analysis study for the complete mooring system is required and it becomes a specific topic for a 2nd generation TEC design.

Because the effort to make a static analysis tool is similar to the required to develop a dynamical one, both analysis have been integrated together. By this way, the manoeuvres analysis have been carried out in a coupled mode with it, and diverse models and tools have been used for comparing and validating simulation results.

3.2. GESMEY Dynamic Models

A moored TEC is considered a device working in movement with some degrees of freedom, in opposition to the devices rested in the sea bottom in which the main dynamical problem is the fluctuation of the forces in the blades due to the effect of the current's turbulence -fluctuation of the inlet velocity in a blade section- caused by the waves in its upper part and the depth variation (shear effect) in its lower part (Bard 2009). This is why in a moored TEC it is necessary to complement the static analysis with the dynamic effects, at early design stages, which could include:

- The study of the turbulence of the current over the rotor and its transmission through the PTO. This field requires more intensive analysis by hydrodynamics specialists.

- The analysis of the kinematic and mechanical behaviours of the mooring system in static and dynamic regime. This is the more specific case for the moored TECs so it will be analyzed in more detail in the next sections.
- The analysis of the dynamic loads generated in all the structural elements, which can be done with the usual FEM methods and S-N curves (DNV 2008).
- The study of the seakeeping when the TEC is over the sea surface. Tests and tools usually used for the study of offshore structures can be directly applied.
- And last but not least the PTO control. This can be analyzed with usual simulation tools like Matlab and Simulink (Somolinos 2010).

3.3. GESMEY Hydrodynamic Model

There are specific commercial tools for the study of the moored systems, as OrcaFlex (Orcina 2011), developed for the offshore industry, extensively used and validated, and homologated by certification entities, as the Ship Classification Societies.

But, for the usage of these commercial tools the following additional operations are required:

- The modelling of the hydrodynamic aspects of the TEC (drag, lift, added mass).
- The development of procedures for the integration of the simulation and control tools.

A valid method to solve the first of the former tasks is the decomposition of the TEC in a series of elements, which usually have a well defined geometry - cylinders, ellipsoidal prisms, spheres, flat plates, ellipsoids, etc.-, then study the behaviour of each of these geometries separately and finally add their effects.

The same type of tools and models can be used for the study of the manoeuvres of the TEC, including the change from the operation situation to the floating situation and vice-versa but adding a module representing the actuator's effect, usually a change in the volume of some ballast tanks.

Therefore, two mathematical sub-models are necessary: the mechanic and the hydrodynamic ones. Both of them can be developed based on similar ones used in naval architecture design.

For the GESMEY TEC the hydrodynamic model developed is based on the segregation of the device structure into different elements and then the computation of their drags as function of their respective speeds, according with equation (1).

$$F_d = 0.5 \cdot C_d \cdot \rho \cdot A \cdot V^2 \quad (1)$$

Where F_d denotes the drag of each element, A is its significant surface, V and ρ represents the water speed and density and C_d is the form coefficient.

The model neglect the lift forces because all the elements are disposed in a symmetric way with respect

to the direction of the flow, and the mooring system let the device automatically orient along the flow direction.

All the volumes are computed, and then, buoyancy forces are applied over each of these elements. On the other hand, all the weights are computed too, and gravitational forces are obtained. The hydrostatic forces are obtained by composing both forces.

Once all the hydrodynamic, hydrostatic and rotor forces have been computed, they are integrated into the mechanical model with the definition of the mooring points and the torque from the rotor.

The dynamic equations of the generator with one mooring point (like in the emersion manoeuvre) are obtained from a model with only a point mass concentrated in device c.o.g. The acting forces over the device (figure 6) put the rope in tension, keeping it straight, and producing a torque about its attachment point at sea ground (G), so the basic equation becomes (2).

$$\Sigma Q = I \cdot d^2 \alpha / dt^2 + dI/dt \cdot d\alpha/dt \quad (2)$$

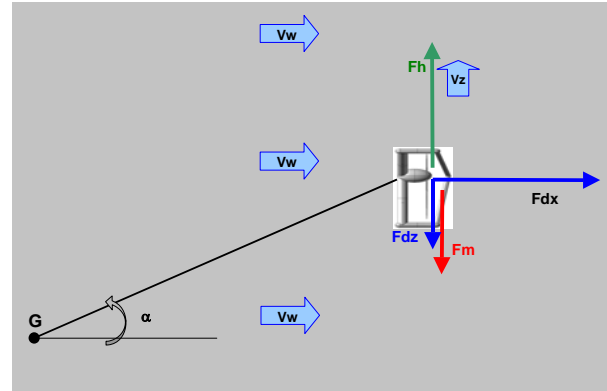


Figure 6. Main Hydrodynamic Forces

Usually dI/dt is very low and then the second term can be neglected, I denotes the device inertia -around G turning point-, α the rope's angle and ΣQ the sum of the turning torques caused by the input forces on the devices. If vertical forces (F_z) and horizontal ones (F_x) are grouped and L denotes the length of the rope. ΣQ can be written as (3).

$$\Sigma Q = -L \cdot F_z(t) \cdot \cos \alpha(t) - L \cdot F_x(t) \cdot \sin \alpha(t) \quad (3)$$

The horizontal force is related with the structure and rotor drag of the device and it can be calculated from equation (4). Vertical forces are equal to the net buoyancy -volume mass minus weight- plus the drag due to the vertical motion as is expressed in (5).

$$F_x = K_{dx} \cdot |V_w - V_x(t)| \cdot (V_w - V_x(t)) \quad (4)$$

$$F_z = M_g \cdot g - V_g \cdot \rho \cdot g - K_{dz} \cdot |V_z(t)| \cdot V_z(t) \quad (5)$$

Where K_{dx} and K_{dz} denotes the hydrodynamic drag coefficients, V_w the water current speed, V_x and V_z the device horizontal and vertical speeds, M_g and

V_g are the generator mass and volume, ρ is the water density and g the Earth gravity.

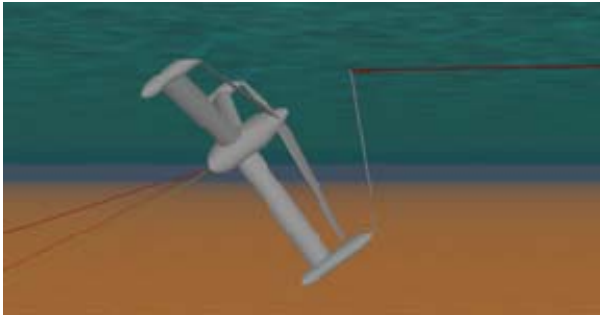


Figure 7. GESMEY Self-rotation State

4. GESMEY SIMULATIONS

4.1. Manoeuvres Aims

As it was said in section 2, one of the keys to the success of the GESMEY generator is the simplicity of their manoeuvres for installation, maintenance and decommissioning. As an example, the emersion process to bring up the generator to the sea surface for the periodical maintenance includes the following steps:

- On operation, the device is maintained on position by the mooring system and by controlling the ballast level.
- When the “stern rope” is detached and some ballast removed, the device goes up to surface smoothly (figure 4).
- When it reaches the sea surface a new change on ballast tanks produce a self-rotation (figure 7).
- Finally the device floats on sea surface with the rotor outside, ready for first level maintenance or transport (figure 5).

4.2. Model Validations

Based on the results of the HACERIC tool, the two-dimensional model resumed in section 3.3 was tested using Simulink (Somolinos 2009), by supposing a virtual rigid fitting cable with negligible hydrodynamic properties, modelling the global mass and the drag forces in the two main directions and neglecting the device turn (minimum in the step 2).

By comparing the results of this first simulation with the obtained with OrcaFlex, the main differences were observed when the TEC was reaching the surface, appearing also differences in the oscillation periods.

That is why it was decided to do a more detailed study of the hydrodynamic behaviour -drag and added mass- (White 2006; Korotkin 2007), of the generator structure and make some tests in a towing tank, with the model showed of figure 5.

Finally, by applying the results of this study to the simulations done for a basic case with Simulink and OrcaFlex, and with some parameter fitting the simplest model was adjusted with a precision of 1% in the

submerged phase and 5% on the surface, so the model can be validated considered, waiting for the final calibration based on experimental tests.

4.3. Simulation Results

A brief comparison between the obtained results from Simulink and OrcaFlex for the emersion process of a simple body, with the refined hydrodynamic model, is shown in figure 8. The added masses for this model are different from those on a ship and their calculation required an important and novelty effort.

It can clearly observed from this figure, that both simulation results begin from the same 50 meters depth at the initial time of $t = 0$ seconds. Both time responses offer a similar rise time of about 72 seconds, been the obtained results from OrcaFlex -when added masses and hydrodynamic effects of the structure are better modelled and considered- a little faster than the obtained results from Simulink (simplest model). The same effect of this slightly faster response can be appreciated if the frequencies of the oscillations part of the whole responses are analyzed.

The reasons of these small discrepancies can be justified in base of the difficulties for modelling partially submerged bodies. The method can be improved by using RAOs methods, but the obtained results are considered of enough precision, because in practice the waves' effects, even for the lower waves, will be more important than the natural oscillations passed the first two or three cycles.

The good agreement between both simulation responses obtained with different tools as Simulink and OrcaFlex justifies the goodness of the dynamic models that have been used, and confirms the feasibility previous to experimental testing with a real prototype. Both responses exhibit non linear oscillations with similar amplitudes and small damping factors which correspond with the water-air interaction of the generator in the final stage of the emersion manoeuvre.

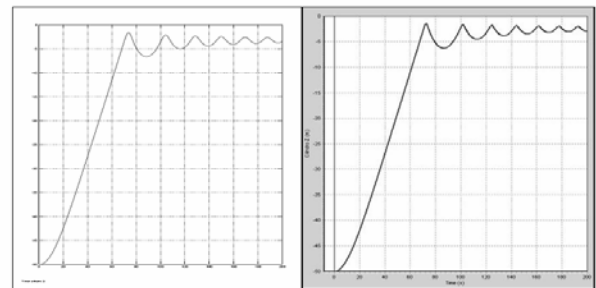


Figure 8. Simulated Dynamic Response from Simulink and from OrcaFlex

Finally in the figure 9 the trim angle -red- and the depth of the centre of gravity -green-, obtained from the simulation of the turning phase, are shown. These curves correspond to the initial tests of the ballast control system, and they show that the self-rotation is

very fast when passed a critical angle with some post-critical point's oscillations.

As result of these simulations, it is necessary to design a robust ballast control system based on the obtained dynamic model that allows performs smooth manoeuvres without any kind of human intervention. Currently a 1/10 scale model for testing at sea -inside the project stage 3- is being built in order to perform validation procedures based on extended experimental tests, and a good match between simulation models of diverse complexity level and real measured responses is expected..

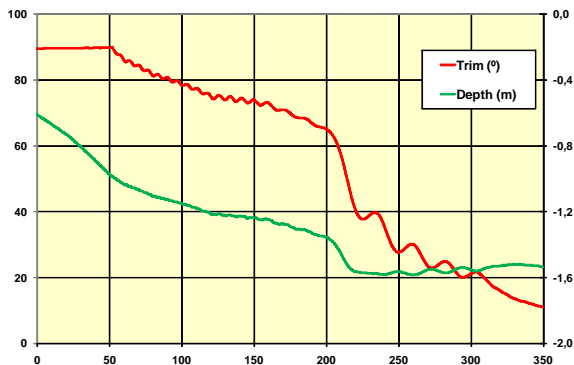


Figure 9. Surface Turn Simulation Results

5. CONCLUSIONS

This work has shown the state of development of the TECs, the tools used for their design and analysis, and how these problems have been solved in the particular case of the GESMEY generator.

Also has showed that it is necessary a new generation of converters capable of extracting energy of currents from sites with a depth over 40 m.

The design of the GESMEY generator properly achieves their proposed objectives, being one of the more promising generators of the second generation.

The study of the dynamic problems is an especially important challenge. More detailed studies are needed, especially in the fields of the hydrodynamic and manoeuvring control.

The preliminary results from GESMEY simulations with different models and tools show a good agreement of manoeuvring results.

It is convenient to complete these studies of the turning movements with higher scale models, in order to check the interaction between the ballast control system and the movements on the surface.

ACKNOWLEDGMENTS

The authors want to express their gratitude to all the R&D team -more than 20 persons- that have been working in the different phases of the project, to the UPM Research Office for their support to this project and the generation of a standing research group on marine renewable energies -GIT-ERM-, to the management team of Soermar, for their financial and organizational support to the development of the

different phases of the project, to the Spanish Ministry of Science and Innovation, to the Spanish Ministry of Industry, Commerce and Tourism, to the Naval Sector Management Office and to the Madrid Regional Administration for the award given to the GESMEY Project.

REFERENCES

- Bard, J. et al., 2009. *Control strategies for dynamic load reduction in marine current turbines*. Wasserbaukolloquium.
- DNV, 2008. *Guideline for Wave and Tidal Energy Certification Schemes*. Det Norske Veritas. Oslo.
- G^a-Lafuente, J. et al., 2010. *Mapa de flujos de energía en el Estrecho de Gibraltar para su aprovechamiento como fuente de energía renovable*. GOFIMA - U. Málaga,.
- IEC, 2010. *Strategic Business Plan*. International Electrotechnical Commission, TC114. December
- King, J.J. et al., 2009. Tidal stream power technology. State of the Art. *OCEANS-2009-Europe Conference*. May, Bremen.
- Korotkin, A.I., 2007. *Added masses of ship structures*. 1st ed. Springer.
- López, A., 2007. Patent n° P200700987, *OEPM Spanish Patent and Trademark Office*. Madrid, April.
- López, A. et al., 2009, Modelado y simulación de la operación de un generador para el aprovechamiento de las corrientes del estrecho de Gibraltar. *XXX Jornadas de Automática, CEA-IFAC*, Valladolid, September.
- Myers, L.E. et al., 2011. Equimar Deliverable D5.2: Device classification template. *Equitable Testing and Evaluation of Marine Energy Extraction Devices in terms of Performance, Cost and Environmental*. COMMISSION OF THE EUROPEAN COMMUNITIES. March.
- Núñez, L.R. et al. 2010. The GESMEY ocean current turbine. A proposal for marine current energy extraction on deeper waters. *3rd International Conference on Oceanic Energy*. Bilbao, October.
- Orcina, 2011. <http://www.orcina.com/> [Accessed on March 2011]
- Somolinos, J.A. et al., 2009. Simulation of the emersion procedure for a new underwater generator. *International Workshop on Modelling and Applied Simulation EMSS'09*. Puerto de la Cruz, Spain, September.
- Somolinos, J.A. et al., 2010. Automatic system for underwater ocean current turbines. Application to GESMEY. *3rd International Conference on Oceanic Energy*. Bilbao, October.
- Southampton University, 2008. Tidal-current Energy Device. Development and Evaluation Protocol. *IEA-OES Guidelines for Development and Testing of Ocean Energy Systems, Task 2.2*. Southampton, December.
- White, F.M., 2006, *Fluid mechanics*. 5th ed. McGraw-Hill.

USING SEMANTIC WEB TECHNOLOGIES TO COMPOSE LIVE VIRTUAL CONSTRUCTIVE (LVC) SYSTEMS

Warren Bizub^(a), Julia Brandt^(b), Meggan Schoenberg^(c)

^(a)Joint Advanced Concepts, Joint and Coalition Warfighting, 116 Lake View Parkway, Suffolk, VA 23435, USA

^(b)General Dynamics Information Technology, 112 Lake View Parkway, Suffolk, VA 23435, USA

^(c)U.S. Navy, Naval Sea Systems Command, 116 Lake View Parkway, Suffolk, VA 23435, USA

^(a)warren.bizub@jcom.mil, ^(b)jbran026@odu.edu, ^(c)meggan.schoenberg@jcom.mil

ABSTRACT

Department of Defense (DoD) closed architectures and proprietary solutions limit ability to provide gaming, semantic reasoning and social networking capabilities employed by industry and available in the open source community. Exorbitant sustainment costs of legacy solutions are unjustifiable and inhibit transition to enhanced LVC solutions. Furthermore, legacy solutions are dependent on an aging workforce of static-centric modelling & simulation (M&S) subject matter expertise (SME) to promote reuse, while budget cuts increase attrition among junior-level technical staff. This paper describes challenges and recommendations for changing the DoD M&S training paradigm to facilitate interoperability, incorporate emerging semantic web technologies, and provide a knowledge base to promote reuse. Two ongoing R&D projects will illustrate innovative strategies and their potential to alleviate legacy system interoperability issues while transitioning to a LVC Defense Training Environment (DTE) where US and Coalition Command and Control (C2) and M&S systems seamlessly interoperate to train as we fight.

Keywords: semantic web, interoperability, LVC, ontologies

1. INTRODUCTION

Reducing time and resources required to deliver effective training through combination of live military assets, virtual reality systems, and other forms of computer models and simulations is paramount. The ability to quickly and easily assemble system components from mixed-architectures to create LVC environments is a key enabler to support test, training and experimentation.

The LVC Architecture Roadmap (LVCAR) Final Report released in December 2009 identified fourteen Summary Focus Areas and nineteen Investment Recommendations. Key findings addressed include the need to:

- Start focusing on semantics of these systems.
- Provide resources to address LVC issues that are not directly architecture-related (e.g., semantic interoperability, conceptual modeling, etc.).

- Lead efforts to standardize or automate translations of data/scenario inputs to simulations and data capture formats.

A team of architects is developing a solution based on semantic web technology and its supporting tools. This approach also provides the ability to develop and implement automated selection, translation, and implementation of components, data, and scenario inputs to LVC systems and environments.

First, a technical framework was created that formalized the description of warfighter missions, the LVC environment, and what is required to implement these missions in a test, training or experimentation domain. Second, was the creation of semantically rich resource descriptions of systems, components, data, data exchange, and object models. This provided the base of a consistent set of artifacts semantically linked through relationships based on the warfighters' language. The linkages were created through the use of the Joint Capability Areas and Uniform Joint Task List. Warfighter linkage through the use of warfighting terminology is critical to enable mixed-architectures to support LVC environments. Finally, this approach allows creation of a resource repository that permits the user to search, compare, select, modify, and assemble training, test and experimentation systems, components, data, data exchange, and object models for LVC environments.

2. UNDERLYING POTENTIAL TECHNOLOGY SOLUTION

The DTE is envisioned as a government enterprise with the goal of providing an environment where C2, Intelligence, Surveillance and Reconnaissance (ISR) systems, training ranges, and simulation systems seamlessly communicate across departmental, agency and multinational boundaries in accordance with security and privacy regulations and laws. The DTE will support multiple domains when realized. DoD net-centric environments have been conceptualized over the past decade (e.g., Global Information Grid (GIG) and Net Centric Data Strategy (NCDS) (DoD 2003)); however, only minor successes have been fully realized. Stable well-tested technologies and ontology resources, which did not previously exist to enable the NCDS, are now readily available in the open source community. What is still missing? DoD acceptance, documented

best practice, open source implementation strategy and trained personnel.

During the past decade, the web continued to develop via open community practices and collaborative efforts. One such development was the Semantic Web. Semantic Web technologies are inherently extensible and modifiable, and collectively provide the common framework for data to be shared and reused. Unfortunately, because information technologies and acquisition processes being utilized in the DoD sufficed, the development and arrival of the Semantic Web and open technology development (OTD) practices largely went unheralded in DoD programs until recently.

Research has shown that many government, Service, and international organizations are beginning to transition to semantically-driven infrastructures, utilizing OTD processes and realizing information exchange in cloud computing environments. The benefits of these technologies and methods include rapid data generation and alignment, collaborative development opportunities, reuse, and distributive interaction. However, common vocabularies and ontologies to share this data are even more critical. If the various practitioners develop vocabularies and ontologies in a vacuum we will end up creating interoperability alignment issues that will plague our ability to build a robust DTE. Similar to the issues we experience today due to the development and integration strategy utilized in our current acquisition approach of major information management systems.

Ontologies and OTD processes enable systems to communicate natively by generating meaningful data that is exchanged machine to machine. Perception data, in the form of ontological representations, improves event fidelity and enables more effective event management.

It is the opinion of the authors of this paper that the Semantic Web and OTD practices offer the most promising direction for a scalable and dynamic LVC solution. A well-managed, open, and transparent approach that incentivizes contributors can be implemented to encourage Services', agencies' and Coalition Partners' cooperation and participation. Adopting semantic technologies, ontologies and cloud computing is the best way to realizing an agile and effective DTE.

The IT Government workforce is dominated by baby boomers and pre-baby boomers who are now in leadership positions. Their reluctance towards the internet evolution and OTD may be a contributing factor to the malaise state as depicted in Figure 1. There has been no significant leap in M&S technology since the mid 1990's when High Level Architecture (HLA) was created. In effect, M&S is still at 1.0 while the internet is moving toward 3.0 and beyond. The internet has continuously reinvented itself through OTD practices, which foster innovation and development through communities of practice (COP).

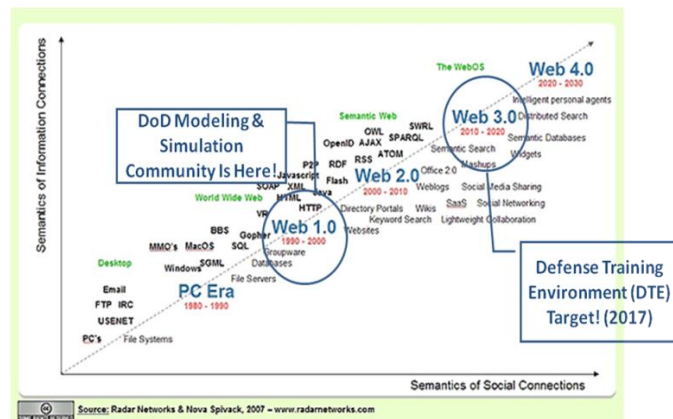


Figure 1 – Internet Evolution

Many of the new technologies and capabilities developed through the internet have been produced by the generation “x” and “y” digital natives. The “Net Gen”, as they are called, is comfortable with OTD and the fast-paced internet evolution, while the baby boomers and pre-baby boomers are not.

In order for the DoD to push past this state of malaise, and move toward the next generation, they must make use of their “Net Gen” talents, institutionalize OTD, and embrace internet technologies – at least until quantum computing makes its debut.

2.1. DTE Development Challenges

To create joint scenarios, multiple types of data must be combined together. Data initialization of disparate systems is complex and lengthy. Currently, interoperability problems arise because of misalignment between multiple kinds of data (geospatial, order of battle, messages, events, etc.) due to:

- No common oversight or lack of documentation as to data formats.
- No common mode of access to relevant content.
- No common framework for data retrieval and reasoning.
- No common authoritative or non-authoritative data.
- No common vocabulary.
- No generally applicable strategy for combination and alignment.

DTE LVC development efforts focus on increasing interagency, intergovernmental, and multinational integration of solutions through developing common approaches, producing and sharing data, applying common standards, and emphasizing re-use. The most promising direction for a scalable and dynamic solution incorporates the effective use of Semantic Web technologies and best practices (e.g., accurate use of ontological representations, robust semantic descriptions, and adherence).

Semantic Web technologies, cloud computing and OTD practices are inherently extensible and modifiable at all levels and during any phase of a system's lifecycle. Data exchange benefits can be realized, and

content can be generated automatically on an as-needed basis to provide an operational perspective. This enables greater agility in development, improves interoperability between diversified systems as required in LVC integration, and provides high-level fidelity of real-world warfighting operations to facilitate rigorous and realistic collective defense training.

“Ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents” (Gruber 1992). The use of a common ontology, or a common suite of ontology modules, along with a well-documented governance process, reduces redundant efforts, and results in highly discoverable, composable data with improved Understandability, Reusability, Extensibility and Discoverability.

The US Government has been progressively adopting Semantic Web technologies, developing ontologies, incorporating cloud computing instantiations and espousing OTD practices to improve net-centric communication. Representative organizations and their work include:

US Government Executive Branch (www.data.gov/semantic). The purpose of data.gov is to increase public access to high value, machine-readable datasets. Data.gov uses web semantic technologies to develop data mashups, the combination of data, presentation, or functionality from two or more sources to create new services. This site provides access to over 3,000 public Resource RDF datasets generated by the Executive Branch.

National Cancer Institute (NCI). NCI is part of the National Institutes of Health (NIH) and is one of eleven agencies that are part of the US Department of Health and Human Services. NCI coordinates the US National Cancer Program, and conducts and supports research, training, health information dissemination, and activities related to the causes, prevention, diagnosis, and treatment of cancer, supportive care for cancer patients and their families, and cancer survivorship.

Networking and Information Technology Research and Development (NITRD). “The NITRD program is the nation’s primary federally funded source for revolutionary breakthroughs in advanced information technologies, such as computing, networking, and software” (NITRD 2011). It is the framework for the collaborative efforts of the fourteen member federal agencies and many other research and development agencies to coordinate efforts.

National Aeronautic and Space Administration (NASA). NASA has a variety of initiatives based on Semantic Web technologies. The most well known is the Semantic Web for Earth and Environmental Terminology (SWEET). SWEET has over 200 ontologies cumulating in over 6,000 earth science concepts.

A prerequisite to moving the DTE to the Semantic Web is to classify, organize, and catalogue data in a machine-readable format. The Semantic Web builds on

the Extensible Markup Language’s (XML) ability to define customized tagging schemes and the Resource Definition Frameworks (RDF’s) flexible approach to representing data. Adopting Semantic Web technologies and practices as part of the LVC DTE infrastructure will result in a more efficient and cost-effective enterprise with significantly improved US-US and US-Coalition interoperability. The Semantic Web is not owned by any single or group of commercial organizations. Much of the technology, standards, and techniques developed are managed through open source and open specification programs and projects. Best practice is usually defined as agreed upon by the leading practitioners of the World Wide Web Consortium (W3C).

2.2. Open Source Discussions

James Carter of the National Security Agency (NSA) was directly involved with NSA’s Security Enhanced Linux (SELinux) project. SELinux was NSA’s initiative to get some of the more critical security enhancements into computer operating systems. Open Source (OS) provided them a means to test and provide security enhancements to both the OS community (i.e., Linux) and proprietary OS developers. On the SELinux effort, James Carter said: “The barriers for us involved the rather long process of getting approval from all the relevant stakeholders, such as our General Counsel and Public Affairs offices. The process we went through was specific to our agency, so you would have to determine who the relevant stakeholders are for your project and determine what they would require for approval. The benefits we have seen are those of any OS project: our project is available to be used and studied (as a research organization, one of our goals), many external developers have contributed code and ideas, and we’ve been able to have far more impact than we would have otherwise.”

Paul Byrne of Sun Microsystems was asked about the additional overhead associated with OS. He said that he doesn’t see a significant amount of overhead. Sun has a policy to develop all of their software, both directly OS and proprietary, with an “open” philosophy even if “no one is listening.” For all intents and purposes, the number of developers working on the core of an OS project is rather limited. Deep in the software stack, there’s little there that directly impacts substantive users of the product and hence the interest of the majority of user-developers will remain at the higher levels of the software stack. This is good because corporate focus (and some good geek help) can be placed at the lower levels of the stack thereby maintaining a semblance of control.

Paul also offered what drives a good OS project. He said leadership was very important and he was careful to point out that “leadership” is probably not one person. He said it’s important for the leadership to be completely frank with the community throughout the process. Additionally, he offered caution on development forks. The threat of a fork helps keep

forks to a minimum. Forks are VERY expensive for the group initiating the fork. Forks, in fact, are a relative rarity in the OS community. He said if the leadership is open and honest, then you're probably not going to see a fork. If you spot another initiative that's close to your own, you need to build a relationship with that other project. They have to understand your vision/goals...and if they do, they are more likely to align with your cause.

Finally, Paul provided some thoughts on benefits of OS. He said OS provides a great mechanism for identifying people to hire. He said that training for newcomers happens 24/7 because everyone in the particular OS community takes an interest. He said that one of the biggest advantages of OS is that there are virtually no delays in testing intermediate releases. Because the source is open, when a new feature, etc., is introduced, the community is very quick to start thrashing on it. Lastly, open source "can't go away." That is, if the sponsoring organization loses interest in that particular project, the project will continue and its products will not be lost on some out-of-date server in corporate bowels.

Brian Newburn of Black Duck Software noted that one of the challenges with OS is that developers will borrow code from one OS project and use it in another. This can cause problems if the licenses for two OS projects are incompatible. Black Duck has developed a number of products to help manage OS projects. Their Code Center product has a database of billions of lines of OS code. It also has the legalese for approximately 1400 different OS licenses. Code Center scans incoming code for OS reuse. If it finds code-in-common, it checks license compatibility and alerts the OS project leadership about the code's inclusion and the legal implications. Black Duck did a quick study in 2008 on the value of OS. They concluded that if OS was a country, its GDP would be the 77th largest in the world (the list had 190 countries on it at the time). Approximately 4.7 million lines of OS code get generated every day.

3. ONTOLOGY BASED STRATEGY

Past efforts in DoD has resulted in lack of coordination in the presentation and handling of data, which has significantly hampered interoperability. An ontology-based approach is recommended to address the complexity of initializing and operating multiple heterogeneous systems as will be required in the LVC DTE. The creation of consensus-based ontologies, which are controlled structured vocabularies that can be used for consistent presentation of data, enables more effective retrieval and reasoning of data.

A four-step strategy for ontology creation using OTD processes, similarly outlined with some adjustments to William Mandrick's C2 Core Ontology Study Report, is recommended.

First, *leverage* previous work in DoD. The foundation is The Universal Core ([UCore](#)) (Wikipedia, 2011) which is a US Federal Government information

sharing initiative that is supported by the US Departments of Defense, Energy, Justice, Homeland Security, Intelligence community, and other national and international agencies. The UCore vision is to improve information sharing by defining and exchanging a small number of important, universally understandable concepts across a broad stakeholder base to improve data interoperability between known and unanticipated users while achieving cost and time savings through standardization, modularity, and reuse. Its current form is the UCore 2.0 that serves as a central hub designed to maintain a broad community perspective. The long-term goal is that these common terms will create a common reference platform allowing data from diverse domains to be understood across various systems. The Army NCDS (ANCDs) (DA 2011) Center of Excellence created UCore SL to supplement the semantics of UCore 2.0. The UCore 2.0 taxonomy does not include relations with domain and range declarations or disjointness, equivalence, and union axioms. These additional logical resources are provided as extensions of UCore. UCore SL employs the W3C's Web Ontology Language (OWL) to enable semantic validation of both individual extensions of UCore as well as the combined set of all extensions. It provides for logical decomposition of terms and definitions, the ability to reason logically on the basis of the content of these definitions, and thereby enhanced support for the creation of consistent extension modules. Finally, [C2 Core](#) (USJFCOM 2010Live) is a DoD ontology currently being developed to provide a level of interoperability between C2 systems unachievable with data dictionaries and custom schemas. It ensures the meaning of information between systems will retain its context and meaning. C2 Core ontology is represented using OWL and is extended from the UCore. The objective of C2 Core is to develop an open standard supporting extensible markup language (XML)-based C2 data exchange. The C2 Core follows the same approach as UCore insofar as it identifies a set of terms that is core across the C2 domain. C2 Core has logical consistency through a top-down extension of UCore 2.0 terms, logically defined using the resources of UCore SL, and applying the result to create a C2 conceptual data model called C2 Core Common Data Model (CDM), which contains over 200 high-frequency terms that define the C2 domain. These terms pertain to situational awareness, structuring a military organization, planning and assigning tasks, decision making, and assessing progress. Examples of potential targets for extensions of the existing C2 Core include sub-domains such as Strike, Unit Readiness, Planning and Operations, and the Military Decision Making Process (MDMP). The DTE must be developed following the operational community's foundational concepts and descriptions.

Second, *develop* a small consensus-based controlled vocabulary to serve as the basis for the description (e.g., tagging) of data. This ontology will use best practices and standard operating procedures for

ontology development, including automatic realization of the net-centric approach since data annotated with an ontology becomes automatically identifiable through the corresponding Uniform Resource Identifiers (URIs). It should rest on a strategy of *maximal realism*: seeking not a *data model*, but a *reality model* that is based on the Joint Operating Environment (JOE). The ontology is based on military doctrine, using the common terms used by the warfighters themselves. It draws wherever possible on existing ontology efforts, and strives for consistency with current initiatives.

Third, *emphasize* an adaptive modular plug-and-play approach. Create custom extensions for specific domains. The suite of extensions will include a generic ontology that consists of terms of common interest to all endeavors, along with more specific extensions ranging across various domains. Core ontology extensions should be created for specific operational domains of interest such as Close Air Support, Human Social Cultural Behavior, Intelligence, Humanitarian Assistance, Logistics, Missile Defense, Counterinsurgency, Incident Management, etc. The goal is to have each COP with a unique data annotation to embrace a single, incremental strategy of synchronized development of extensions. Establish a governance process to ensure change management, coordination, availability of authoritative data sources, and to provide dedicated cross-community training and pilot testing initiatives.

Fourth, *incentivize and socialize* the use of the ontology and its network of extensions. The goal is to create a situation where use is by all major participants along the data chain. Leaders of the relevant communities must be incentivized to contribute to the maintenance of the ontology (because coherence of one's own work depends upon it being of high quality, including needed terms, and being up-to-date). The assumption is that, as the benefits of the core and extensions approach become manifest, more resources will accrue to the project.

4. BENEFITS OF AN ONTOLOGY-BASED APPROACH

The availability of a C2 Core Ontology and of an expanding set of authoritative data sources will allow realization of the DTE. Developers and integrators of the LVC DTE must collaborate with ontologists in creating a DTE-C2 Core Ontology, the common platform for the new approach to data initialization, scenario development, and event execution made possible by current and future net-centric and internet technology, in which the continual need for investment of manual effort in data preparation and exchange will be substantially reduced. Certain factors must align for this to happen. There needs to be a concerted effort to enhance coordination for effective ontology development work and description of data across a large population of domains. The division of expertise must be exploited. A strategy of orthogonal modules will allow exploitation of the division of expertise on the

part of different Communities of Interest (COIs) and SMEs that ensures consistent interoperability of the whole. Training of the workforce is paramount. The ontology-based approach provides more effective use of resources in the creation and application of software as the standard operating processes for ontology development and application for data use is similar and can be adopted across domains. Personnel can be trained once, and their expertise used multiple times.

Following best practices in the creation and application of ontologies will facilitate a LVC DTE solution that can rely on software resources that are standards-based, lightweight, scalable, secure, and deterministic that utilizes efficient development, integration, test, and configuration resources. The ontology-based approach provides an incremental strategy for quality improvement of the data flowing from the warfighting community to the training community. Annotation with common ontologies allows authoritative data to be maintained in ways that make it discoverable, retrievable, and useable in the DTE. Data silos (or data cemeteries) is avoided because the ontologies themselves are based on doctrine, are well disseminated, and are used at every stage in the data pipeline. Reuse of data across multiple domains results in enhanced realism because ontologies are based directly on operations-based ontologies, the approach will bring greater realism to the DTE. XML Schema, often used in DoD message standards, is better suited for specifying the format and structure in which data is exchanged (data model) than specifying the meaning of the data (reality model). The ontologies provide a cleaner separation of issues of presentation (data models) from issues of meaning (reality models). Ontologies developed to support the DTE will be thoroughly net-aware and made available through industry best practice web services, thus presenting data in terms of DTE ontologies guarantees an automatic adoption of the net-centric approach. The ontology-based approach allows for more effective governance of the creation and use of the authoritative data sources formulated in their terms. Finally, the easy combinability of ontologies and data resources will create, for the training domain, an environment in which plug-and-play modules for different types of scenarios can be developed and reused through automation, significantly reducing manual input. This creates greater flexibility, and a more rapid response in addressing mission planning and rehearsal needs.

5. OTD SUCCESS STORIES

Ongoing DoD R&D projects described in the following section illustrate innovative strategies, OTD practices, and their potential to alleviate many legacy system interoperability issues for transitioning to the next generation DoD DTE infrastructure.

5.1. Coalition Battle Management Services (CBMS)

CBMS is a technical infrastructure that enables the exchange of resources between Command and Control

(C2) and M&S systems, and robotic forces. Initial use cases include: exchange of orders, reports, and requests between fielded legacy C2 and M&S systems; After Action Review (AAR) support and visualization capability to support a Common Operational Picture (COP); data distribution management; persistent store (XML data store) with respective metadata to provide resend/replay capability; time management to track and synchronize message passing for improved situational awareness; and parametric search/filtering to locate and provide relevant-only information.

CBMS uses an open architectures/OSS design philosophy. It will be accessible via any commercially available web browser and uses only next generation XML-based technologies in its implementation as depicted in Figure 2. It is system-independent, allowing each consumer or producer system to map their respective system language to another language.

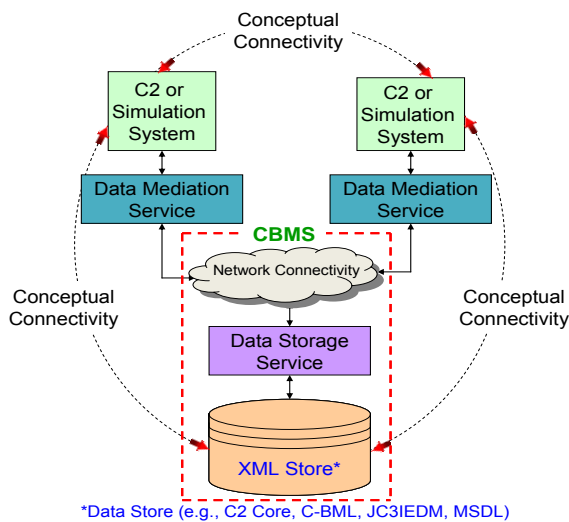


Figure 2 – CBMS Diagram (Diallo 2011)

The Simulation Interoperability Standards Organization standards committee is currently reviewing the CBMS enterprise architecture. CBMS leverages only open source web technologies:

- Xbase for document persistence and XQuery processing.
- Atmosphere framework for HTTP-based messaging.
- Jersey for RESTful web servicing.
- xLightweb for client-side HTTP processing.

In contrast to the kluge of disparate architectures, bridges, gateways, and data sharing strategies currently used by the DoD M&S community, CBMS is a decoupled collection of composable web services that can be orchestrated to support the needs of a particular federation. This design facilitates rapid technology refresh and encourages reuse.

Currently, the CBMS suite of tools is being further developed and refined to accommodate an OTD environment with Coalition partners. Executable code and documentation have been provided to Coalition partners through The Technical Coalition Program and

the NATO Modelling and Simulation Group to facilitate peer review and product feedback.

5.2. Live Virtual Constructive Framework (LVCAF)

LVCAF is a framework that supports search, discovery, and composition of federation components from multiple architectures while providing linkages to functional mission capabilities. It uses ontologies as a common vocabulary to facilitate machine-to-machine communication and a knowledge base to simplify reconciling models and promote reuse. LVCAF translates object models between the following disparate DoD M&S architectures:

- High Level Architecture (HLA) Federation (1.3, 1516 and 1516 evolved)
- Test and Training Enabling Architecture (TENA) Logical Range Object Model (LROM)
- Distributed Interactive Simulation (DIS) Protocol Data Units
- Common Training Instrumentation Architecture (CTIA) object models

LVCAF stores semantically matched components in composed data exchange models (DEM) with linkages to mission threads to facilitate event execution as shown in Figure 3.

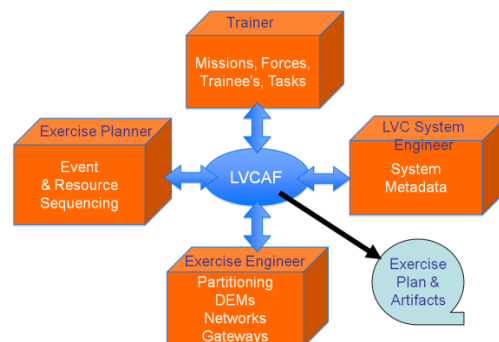


Figure 3 – LVCAF Diagram

LVCAF utilizes standard Semantic Web formats:

- eXtensible Markup Language (XML)
- Resource Description Framework (RDF)
- Web Ontology Language (OWL)

LVCAF leverages only open source web tools and development techniques:

- Archiva for repository management.
- OntoWiki for project collaboration.
- Mercurial for code versioning.
- Protégé for the inference engine and knowledge representation.
- JIRA w/ Greenhopper for project management, managing scrum backlog, planning sprints, and release tracking.
- Agile software development methodology with 30-day sprints and daily scrums.

LVCAF provides an object model comparison capability to reduce manual resource intensive object model reconciliation by the DoD M&S SME. LVCAF

is finalizing the OTD process to launch the tool suite for public development and consumption.

6. DTE STRATEGY MOVING FORWARD

The strategy for improving the ability of DTE software practitioners to realize the future LVC environment and integrate data from disparate sources into an internally consistent and well-formatted package is to initiate a three-pronged approach. The community should focus on the data initialization process, a robust runtime architecture framework, and coordinated Governance implementation.

The following is offered as a potential solution to bring about the following:

- Identify and align authoritative data that can be annotated (i.e., described, tagged) using the common suite of C2 related ontologies and DoD M&S Steering Committee (MSSC) Rapid Data Generation (RDG) High Level Tasks (HLT) Open Source (OS) development activities.
- Develop a common suite of realistic C2 and related ontologies for describing warfighting situations in a common architecture framework through OTD practices that facilitate development of the DTE.
- Implement a governance process that allows ontologies to evolve and expand in a maximally consistent and useful way across all domains through an open and transparent process.

The first issue to be addressed is one of rapid data generation for initialization. “How can we integrate as rapidly as possible the many different kinds of data needed within a given scenario?” Currently, these data cause problems for the scenario designer, because they are derived from disparate sources, differ in format, rely on heterogeneous and often poorly structured vocabularies, are redundant, and require ad hoc manual resolution. This limitation produces independent data solutions that increase cost, and inhibit discovery, reuse, visibility, and interoperability. Consequently, current scenario data integration efforts are time and resource intensive, lasting from months to years.

A Service-Oriented Architecture (SOA) approach to assist with the discovery, use, and re-use of data will benefit developers by reducing both costs and timelines for data development and integration. The objective of the RDG HLT is to reduce the time and cost of producing and sharing (reuse) of high-quality (well-maintained) data to initialize systems across the DoD M&S enterprise. The RDG program will implement a cross-cutting Common Data Production Environment (CDPE) that aligns with DoD enterprise processes to progress towards a Department level enterprise solution that provides the means to rapidly produce ready data. The RDG effort will use a common set of ontologies, maintained by domain experts committed to the acceptance of tested best practices and vetted by a community of authorities in a well-documented

governance process, which reduce the overall cost of integration efforts and significantly reduce initialization time and increase flexibility and realism. The objective is to improve understandability, reusability, extensibility, and discoverability. The intent is to use a common suite of ontology modules designed for interoperability – along with an effective governance process that brings about a network effect (Weber 2004) where the value of the ontology exponentially increases as more people use it to describe their respective data. When combined with the employment of open source technologies and practices – a legal and technical framework to reduce cost and waste – the result is a web-oriented architecture within which data services, tools and resources of importance to data initialization become more discoverable, composable, and increasingly re-used.

Moreover, to achieve a more efficient, well-tested, and sustainable interoperability solution the training community must transition from static to dynamic data management. DoD should employ Governance, Resourcing, Education, Architecture, Incentivization, and Training (GREAT – pronounced GREAT):

- Governance – Responsibility for developing semantic representations (i.e., ontologies) of core domains must come with authority; otherwise, a new problem is introduced through proliferation of competing models. Authority is needed at both the design and authoring stage of these resources and at successive stages of version management.
 - Ensure reference and domain ontologies are based on need, grounded in representations of the real-world JOE, and align in order to avoid the creation of information stovepipes.
 - Coordinate the development of consistent extensions, identify and document best practices, and ensure that they are being used through control for quality, relevancy, and usability of products.
 - Standardize and facilitate Community of Interest (COI) activities, establish governance to review proposals for change, and ensure dissemination of the ontology and best practices.
- Resourcing – To guarantee quality, organizations must be identified and resourced to develop and maintain ontologies.
- Education – DoD expertise in open source development processes, Semantic Web technologies and best practices and cloud computing is minimal. Education programs must be designed and institutionalized to develop career and vocational specialists to effectively support the DTE needs.
- Architecture – DoD should be provided a best practice architecture that demonstrates how consistent and high availability runtime systems can be implemented to support DoD missions.

- Incentivization – DoD should incentivize organizations to follow OTD and Semantic Web best practices, manage a domain, maintain solutions, and coordinate with other domains in an open and transparent collaborative process.
- Training – Institutionalized training in OTD practices, cloud computing and semantic principles, technologies, and best practices should be established to increase awareness at all management levels and encourage revolutionary transformation of the DTE. To mitigate consequences of multiple contracting agencies and software service contractors working independently on the DTE, it is important to create a cadre of software engineers who share a common understanding of tested best practices in semantic interoperability who will work with the Services, Agencies, and Multinational partners in coordinating activities. A training program should be initiated to address this short-term need. Long term, as semantic interoperability in particular and the NCDS in general, become more ubiquitous to data operations in DoD, we believe that the DoD should create an academic center for OTD, ontology and cloud computing training.

7. SUMMARY

Today's training environment is predominantly a paradigm of fixed training sites and large-scale integrations. This paper discusses the potential for Semantic Web technologies, cloud computing and industry best practices to enhance operational and training data and systems interoperability into a harmonized dynamically evolving framework. The next generation DTE will need to look at new modes of LVC integration, including further exploitation of W3C standards to align M&S and C2 architectures (beyond data exchange to semantic understanding).

The movement of the DoD to Semantic Web is inevitable. All of the services have their own initiatives underway; however, there does not appear to be any department wide coordination or single view of the progress towards a common semantic vision. It is imperative that leadership provide direction and governance to ensure that efforts are accomplished in a coordinated, transparent, and visible way. The organizations that develop capability to form the next generation LVC DTE need to take advantage of the techniques and technologies made available through the W3C and the many communities of interest that are developing technologies and data products that have direct applicability to not only training but also operations. The end state must be a common C2 and M&S enterprise grounded in the operational domain. In some cases, the best of breed efforts may require incentivization to ensure supportability and continuity.

ACKNOWLEDGMENTS

The authors would like to acknowledge Mr. Mark Phillips (Lockheed Martin, Inc.), Dr. Barry Smith (University at Buffalo), Mr. Scott Streit (Intervise, Inc), and Dr. Lowell Vizenor (Alion Science and Technology) for their invaluable input through their work on the *Independent Study of the Semantic Web, Ontologies, and Initiatives for Realizing the Next Generation Defense Training Environment* published 21 February 2011, sponsored by the United States Joint Forces Command. In addition, we would like to thank Ms. Capri Huzi for her superb technical editing efforts.

REFERENCES

- Chief Information Officer. Assistant Secretary of Defense (Networks & Information Integration). http://cio.nii.defense.gov/initiatives/netgenerationuide/current_environment.html
- Department of the Army (DA). "Army Net-Centric Data Strategy". Accessed May 2011. <http://data.army.mil/>
- DOD. 2003. "DOD Net-Centric Data Strategy (NCDS)".
- DOD. 2010. "Strategic Plan for the Next Generation Training for the Department of Defense". DOD Office of the Under Secretary of Defense (Personnel & Readiness)
- Gruber, Tom. 1992. "What is an Ontology?". Accessed April 2011, <http://www.ksl.stanford.edu/kst/what-is-an-ontology.html>
- Jacobs, I. 2010. "Description of W3C Technology Stack Illustration." Accessed May 2011. <http://www.w3.org/consortium/techstack-desc.html>.
- Mandrick, William. 2010. *C2 Core Ontology Study Report*.
- NITRD. 2011. "About the NITRD Program". Accessed May 2011. http://www.nitrd.gov/About/about_nitrd.aspx
- Ruttenberg, A. 2009. "The realist approach to building ontologies for science." Accessed May 2011. http://www.stateofthesalmon.org/agencypartnerships/downloads/SalDAWG_ppts_1109/Ruttenberg-realistapproach.pdf
- Smith, B. et al. 2007. "The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration." *Nature Biotechnology*, 25 (11). 1251–1255. Cambridge, MA: Harvard University Press. Accessed February 2011, <http://www.nature.com/nbt/journal/v25/n11/full/nbt1346.html>
- Stenbit, J. P. 2003. Department of Defense Memorandum, DoD Net-Centric Data Strategy.
- Tolk, A. S.Y. Diallo, K. Dupigny, B. Sun, C.D. Turnitsa. 2005. "A Layered Web Services Architecture to Adapt Legacy Systems to the Command & Control Information Exchange Data Model (C2IEDM)". European Simulation Interoperability Workshop, Paper 05E-SIW-034, Toulouse, France.

- USJFCOMLive. 2010. "C2 Core approved for piloting". Accessed May 2011.
<http://usjcom.dodlive.mil/2010/11/17/c2-core-approved-for-piloting>
- Weber, S. 2004. The Success of Open Source. Cambridge, MA: Harvard University Press. Accessed January 2011.
<http://www.hup.harvard.edu/catalog.php?isbn=9780674018587>
- Wikipedia. 2010. "Ontology (information science)". Accessed May 2011.
[http://en.wikipedia.org/wiki/Ontology_\(information_science\)](http://en.wikipedia.org/wiki/Ontology_(information_science))
- Wikipedia. 2011. "UCore". Accessed May 2011.
<http://en.wikipedia.org/wiki/UCore>

AUTHORS BIOGRAPHY

WARREN BIZUB is the Program Director, Joint Advanced Concepts for the Joint and Coalition Warfighting Center. He served as the Technical Management Division Chief of the Joint Training Development Solutions Group and Director, Joint Advanced Training Technologies Laboratory at USJFCOM. Mr. Bizub holds a BS in Ocean Engineering, is a graduate of the Navy's Senior Executive Management Development Program, and is a MIT Fellow in Foreign Politics, International Relations, and The National Interest.

JULIA BRANDT is a Senior Technical Project Manager with General Dynamics Information Technology at the Joint and Coalition Warfighting Center. She has accumulated over a dozen years of experience in managing M&S R&D projects in the DoD training domain at Alion Science and Technology and BMH Associates, Inc. She also managed an enterprise pilot project for an online live instructor-led training for the US Coast Guard, Performance Technology Center. She received a BS of Occupational and Technical Studies with a Training Specialist Emphasis from Old Dominion University.

MEGGAN SCHOENBERG is a Scientist for the Navy Combat Direction Systems Activity and the Joint National Training Capability (JNTC) Program Information Officer at the Joint and Coalition Warfighting Center. She has 20 years of experience developing and managing C4ISR, Modeling and Simulation (M&S), and business automation software systems. Mrs. Schoenberg holds BS degrees in both Mathematics and Computer Science, and an MS in Systems and Industrial Engineering, Engineering Administration.

SIDE DIFFERENCES IN MRI-SCANS IN FACIAL PALSY: 3-D MODELING, SEGMENTATION AND GRAY VALUE ANALYSIS

Paolo Gargiulo^{(a),(b)}, Carsten Michael Klingner^(c), Egill A. Friðgeirsson^{(a)(b)}, Hartmut Peter Burmeister^(d), Gerd Fabian Volk^(e), Orlando Guntinas-Lichius^(e)

^(a) Department of Development and Consultancy HUT, University Hospital Landspítali;

^(b) Department of Biomedical Engineering, University of Reykjavik,

^(c) Hans Berger Clinic for Neurology University Hospital Jena Friedrich-Schiller-University

^(d) Institute of Diagnostic and Interventional Radiology University Hospital Jena Friedrich-Schiller-University

^(e) Department of Otorhinolaryngology University Hospital Jena Friedrich-Schiller-University Jena

^{(a),(b)} paologar@landspitali.is, ^(c) carsten.klingner@med.uni-jena.de, ^{(a),(b)} egillf05@ru.is, ^(d) hartmut.burmeister@med.uni-jena.de, ^(e) fabian.volk@med.uni-jena.de and ^(e) orlando.guntinas@med.uni-jena.de

ABSTRACT

In this paper, we describe a method to analyze facial muscles and display structural changes occurring under normal and pathological conditions from segmented MR images. Orbicularis oculi and zygomaticus muscles are isolated from the surrounding tissues and their gray values analyzed.

We use 3 Tesla magnetic resonance imaging (MRI) and special image processing tools to produce high quality images and 3-D models of human heads from patients suffering from peripheral facial palsy as well as from healthy probands.

Although peripheral facial palsy is the most common pathology of the cranial nerves with an incidence ranging from 20 to 30 cases per 100,000 people, only a minority of the patients need a far-reaching surgical treatment. A profound diagnostic is essential before deciding for a drastic reconstructive surgery of the face due to the large variety of surgical techniques. This does not only include the classification of the etiology but also of the degree and distributions of the damage of the nerve and of the effected muscles are essential.

Thus, we propose a method to segment and analyze facial muscles from MR data, and to distinguish palsy and normal facial side by measuring local gray values and calculating gradient distributions.

Keywords: MRI, Segmentation, 3-D Modeling, Numerical methods, facial palsy.

1. INTRODUCTION

The term facial palsy summarizes incomplete loss (paresis) as well as the complete loss (paralysis) of facial nerve function. The distinction is highly important, as the indication for surgical reconstruction in patients with incomplete facial palsy has to be assessed much more critically. On the other hand, the reconstruction in case of a complete functional deficit is more complex. Permanent facial palsy and no transient functional deficits are the main indication for surgical reconstruction of facial nerve function.

Depending on the localisation of the lesion site, peripheral facial nerve lesion is separated from central facial nerve lesion: in peripheral palsy, the facial nerve fibres or the motoneurons in the brainstem nucleus are damaged. In contrast, the lesion site in central palsy is located central to the nucleus (supranuclear lesion) in the course of the corticonuclear tract. The otolaryngologist or head and neck surgeon is mostly confronted with patients with peripheral nerve lesion. However, sometimes the exact localisation of the lesion might be unclear, for instance in patients after brainstem astrocytoma surgery.

The type of palsy must be clarified before a reconstructive surgery because any kind of direct facial nerve reconstruction is not effective in patients with central palsy.

From the functional point of view, two different situations have to be distinguished: Firstly, patients without any sign of facial nerve regeneration due to complete interruption of re-sprouting of axons proximal to the lesion site are candidates. Second, patients who have developed spontaneous axonal sprouting but have a functionally hindering defective healing not compensated by central brain plasticity are also candidates for surgical rehabilitation. Defective healing without spontaneous regeneration is impossible. The most important clinical signs of facial nerve defective healing are: a) dyskinesia, i.e. abnormal mimic movements during voluntary action, b) synkinesia, i.e. involuntary synchronous mimic movements while the patient is performing another voluntary movement, and c) autoperalytic syndrome as a special form of synkinesia characterized by synkinetic activity of antagonistic muscles. The synchronous antagonistic movements are detectable using electromyography but the clinical result is a decreased or not visible muscle activity of the intended mimic movement. Dyskinesia and synkinesia can lead to d) hyperkinesia, i.e. an abnormal much stronger movement than physiologically used.

An exact classification of the individual facial palsy due to the above mentioned criteria is mandatory prior to surgical decision making. In addition, the mimic musculature itself, the cerebral cortex and the other cranial nerves have to be examined for pathologies. There have been attempts to identify these nerve changes in MRI; Sartoretti-Schefer S (1998) studied correlation between T2 weighted three dimensional fast spin echo MRI and intraoperative findings for facial nerves in peripheral facial nerve palsy. Correlations between the swelling of the facial nerve and visualization of an enhanced segment by MRI have been studied by In Sup Kim (2007) and Burmeister (2011).

Facial muscles have also been identified and measured in previous studies. MRI measurements of individual muscles and with the mean muscle dimensions was made to assess muscle wasting in facial and tongue muscles in patients with myasthenia gravis (Farrugia 2006). Moreover orbital magnetic resonance imaging (MRI) was used to investigate the structural basis of motility abnormalities in congenital fibrosis of the extraocular muscles (Joseph 2010).

The aim of our study is to gain more appropriate information about the mimic muscles innervated by the facial nerve before, but also after the surgery during recovery. By this, not only the decision for a reconstructive procedure can be supported, but also the postoperative benefit of the procedure can be quantified.

For this purpose, 3-D modelling and segmentation techniques are employed to isolate specific regions of interest from MR data and special computational tools are developed to qualify and quantify facial muscle morphology.

2. MATERIAL AND METHODS

Specific gray value information, 3-D modeling and segmentation techniques have been applied to monitor quadriceps femoris in paraplegic patients undergoing electrical stimulation as described in (Helgason 2005) and (Gargiulo 2010). Similar techniques are being used here.

In this work we develop high (0.67 to 1.0 mm) resolution human head models from segmented MR images. These models have detailed 3-D representation of major tissue surfaces. The gray values distribution within the segmentation masks may offer indications on the muscle conditions and account the differences between denervated and innervated side. For this purpose, we measure mean gray values from the segmented muscles (particularly zygomaticus major and minor and orbicularis oculi muscles), its changes along the scanning axes (z-axis) and finally the gradient distribution.

2.1. MR data

We analyse 17 subjects with unilateral palsy. MRI images were acquired on a 3 Tesla MRI between years 2006 and 2010 (Magnetom Tim Trio, Siemens, Erlangen, Germany) of the head and the face in the department of radiology I in Jena, Germany (Volk

2010). The entire MR imaging was performed using a dedicated 12-channel head coil provided by the manufacturer. All patients were examined in supine position. The imaging protocol of all patients included a sagittal T1-weighted sequence (TR 2300ms, TE 3.03ms, flip angle 9°, voxel size 1mm × 1mm × 1mm [=1mm³], matrix 256 x 256, TA 5:21 min) covering the whole head including the face. Each slice has 256 × 256 pixels, and each pixel has a gray value within the 4096 gray-scale values, meaning that it is represented with a 12-bit value. The contiguous slice thickness was 1.0 mm and pixel resolution was 1.0 mm. A total data set from a single scan of 192 slices is therefore (256 × 256 × 12 × 192)/8 = 1,9 * 10⁷ bytes, considering that 2 bytes are needed for 12 bit representation then a data set of this type is approximately 36 MB. This data set gives a complete 3D description of the tissue within the head.

During our study we developed an improved coronal T1-weighted fast low angle shot 3D-sequence with high spatial resolution (TR 5.67ms, TE 2.48ms, flip angle 11°, voxel size 0.67 mm × 0.67 mm × 0.67 mm [=0.3mm³], matrix 384 x 384, TA 9:35 min), focused on the facial muscles only including the face and the forehead. Slice orientation was tilted by 90° compared to the orthogonal plane running parallel to the hard palate. In this case, a typical data set with 192 slices will be approximately 81 MB. From 2010 on, we used both sequences to get an optimum of information about the facial muscles.

Figure 1 shows a sagittal view from the two MR protocols used in this work. The difference between figure 1A (protocol 1) and B (protocol 2) is clear; the image resolution and contrast are much better in the latest protocol (figure 1B), this improvement facilitates the segmentation work.

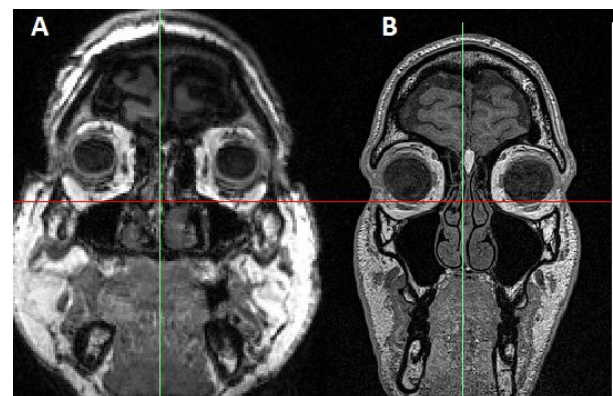


Figure 1: Comparison between sagittal views from two different MR sequences: protocol 1, T1-weighted sequence (A); protocol 2, T1-weighted fast low angle shot 3D-sequence with high spatial resolution (B).

2.2. Image processing

In order to isolate the single muscle segment and measure the growth, MR data are imported into a special image processing and editing computer program called MIMICS (www.materialise.com). In this software environment, the 3-dimensional form of the

facial muscles is reconstructed and specific regions of interest (ROI) are extracted and isolated.

Segmentation process begins by establishing a threshold, which discriminates the region of interest from the rest by selecting appropriate ranges of gray values (GV). From the visual point of view thresholding allows highlighting (for example in different colors) pixels with certain gray values from the others. In an $M \times N$ slice, each element in the image matrix $a[m,n]$ displays a level of brightness coded by a grey value which in medical imaging varies from 0 to 4095 ($= 2^{12} - 1$). For example, the region of interest to be visualized is between GV_{min} to GV_{max} then the threshold test condition will appear as in (1):

$$\begin{aligned} \text{If } GV_{min} \leq a[m,n] \leq GV_{max} \text{ then } a[m,n] = \text{object} = \text{color} \\ \text{Else } a[m,n] = \text{background} \end{aligned} \quad (1)$$

After thresholding further segmentation tools are usually required to isolate ROI from surrounding areas. Region growing for example is a segmentation tools used to eliminate floating pixels from belonging to the selected threshold. It is often used when determining pixel class membership: pixels that belong to the same region are connected. Other operations such as Boolean and morphological are possible on defined pixel classes in order to improve the segmentation work.

MRI show higher detail in the soft tissues such as tendons and ligaments but very often MR data require more difficult and time consuming segmentation work compared to CT because of the absence of defined thresholds for specific tissues such as muscle, bone, fat, etc. (Gargiulo 2011). In this work, depending on the quality of the MRI data, automatic segmentation techniques are combined with manual editing to correct misclassified pixels.

2.3. Facial muscle segmentation

3-D representation of major muscle tissues within face including orbicularis oculi, zygomaticus major and minor, nasalis and levator labii superior muscles were developed here. We used E-anatomy templates (<http://www.imaios.com/en/e-Anatomy>) to localize the different facial muscles. As a control, muscles of the non-injured contralateral sides were segmented and analysed.

The facial muscle threshold for the patients scanned with the first protocol (Fig.1A) is [300, 600] GV while the muscle threshold for the patients scanned with the second protocol (Fig.1B) is [50, 300] GV. Other tissues surrounding the facial muscle are displayed within the same gray value intervals. For this reason editing tools applied slice by slice were used to isolate the muscles.

The process starts from a cross section where the selected tissue boundaries are well visible. A contour is manually drawn around the region of interest (ROI) and projected to the next cross sections in both directions. If the contour fits well the new cross sectional area is then projected unchanged forward to the next slice, otherwise it is adapted using manual editing and then projected to the next slice. The process continues until

all cross sections containing the selected ROI are covered. The contour areas are then erased creating a gap between the ROI and surrounding.

Finally, the facial muscle mask is created applying a region growing procedure which separates the edited ROI which is no longer connected to the surrounding (tissue). The result of the segmentation process applied on the two protocols is shown in figure 2.

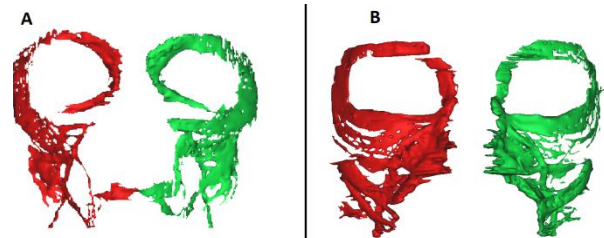


Figure 2: 3-dimensional reconstruction from facial muscle segmentation (right –red mask, left-green mask) from protocol 1 (A) and protocol 2 (B).

Immediately available after segmentation are the mask volumes and mean gray values. The distribution of mean gray values for zygomaticus and orbicularis oculi from all the patients included in the study is shown in figure 3. Red bars indicate mean gray values from palsy side while blue bars indicate the healthy contralateral side, in figure 3A are seen zygomaticus GVs and in figure 3B orbicularis oculi GVs.

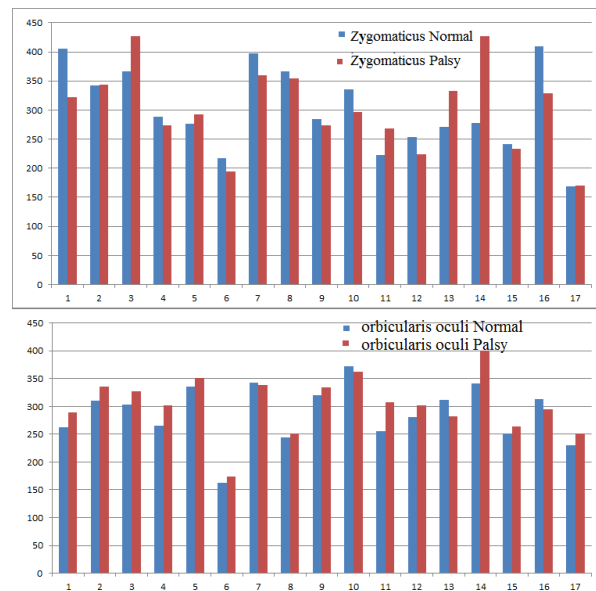


Figure 3: Mean gray values from segmented mask: Zygomaticus major and minor (A), and orbicularis oculi (B).

The segmentation mask is exported as matrix of dimension: $np \times 4$, where np is the number of pixels contained in the mask and the four columns account respectively the voxel coordinate x, y, z and the relative GV. These data are further processed in Matlab (Matworks Inc).

2.4 Mean Gray values along the cross section

Qualitative and quantitative information concerning the segmentation mask are obtained calculating, the mean GV on each cross section along the scanning axes (z). In this way the segmentation masks representing facial muscles from palsy and normal sides can be evaluated and compared (Fig. 4). The facial muscle lengths are between 60 and 100 mm depending from patient anatomy (in our study the region of interest start above the orbicularis oculi and end above the labialis) the slice increment is 0.1 mm therefore the number of mean values is between 60 and 100. In figure 4 the right (palsy) profile is 78,95mm while the left (normal) profile is 87,95 mm long.

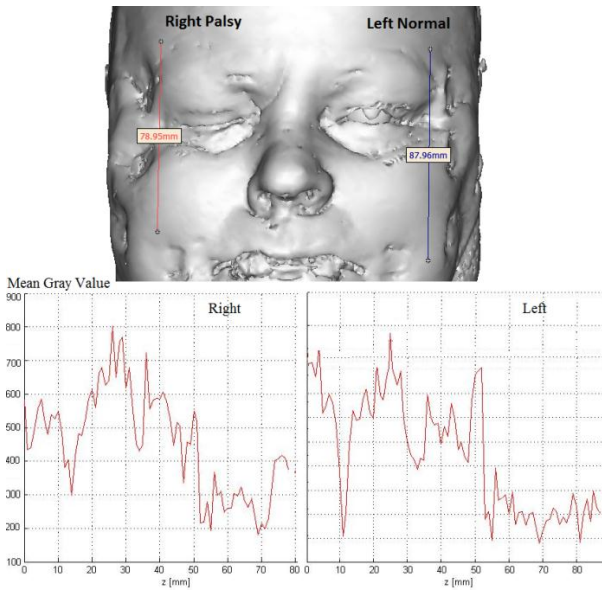


Figure 4: Patient with right palsy and associated mean GV profile along the z axes.

2.5 Gradient distribution

Consider a cross section from the MR data in which the tissue composition is given by a scalar field, GV (associated to each pixel), so at each point (x,y,z) the tissue is displayed with $GV(x,y,z)$. At each pixel in the image, the gradient of GV at that point will show the direction the tissue GV changes most quickly. The magnitude of the gradient will determine how fast the GV rises in that direction. The algorithm to calculate and display the gradient distribution from the segmented muscle was developed in Matlab (Matworks Inc).

First, the gradients were computed using a central finite difference approximation for each voxel. So for each point $f(x, y, z)$ each gradient component is found by (2):

$$\begin{aligned} \frac{\partial f}{\partial x} &= f\left(x + \frac{h_x}{2}\right) - f\left(x - \frac{h_x}{2}\right) \\ \frac{\partial f}{\partial y} &= f\left(y + \frac{h_y}{2}\right) - f\left(y - \frac{h_y}{2}\right) \\ \frac{\partial f}{\partial z} &= f\left(z + \frac{h_z}{2}\right) - f\left(z - \frac{h_z}{2}\right) \end{aligned} \quad (2)$$

Where h_x , h_y and h_z are the separation between the adjacent points in x, y and z directions, respectively. This gradient magnitude distribution was then visualized using isosurfaces and color coding as can be seen in figure 5 (same patient of figure 4).

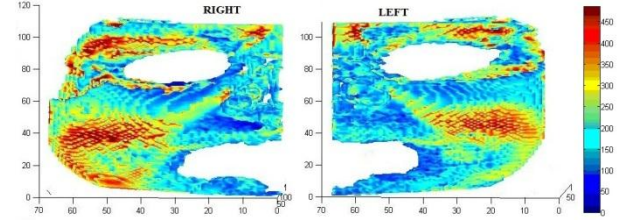


Figure 5: 3-D gradients distribution and color coding from a patient with right palsy: high gradients area is colored in red, low gradient area in blue.

3. RESULTS

Muscle volumes are calculated from the segmented data. The result from such segmentation process depends strongly upon the MR data quality. So far, our volume measurements are not reliable in term of absolute value; indeed the segmentation accuracy and the resolution of the images are not yet optimal. What we are actually using are the differences between volumes: within the same data set in order to compare healthy from palsy side and from time to time between different MR data to monitor recovery process in patient undergoing surgical treatments. In figure 6 volume changes are displayed in a patient with idiopathic facial palsy on the right side with good recovery in less than 3 month. The patient is scanned in the first week after onset of palsy (fig.6.A) and then again after 3 months (fig.6.B). Here it can be noticed that the difference between left and right muscles volume is significantly reduced in the first scan on the parietic side.

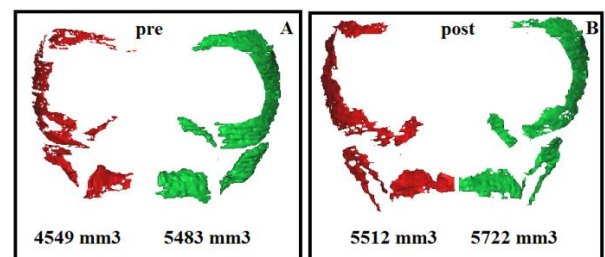


Figure 6: Muscle 3-D reconstruction and volume changes in a patient with idiopathic facial palsy on the right: 1st scan in the first week after onset of palsy (A), 2nd scan after complete recovery after 3 months (B).

The analysis of the mean gray value from the segmented mask (see figure 3) shows that orbicularis oculi GV on the palsy side tend to have higher values compared to normal: 13 from 17 patients measured. Vice versa zygomaticus muscles GV on the palsy side tend to have lower compared to normal. The measurements from the same 17 patients show that: 10 patients have lower GV value, 2 times approximately the same value, 5 patients have higher GV on palsy

side. This fact can be seen also in figure 4 where the patient gray value along the cross section show higher GV in the region of orbicularis oculi (first 30 mm of the GV profile) and generally lower in the region of zygomaticus muscles.

The analysis of the gradient shows also some interesting data. Figure 7 gather the gradient distributions from 4 patients and 2 controls. The histograms highlights that left and right gradient distributions are never completely symmetric even in control subjects. Moreover, our preliminary results show a strong tendency of having higher gradient on the palsy side meaning that there is more variability between gray values in the regions where the muscles are denervated. Likewise on the healthy side, there is noticeable higher number of pixels with low gradient meaning that in these volumes there are small differences between pixel gray values.

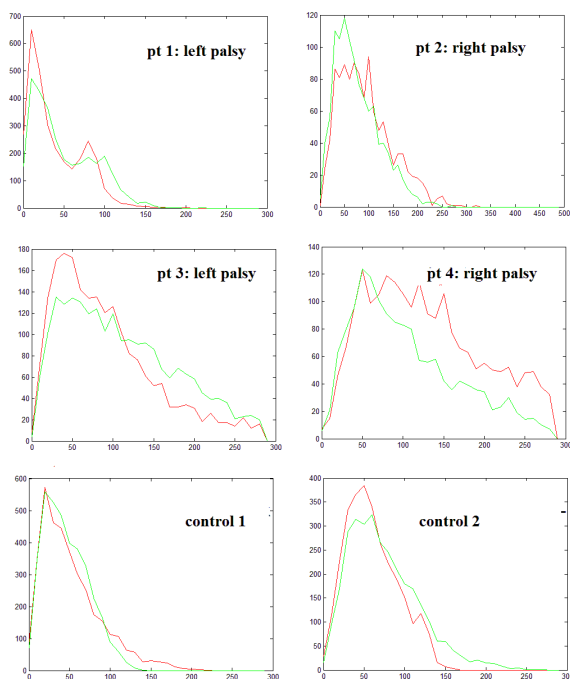


Figure 7: histograms showing gradient distributions: in red the right side and in green the left side.

Finally for the patient depicted in figure 4 (patient with idiopathic facial palsy on the right) we develop the 3-D gradient distribution to observe local changes from time to time. Figure 8 shows that in the first scan (fig.8.A) the right side (with acute idiopathic palsy) has higher gradient, above 300 GV, especially in the region of orbicularis oculi muscle. The situation changes in the second scan when the patient recovers (fig.8B). Here the high gradient areas are similar on both side of the face.

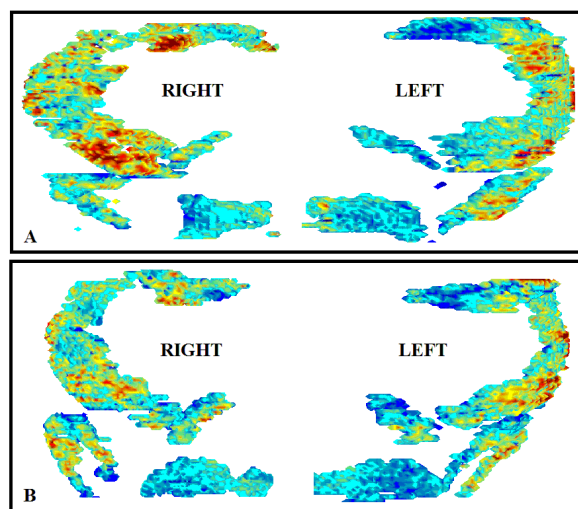


Figure 8: 3-D gradient distribution before (A) and after palsy recover (B)-on the right side.

4. CONCLUSION

The technique introduces the possibility to determine structural changes in individual facial muscles and muscle groups selectively. Comparisons between both sides and also comparisons over time are possible. Therefore, using these techniques, even before opening the skin, the surgeon has already information about the quality and quantity of the facial muscles. Considering these information can help to choose the optimal concept in facial reconstruction for the individual patient.

During the long time of recovery after a reconstruction of the facial nerve, the technique can help to quantify the process of reinnervation. This information can not only improve individual concepts of physiotherapy but also help to understand principal concepts of de and re-innervated muscles. So we hope to develop a valuable tool for facial surgeons, physiotherapists but also for basic research.

ACKNOWLEDGMENTS

This work has been supported by MED-EL research fund. (<http://www.medel.com/>)

REFERENCES

- Burmeister HP, et al 2011 "Evaluation of the early phase of Bell's palsy using 3 T MRI" European Archives of Oto-Rhino-Laryngology.0937-4477
- Farrugia M E "MRI and clinical studies of facial and bulbar muscle involvement in MuSK antibody-associated myasthenia gravis" Brain (2006), 129, 1481-1492
- Gargiulo P, et al 2010. "Quantitative colour 3-dimensional computer tomography imaging of human long-term denervated Muscle" Neurological Research. 32:13-20
- Gargiulo P, et al 2011 "Monitoring of Muscle and Bone Recovery in Spinal Cord Injury Patients Treated with Electrical Stimulation Using Three-Dimensional Imaging and Segmentation Techniques: Methodological Assessment" Artificial Organs 35(3):275-281

- Helgason T, et al 2005 “*Monitoring Muscle Growth and Tissue Changes Induced by Electrical Stimulation of Denervated Degenerated Muscles With CT and Stereolithographic 3D Modeling*” *Artificial Organs* 29(6):440–443,
- In Sup Kim 2007 “*Correlation between MRI and Operative Findings in Bell’s Palsy and Ramsay Hunt Syndrome*” *Yonsei Med J* 48(6):963 - 968, 2007
- Joseph L. Demer 2010. “*Evidence of an Asymmetrical Endophenotype in Congenital Fibrosis of Extraocular Muscles Type 3 Resulting from TUBB3 Mutations*” *IOVS*, September 2010, Vol. 51, No. 9.
- Sartoretti-Schefer S, 1998. “*T2-weighted three-dimensional fast spin-echo MR in inflammatory peripheral facial nerve palsy*”. *AJNR Am J Neuroradiol* 1998;19:491-5.
- Volk GF, et al 2010. “*Modern concepts in facial nerve reconstruction*” *Head & Face Medicine*. 1;6:25.

AUTHORS BIOGRAPHY

Paolo Gargiulo is assistant professor at Reykjavik University and works as biomedical engineer and researcher at the University Hospital of Iceland. He studied electrical engineering at University Federico II in Naples, and obtained a PhD at Technical University of Vienna, Austria. His research interests are in the field of electrical stimulation medical modeling and rapid prototyping for clinical applications.

Hartmut Peter Burmeister studied medicine in Würzburg, Vienna, and Lübeck. He works as postdoctoral researcher and radiologist at the Institute of Diagnostic and Interventional Radiology of the University Hospital in Jena, Germany. He is specialized in Head and Neck magnetic resonance imaging, in particular imaging of cranial nerves and human sense organs.

Egill Axfjord Fridgeirsson is a master student at Reykjavik University Iceland in Biomedical Engineering. His research interests are Medical modeling, Biomagnetism and clinical applications of medical models. He currently works in the Clinical Engineering department of Landspítali University hospital.

Gerd Fabian Volk works as medical doctor at the Department of Otorhinolaryngology of the University Hospital in Jena. He studied medicine at the University of Münster. His scientific focus is on regeneration of nerves and imaging procedures to track this process.

Orlando Guntinas-Lichius is head of the Department of Otorhinolaryngology of the University Hospital Jena Friedrich-Schiller-University Jena. He studied medicine and completed his specialist training as otorhinolaryngologist at the University of Cologne, Germany.

EXPLOITING VARIANCE BEHAVIOR IN SIMULATION-BASED OPTIMIZATION

Pasquale Legato^(a), Rina Mary Mazza^(b)

^(a)Dipartimento di Elettronica, Informatica e Sistemistica (DEIS), Università della Calabria
Via P. Bucci 41C 87036 Rende (CS), Italia

^(b)Dipartimento di Elettronica, Informatica e Sistemistica (DEIS), Università della Calabria
Via P. Bucci 42C 87036 Rende (CS), Italia

^(a)legato@deis.unical.it, ^(b)rmazza@deis.unical.it

ABSTRACT

The methodological contribution proposed herein arises from considering the integration of stand-alone optimization techniques in discrete-event simulation, in order to model dynamic logistic processes, under realistic conditions of uncertainty. A simulation-based optimization approach is investigated to optimize overall system performance measures. A special focus is laid on evaluating alternative system solutions when they are either known *a priori* or revealed at run-time. To this end, a variance-guided statistical technique for the ranking and selection of candidate solutions has been devised and integrated into a solution generating algorithm on which the search process for the best solution may be centered. The findings returned from this work have been coupled with a queuing network model developed and applied in container stacking/retrieval operations via *dts - direct transfer system* on the yard of a real maritime container terminal for pure transshipment.

Keywords: discrete-event simulation, optimization, simulation-based optimization, metaheuristics, logistics

1. INTRODUCTION

Many modern day systems providing products and services in popular fields such as logistics, manufacturing, transportation, network-centric computing, etc., are studied with the objective of carrying-out performance analysis and optimization. The greater the complexity of similar systems, the more the common approach to problem design and solution is based on decomposing the original problem into several smaller models. However, when dealing with dynamic and random-based activities, to deliver overall optimized system performance, a more satisfactory contribution could spring from the combination of stand-alone algorithms used for optimum-seeking with discrete-event simulation used for performance evaluation. This awareness has led to the introduction of an integrated methodology which significantly aids decision-making under uncertainty: Simulation-based Optimization (SO).

In (Fu 2005), the author divides the types of SO techniques in the following main categories:

- statistical procedures (e.g. ranking & selection procedures and multiple comparison for the comparison of two or more alternative system configurations);
- metaheuristics (methods directly adopted from deterministic optimization search strategies such as simulated annealing);
- stochastic optimization (random search, stochastic optimization);
- other (including ordinal optimization and sample path optimization).

Here we focus on procedures, included in the first two categories, to generate and estimate the best among a set of alternative solutions, whether they are all known in advance or actually revealed during a simulation run. To select the best system, we devise a decisional mechanism based on variance estimation with the purpose of guiding the sampling activity required to perform the analysis of simulation output. We then integrate the SO models proposed into a computational framework and exploit this unifying structure with reference to the container stacking/retrieval process occurring in the container terminal of Gioia Tauro in Southern Italy. The final objective amounts to selecting the best among a set of different policies adopted by the operation manager to transfer yard cranes from one block of the container storage area to another.

2. SIMULATION-BASED OPTIMIZATION

2.1. The Methodology

Simulation-based optimization consists in searching for the settings of controllable decision variables that yield the maximum (minimum) expected performance of a stochastic system that is represented by a simulation model (Fu and Nelson 2003). Formally,

$$\max (\min) E[f(\theta)] \quad (1)$$

where θ is the vector of decision variables and $E[f(\theta)]$ the mathematical expectation of the performance measure of interest which should be estimated by statistics on random variates returned from simulation-generated sample paths.

As we will later see, the alternative systems to be simulated can either be a limited number and all known in advance or a great, but countable number and generated by a properly designed optimization procedure. Whatever the case, the simulation component of the SO solution effort calls for the following considerations.

Performance evaluation is based on observations that are random variates returned by a simulation process. Thus, one may or may not select the system solution which is truly representative of the best solution. To deal with this, we consider an indifference-zone (IZ) ranking and selection (R&S) procedure and give some background information for this approach.

In terms of notation, let

k	the number of alternative simulated system solutions ($i=1..k$),
n	the number of observations sampled from each system solution ($j=1..n$),
$\mu_1, \mu_2, \dots, \mu_k$	the unknown k expected values of the performance measure of interest,
$\mu_{[k]} \geq \dots \geq \mu_{[1]}$	the ordered unknown k expected values of the performance measure of interest,
$\bar{X}_k, \dots, \bar{X}_1$	the sample means of the performance measure of interest for each system solution,
$P\{CS\}$	the probability of correct selection,
δ	the indifference zone chosen by the experimenter.

An IZ procedure is statistically indifferent to which system solution is chosen among the k competing alternatives when all these alternatives fall within a fixed distance δ from the best solution. In a maximization problem the probability of performing a correct selection with at least level of confidence P^* is

$$P\{CS\} \triangleq P\{\mu_k > \mu_i \forall i \neq k \mid \mu_k - \mu_i \geq \delta\} \geq P^*. \quad (2)$$

Under the hypothesis of normality of the statistics involved, this probability was first computed by Rinott in (Rinott 1978) starting from the following inequality

$$P\{CS\} \geq \int_{t=0}^{\infty} F_{T_{k-1}}(t+h) f_{T_k}(t) dt \quad (3)$$

where

$$T_k \triangleq \frac{\bar{X}_k - \mu_{[k]}}{\delta/h} \quad \text{and} \quad T_{k-1} \triangleq \frac{\bar{X}_{k-1} - \mu_{[k-1]}}{\delta/h} \quad (4)$$

are distributed according to Student's law. The above integral is set equal to P^* and solved numerically for h , for different values of n . Numerical values for h , which is also known as Rinott's constant, are tabled in (Wilcox 1984).

In conclusion, when simulating k alternative system solutions, IZ procedures guarantee the selection of the "best" solution or a "near best" according to a pre-specified probability. From a practical point of view, considering a large number of simulation replications for each solution reduces sampling errors; on the other hand, the computational expense of even one single replication of any simulation model is likely to be cumbersome. Bearing in mind these conflictual objectives, pioneering two-stage indifference-zone ranking and selection (R&S) procedures (Rinott 1978, Dudewicz and Dalal 1975) have been followed by more recent and advanced procedures based on an n -stage logic, with $n > 2$ (Kim and Nelson 2001, Chen and Kelton 2005). In our SO approach we also exploit an n -stage IZ R&S procedure where the idea of "efficient" sampling is pursued by basing the number of output observations to be taken from each system on the corresponding variance behavior (i.e. how variance changes as the sample from simulation output grows), given a fixed computing budget. Thus, for our enhancement, it is necessary to establish how such variance should be estimated.

If for system i ($i=i..k$) the n elementary output observations $X_i \triangleq \{X_{ij}, j=1..n\}$ returned from a simulation run are independent and normally distributed, one may pursue variance estimation by simply using classical statistics and computing the sample mean

$$\bar{X}_i = \frac{1}{n} \sum_{j=1}^n X_{ij}, \quad (5)$$

followed by the unbiased sample variance which is used as variance estimator

$$VAR[X_i] = S_i^2 = \frac{1}{n-1} \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2. \quad (6)$$

Should this not be the case - as customary in a simulation-based study of practically any real-life system - then one must start from the output stochastic process, organize its data and compute the process variance.

For example, for system i let $\{X_1, \dots, X_j, \dots, X_n\}$ be a *weekly dependent* stationary output process with

mean μ_X and variance σ_X^2 . This process is said to be *weakly dependent* if the lag-j covariance

$$\gamma_j \triangleq \text{Cov}[X_i, X_{i+j}], \quad j = 0, \pm 1, \pm 2, \dots \quad (7)$$

satisfies $\gamma_j \rightarrow 0$ as $|j| \rightarrow \infty$ (Billingsley 1995).

If one chooses to organize this data in batches of size k , the sample mean for batch i is given by:

$$\bar{X}_i(k) \triangleq \frac{1}{k} \sum_{j=i+1}^{i+k} X_j \quad (8)$$

and according to the Central Limit Theorem

$$\bar{X}_i(k) \xrightarrow{D} Z(\mu_X, \sigma^2(k)/k), \quad k \rightarrow \infty \quad \forall i \quad (9)$$

where

$$\sigma^2(k) = \sigma_X^2 + 2 \sum_{j=1}^{k-1} \left(1 - \frac{j}{k}\right) \gamma_j. \quad (10)$$

Furthermore, the variables in the following set

$$\{\bar{X}_1(k), \dots, \bar{X}_i(k), \dots, \bar{X}_n(k)\} \quad (11)$$

become independent as $k \rightarrow \infty$ and

$$\lim_{k \rightarrow \infty} \sigma^2(k) = \lim_{k \rightarrow \infty} k \text{Var}[\bar{X}_i(k)] = \sigma_X^2, \quad \forall i. \quad (12)$$

By (Hogg and Craig 1978)

$$\frac{S_{\bar{X}}^2(n, k)}{\sigma_X^2 / k} \approx \frac{\chi_{n-1}^2}{n-1}. \quad (13)$$

Applying the mathematical expectation to the above formula

$$E \left[\frac{S_{\bar{X}}^2(n, k)}{\sigma_X^2 / k} \right] = E \left[\frac{\chi_{n-1}^2}{n-1} \right] = 1 \quad (14)$$

and thus

$$E[k \cdot S_{\bar{X}}^2(n, k)] = \sigma_X^2 \quad (15)$$

where

$$S_{\bar{X}}^2(n, k) \triangleq \frac{k}{n-1} \sum_{i=1}^n \left(\bar{X}_i(k) - \bar{\bar{X}}(n, k) \right)^2 \quad (16)$$

is the estimator of the (output) process variance.

This stated, our procedure uses a variance-weighted decisional mechanism based on the variance estimator described above to guide the sampling activity on the number of additional simulation output observations to be taken from each system. Practically speaking, when process variance decreases this multi-stage procedure is expected to terminate faster than classical two-stage R&S algorithms because of its auto-adaptive control. In every other case, the number of iterations during which the sample variance either remains constant (during the last x runs) or increases is controlled by an upper bound (UB) on the number of additional simulation runs to be carried-out which is given by the well-known formula based on Rinott's constant

$$\text{additional runs} = \left(h^2 S_i^2 / \delta^2 \right) \quad (17)$$

The following pseudo-code provides a high-level description of our approach when considering a maximization problem:

Table 1: Our IZ R&S Procedure

1	$P^*, \delta, n_0, h, x, UB \leftarrow$ select procedure settings
2	for $i = 1$ to k do
3	for $j = 1$ to n_0 do
4	$X_{ij} \leftarrow$ take a random sample of n_0 from each of the k systems
5	end for
6	$\bar{X}_i \leftarrow$ compute an estimate of the sample mean of the performance index of interest for system i
7	update <i>stopping condition</i> [n]
8	end do
9	$N_i = \max(n_0, h^2 S_i^2 / \delta^2) \leftarrow$ determine the sample size to take from each system
10	if $n_0 \geq \max_i N_i$ then
11	$\max_i \bar{X}_i \leftarrow$ select system with greatest sample mean as best and stop
12	Else
13	For $i = 1$ to k do
14	while $N_i \leq UB$ do
15	$X_{ij} \leftarrow$ take one additional random sample for system i
16	$\bar{X}_i \leftarrow$ compute an estimate of the sample mean of the performance index of interest for system i
17	$S_i^2 \leftarrow$ compute a run-weighted estimate of the sample variance of the performance index of interest for system i
18	$N_i = \max(n_0, h^2 S_i^2 / \delta^2) \leftarrow$ determine the new sample size for system i
19	if $N_i \leq n_0$ or $S_i^2 = \text{constant}$ in the last x runs then

20	stop sampling for system i
21	end while
22	End for
23	$\max_i \bar{X}_i \leftarrow$ select system with greatest sample mean as best

So doing, our approach avoids relying on too much information obtained in just one stage and, at the same time, allows to save on computing budget.

2.2. The Framework

The simulation-based optimization framework now proposed in Table 2 serves a double purpose. On one hand, it offers a common ground where to define and compare the different IZ R&S techniques that, in turn, are recalled throughout this work or in companion papers (Legato, Canonaco and Mazza 2009). On the other, it shows how a simulation engine inserted in an optimization algorithm is often the only practical solution method available when dealing with difficult-to-solve combinatorial problems, embedded in realistic dynamic logistic processes characterized by several elements of randomness.

Table 2: SO Framework for Solution Generation and Evaluation

1	$k, n=0, \text{stopping condition}[0] \leftarrow$ select procedure settings
2	$i^* = i \leftarrow$ set best solution = initial solution
3	while $\text{stopping condition}[n] = \text{false}$ do
4	$n = n + 1$
5	$i_1(n), i_2(n), \dots, i_k(n) \leftarrow$ at iteration n take/generate k alternative solutions
6	$i^* = \text{best}\{i^*, [i_1(n), i_2(n), \dots, i_k(n)]\} \leftarrow$ compare the k alternative solutions at iteration n with current best and, eventually, update the best
7	update $\text{stopping condition}[n]$
8	end do
9	$i^* \leftarrow$ return best solution

As one may observe, on line 5 solutions are either taken or generated. In the latter case, a metaheuristic approach based on a variant of the well-known Simulated Annealing (SA) algorithm (Alrefaei and Andradóttir 1999) has been adopted. Besides discarding the basic assumption according to which the temperature $Temp_k \rightarrow 0$ as $k \rightarrow \infty$ by assuming $Temp_k = Temp \forall k$, this approach bears two possible ways of estimating the optimum solution. It either uses the most visited solution or selects the solution with the best average estimated value of the objective function. The effectiveness of this constant temperature approach is not yet consolidated for complex and large practical applications. (Mazza 2008) discusses this issue and introduces a guided-search refinement in the SA

algorithm based on choosing the candidate solution j among m neighboring solutions j_1, j_2, \dots, j_m of the current solution i .

As for solution comparison and selection, the procedure reported in Table 1 is inserted on line 6 of the above schema.

3. APPLICATIONS IN PORT LOGISTICS

Container terminal logistics have received great interest in the scientific literature from both the theoretical and practical standpoint (Stahlbock and Voß 2008). The reason for such concern is straightforward if one considers the number and random nature of operational activities carried-out in these facilities: vessel arrival and berthing, resource assignment and scheduling, container transfer and handling, emergency management (e.g. equipment failure, congestion phenomena, weather conditions) and so on. In a maritime container terminal many different company-based rules, regulations and practices can be the grounds of application for the simulation-based optimization framework previously described. Real case studies are given in companion papers (Legato, Mazza and Trunfio 2008; Legato, Mazza and Trunfio 2010). Here we consider the yard and some organizational and operational issues pertaining to its role within the terminal. We then propose to manage the yard activity with respect to policies and equipment employed for container stacking/retrieval by applying the SO approach.

3.1. Problem Description

The purpose of a stacking yard in a terminal is to provide storage space for containerized cargo during import, export or transshipment operations. Whether dedicated or shared among different shipping companies, suitably-sized lots of the yard are generally assigned to each company and equipped with technological means in order to enable the stacking/retrieval of container batches (i.e. a set of containers sharing some common properties).

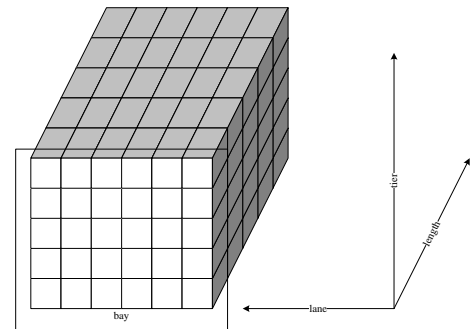


Figure 1: Definition of a Yard Block.

A yard is typically organized in *zones* that, in turn, are divided into *blocks*. As shown in Figure 1, the size of a block is defined by three dimensions: *i*) number of *lanes* or *rows* (e.g. 6 or 13, along with an extra lane if internal trucks are used to perform container transfer);

ii) number of container tiers or stack height for each lane (e.g. 5); iii) number of containers in length (e.g. 20). A vertical section of a block (e.g. 5 tiers * 6 lanes) is normally referred to as bay.

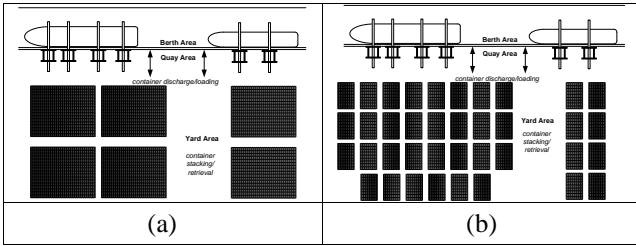


Figure 2: Two Alternative Yard Organizations.

It is worth observing that both the number and size of blocks in a yard affect the average travel time of shuttle vehicles cycling between the quay and the yard areas, as well as the container handling time on the yard. For instance, in the yard organization depicted by Figure 2.(a), the average distance to be covered in order to reach a container is greater than the average distance deriving from the solution portrayed in Figure 2.(b). On the other hand, more container handling equipment can be concentrated in a specific area in the former case, thus returning a smaller service time, whereas this possibility is prevented in the latter case due to potential interference between container movers meant to operate on adjacent yard bays

If container stacking/retrieval on the yard is performed by transfer cranes, such as rail-mounted gantry cranes (RMGCs) or rubber-tired gantry cranes (RTGCs), then a common operational issue actually consists in periodically deciding how many and which cranes are to be assigned to a block. This decision usually depends on the expected daily workload in each block and, therefore, on the total crane capacity (measured in time units) required to complete container stacking/retrieval operations.

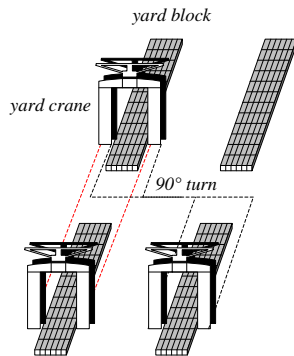


Figure 3: Possible Intra-block Crane Transfer.

To do so, cranes must be transferred from one block to another. If we consider RTGCs, these cranes can travel between adjacent yard blocks without any turning motion or by changing lanes. In the former case, crane transfer can take about 10 minutes; in the latter, about 5 additional minutes are required to perform 90 degree

turns (see Figure 3). These movements are exclusively referred to inter-block (and not inter-zone) crane transfer.

In our study, we focus on the new operational scenarios generated by five alternative management policies - all known a priori - for assigning yard cranes to yard blocks and accounting for order, times and routes of the crane transfer. The objective is to select, by way of the SO framework, the policy which allows us to minimize the maximum average time to complete stacking/retrieval operations of suitable batches of containers in the yard.

3.2. Numerical Experiments

To perform the comparison of five alternative system solutions we consider the corresponding variance patterns with respect to a hypothetical operational scenario in which average container traffic in yard blocks is at a medium level (e.g. not many shipping lines stack/retrieve containers in that area) and average crane transfer times between blocks are high (e.g. in an extensive yard area). Figure 4 illustrates an example of how variance changes as the samples taken from system simulation under different policies grows. Observe that for the first three policies variance behavior is stable, meaning that there are no significant changes in variance estimation as the sampling procedure progresses. Thus the algorithm continues adding single observations (or batches or simulation replications) as required by the "stable" variance estimate until the upper bound provided by Rinott's two-stage procedure is reached (Legato and Mazza 2008). When the variance pattern increases, as for policy n°4, the upper bound is still provided by Rinott's procedure.

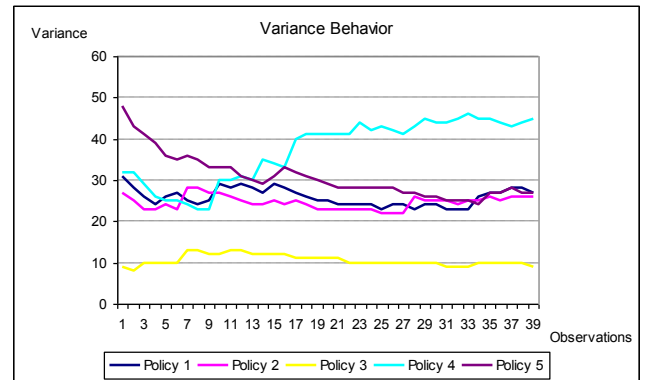


Figure 4. Sample Paths of Variance Behavior

Instead, in policy n°5 the variance estimate has a decreasing trend and, thus, the algorithm is expected to terminate faster. This expectation is justified by the auto-adaptive control of the procedure which can be monitored according to a step-by-step logic. In this sense, Table 3 provides a trace of the variance behavior for policy n°5. As one may observe, after setting $n_0 = 10$, $P^* = 0.90$, $\delta = 5$ and, thus, $h = 3.137$, according to Rinott's procedure the number of runs to consider for system i are

$$N_i = \max(n_0, h^2 S_i^2 / \delta^2) \\ = \max(10, 3.137^2 * 120.30 / 5^2) = 48 \quad (18)$$

So, $N_i - n_0 = 38$ additional runs must be added to guarantee the predefined probability of correct selection $P^* = 0.90$. Alternatively, as shown in Table 3, our procedure after only one supplementary run at step 11, returns

$$N_i = \max(11, 3.137^2 * 108.30 / 5^2) = 43 \quad (19)$$

meaning 32 additional runs (i.e. 43 – 11 previous runs). It, thus, realizes a gain of 6 runs after one single run.

Table 3: Step-by-step Trace of Variance Behavior for Policy n°5

Step	N° of observations for policy i=5		N _i
	Sample mean	Sample variance	
10	92.34	120.30	48
11	92.39	108.30	43
12	92.96	102.34	41
13	92.48	96.85	39
14	92.20	90.49	36
...

In numerical terms, given that both procedures choose policy n°3 as best, in the worst case our procedure returns the same results as Rinott's two-stage procedure ($\Delta = 0\%$), while for decreasing variance behavior our procedure is more efficient by 31,25%, as illustrated in Table 4.

Table 4: Comparison of Observations Required by Rinott's Procedure (RP) and Our R&S procedure

Alternatives	N° of observations		Our Performance ($\Delta\%$)
	RP	Ours	
policy 1	31	31	0%
policy 2	27	27	0%
policy 3	9	9	0%
policy 4	32	32	0%
Policy 5	48	33	+31.25%

4. CONCLUSIONS

An n -stage indifference-zone based ranking and selection procedure has been proposed to "hopefully" deliver more efficient sampling than classical two-stage algorithms. Its performance has been tested by some numerical experiments. Rather than just using a classical sample mean, it appears that tracking the variance behavior reveals improvement margins when the variance pattern is decreasing. In the future, a further possibility may lie in investigating how to use an

estimate of the skewness of the sample mean distribution, given that the normality assumption is approximately verified only after a large number of simulation runs - a condition one should avoid, due to the computational burden it is bound to bear.

REFERENCES

- Alrefaei, M.H. and Andradóttir, S., 1999. A simulated annealing algorithm with constant temperature for discrete stochastic optimization. *Management Science* 45 (5), 748-764.
- Billingsley, P., 1995. *Probability and Measure*. Third Edition. John Wiley & Sons, Inc.
- Chen, E.J. and Kelton, W.D., 2005. Sequential selection procedures: using sample means to improve efficiency. *European Journal of Operational Research* 166, 133-153.
- Dudewicz, E.J. and Dalal, S.R., 1975. Allocation of observations in ranking and selection with unequal variances. *Sankhya* B7, 28-78.
- Fu, M.C., 2001. Simulation optimization. In: *Proceedings of the 2001 Winter Simulation Conference*, Peters, B.A., Smith, J.S., Medeiros, D.J., and Rohrer, M.W., Eds, pp. 53-61. December 9-12, Arlington (Virginia, USA).
- Fu, M. and Nelson, B., 2003. Guest Editorial. *ACM Transactions on Modeling and Computer Simulation* 13(2), 105-107.
- Hogg, R.V. and Craig, A.T., 1978. *Introduction to mathematical statistics*. Fourth Edition. Macmillan Publishing Co., Inc., New York.
- Kim, S.-H. and Nelson, B.L., 2001. A fully sequential procedure for indifference-zone selection in simulation. *ACM TOMACS* 11, 251-273.
- Legato, P., Canonaco, P. and Mazza, R.M., 2009. Yard crane management by simulation and optimization. *Maritime Economics and Logistics* 11(1), 36-57.
- Legato, P. and Mazza, R.M., 2008. Selecting the Optimum by Searching and Ranking Procedures in Simulation-based Optimization. In: *Proceedings of the 20th European Modeling and Simulation Symposium (Simulation in Industry)*, pp. 561-568. September 17-19, Campora S. Giovanni, (CS) Italy.
- Legato, P., Mazza, R.M. and Trunfio, R., 2008. Simulation-based optimization for the quay crane scheduling problem. In: *Proceedings of the 2008 Winter Simulation Conference*, Mason, S.J., Hill, R., Moench, L., and Rose, O., Eds, 2717-2725. December 7-10, 2008. Miami (Florida, USA).
- Legato, P., Mazza, R.M. and Trunfio, R., 2010. Simulation-based Optimization for discharge/loading operations at a maritime container terminal. *OR Spectrum* 32 (3), 543-567.
- Mazza, R.M., 2008. Simulation-based optimization in port logistics. Thesis (Ph.D). Università della Calabria.

- Rinott, Y., 1978. On two-stage selection procedures and related probability-inequalities. *Communications in Statistics - Theory and Methods* A7(8), 799-811.
- Stahlbock, R. and Voß, S., 2008. Operations research at container terminals: a literature update. *OR Spectrum* 30(1), 1-52.
- Wilcox, R.R., 1984. A table for Rinott's selection procedure. *Journal of Quality Technology* 16(2), 97-100.

AUTHORS BIOGRAPHY

Pasquale LEGATO is an Associate Professor of Operations Research at the Faculty of Engineering (Università della Calabria – Rende, Italia), where he teaches courses on simulation for system performance evaluation. He has published on queuing network models for job shop and logistic systems, as well as on integer programming models. He has been involved in several national and international applied research projects and is serving as reviewer for some international journals. His current research activities focus on the development and analysis of queuing network models for logistic systems, discrete-event simulation and the integration of simulation output analysis techniques with combinatorial optimization algorithms for real life applications in Transportation and Logistics. His home-page is <<http://www.deis.unical.it/legato>>.

Rina Mary MAZZA went to the Università della Calabria, Rende (Italia), where she received her Laurea degree in Management Engineering and a Ph.D. degree in Operations Research. She is currently Head of the Research Project Office at the Dipartimento di Elettronica, Informatica e Sistemistica (DEIS, Università della Calabria). She is also a consultant for operations modeling and simulation in terminal containers. Her current research interests include discrete-event simulation and optimum-seeking by simulation in complex logistic systems. Her e-mail address is: <rmazza@deis.unical.it>.

ACCELERATED FULLY 3D ITERATIVE RECONSTRUCTION IN SPECT

Werner Backfrieder^(a), Gerald A. Zwettler^(b)

^{(a), (b)}University of Applied Sciences Upper Austria, Institute of Software Engineering, Campus Hagenberg, Austria

^(a) Werner.Backfrieder@fh-hagenberg.at, ^(b) Gerald.Zwettler@fh-hagenberg.at

ABSTRACT

Image quality in single photon emission computed tomography (SPECT) is substantially influenced by scatter and a finite volume of response associated with single detector elements. These effects are not restricted to the image plane, implying a shift in the tomographic imaging paradigm from 2D to 3D. The application of a 3D reconstruction model suffers from huge numerical efforts, affording for high performance computing hardware. A novel accelerated 3D ML-EM type reconstruction algorithm is developed by the implementation of a dual projector back-projector pair. An accurate 3D model of data acquisition is developed considering scatter and exact scanner geometry in opposite to a simple pencil-beam back-projection operator. This dual concept of projection and back-projection substantially accelerates the reconstruction process. Speed-up factors achieved by the novel algorithm are measured for several matrix sizes and collimator types. Accuracy of the accelerated reconstruction algorithm is shown by reconstruction of data from a physical Jaszczak phantom and a clinical endocrine study. In both cases the accelerated 3D reconstruction method achieves better results. The novel algorithm has a great potential to scale fully 3D reconstruction down to desktop applications, especially with the new possibilities employing massive parallel graphics hardware. The presented work is a step towards establishing sophisticated 3D reconstruction in a clinical workflow.

Keywords: emission tomography, fully 3D reconstruction, nuclear medicine, high-performance computing

1. INTRODUCTION

Nuclear medicine imaging modalities show the distribution of radioactive tracer providing diagnostic information. Main fields of application are tumor diagnostics and in vivo assessment of metabolism. Therapeutic applications are limited to therapy with beta-emitters, e.g. radioiodine therapy of the thyroid. In nuclear medicine imaging the kinetics of radioactive tracer particles within the human body is the basis of diagnostic information. After intravenous application specialized radiopharmaceuticals distribute within the body and finally accumulate in targeted morphological

regions. In tumor diagnosis tissue pathologies are imaged as hot spots. The amount of activity uptake and the size of the lesion are an important measure for the progress of the tumor disease. Both, distribution and kinetics of the radiotracer are subject to functional imaging, e.g. perfusion images of the human brain after stroke or assessment of the clearance rate in kidneys.

Single photon emission computed tomography (SPECT) is a volume imaging technique, visualizing the human body as a series of transversal slices. The photons generated during disintegration of a short lived radionuclide, e.g. Tc-99m, are registered by a gamma-camera as projection images. There is a great variety of algorithms for the reconstruction transversal slices from projection data. Filtered back-projection (FBP) in combination with specific filter windows is, due to its high performance, the main method in clinical practice (Herman 2009). With increasing computational power iterative methods, allowing a more accurate modeling of geometrical and physical properties of the imaging process, were introduced into clinical environments. The maximum likelihood expectation maximization (ML-EM) algorithm (Shepp and Vardi 1982) is the foundation of a series of optimized algorithms in emission tomography (Hudson and Larkin 1994).

The slice topology of the reconstruction algorithms is a major limitation in image quality of emission tomography. In contrast to x-ray computed tomography the scanner hardware allows major interferences from adjacent slices. Collimator geometry defines a conic volume of response to a single detector position; low count rate and scatter are major deteriorating effects in SPECT imaging. Fully 3D image reconstruction accomplishes the simultaneous reconstruction of the whole image volume, but at the cost of high computational burden (Backfrieder et al. 2002, Backfrieder et al. 2003a, Backfrieder et al. 2003b, Benkner et al. 2004).

In the following a fully 3D iterative reconstruction algorithm is described implementing a dual projector back-projector pair for accelerated reconstruction. The newly developed algorithm is based on the OS-EM family providing accelerated convergence (Hudson and Larkin 1994).

2. MATERIAL AND METHODS

The imaging equation in tomography reads

$$y_i = \sum_j a_{ij} \cdot x_j + e_i, \quad (1)$$

it describes the relation between the pixels of the source distribution (=image) x and a single projection values y . Both, the image and projection array are two-dimensional, i.e. x -and y -direction in the image, angle and lateral distance in projections, but are represented by a single linear index. A value of the system matrix a_{ij} describes the contribution of pixel x_j to the projection value y_i . This allows the accurate modeling of

- scanner geometry
- photon attenuation
- detector response
- scatter

Figure 1 shows a sketch of the image plane and the pixels summing up a single projection value.

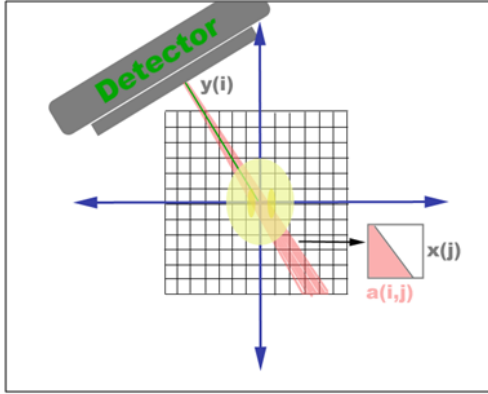


Figure 1: Sketch of SPECT scanner geometry.

Each measured value contains an error term e_i . In the case of radioactive decay and detection of photons this error term is Poisson-distributed. Under the constraint of a Poisson-distributed random process image reconstruction is formulated as a maximization problem of the likelihood L of measured data y

$$L(y | x) = \prod_i \frac{\left(\sum_j a_{ij} \cdot x_j \right)^{y_i} e^{-\sum_j a_{ij} \cdot x_j}}{y_i!}, \quad (2)$$

where the sum represents the expectation value of the respective measured projection value y_i . The algorithm aims in maximizing the term in Eqn. 2 by choosing a proper image-vector \mathbf{x} . The solution is the iterative ML-EM algorithm for tomography by Shepp and Vardi, 1982

$$x_j^{(n+1)} = x_j^{(n)} \sum_i a_{ij} \frac{y_i}{\sum_{j'} a_{ij'} x_j^{(n)}}. \quad (3)$$

A sequence of intermediate images $x^{(n)}$ is calculated until a stopping criterion is satisfied. During the n^{th} iteration each pixel x_j is updated by a multiplicative factor. This factor is the weighted sum of all projection values y_i affected by the pixel x_j . The correction term depends on the quotient of the measured projection value and the calculated pseudo-projection

$$y_i^p = \sum_{j'} a_{ij'} x_j^{(n)}. \quad (4)$$

The iteration steps in Eqn. 3 converge to a feasible solution, representing a maximum entropy solution to the imaging equation. To further accelerate the convergence of the algorithm ordered subsets are implemented.

The discussed reconstruction model describes the reconstruction of a single image slice. Spatial activity distribution out of the slice is not considered by this model. Finite collimator aperture and scatter have significant contribution from pixels out of the considered slice on the projection values, necessitating a three dimensional (3D) approach to the reconstruction problem for further improvement of image quality. In contrast to FBP the iterative approach allows a simple extension to 3D by covering the whole image volume and projection values of all slices by respective vectors. As a consequence the system matrix A grows $o(N^6)$ with the lateral length (N pixels) of the image cube.

2.1. Modeling of the system matrix

Each line of the system matrix defines the weights of all voxels to a specific projection value. In a conventional SPECT study the image volume consists of 128 slices, with a matrix size of 128x128 pixels, each. A row consists of $128^3=2.097.152$ elements. The number of projection values, i.e. the number of lines of the matrix, is calculated from the size of the projection matrix and the angular increment of the detector head, i.e. 128x128x120 for a 3 degrees increment on a circular orbit. In total the system matrix contains 4.12×10^{12} elements. Even dedicated high-performance-computer (HPC) systems cannot hold this huge amount of data in memory.

Since a line of the system matrix considers all elements of the image volume, most of the entries are zero. With careful modeling of the geometrical and physical properties of data acquisition, this leads to a significant reduction of data.

Each projection value is related to a flat rectangular region of the detector surface, i.e. the field of view (FOV) divided by the number of elements of the projection matrix. For assessment of the contribution of each voxel to a specific projection value, a point source is positioned at the center of a voxel and the fraction of radiation reaching the detector element is calculated. This corresponds to the ratio of the surface of a sphere, with origin in the voxel and the radius is the distance to the detector element, and the projection of this detector element onto this sphere. This simple geometrical

consideration leads to a model of the volume of response as a cone targeting to the detector surface. With increasing distance to the detector the cone-width increases and the weight of voxels decreases. The voxel weights at the level of the central slice of the volume of response (VOR), i.e. at the level of the projection value, are shown in Fig. 2.a. The VOR has circular symmetry.

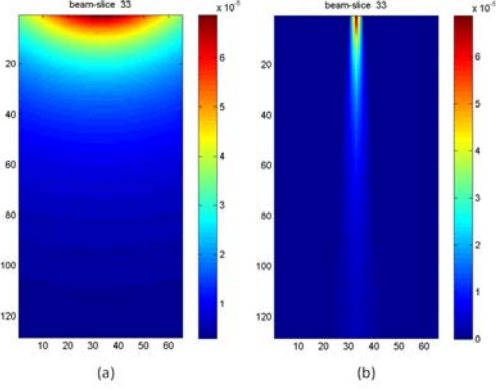


Figure 2: Volume of response of a 128x128 projection matrix (a) and its amplification by a LEGP collimator (b).

The camera head is equipped with a lead collimator, to limit the viewing direction approximately to bars normal to the detector surface. The design consists of a regular pattern of thin lead septa, arranged as long thin bore holes or as a honeycomb grid. The modulation factor is the cast shadow of the collimator septa depending on the detector thickness and the ratio of wall thickness of septa and their aperture. Scatter is a further amplification of the voxel weights; usually it is modeled by a zero centered Gaussian distribution. The total pixel weight reads

$$a_{ij} = \Theta_{scatter} \alpha_{coll} \Phi_{geom}, \quad (5)$$

where the geometrical form factor is Φ , the attenuation factor of the collimator is α and the contribution of photon scatter by human tissue is Θ . Figure 2.b shows the application of the collimation factor to the VOR.

2.2. Dual projector back-projector pair

In the previous section the OS-EM algorithm and the modeling of the system matrix is discussed in detail. With the generalization of the reconstruction problem to 3D the computational effort increases substantially, affording for HPC hardware to achieve suitable performance for image reconstruction, to establish it in a clinical environment.

The high cost of the ML-EM algorithm is caused by a series of projections and back-projections during each iteration step, cf. Eqn. 3. The sum over all projection values containing the actual pixel can be considered as back-projection. From each intermediate image $x^{(n)}$ pseudo-projections are calculated. The number of

numerical operations is proportional to the non-zero elements of the system-matrix \mathbf{A} . To achieve most accurate physical and geometrical modeling the forward projection is implemented by the modeled weights according to Eqn. 4. The accurate assessment of pseudo-projections is crucial, since its ratio to measured projection values y_i defines the amount of the correction term. The back-projection operator comprises the projection values considered for the update of a specific pixel. In this novel approach not all elements, as defined by the above model of the system matrix, are considered, but only a subset defined by orthogonal projection onto the detector surface. The lateral distance from the center of the profile is

$$l = x \cdot \cos \mathcal{G} + y \cdot \sin \mathcal{G}, \quad (6)$$

where x, y are the coordinates of the updated voxel and \mathcal{G} is the rotation angle of the detector head. Only projection values within the slice are considered. The ML-EM algorithm with dual projector and back-projector pair reads

$$x_j^{(n+1)} = x_j^{(n)} \sum_i l_{ij} \frac{y_i}{\sum_{j'} a_{ij'} x_{j'}^{(n)}}. \quad (7)$$

The coefficients l_{ij} denote the reduced set of back-projection values. The speed up factor is linear to the reduction of the l_{ij} coefficients in relation to the total number of entries in the system matrix entries a_{ij} . For a standard 128x128 matrix and a LEGP parallel collimator the speed-up factor is 218.53. This speed up of fully 3D reconstruction implemented together with the ordered subsets concept, the newly developed algorithm is called 3D accelerated ordered subsets expectation maximization (3D-AOS-EM).

2.3. Physical phantom and patient data

Data are collected from a circular clinical standard Jaszczak SPECT phantom on a three headed Philips IRIX camera. The phantom was filled with 600 MBq Tc-99m. Acquisition parameters were: 128 by 128 projection matrix, pixel size 4.4mm, 120 projections on a full circular orbit of 360 degrees and 20s acquisition times in stop and go mode.

On the same camera data from a clinical endocrine study, 55MBq I-131 applied activity, were acquired on a 64 by 64 projection matrix over a 565mm FOV, with 60 projections on a circular orbit, and 30s acquisition time per projection.

3. RESULTS

Results are shown for acceleration of the algorithm in contrast to 3D ML-EM, a comparison of reconstruction methods applied to physical phantom data and a clinical study.

3.1. Speed up factors

The speed-up factors - as a consequence of the implementation of the accelerated back-projection operator - are shown in Figure 3. Results are shown for two collimator types, a low energy general purpose (LEGP) and a high energy high resolution (HEHR) collimator. The speed up factor directly relates to the reduction of entries in the back-projection matrix compared to those in the respective projection matrix, as shown by different factors for the collimators used during the studies. The HEHR collimator has a significantly smaller VOR thus the speed-ups are smaller than those of the LEGP collimator, cf. Fig. 3.

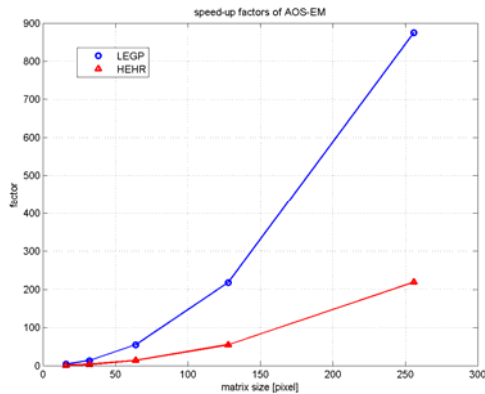


Figure 3: Speed-up factors

3.2. Phantom data

A slice of the Jaszczak phantom comprising 6 sectors of cold rods with increasing diameter is shown in Fig. 4. Slices were reconstructed using FBP, the clinical standard, and the accelerated fully 3D reconstruction with dual projection back-projection (3D-AOS-EM). During iterative reconstruction 15 iterations with 4 subsets were performed. Compared to FBP the contrast of cold spots is significantly increased with 3D-AOS-EM. In sector 4 (numbered in order of decreasing diameter) rods are still distinguishable, especially in the distal part of the phantom, since with FBP the whole sector is blurred out.

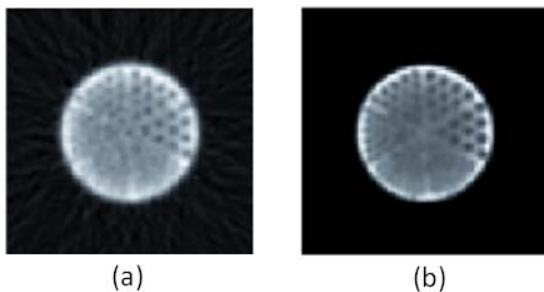


Figure 4: Reconstruction of a standard Jaszczak phantom using (a) FBP and the novel 3D-AOS-EM algorithm.

3.3. Clinical data

Data from the clinical study show a transversal slice through the thyroid, cf. Fig. 5. FBP suffers from low

signal intensity, manifested by substantial star-artifacts centered at the lesion. The hot lesion is a connected oval region with a small tail at the lower left. This image data cannot clearly support the decision, if this tiny image structure is a real pathology or an artifact. 2D-ML-EM reconstruction shows a clearly manifested hot lesion in this part of the image. Reconstruction of the image using the newly developed 3D-AOS-EM yields two clearly distinguishable hot lesions.

4. DISCUSSION

Fully 3D image reconstruction is the most accurate reconstruction model for nuclear medicine emission tomography. The direct implementation of the 3D data model suffers from high computational complexity resulting in long reconstruction cycles, hardly to establish in a clinical workflow. The substantial acceleration of the algorithm by introduction of a dual projector back-projector pair has high potential to scale down the problem from HPC platforms, as already implemented on PC-clusters (Backfrieder et al. 2003b), to desktop hardware. The actual algorithm is implemented as a MATLAB prototype, thus the evaluation of the performance is done on basis of speed-up factors. The newly introduced programming interface CUDA to the highly parallel architecture of the graphics-subsystem offers new perspectives to solve computationally intensive numerical problems. In ongoing work the 3D-AOS-EM algorithm will be implemented in the C-CUDA framework.

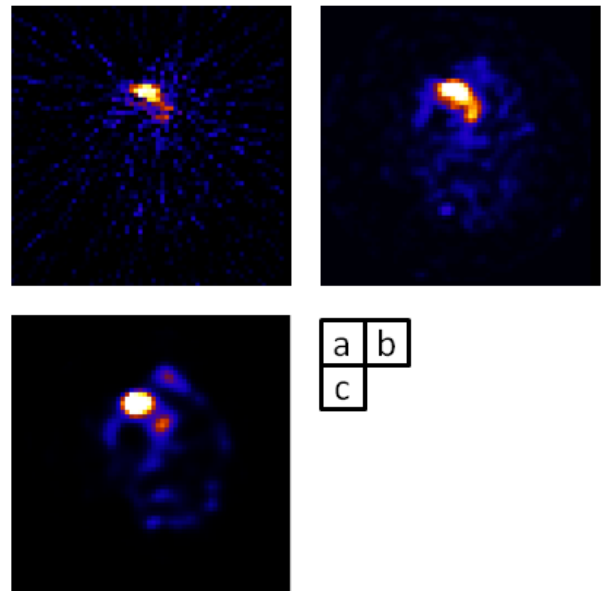


Figure 5: Endocrine study reconstructed FBP (a), 2D-ML-EM (b) and 3D-AOS-EM.

The acceleration of the fully 3D reconstruction together with its implementation on desktop systems is a further step towards sophisticated image processing supporting clinical diagnostics.

ACKNOWLEDGMENTS

Authors want to thank Univ.-Prof. Dr. Michael Gabriel and the radiological technologist from the Institute of Endocrinology and Nuclear Medicine of the General Hospital Linz, Austria, for providing SPECT data.

REFERENCES

- Backfrieder, W., Forster, M., Benkner, S., Engelbrecht, G., Terziev, N., Dimitrov, A., 2002. Accurate attenuation correction for a fully 3D reconstruction service. *Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS '02)*, Las Vegas, Nevada, USA, June 24 – 27, 2002: 680-685
- Backfrieder, W., Forster, M., John, P., Engelbrecht, G., Benkner, S., 2003a. Fully 3D Iterative SPECT Reconstruction in A High Performance Computing Environment. *J. Nucl. Med. Technol. Vol.31 No.6*, pp. 201
- Backfrieder, W., Forster, M., Engelbrecht, G., Benkner, S., 2003b. Locally variant VOR in fully 3D SPECT within a service oriented environment. In F. Valafar, H. Valafar (Eds.), *Proc. Int. Conf. on Mathematics and Engineering Techniques in Medical and Biological Sciences (METMBS)*, ISBN 1-932415-04-1, (2003) pp. 216-221
- Benkner, S., Engelbrecht, G., Backfrieder, W., Berti, G., Fingberg, J., Kohring, G., Schmidt, J.G., Middleton, S.E., Jones, D., Fenner, J., 2004. Numerical Simulation for eHealth: Grid-enabled Medical Simulation Services. In G.R. Joubert, W.E. Nagel, F.J. Peters, W.V. Walter (Eds.), *Software Technology, Algorithms, Architectures and Applications, included in series: Advances in Parallel Computing*, Elsevier (2004)
- Herman, G. T., 2009. *Fundamentals of computerized tomography: Image reconstruction from projections*, 2nd edition, Springer, 2009
- Hudson, H.M., Larkin, R.S., 1994. Accelerated image reconstruction using ordered subsets of projection data. *IEEE Trans Med Imaging*. 1994;13:601–609.
- Shepp, L.A., Vardi, Y., 1982. Maximum likelihood estimation for emission tomography. *IEEE Trans Med Imag*. 1982;MI-1(2):113–121.

AUTHORS BIOGRAPHY

Werner Backfrieder received his degree in technical physics at the Vienna University of Technology in 1992. Then he was with the Department of Biomedical Engineering and Physics of the Medical University of Vienna, where he reached a tenure position in 2002. Since 2002 he is with the University of Applied Sciences Upper Austria at the division of Biomedical Informatics. His research focus is on Medical Physics and Medical Image Processing in Nuclear Medicine and Radiology with emphasis to high performance computing. Recently research efforts are laid on virtual

reality techniques in the context of surgical planning and navigation.

Gerald A. Zwettler was born in Wels, Austria and attended the Upper Austrian University of Applied Sciences, Campus Hagenberg where he studied software engineering for medicine and graduated Dipl.-Ing.(FH) in 2005 and the follow up master studies in software engineering in 2009. In 2010 he started his PhD studies at the University of Vienna at the Faculty of Computer Sciences. Since 2005 he is working as research and teaching assistant at the Upper Austrian University of Applied Sciences at the school of informatics, communications and media at the Campus Hagenberg in the field of medical image analysis and software engineering with focus on computer-based diagnostics support and medical applications.

STUDY ON THE SERVICIALIZATION OF SIMULATION CAPABILITY

Y L Luo^(a), L Zhang^(a), F Tao^(a), Y Bao^(a), L Ren^(a)

^(a) School of Automation Science and Electrical Engineering, Beihang University,
Beijing 100191, P. R. China

yongliang2002@gmail.com, zhanglin@buaa.edu.cn,
ftao@buaa.edu.cn, baoyuan_0426@163.com, lei_ren@126.com

ABSTRACT

Manufacturing capability (MC) servilization is a key to realize on-demand use, dynamic collaborative work, and circulation of manufacturing resources and capability in the cloud manufacturing (CMfg) system. This paper emphasizes the servilization of simulation capability (SC), which is a very important kind of MC. According to task demands and characteristics of complex product's simulation process, concepts and state of the art related to MC were systematically analyzed and summarized firstly in this paper, then a conceptual model of SC were presented, A application model of SC service life-cycle in CMfg system is proposed. Then the framework for simulation capability servilization is investigated, as well as several key issues involved in the servilization process such as elements of simulation capability, modeling and decription of simulation capability, and so on. Finally, an application example analysis of SC was presented.

Keywords: cloud manufacturing, manufacturing capability, simulation capability, servilization

1. INTRODUCTION

Cloud manufacturing (CMfg) is a new service-oriented, highly efficient, lowly consumption knowledge based, and intelligent networked manufacturing model (Bohu Li and Lin Zhang *et al.* 2010). It is combined with advanced manufacturing and information technologies organically (e.g. cloud computing, the internet of things, semantic web, and information system integration) in order to achieve virtualization and servilization of manufacturing resources and capability, CMfg provides users with application services which are on-demand using, safe and reliable in the whole life-cycle of products through network (Bohu Li, Lin Zhang *et al.* 2010). CMfg aims to achieve agile, service-oriented, green and intelligent manufacturing, is a new phase of networked manufacturing, and is the materialization of service-oriented manufacturing (Lin Zhang and Yongliang Luo *et al.* 2011). Therefore, CMfg can provide theoretical and technical supports for the transformation from production-oriented manufacturing to service-oriented manufacturing.

Manufacturing capability (MC) servilization is one of the most important innovative points of CMfg,

however. Because of MC is a complex concept and its correlative research is less, as a result there is no clear definition currently. At present, there are two different understandings on the concept and connotation of MC, some people argue that MC reflects the performance of enterprise from a macro points of view, i.e., Skinner first proposed the MC in 1969, he holds that MC includes many elements such as cost, delivery time, quality, and the relationship between these elements. MC reflects the completion of manufacturing objective, and it is a performance level of the standard which is pre-sat by working organization (Mattias Hallgren. 2006). Guan (2004) commented that MC is the core part of enterprise innovation capability, it is conversion capability of results which meet market demand, design requirements of product and mass-produced. The relationship between MC and enterprise performance is discussed from the perspective of achieving low operating costs and high product quality (Siri Terjesen, Pankanj C. Patel *et al.* 2011). The other comment that MC is a integration of manufacturing resources based on microscopic pint of view, i.e., Richard (1973) considered that capability includes knowledge, skills, and experience of enterprise. MC reflects the performance of completing setting function based on manufacturing resources in order to support the operation of enterprise activities (Cheng Yun, Yan Junqi. 1996) Keen(2000) commented that MC is the integration of intangible resources and tangible resources, where the tangible resources include labor, capital, facilities and equipment, simultaneously, and the intangible resources include information, procedures, equipment and the organizational system. Khalid(2002) concluded that MC is the effective integration of related resources in the process of achieving expected target task. Cheng (2009) gave the definition that MC is a set of elements involved in the implement process of manufacturing enterprise's strategy. Zhang Lin (2010) commneted that MC is an intangible and dynamic resource in CMfg model, it is the subjective condition of production-related goals.

Combined with above views about MC, several problems are systematically summarized as follows:

- The current research about MC have been widely studied from a management point of view, because mostly based on qualitative

analysis, lack of supports for quantitative description on MC;

- Lack of the logical relationship analysis between construction elements of MC

In this paper, is considered as a subjective condition, what manufacturing enterprises needed to complete one task or objective. It is a intangible and dynamic resources form. And it is a kind of capability which can be represented in the manufacturing activities. MC including design capability (DC), simulation capability (SC), product capability (PC), and many other capabilities related to life-cycles of complex products. MC is tightly linked to manufacturing activities and manufacturing resources, it can't be reflected without concrete activity tasks and resources elements. According to task demands and resources characteristics of complex products' simulation process, this paper studies MC based on simulation. In order to comprehensively understand SC, concepts and state of the art related to MC were systematically analyzed and summarized firstly, then a conceptual model of SC was presented. An application model of SC service life-cycle in CMfg system is proposed, and the key technologies involved in each process were investigated. On this basis, a new simulation capability servilization (SCS) framework is presented, several issues related to SCS are discussed in detail.

2. THE CONCEPT AND CONNOTATION OF SIMULATION CAPABILITY

SC is an important kind of MC in CMfg system, is a simulation process, which reflects a capability of complete a simulation task or experiment supported by related resources and knowledge. Through SCS, it can not only realize the function sharing of resources, but also share the experience and knowledge in the simulation process, such as simulation flow, simulation data, experience of simulated staff, and so on.

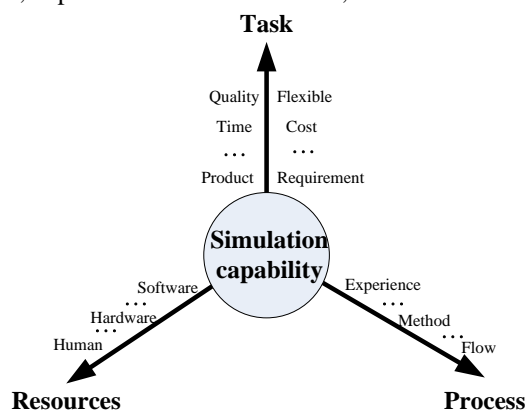


Fig.1 The conceptual model of simulation capability

The conceptual model of simulation capability is illustrated in fig1, which primarily consists of three dimensions:

1) Resources dimension(R)

SC is the integration of all kinds of simulation resources related to perform some tasks or activities, such as simulation software,

simulation equipment and so on. Resources are the foundation of forming SC; it is the subject of SC servilization as well. Resources can be divided into two kinds, one is subject resources, which is the carrier of SC performance, for example, the subject resources of software's SC is the simulation software. The other is auxiliary resources such as material, which is support for product and related goal. In addition, a SC possibly refer to several subject resources

2) Process dimension(P)

SC is a kind of activity process; it contains a knowledge set generated in the implementation process of task and goal realization, such as constraint condition, simulation method, and simulation experience and so on. In addition, knowledge is the effective carrier of procedure representation.

3) Task dimension(T)

It contains two aspects information, one part is about the simulation task, and the other part is about the completion of the simulation task target, which include many objective factors and evaluation of user satisfaction factors, elements, i.e., delivery time, cost, quality, innovation, service, et al. This dimension is the most important selection basis for SC users in cloud manufacturing service platform.

The relationship among resources dimension, process dimension and task dimension is investigated as follows, resources is the basic of achieving SC. Task dimension shows us the result of SC, it is the most important basis for user optimize selecting in CMfg system. Process dimension is the method of SC's forming.

3. SIMULATION CAPABILITY SERVICE LIFE CYCLE MODEL

As shown in figure 1, the life cycle of simulation capability service can be divided into the following four parts:

(1) Simulation capability publication (SCP)

SCP is a servilization process of SC; it is the basic of cloud manufacturing service platform to realize on-demand use and sharing of simulation capability. It combines the characteristics of simulation resources with simulation capability classification in the CMfg model. Elements of SC are extracted and analyzed firstly, and the unified semantic description model of SC is presented. Secondly, in order to achieve the formal description of SC, the existing services description language will be expanded or improved, then SC will be released in the form of service in CMfg system. It will support the trading and distribution of SC for users through network. Some key technologies involved in this process, such as simulation capability classification, simulation capability modeling, manufacturing capability description language expansion and so on.

(2) *Simulation capability discovery (SCD)*

SCD is responsible for achieving semantic searching and dynamic composition of SC services in CMfg system. According to the characteristics of SC, such as relative, complexity, dynamic and so on, in order to reflect multi-dimensional attributes of SC fully and clearly, all kinds of SC description information should be classified, fused and normalized firstly, and

then construct the ontology of SC to support semantic search. The ontology can improve searching accuracy. At last SCD can support the sharing of SC through the network. The process of SCD includes several key technologies, such as domain ontology construction, semantic matching and dynamic composition of SC, services of SC sharing and so on.

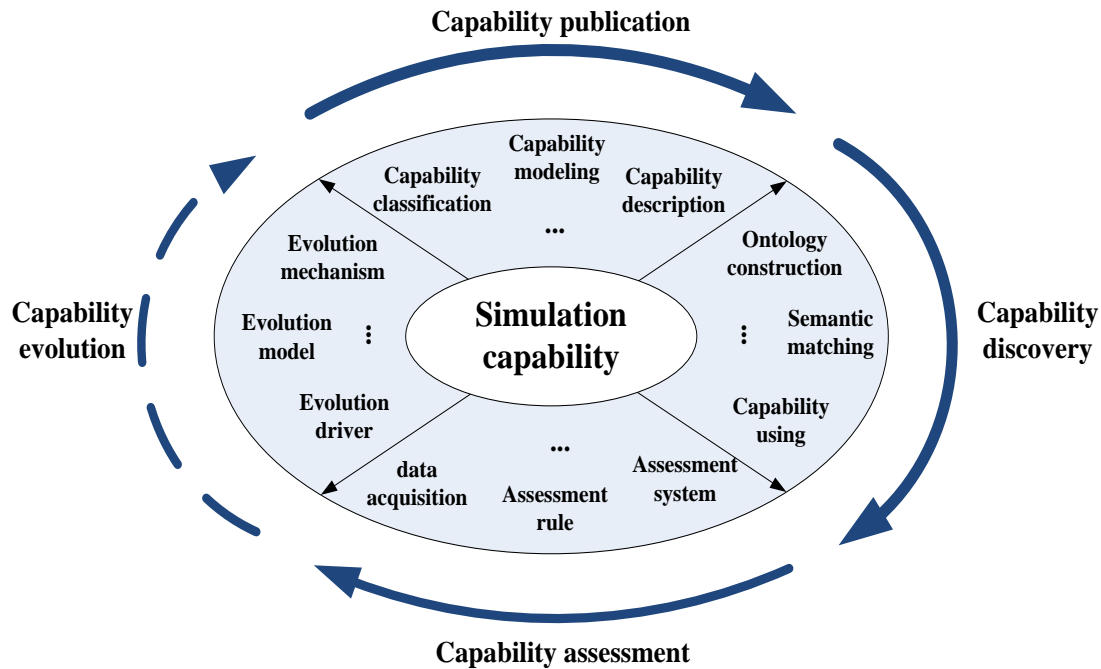


Fig.2 The life cycle of simulation capability service

(3) *Simulation capability assessment (SCA)*

Combined user feedback with operation of the SC, SCA will realize the comprehensive utility evaluation of simulation capability. Due to the complexity and dynamic of SC, how to measure SC is an important issue of SCS, but in order to achieve the assessment of SC, we need to provide comprehensive assessment system establishment and suitable assessment methods based on description system and evaluation factors of simulation capability, where the assessment system should be level and systemic, it can reflect the overall characteristic of simulation capability, and all level and dimensions attributes which contains quantitative and qualitative. The assessment methods of SC should take full account of dynamic changes during the process of the SC's using. Some key technologies involved in this process, such as capability assessment system construction, SC assessment of data acquisition, SC assessment method, operation monitoring of SC and so on.

(4) *Simulation capability evolution (SCE)*

It is a response and adjustment process of manufacturing enterprises or systems in the face of

changing external environment. On the basis of enterprise evolution theory and dynamic capability theory, driver attributes of simulation capability evolution is systematically analyzed in the CMfg mode firstly, combined with assessment index system of SC, form, process and mechanism of simulation capability evolution are deeply discussed from the qualitative point of view, and then an empirical analysis on the process of SCE is done by mapping the qualitative to quantitative. SCE will provide support for dynamic maintenance and intelligent update of SC to CMfg system. Key technologies of SCE include evolution driving factors, evolution mechanism and method of simulation capability, evolution model construction, evolution procedural knowledge representation and so on.

4. THE FRAMEWORK OF SIMULATION CAPABILITY SERVICIALIZATION

SCS plays an important role in the process of achieving on-demand use of SC. The process of SCS is shown in fig3, it can be divided into the following five parts:

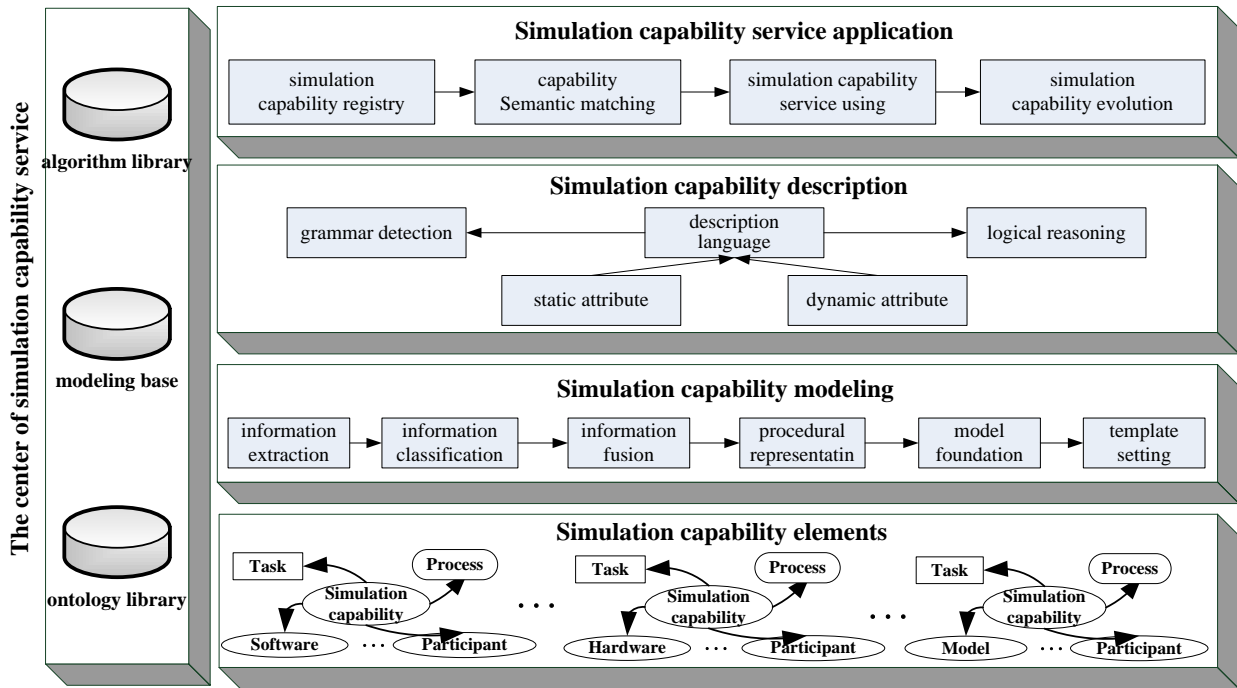


Fig.3 The framework of simulation capability servilization

➤ **Simulation capability elements**

Most of the current researches about capability elements refer to the performance of manufacturing capability. SC elements mainly contain various resources and related assessment factors in the construction process of SC, it is a comprehensive and integrated reflection of SC, and will provide support for construction and formal description of SC model. In order to achieve on-demand use and dynamic collaborative of SC, the display content of SC oriented

to users should be analyzed and classified firstly. Then according to the actual requirements related to product and goals, elements of SC will be summarized. For example, SC elements can be divided into six parts ,as shown in fig4: major resources, product and business, participator, process knowledge, SC assessment information, enterprise integrated information. And each part is also consisted by many related elements in detail, e.g., process knowledge contains design model, experience knowledge, simulation method and so on.

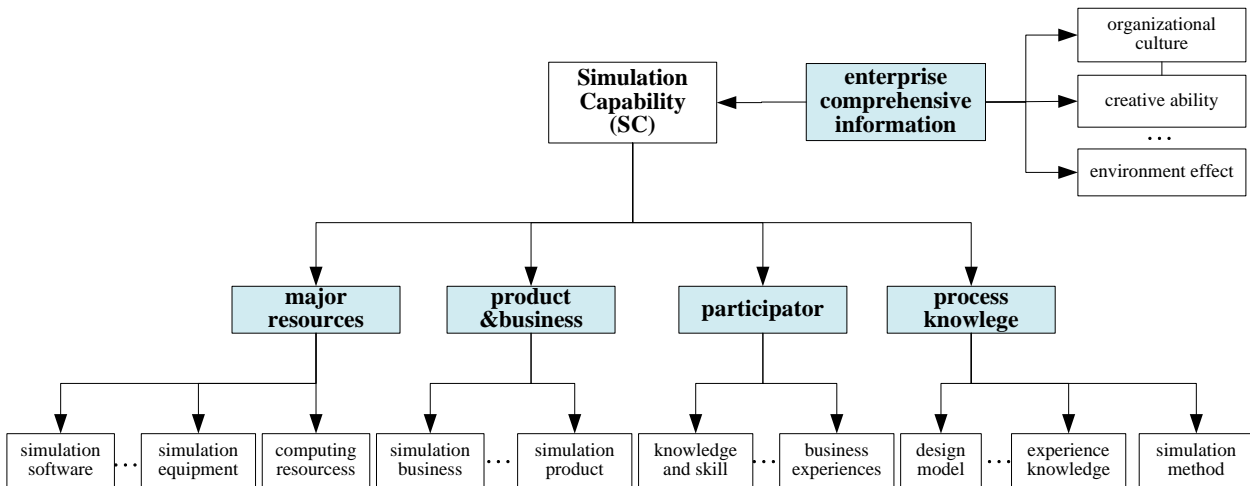


Fig.4 Elements of simulation capability

➤ **Simulation capability modeling**

According to the characteristics and connotation of SC, combined with the above introduction about SC elements, description model of simulation capability (DMSC) is abstractly represented. The transformation from qualitative to quantitative elements of SC is realized. And related key technologies involved in the model as shown in fig4, such as information extraction, information classification, information fusion, process

representation, SC description template construction and so on.

The model can be divided into two parts as follows:
 $DMSC=(T, R, P, K,E) + fun(T, P, R, K,E)$

In above description model, the first part is a four-quad, where *T* is the simulation task and the objective needed to realize a simulation activity. *R* is simulation resources elements involved in the simulation process, i.e., hard-resource (including simulation equipments,

materials) software, et al). P is the participants involved in the simulation process, i.e., the human resource in the simulation resources elements including personal and organization. K mainly includes all kinds of knowledge possessed by resources elements and experiences accumulated in simulation process, i.e., production flow, burden scheme, et al. E reflects the comprehensive information of enterprise, such as organizational culture, creative ability and so on. The second is SC evolution function - $fun(T, P, R, K, E)$, it expresses the logical relationship between each elements of SC, for instance, the completion of task will be affected by enterprise reputation and organizational culture.

➤ **Simulation capability description**

Based on the above simulation capabilities model, in order to achieve SC servilization, we need to select a proper way to realize the formal description of SC, as far as the existing service description language include web ontology language (OWL), OWL for service (OWL-S), simple HTML ontology extension (SHOE), can't fulfill the actual requirements. According to the above problems and resources characteristics, many extended service description languages are proposed, for example, OWL-SP is presented based on OWL-S by adding dynamic logical operator. Due to the complexity, uncertainty and knowledge of SC, existing service description languages are not able to meet the requirement of SC servilization. Based on the grammar and lexical of existing service description language, characteristics and servilization requirement of SC were systematically analyzed and summarized, then achieved the expansion of service description language, expanded content main includes the description of the formation of SC, the logical relationship between the elements and related reasoning.

➤ **Simulation capability application**

It is responsible for the application of SC servilization. On the basis of above proposed methods and related technologies, a prototype of SCs system will be developed, its main functions include SC servilization publication, capability service semantic matching, SC transaction and assessment, SC evolution and so on. Then SC can be provided to user in term of cloud services which are stored in CMfg system through the network. In addition, corresponding methods are taken according to different simulation resources in order to achieve resources intelligent accessing to CMfg system. For example, virtualization technology is adapted to simulation software access, but to simulation equipment, related technologies of the internet of things will be used for that.

➤ **The center of simulation capability service**

The center of simulation capability service is the foundation of application related to SCS, it is responsible for field ontology library, modeling base and algorithm library. Due to knowledge is the basis of SC formation, so how to store the formal knowledge, is an important issue to SC servilization. Furthermore, description information of SC servilization need to be stored and classified with related rules, it can provide

support for user achieving on-demand use of simulation capability service.

5. CONCLUSIONS

MC servilization is a core of CMfg philosophy, and is the key to achieve on-demand use and dynamic collaborative of manufacturing resources and capability. This paper discusses the application process of MC services' life cycle in the complex product's simulation stage, and then elaborates the simulation capability servilization in detail. Simulation capability servilization is helpful to provide user with simulation capability service by network. In the future, a prototype of cloud service platform for complex product simulation will be developed according to above proposed methods and related technologies.

ACKNOWLEDGMENTS

This paper is partly supported by the NSFC (National Science Foundation of China) Projects (No.61074144 and No.51005012), the Doctoral Fund of Ministry of Education of in China (20101102110009), and the Fundamental Research Funds for the Central Universities in China.

REFERENCES

- Skinner W, 1969. Manufacturing –missing link in corporate strategy. *Harvard business review*. 6-9, 136-145
- Richard G B, 1972. The organization of industry. *Economic Journal*. 82, 883-896
- Skinner W. 1974. The focused factory. *Harvard business review*, 5-6, 113-121
- Hayes R, Wheelwright S. 1984. Restoring our competitive edge: Competing through manufacturing. *John Wiley & Son. New York*, 17-21
- P T Ward, R Duray, G K Leozzg, C sum. 1995. Business environment, operations strategy and performance: an empirical study of Singapore manufacturers. *Journal of operations management*. 13, 99-115
- Morgan Swink, W. Harvey Hegarty. 1998. Core manufacturing capabilities and their links to product differentiation. *International journal of operations & production management*, 18(4), 374-396
- Keen P, Mcdonald M, 2000. *The process edge: creating customer value and business wealth in the internet era*. McGraw-Hill. New York
- Toni M. Somers, Klara G. Nelson, 2003. The Impact of strategy and integration mechanisms on enterprise system value: empirical evidence from manufacturing firms. *European Journal of operational research*, 146,315-338
- Guo Haifeng, Tian Yezhuang, Liang Zhandong. 2006. An empirical analysis on relationships of manufacturing practices and manufacturing

capabilities. *IEEE International conference of management of innovation and technology*, pp: 969-972

- Rajiv D. Banker. 2006. Plant information systems, manufacturing capabilities & plant performance. *MIS Quarterly*, 30(2), 315-337
- Robert Sarmiento, Mike Byrne, Nick Rich. 2007. Delivery reliability, manufacturing capabilities and new models of manufacturing efficiency. *Journal of manufacturing technology management*, 18(4), 367-386
- LI Bohu, ZHANG Lin, WANG Shilong, TAO Fei, et al. 2010. Cloud manufacturing: a new service-oriented manufacturing model. *Computer Integrated Manufacturing Systems*, 6(1), 1-7.
- ZHANG Lin, LUO Yongliang, TAO Fei, REN Lei, et al. 2010. Study on the key technologies for the construction of manufacturing cloud. *Computer Integrated Manufacturing Systems*, 16(11), 2510-2520
- ZHANG Lin, LUO Yongliang, FAN Wenhui, TAO Fei, REN Lei, 2011. Analyses of cloud manufacturing and related advanced manufacturing models. *Computer Integrated Manufacturing Systems*, 17(3), 0458-0468.
- TAO Fei, CHENG Ye, ZHANG Lin, et al, 2011. Cloud manufacturing. *Advanced materials Research*, 201-203, 672-676
- LUO Yongliang, ZHANG Lin, HE Dongjing, REN Lei, TAO Fei. 2011. Study on multi-view model for cloud manufacturing. *Advanced materials Research*, 201-203, 685-688
- Siri Terjesen, Pankanj C. Patel, Jeffrey G. Covin. 2011. Alliance diversity, environmental context and the value of manufacturing capabilities among new high technology ventures. *Journal of operations management*, 29,105-115
- CHENG Yun, YAN Junqi, FAN Minlun. 1996. Manufacturing environment modeling based on object-oriented and step technology. *Chinese journal of mechanical engineering*, 32(4),5-10
- Khalid Hafeez, YanBing Zhang, Naila Malak. 2002. Determining key capabilities of a firm using analytic hierarchy process. *International journal of production economics*, 76, 39-51
- Andreas Grobler. 2006. An empirical model of the relationships between manufacturing capabilities. *International journal of operations & production management*, 26(5),458-485

AUTHORS BIOGRAPHY

Yongliang Luo received the B.S. degree and the M.S. degree in 2006 and 2009 from the Department of Computer Science at Shandong University of Science and Technology, China. He is currently working for the Ph.D. degree in Modeling simulation theory and technology at Beihang University. His research interests include service-oriented manufacturing and integrated manufacturing systems.

Lin Zhang received the B.S. degree in 1986 from the Department of Computer and System Science at Nankai University, China. He received the M.S. degree and the Ph.D. degree in 1989 and 1992 from the Department of Automation at Tsinghua University, China, where he worked as an associate professor from 1994. He served as the director of CIMS Office, National 863 Program, China Ministry of Science and Technology, from December 1997 to August 2001. From 2002 to 2005 he worked at the US Naval Postgraduate School as a senior research associate of the US National Research Council. Now he is a full professor in Beijing University of Aeronautics and Astronautics. He is an Editor of “*International Journal of Modeling, Simulation, and Scientific Computing*”, and “*Simulation in Research and Development*”. His research interests include integrated manufacturing systems, system modeling and simulation, and software engineering.

Fei Tao is currently an associate professor at Beihang University since April 2009. His research interests include service-oriented manufacturing, intelligent optimization theory and algorithm, resource service management. He is author of one monograph and over 20 journal articles of these subjects. Dr Tao was awarded the Excellent Doctoral Dissertation of Hubei Province, China and was elected to be a research affiliate of CIRP in 2009. He is the founder and editor-in-chief of *International Journal of Service and Computing-Oriented Manufacturing* (IJSCOM).

Yuan Bao received the B.S. degree in 2010 from the Department of Information Engineering at Tianjin University of Commerce, China. He is currently working for the M.S. degree in Modeling simulation theory and technology at Beihang University. His research interests include service-oriented manufacturing and integrated manufacturing systems.

GYRUS AND SULCUS MODELLING UTILIZING A GENERIC TOPOGRAPHY ANALYSIS STRATEGY FOR PROCESSING ARBITRARILY ORIENTED 3D SURFACES

Gerald Zwettler^(a), Werner Backfrieder^(a,b)

^(a)Bio- and Medical Informatics, Research and Development Department,
Upper Austria University of Applied Sciences, Austria

^(b)School of Informatics/Communications/Media, Upper Austria University of Applied Sciences, Austria

^(a)gerald.zwettler@fh-hagenberg.at, ^(b)werner.backfrieder@fh-hagenberg.at

ABSTRACT

Accurate and robust identification of the gyri and sulci of the human brain is a pre-requisite of high importance for modelling the brain surface and thus to facilitate quantitative measurements and novel classification concepts. In this work we introduce a watershed-inspired image processing strategy for topographical analysis of arbitrary surfaces in 3D. Thereby the object's topographical structure represented as depth profile is iteratively transformed into cyclic graph representations of both, the lowest and the highest characteristics of the particular shape. For graph analysis, the surface elements are partitioned according to their depth value. Neighbouring regions at different depth levels are iteratively merged. For region merging, the shape defining medial axes of the involved regions have to be connected by the optimum path with respect to a fitness function balancing shortness and minimal depth level changes of the solution.

Keywords: topographical surface analysis, cyclic graph representation, sulcus and gyrus classification

1. INTRODUCTION

The accurate quantification of metabolic processes from functional emission tomography imaging modalities like positron emission tomography (PET) and single photon emission computed tomography (SPECT) for diagnosis of neurodegenerative diseases necessitates a precise and patient-specific segmentation and classification of the brain. For segmentation and classification tasks, morphological image modalities as magnetic resonance imaging (MRI) have to be fused with the data acquired by functional emission imaging. Thus, the segmentations and classifications evaluated based on the anatomically-precise imaging modalities can be applied to the emission data, facilitating quantitative analysis of the metabolic activity with respect to pre-classified anatomical regions. The classification concept addressed in this work is the partitioning of gray and white brain matter according to the gyrus and sulcus characteristics.

Any computer-based functional or anatomical classification requires binary segmentation as pre-

processing. Utilizing T1-weighted brain MRI data, segmentation of gray and white matter can be achieved, using *k-means* clustering (Kanungo et al. 2002; Ibanez et al. 2005) for determination of the tissue types to discriminate and region growing for ensuring connectedness.

Based on a binary segmentation of the brain surface, several strategies for sulcus and gyrus classification have been presented and published in the past. A morphologic closing operation, i.e. dilation followed by erosion, with subsequent subtraction of the original MRI data allows processing of the sulcus volume via 3D skeletonization for extraction of the sulcus and gyrus folds (Lohmann 1998). In contrast to applying Euclidean distances, the use of a geodesic depth profile accounts for complexity and partial occlusion of the sulcus folds (Kao et al. 2006).

Besides these morphologic concepts, curvature analysis of a surface mesh calculated from gray and white matter can be utilized for detection of the gyrus and sulcus course (Vivodtzev et al. 2003) with respect to convexity and concavity.

In this paper we present a generic strategy for topographic analysis of arbitrary shapes and transformation of the depth profiles into cyclic graph representations. Thereby we account for imbalances in the local depth profiles due to asymmetries and deformations of the brain. The minimum graph connecting all local maxima and minima respectively is calculated, normalizing the local depth levels similar to the watershed segmentation concept. Our strategy is perfectly applicable for the task of gyrus and sulcus modelling as concave and convex paths can be identified. Based on the graph representations of the sulcus and gyrus courses, modelling of the brain surface can be easily achieved via distance-based classification utilizing morphologic operators as presented and discussed in the following sections.

2. MEDICAL BASICS

Classification of the human brain can be accomplished at different levels of granularity. At a top level, the main anatomical structures, like *cerebrum*, *cerebellum* and the *brain stem* can be identified. The cerebrum is

subdivided into two hemispheres and the main anatomical components, like *white matter*, *gray matter*, *cerebro-spinal fluid (CSF)*, *ventricle*, *fat*, *bones* and the arterial and venous *vessel systems* are demarcated. The brain tissue composed of white and gray matter is sub-classified into *frontal lobe*, *parietal lobe*, *occipital lobe* and some more, all specific areas responsible for diverse neurological functions of the body (Pschyrembel 2002). Each lobe comprehends several *gyrus* and *sulcus* areas, forming the brain surface. Thereby the gyri refer to the convex bulgs on the brain surface that are delimited by convex trenches, the so called sulci. The notable main sulci and gyri are named, listed and charted in anatomical atlases (Ono et al. 1990).

The topography of gyrus and sulcus characteristics is highly applicable for registration tasks in case of multi-modal image processing or follow-up examinations. Furthermore, modelling of the gyrus segments facilitates the quantitative analysis of metabolic activities with respect to defined anatomical structures.

3. DATA

For testing of the gyrus and sulcus modelling concept, $n=20$ T1-weighted MRI datasets of simulated brainweb database (Cocosco et al. 1997; Kwan, Evans, and Pike 1999) and associated reference segmentations are used.

Further test runs and validations will be performed utilizing $n=12$ anonymous multi-modal patient studies comprising morphologic image acquisitions (T1, T2, PD, ...) as well as functional images (SPECT, PET).

4. METHODOLOGY

Prior to performing the analysis process, a binary representation of the targeting object's surface, not addressed in this work, and a 3D depth profile must be pre-processed.

4.1. Estimation of the Reference Shape

For calculation of the depth profile of an arbitrarily shaped object, the reference shape, i.e. the smoothed shape without the vales and ridges, must be estimated.

Processing a solid body with an approximately spherical shape, like the human brain, calculation of the 3D convex hull (Barber, Dobkin and Huhdanpaa 1996; Sonka, Hlavac and Boyle 2007) as reference shape is highly feasible, see Fig. 1.

For other more complex shapes, where a spherical approximation would be too imprecise, an alternative calculation of the reference shape is feasible. When calculating a winged-edge isosurface of a binary 3D body (Baumgart 1972; Baumgart 1975; Ritter 2007; MeVis 2011), utilizing the quality factor, allows steering of the smoothing effect by polygonal reduction, i.e. up to which level, vales and ridges should influence the depth profile calculation, see Fig. 2 (b). The resulting isosurface is projected back to regular 3D voxel grid for further processing. Furthermore, morphologic closing operations as dilation followed by

erosion can be utilized for smoothing the surface and calculation of the reference shape, see Fig. 2 (c).

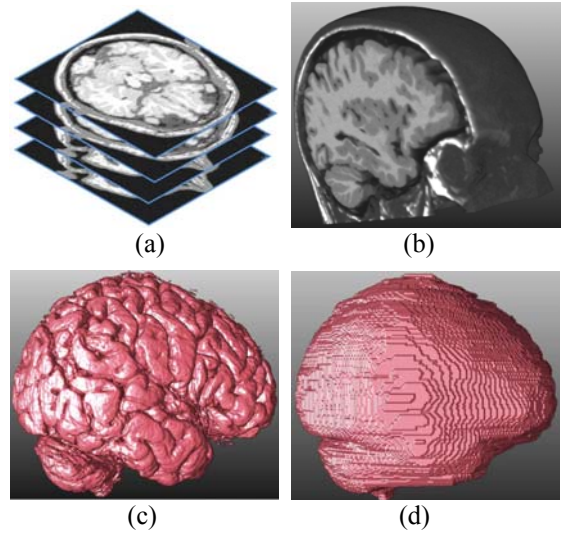


Figure 1: The stack of 2D MRI slices (a) assembles a 3D volume of the brain (b). After binary segmentation (c), the calculated convex hull (d) is the reference shape for depth profile calculation.

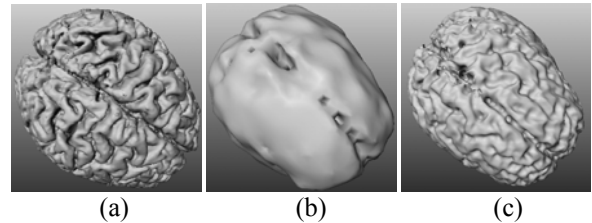


Figure 2: The precisely calculated surface model (a) can be smoothed for use as reference shape via rough isosurface calculation (b). As an alternative reference shape calculation strategy, a morphological closing operation with dilation kernel size $7 \times 7 \times 7$ followed by erosion $6 \times 6 \times 6$ can be applied (c).

4.2. Calculation of the Depth Profile

The depth profile is calculated as the minimum Euclidean distance between the surface of the object and the reference shape, see Fig. 3 as illustration of 2D depth profile calculation. A distance map calculation is used to represent the depth profile in 3D with the Euclidean neighbourhood weights for the adjacency constellations N_6 , N_{12} and N_8 in $3 \times 3 \times 3$ neighbourhood according to the distance from the hot spot as

$$w(N_6)=1.0; w(N_{12})=\sqrt{2}; w(N_8)=\sqrt{3}. \quad (1)$$

Starting at the convex hull, the distance weights are propagated to the particular neighbours to set or update their values. Whenever the depth value of a neighbour gets adapted, the change is recursively propagated to all adjoined neighbours. Calculation of the depth profile is finished, when the depth value of each voxel refers to the minimal Euclidean distance from the convex hull.

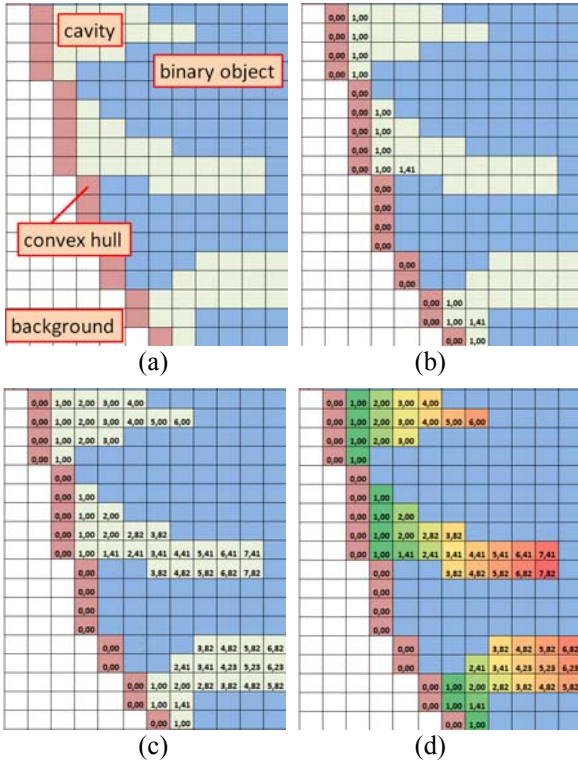


Figure 3: Calculation of the 2D depth profile for cavities between convex hull and the object's surface (a). In 2D case the neighbour weights are defined as $w(N_d)=1.0$ and $w(N_D)=1.41$. The depth profile is iteratively propagated until all depth values are calculated and convergence is reached (b-c). A colour-encoded representation of the final depth profile is plotted in (d).

The depth profile is calculated utilizing an Euclidean distance transform (Sonka, Hlavac and Boyle 2007). For calculation of the distance metrics, morphologic propagation of the outer surfaces, similar to the concept of grassfire transform (Blum 1967) is applied for fast approximate calculation of the Euclidean distance map from the object's borders. Results of the depth profile calculation are presented in Fig. 4. For the depth profile only the outer depth values below a threshold t are of relevance for gyrus and sulcus modelling. Depth profile values in the inner ventricular area are to exclude.

4.3. Watershed-Inspired Topography Analysis

Based on the calculated depth profile, the shortest graph interconnecting all local depth minimums is calculated, as well as a graph for connection of all local depth maxima. In the following delineation of the method, only the graph creation for the local maxima is addressed. The graph modelling for the local minimums can be derived by changing the leading signs and processing order.

The iterative process of graph construction is outlined in pseudo-code listing 1 and explained in detail in the following paragraphs.

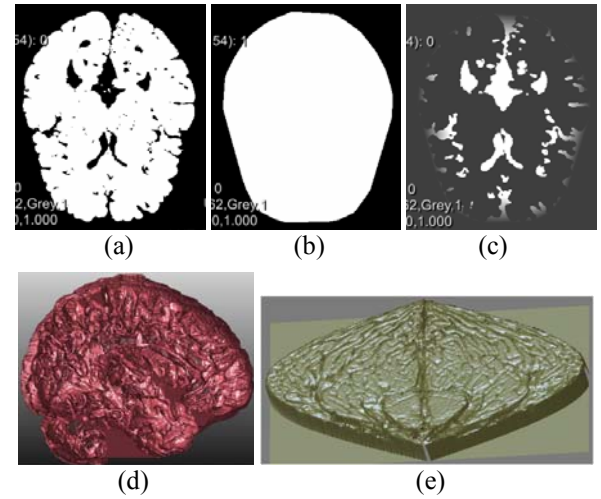


Figure 4: Planar representation of binary brain (a), convex hull (b) and depth profile (c) of mid slice transversal (axial) view. The 3D depth profile (d) as 2D map via Mollweide projection (Snyder 1987) is shown in (e).

The depth profile is processed at each particular depth level, starting at the maximum depth values and ending with the lowest depth values. At each depth level, voxels adjoined in N_{26} neighbourhood having matching depth values, are assembled together to recursively build up larger connected regions. For each of these constructed regions, the adjoined neighbourhood regions must be identified, differentiating between the following three constellations for the neighbourhood count N :

1. $N=0$: there are only background voxels or voxels at lower depth level not yet processed in the neighbourhood \rightarrow local maximum detected.
2. $N=1$: region is adjoined to one neighbouring region processed before with a higher depth value \rightarrow current region will be merged with the existing one.
3. $N>2$: there are several adjoined regions with higher depth values. The region to process is merged with the first neighbour region. Then the other neighbours are iteratively merged to one remaining cumulated region. Thereby, the skeleton sub-graphs must be interconnected.

For the autonomous regions to be interpreted as local maxima (condition 1), the first part of the graph is calculated via skeletonization. Below a region count of $R=50$ voxels, the region element closest to the centre of mass is taken as starting sub-branch of the graph, whereas for larger regions the medial axis, precisely extracted via skeletonization (Jonker 2002; Zwettler et al. 2009), is applied as starting sub-branch of the graph.

Regions with only one adjoined neighbouring region (condition 2), necessitate no discrete skeleton calculation. Instead, the region elements are just merged

with the neighbouring region, already defining a sub-branch.

If the current region is surrounded by more than one neighbour region at higher depth profile values (condition 3), besides a merge with the first neighbour region, all involved adjoined regions need to be cumulated. As all of the involved neighbouring regions have an already defined sub-part of the graph, these segments must be iteratively linked together. This link operation, described in the following sub-chapter, is a crucial task as it significantly influences the resulting graph after processing all depth levels from the deepest to the lowest profile values, see Fig. 5 for illustration of described iterative topography analysis. Starting at the deepest values with autonomous regions and the first skeleton parts Fig. 5 (c), the regions are enlarged whenever adjoined new regions at lower depth profile values are reached Fig. 5 (d-e). In case of reaching regions with already defined skeletons, the optimum connective path must be found Fig. 5 (f-g) utilizing detection of the optimal path, see Fig. 6. That way the topography describing path can be iteratively constructed until one final region remains Fig. 5 (h).

```

regions;
for(depth=maxVal; depth>=minVal; depth--)
  currRegions;
  for(xIdx=0; xIdx < sizeX; xIdx++)
    for(yIdx=0; yIdx < sizeY; yIdx++)
      for(zIdx=0; zIdx < sizeZ; zIdx++)
        if((distanceMap[xIdx][yIdx][zIdx] ==
            depth)&&
            (!classified(xIdx,yIdx,zIdx)))
          currRegions.Add(
            getRegionAt(xIdx,yIdx,zIdx));
  for(region : currRegions)
    neighbourRegions = getNeighbours(region);
    if(neighbourRegions.size == 0)
      region.CalculateSkeleton();
      regions.Add(region);
    else if(neighbourRegions.size == 1)
      merge(neighbourRegions.first, region)
    else
      merge(neighbourRegions.first, region)
      for (nRegion : neighbourRegions)
        if(nRegion != neighbourRegions.first)
          link_merge(
            neighbourRegions.first, nRegion)

```

Code Listing 1: Illustration of the topography analysis algorithm implemented in pseudo code. Regions at the same depth level are grouped together via region growing and stored in `currRegions`. Then for each region in `currRegions`, the neighbouring regions are identified. In case of neighbourhood *condition 1* with $N=0$, a seed region has been detected and the first skeleton is calculated. Seed regions are added to the global region container `regions`. If there is exactly one neighbouring region at higher depth level (*condition 2*), a region merging is performed. In case of additional neighbours (*condition 3*), the shortest skeleton path linking the involved regions is calculated prior to performing the merge operation.

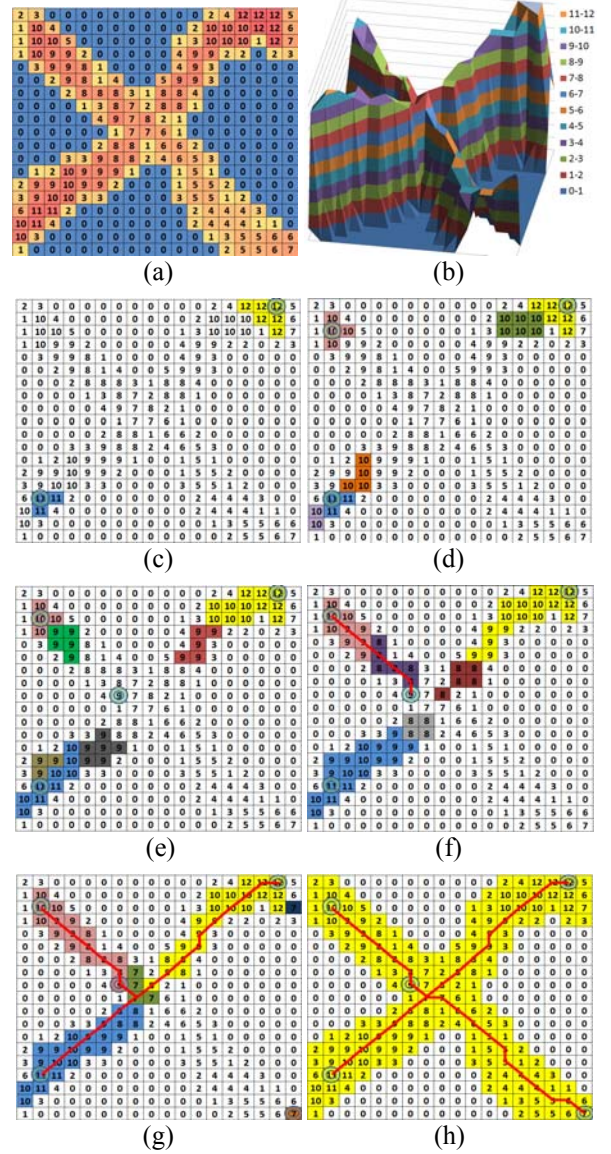


Figure 5: Illustration of the topography analysis on an cross-shaped 2D terrain.

4.3.1. Assembling Graph from Sub-Segments

When assembling two regions r_1 and r_2 with pre-calculated skeletons, i.e. graph sub-segments, the optimum path between the two skeletons $skel_1$ and $skel_2$ must be found. The search for the optimal connection path p_{min} is defined as minimization problem of a fitness function F , accounting for the Euclidean distance $dist()$ of the path and depth gradients of the path as $depth()$ with respect to the maximum profile depth $depth_{max}$, as

$$F(p) = \sum_{i=2}^{length(p)} (dist(p, i-1, i) \cdot (1 + depth_{max} - depth(i))) \cdot (2)$$

Thereby the target path p_{min} is that path of all possible non-cyclic connections between $skel_1$ and $skel_2$ with a minimal value for F .

For calculation of the optimal path, the metric value F of the current sub-path is propagated to its neighbours, starting at the skeleton elements $skel_1$ of

region r_j . The sub-path fitness value is recursively propagated to the particular neighbours and updated for each added path element. Whenever the first element of $skel_2$ is reached, an upper border for the optimal path fitness value is given. That allows reduction in the calculation complexity as all sub-paths exceeding this upper border fitness value can be aborted.

After convergence of the fitness values is reached, the optimum path can be traced back starting at the element e_2 of skeleton $skel_2$ with an adjoined neighbour showing the lowest fitness. Starting at e_2 the way to skeleton $skel_1$ is traced back by selecting the element with lowest fitness in each particular neighbourhood. The optimal path is finished, when the first neighbour of $skel_1$ has been reached. The described algorithm is illustrated in Fig. 6.

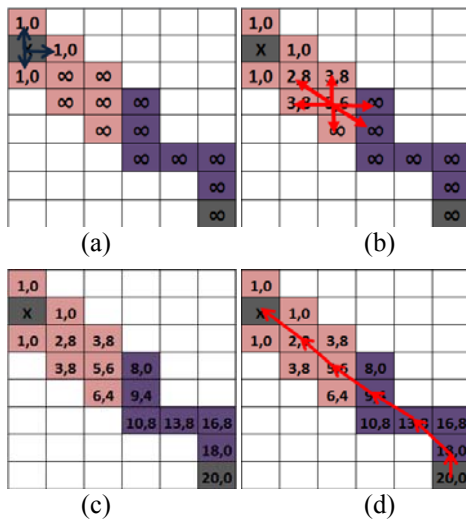


Figure 6: Detection of the optimal path between the two regions in Fig. 5 (f). Starting at the skeleton of the first region, all neighbours in N_8 are set or updated with the minimal path value according to the metrics defined in Eq. 2. Whenever a path value gets changed, all adjoined neighbours must be checked for propagation of the new value (b). Finally the path connecting the skeletons $skel_1$ and $skel_2$ can be traced back by picking each neighbour with the lowest value until skeleton of r_j is reached (d).

5. RESULTS

In this chapter first results of the discussed method are presented and discussed.

5.1. Results of Sulcus Classification

The graph model resulting from sulcus classification highly correlates with the main anatomical folds, see Fig. 7. Some pruning and smoothing operations on the graph can be utilized in future to remove dispensable bifurcations and redundant node elements.

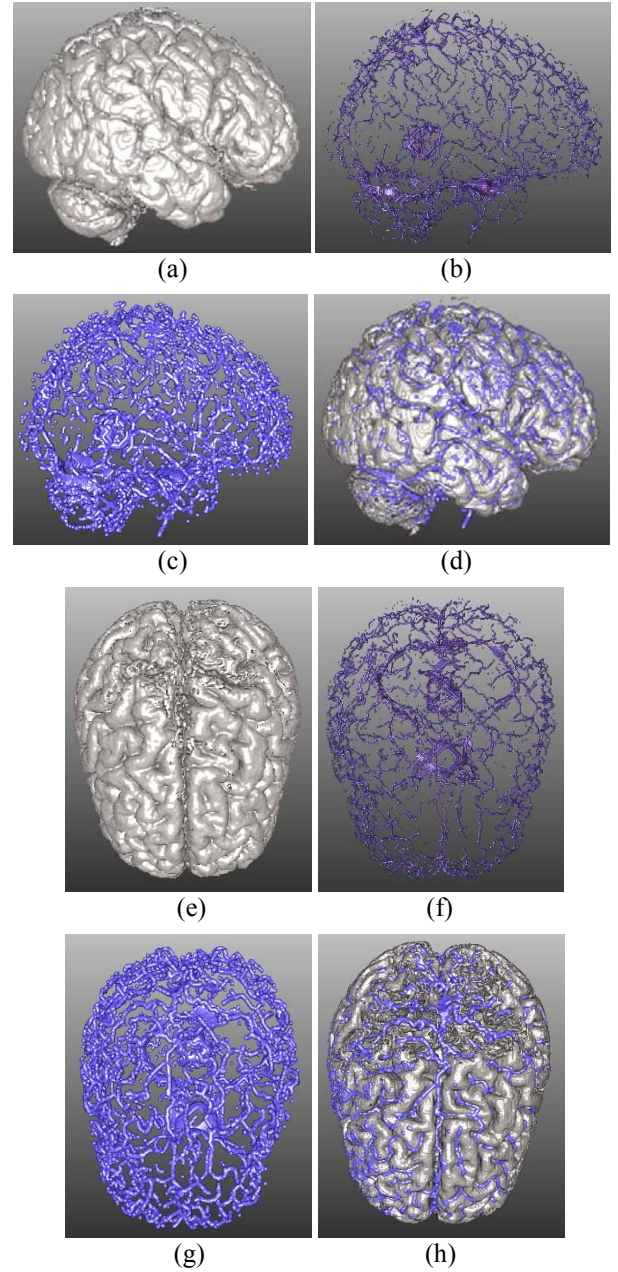


Figure 7: Visualization of sulcus analysis results in sagittal view (a-d) and axial view (e-h). The resulting tree model for the sulcus folds (b)(f) is amplified via morphologic dilation operation utilizing a $3 \times 3 \times 3$ structuring element for better visibility (c)(g). Correlation of original brain surface (a)(e) and the sulcus models is presented in (d)(h) via overlay.

As our presented algorithm only interconnects adjoined region skeletons bridging shallow sections, foreshortening could arise if the depth profile is strictly monotonic increasing. That problem can only be exploited by processing artificial testing data. As the first results show, even minimal deflection from strict monotonic behaviour prevents any foreshortening of the particular sub-branch, processing real-world medical image data.

6. DISCUSSION AND CONCLUSIONS

In this paper we introduce a novel concept for classification of the gyrus and sulcus folds and representation as graph model.

The presented iterative topography analysis with step-wise assembling and interconnecting the final graph allows handling of local imbalances of the depth profile, e.g. due to deformation or asymmetries of the brain. Besides roughly estimated upper threshold value for the profile depth to process with respect to imaging resolution, no further parameterization is required.

Linking the final graph representations and the depth profile allows later restrictions on the granularity of sulcus and gyrus representation.

Utilizing the presented sulcus and gyrus graph models, classification of the brain areas can be easily achieved by iterative assigning the gray matter and white matter voxels to the closest neighbouring subtrees, similar to vascularization-based anatomy classification (Zwettler, Backfrieder and Pichler 2011). Thereby the gyrus course can be used for iterative classification of the surrounding tissue and the sulcus lines are incorporated as additional barriers to prevent invalid merging of adjoined brain sections.

ACKNOWLEDGMENTS

Thanks to our medical partners from the Wagner-Jauregg state mental hospital, Linz, Upper Austria, at the institute for neuro-nuclear medicine headed by MD Robert Pichler for providing medical image data and valuable discussion.

This research is part of the INVERSIA project (<http://inversia.fh-linz.at>) which was funded by the European Regional Development Fund (ERDF) in cooperation with the Upper Austrian state government (REGIO 13).



REFERENCES

- Barber, C.B., Dobkin, D.P., Huhdanpaa, H., 1996. The Quickhull Algorithm for Convex Hulls. In *ACM Transactions on Mathematical Software (TOMS)*.
- Baumgart, B.G., 1972. Winged-Edge Polyhedron Representation, *Stanford University Artificial Intelligence Report No. CS-320*.
- Baumgart, B.G., 1975. Polyhedral Representation for Computer Vision. In *Proc. of the National Computer Conference*, 589-596.
- Blum, H., 1967. A Transformation for Extracting New Descriptors of Shape. In Wathen-Dunn, W., eds. *Models for the Perception of Speech and Visual Form*. MIT Press, Cambridge, 362-380.
- Cocosco, C.A., Kollokian, V., Kwan, R.K.-S., Evans, A.C., 1997. BrainWeb: Online Interface to a 3D MRI Simulated Brain Database. *Proceedings of the 3-rd International Conference on Functional Mapping of the Human Brain*, 5(4):425.
- Ibanez, L., Schroeder, W., Ng, L., Cates, J., 2005. *The ITK Software Guide*. Kitware Inc.
- Jonker, P.P., 2002. Skeletons in N dimensions using shape primitives. *Pattern Recognition Letters*, 23:677-686.
- Kanungo, T., Mount, D.M., Netanyahu, N., Piatko, C., Silverman, R., Wu, Y.A., 2002. An efficient k-means clustering algorithm – Analysis and implementation. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:881-892.
- Kao, C.-Y., Hofer, M., Sapiro, G., Stern, J., Rottenberg, D., 2006. A Geometric Method for Automatic Extraction of Sulcal Fundi. In: *IEEE International Symposium Biomedical Imaging*, 1168-1171.
- Kwan, R.K.-S., Evans, A.C., Pike, G.B., 1999. MRI simulation-based evaluation of image-processing and classification methods. *IEEE Transactions on Medical Imaging*, 18(11):1085-1097.
- Lohmann, G., 1998. Extracting Line Representations of Sulcal and Gyral Patterns in MR Images of the Human Brain. In: *IEEE Transactions on Medical Imaging* 17(6):1040-1048.
- MeVis, 2011. *MeVisLab – medical image processing and visualization*, MeVis Medical Solutions, Bremen, Germany. Available from [http://www.mevislab.de/developer/documentation/\[04/2011\]](http://www.mevislab.de/developer/documentation/[04/2011]).
- Ono, M., Kubick, S., Abernathy, C.D., 1990. *Atlas of the cerebral sulci*. Thieme publisher
- Pschyrembel, W., 2002. *Pschyrembel klinisches Wörterbuch*. De Gruyter, Berlin, Germany.
- Ritter, F., 2007. Visual Programming for Prototyping of Medical Applications. In *IEEE Visualization 2007 workshop*.
- Snyder, J.P., 1987. Map Projections – A Working Manual. In: *U.S. Geological Survey Professional Paper 1395*, Washington DC., U.S. Government Printing Office.
- Sonka, M., Hlavac, V., Boyle, R., 2007. *Image Processing, Analysis, and Machine Vision*. Cengage Learning, 3rd edition.
- Vivodtzev, F., Linsen, L., Bonneau, G.-P., Hamann, B., Joy, K.I., Olshausen, B.A., 2003. Hierarchical Isosurface Segmentation Based on Discrete Curvature. In: *Joint EUROGRAPHICS – IEE TCVG Symposium on Visualization*, 249-258.
- Zwettler, G.A., Swoboda, R., Pfeifer, F., Backfrieder, W., 2009. Fast Medial Axis Extraction Algorithm on Tubular Large 3D Data by Randomized Erosion. In: Ranchordas, A.K., Araujo, H.J., Pereira, J.M., Braz, J., eds. *Computer Vision and Computer Graphics – Communications in Computer and Information Science*. Springer Publisher, 24:97-108.
- Zwettler, G., Backfrieder, W., Pichler, R., 2011. Diagnosis of Neurodegenerative Diseases based on Multi-modal Hemodynamic Classification of the Brain. In *Proceedings of the International Conference on Computer Aided Systems Theory EUROCAST 2011*, 363-365.

AUTHORS BIOGRAPHY

Gerald A. Zwettler was born in Wels, Austria and attended the Upper Austrian University of Applied Sciences, Campus Hagenberg where he studied software engineering for medicine and graduated Dipl.-Ing.(FH) in 2005 and the follow up master studies in software engineering in 2009. In 2010 he has started his PhD studies at the University of Vienna at the Institute of Scientific Computing. Since 2005 he is working as research and teaching assistant at the Upper Austrian University of Applied Sciences at the school of informatics, communications and media at the Campus Hagenberg in the field of medical image analysis and software engineering with focus on computer-based diagnostics support and medical applications. His e-mail address is gerald.zwettler@fh-hagenberg.at and the research web page of the Research & Development department at campus Hagenberg, he is employed at, can be found under the link <http://www.fh-ooe.at/fe/forschung>.

Werner Backfrieder received his degree in technical physics at the Vienna University of Technology in 1992. Then he was with the Department of Biomedical Engineering and Physics of the Medical University of Vienna, where he reached a tenure position in 2002. Since 2002 he is with the University of Applied Sciences Upper Austria at the division of Biomedical Informatics. His research focus is on Medical Physics and Medical Image Processing in Nuclear Medicine and Radiology with emphasis to high performance computing. Recently research efforts were laid on virtual reality techniques in the context of surgical planning and navigation.

FAST MARCHING METHOD BASED PATH PLANNING FOR WHEELED MOBILE ROBOTS

Gregor Klančar ^(a), Gašper Mušič^(a)

^(a)University of Ljubljana
Faculty of Electrical Engineering
Tržaška 25, Ljubljana, Slovenia

^(a)gregor.klancar@fe.uni-lj.si, gasper.music@fe.uni-lj.si

ABSTRACT

The paper presents a path planning approach for wheeled mobile robots in obstructed environments. The trajectories of moving objects have to be carefully planned in order to obtain a near-shortest smooth path at still acceptable computational complexity. The combined approach is therefore proposed which utilizes search algorithm A* as well as methods of numerical solving of a particular form of partial differential equation - an eikonal equation. The use of related fast marching method enables to derive smooth trajectories within the shortest path corridor identified by the heuristic search algorithm while keeping the on-line computational burden relatively low. To illustrate the basic idea our investigation is limited to situation with static obstacles, e.g. buildings in the area which is crossed by autonomous vehicles. The proposed approach operation is validated by experimental results on a differential mobile robot.

Keywords: mobile robots, path planning, quadtrees, triangulation, fast marching method

1. INTRODUCTION

In the obstructed environments autonomous moving vehicles need to plan collision safe paths. With the given map of the environment and the target location the path planning aims to determine a trajectory, which will lead the object from the starting position to the target position. In general the planning involves two stages: (i) a suitable representation of the environment where the path has to be planned, and (ii) a search algorithm, that is capable of finding (sub)optimal path from the initial to the target position. The planning can involve a third stage where the path is optimized taking into account dynamic constraints of a moving object. This often leads to a requirement for smooth moving paths free of sharp turns.

The path planning methods initially transform the environment where the object of interest resides into a structure adapted to the requirements of path planning. Such representations include generalized Voronoi diagrams, various methods of space triangulation, regular grids and quadtrees, among others. Particular path planning algo-

gorithms may be used in combination with different representations of the environment, although certain representations are more suitable for some algorithms. Most generally used algorithms include standard A* and its derivatives D*, D* Lite, E*, algorithms based on fast marching method and others.

The literature review indicates environment segmentations based on Delaunay triangulation (dual to Voronoi diagram) (Hongyang, *et al.* 2008, Kallmann 2005) and framed quadtrees (Davis 2000) among the most promising approaches, and A* based path search algorithms as a suitable way of path optimization within the segmented environment. The computational demand of planning in complex environments is reduced by two or three level planning and exploitation of a-priori knowledge or heuristics (Botea, Müller and Schaeffer (2004)). An approach to improve path planning computational complexity by the use of interpolation and D* search algorithm is proposed in (Ferguson and Stentz 2006). The incremental search method that finds shortest paths for similar path -planning problems faster than uninformed search methods is presented in (Koenig 2004). Alternative approaches, e.g. bug algorithms, are also investigated but their application is questionable as the results are often not predictable, calculated path may be far from optimal or the target is not reached. A good overview of path planning approaches can be found in (LaValle 2006).

In the presented work the focus is given to the efficiency of path planning algorithm which is applicable to static or slowly changing dynamic environments. An A* path planning based algorithm is introduced with the quadtree and triangular representations of the search space. For fast changing environments the approach could be upgraded using D* algorithm as proposed in (Ferguson and Stentz 2006). As an alternative, the Fast Marching Method (FMM) based path planning was also investigated. A novel combination of path planning algorithms has been proposed with a suboptimal but efficient corridor calculation and an advanced FMM based path optimization within the corridor. The main goal of the presented approach is to obtain shortest smooth path between the obstacles, which complies to given kinematic constraints and can be calculated at acceptable computational complexity.

2. SEARCH SPACE SEGMENTATION BASED PATH PLANNING METHODS

Two space segmentation methods were considered in combination with an A* based path optimizer: quadtree segmentation and Delaunay triangulation.

2.1. Quadtrees space segmentation

Quadtrees (QT) enable a decomposition of the space map into quadratic cells (with the varying dimension) which are either free or occupied by obstacles. An essential step in quadtree generation is the cell occupancy test which should be carefully designed to retain computational efficiency.

QT segmentation results in compact environment map presentation and enables efficient query about occupancy of some position or area in the environment (Botea, Müller and Schaeffer (2004), Davis 2000).

To shorten the computational time of A* path finding algorithm the obtained quadtree is extended with visibility graph which indicates possible paths among free cells. To each free leaf cell an array of its free neighbour indices is adjoined. This enables easier and computationally efficient moves of A* algorithm among the free quadtree leaves. The visibility graph is computed as follows:

- Query for the area that is a little bigger (1/2 size of the prescribed minimal cell size) than current leaf size.
- Find leaves that are visible (accessible) from current leafs. Store their indices to the current leaf visibility array.
- Calculate distances to the visible neighbor leafs and store them in an array.

By this stage the obtained quadtree structure is prepared for A* path finding algorithm. The pathfinding algorithm based on A* search strategy guarantees the shortest path between start and end point if an optimistic heuristic is used. This means that predicted path cost (length) must always be smaller (shorter) or at least equal to the real path cost (length). Choosing line of sight for the predicted path length always fulfils this condition.

A corresponding example of the space segmentation and a calculated sample path are shown later in the section with experimental results (Fig. 1).

2.2. Space segmentation by constrained Delaunay triangulation

Triangulation is an important tool for representation of planar regions and is used in several application. The planar region is divided into a number of sub-regions of triangular shape. Commonly a set of points in the plane is given and the vertices of sub-regions must match the given set. This can be achieved in several ways, one of the possible triangulations is the Delaunay triangulation (DT).

The Delaunay triangulation assures a minimal number of narrow triangles (with small internal angles) which is a required property in many applications. The algorithms of Delaunay triangulation are also well explored and efficient algorithms yield computational complexity of $n \log n$ where n is the number of given points.

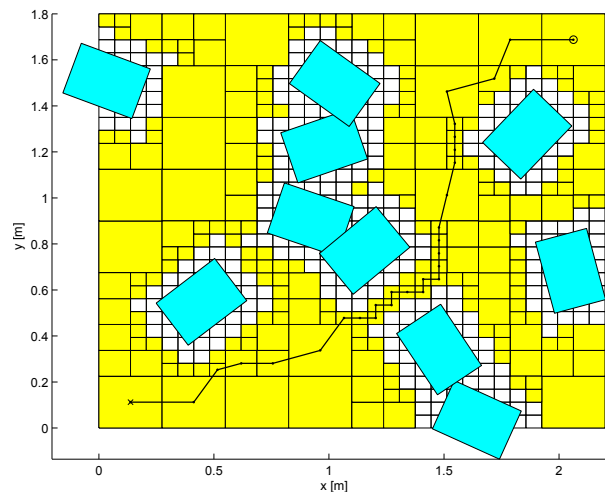


Figure 1: A* path planning based on QT segmentation for the given example.

When certain edges are fixed in advance and the rest of the edges are determined in the DT way, the triangulation is called Constrained Delaunay Triangulation (CDT). The pre-defined edges are constraints during triangulation. CDT is useful in cases where there are objects in the plane which should be taken into account during plane segmentation. Such is the case of path planning in a plane with obstacles.

The triangulated plane can be used for path planning in a similar manner as the quadtree based segmentation. A connectivity graph is built which contains information about the edges that can be crossed, i.e. the edges that do not belong to obstacles. Based on the triangulation and the connectivity graph, a path can be searched for by an A* type algorithm. A corresponding example for the same environment configuration as in Figure 1 is shown in Figure 2. The additional thin edges drawn within the triangles show the alternative paths explored by the algorithm. It can be observed that the built-in heuristic helps the algorithm to search only a part of the whole space.

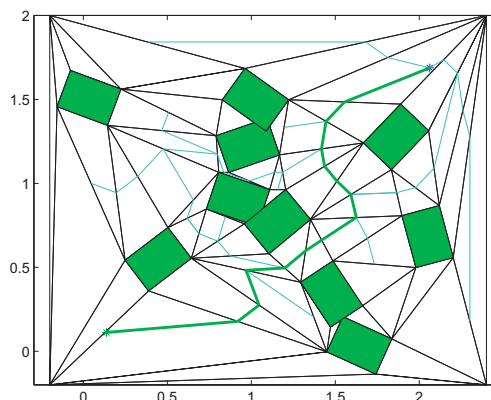


Figure 2: Example of path planning based on triangular segmentation

2.3. Path smoothing and shortening

As can be observed from Figures 2 and 1 the obtained path passes through centers of the cells in quadtree space representation or crosses midpoints of the triangle edges in triangulation. None of the two approaches gives a smooth or straight path. The path is cornered even in the areas with no obstacles, which is not desired from the practical point of view, since the movement of an autonomous object within the environment would be unnatural in such cases.

This can be improved by considering the fact that the array of cells which are crossed by the calculated path actually defines a channel between the starting and the target point. This channel will be termed a path corridor and consists of a set of neighbouring square shaped cells when using quadtrees. When using CDT the corridor consists of a set of neighbouring triangular cells.

In both cases, the path within the corridor can be optimized by applying the funnel algorithm (Kallmann 2005), which finds the shortest path within the corridor. The basic steps of the algorithm are briefly described as follows:

- The corridor is defined by two sets of points: a set of points defining the upper corridor border and a set of points defining the lower corridor border. In case of triangulation the corridor border follows the edges of participating triangles (Fig. 3). In case of quadtree representation, the size of the neighbouring cells may differ, and in such a case only the overlapping part of the edges between the two cells is considered when defining lower and upper corridor border (see Fig. 9).
- The start and the target point are linked to the upper and lower corridor border by additional edges.
- Let p be the starting point and let u and v be the points on the upper and lower border of the corridor, respectively. The shortest path from p to u and from p to v (not leaving the corridor) may overlap up to some point a . At a the paths diverge and are concave until they reach u and v . The a is called apex and the region delimited by path segments from a to u , a to v and uv segment is the funnel.
- The algorithm iteratively adds points on the corridor borders narrowing the funnel and discarding the points falling out of the narrowed funnel. When the top of the funnel shrinks down to a line, the shrunk part of the funnel defines a new segment of the shortest path and the new apex is set at the end of the new segment. The procedure stops when the target is reached. For more details, see (Kallmann 2005).

As an example, the planned path from Fig. 2 for the triangular representation is improved by funnel algorithm as shown in Fig. 3.

2.4. Experimental results

The initial requirement of this investigation was to determine a path planning method which would enable fast, real-time path planning for moving objects within the obstructed environments. The primary application area are wheeled mobile robots moving at relatively slow speeds.

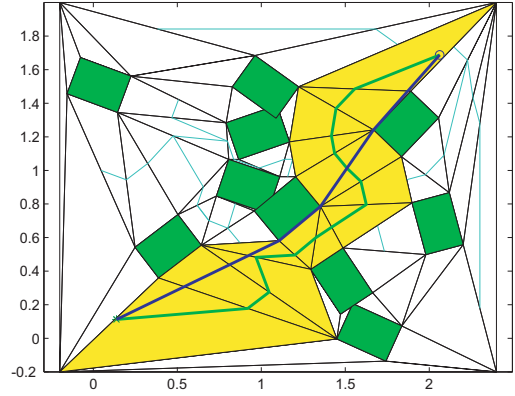


Figure 3: Example of the optimal path (thick dark line) determination inside the corridor based on the triangular representation.

Real-time in this case means a time, which is short enough for the robot to plan a journey without a considerable delay. E.g., delays up to 0.5 s may be tolerated. Furthermore, we limited our investigation to situation with static obstacles, e.g. buildings in the area which is crossed by autonomous vehicles.

All mentioned algorithms were implemented in Matlab m-code and both segmentation methods were tested in combination with environments of different complexities.

Tables 1 and 2 show results of a set of experiments with varying environment size and varying number of obstacles within the environment on a 2.4 GHz PC. Within the tables, the first column shows the number of obstacles, next is the min/max size of segmentation cells, N_{cells} is the number of cells after the segmentation, t_{QT} and t_{DT} are the segmentation computing times, and the remaining columns show computing times and corresponding standard deviation of path search algorithm, corridor boundaries calculation, and the funnel algorithm, respectively. The computing times were obtained by averaging eight consequent runs of the path planning algorithm with different start and target points.

The obtained results show that both methods can be used for path planning in real-time for moderately complex

Table 1: Computation times - quadtrees

Environment size 10 x 10 m, maximal obstacle size 1 x 1 m									
N_{obst}	$MinDim$ [m]	N_{cells}	t_{QT} [s]	t_{A^*} [ms]	σ_{A^*} [ms]	t_{CO} [ms]	σ_{CO} [ms]	t_{FU} [ms]	σ_{FU} [ms]
10	10/32	273	0.2	9	6	4	2	6	3
	10/64	577	0.5	11	6	2	1	4	2
	10/128	1301	1.0	12	11	2	1	4	2
20	10/32	405	0.4	12	7	4	2	6	3
	10/64	877	0.8	11	8	3	1	5	2
	10/128	2189	1.8	14	12	3	1	5	2
50	10/32	797	1.0	13	9	4	1	6	2
	10/64	1921	2.6	23	23	3	2	5	2
	10/128	4817	6.2	49	56	4	3	7	4
100	10/64	3345	6.0	47	37	5	3	9	5
	10/128	9577	16.7	83	114	4	3	7	4
Environment size 100 x 100 m, maximal obstacle size 5 x 5 m									
100	100/256	12265	25	196	192	5	2	8	4

Table 2: Computation times - triangulation

Environment size 10 x 10 m, maximal obstacle size 1 x 1 m									
N_{obst}	MaxDim [m]	N_{cells}	t_{DT} [s]	t_{A^*} [ms]	σ_{A^*} [ms]	t_{CO} [ms]	σ_{CO} [ms]	t_{FU} [ms]	σ_{FU} [ms]
10	/	82	0.2	35	22	/	/	4	7
	1.25	244	0.4	63	54	/	/	0	0
	0.67	594	1.3	147	169	/	/	0	0
	0.4	1434	6.6	369	350	/	/	14	6
20	/	168	0.3	53	36	/	/	6	8
	1.25	328	0.6	96	67	/	/	0	0
	0.67	678	1.7	160	145	/	/	6	8
	0.4	1518	6.0	465	447	/	/	10	8
50	/	432	0.9	107	36	/	/	6	8
	1.25	606	1.6	133	91	/	/	4	7
	0.67	956	3.3	348	280	/	/	8	8
	0.4	1796	10.1	797	539	/	/	12	7
100	/	958	3.5	199	88	/	/	6	8
	1.25	1076	4.0	217	136	/	/	8	8
	0.67	1426	6.2	348	188	/	/	6	8
	0.4	2266	13.8	687	513	/	/	10	8
Environment size 100 x 100 m, maximal obstacle size 5 x 5 m									
100	/	838	2.8	352	220	/	/	10	8

environments provided the space segmentation is done in advance. So only the path search, corridor calculation, and funnel algorithm need to be computed in real-time.

The main drawbacks of the two segmentation approaches are non-smooth paths which consists of straight line segments. Funnel method yield smooth paths, but tend to follow the edges of the obstacles. Therefore an alternative way of path planning was considered, which is described in the next section.

3. FAST MARCHING METHOD BASED PATH PLANNING

Fast marching method (FMM) is based on numerical analysis of viscous matter and is a method of numerical solving a particular form of nonlinear partial equation, i.e. an Eikonal equation.

Simplified, the method gives a description of wavefront propagation through nonhomogeneous medium, where the propagation is represented by a wavefront arrival time for every point in the space (Sethian 1999).

When the propagation velocity for a point in the space is defined by F (which is always non-negative), the arrival time function T is determined by the solution of equation

$$|\nabla T|F = 1 \quad (1)$$

at given border conditions, i.e. at a condition of zero value of T in the starting point. If F depends only on space coordinates, the above equation is an Eikonal equation.

The numerical solution of the equation is based on a space grid, approximation of the gradient by the values in the neighbouring points and an efficient strategy of point calculation order.

Figure 4 shows a point in a 2D grid and its neighbouring points. In 2 dimensions the gradient $|\nabla T(x, y)| = |\nabla T(i, j)|$, where $x = i\Delta x$ in $y = j\Delta y$ can be substituted by an approximation (2)

$$\left(\begin{array}{l} \max(D_{ij}^{-x}T, 0)^2 + \min(D_{ij}^{+x}T, 0)^2 + \\ \max(D_{ij}^{-y}T, 0)^2 + \min(D_{ij}^{+y}T, 0)^2 \end{array} \right) = \frac{1}{F_{ij}^2} \quad (2)$$

where

$$D_{ij}^{-x}T = \frac{T_{i,j} - T_{i-1,j}}{\Delta x} \quad (3)$$

$$D_{ij}^{-y}T = \frac{T_{i,j} - T_{i,j-1}}{\Delta y}$$

and

$$D_{ij}^{+x}T = \frac{T_{i+1,j} - T_{i,j}}{\Delta x} \quad (4)$$

$$D_{ij}^{+y}T = \frac{T_{i,j+1} - T_{i,j}}{\Delta y}$$

In (Sethian 1999) is proposed that the gradient is substituted by a simpler, less accurate approximation

$$\begin{aligned} \max(D_{ij}^{-x}T, -D_{ij}^{+x}T, 0)^2 + \\ \max(D_{ij}^{-y}T, -D_{ij}^{+y}T, 0)^2 = \frac{1}{F_{ij}^2} \end{aligned} \quad (5)$$

Considering (3) and (4) the last equation can be modified to

$$\begin{aligned} \max\left(\frac{T_{i,j} - \min(T_{i-1,j}, T_{i+1,j})}{\Delta x}, 0\right)^2 + \\ \max\left(\frac{T_{i,j} - \min(T_{i,j-1}, T_{i,j+1})}{\Delta y}, 0\right)^2 = \frac{1}{F_{ij}^2} \end{aligned} \quad (6)$$

By setting new labels

$$\begin{aligned} T &= T_{i,j} \\ T_1 &= \min(T_{i-1,j}, T_{i+1,j}) \\ T_2 &= \min(T_{i,j-1}, T_{i,j+1}) \end{aligned} \quad (7)$$

the equation becomes

$$\max\left(\frac{T - T_1}{\Delta x}, 0\right)^2 + \max\left(\frac{T - T_2}{\Delta y}, 0\right)^2 = \frac{1}{F_{ij}^2} \quad (8)$$

Assuming F is always positive, T is monotonically increasing. The solution in a given point is only influenced by solution values in those points where the solution value is smaller. The fast marching method is based on the information propagation in one direction, from smaller values of T to larger values (Baerentzen 2001, Farag and Hasouna 2005).

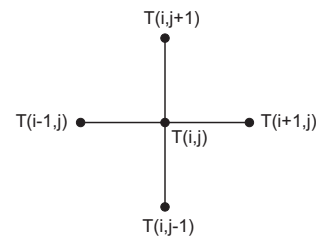


Figure 4: A point in the grid and its neighbours

3.1. FMM method and path planning

The method can be easily applied to shortest path planning as the time of arrival into a point in the space is always the earliest possible time, and the known obstacles are simply taken into account during the calculation of the wavefront propagation. It suffices to set the propagation velocity to zero for any point inside the obstacle, which prevents wavefront from entering.

Once the arrival time function is calculated, the shortest path can be reconstructed by following the largest gradient. This can be done by simple Euler's method or by more precise Heun's (modified Euler's) method. The path can be calculated in both directions, i.e. from the wavefront starting point into any target point in the space or reversed, from a set of starting points to a fixed target point, which is the wavefront starting point. Such an example for an environment of $2.2 \text{ m} \times 1.8 \text{ m}$ with 10 A4 size obstacles is shown in Figure 5. The time arrival function and the calculated paths from 5 points are shown in 3D view first, and then the projection to the x-y plane is shown. Note that the time arrival function value is not defined for the interior of the obstacles, but was fixed at 300 for visualization purposes.

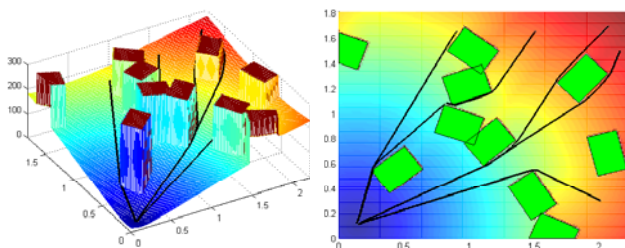


Figure 5: Example of shortest path planning by FMM method and a fixed target point

3.2. Smooth path planning and FMM method

The arrival time function calculated by FMM method serves as a potential field which directs the moving object toward the target point. If additional information is included into this field, this can influence the calculated path. Such an information is the information about the distance to the obstacles. This way the path may be pushed away from the borders of the obstacles.

The distance to the obstacles can be included by the use of Extended Voronoi Transformation (EVT), which is used in digital image processing (and called Distance Transform therein). If EVT information is included into FMM based path planning, the obtained paths are driven away from the obstacles and smooth at the same time (Garrido, Moreno and Blanco 2009). An example is shown in Figure 6.

4. THE COMBINED APPROACH

While in general the FMM method gives better results in terms of optimal lengths of the planned paths and can also be adapted to smooth down the paths as described above, its limitation in the substantial computational burden compared to segmentation based methods. E.g. in the example

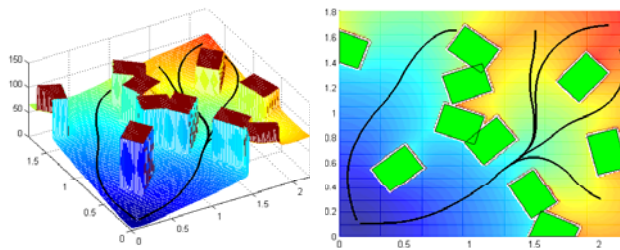


Figure 6: Example of smooth path planning by FMM method and a fixed target point

shown above the size of environment was $2.2 \times 1.8 \text{ m}$, grid points were 0.01 m apart, the EVT calculation took approx. 12 s , but fortunately needs to be performed only once for a fixed configuration of obstacles. FMM algorithm takes around 3.5 s and then backtracking by Heun's method another 0.01 to 0.1 s , depending on the step size. All times were obtained on a 2.4 GHz PC by implementation of algorithms in Matlab m-code. The code is not optimal, nevertheless, the listed computation times indicate that the use of the method in real-time is not feasible for any complex environment.

The computational time for the calculation of the arrival time function strongly depends on the length of the propagating wavefront. It is relatively short when the wavefront can only advance in a narrow corridor. This implies the feasibility of a combined approach, where segmentation of the space is first applied to coarse initial path planning, then a suitable corridor is defined within which a finer path planning is performed by the FMM method.

The corridor used for FMM can be exactly the same that is used by the funnel algorithm. Given a set of obstacles, the EVT transformation is computed first. This could be done for points within the corridor only, but also the obstacles away from the corridor should be taken into account. Due to large computing time of the EVT it is more convenient to compute it for all the points in advance and then mask the points outside of the corridor by setting their EVT value to zero.

More precisely, the EVT transformation operates on a bitmap image and calculates the Euclidean distance from every pixel to the nearest pixel carrying a specific property, in our case a pixel that belongs to the obstacle. This is a concept related to Voronoi graph, with the difference that the information is attached to every picture element and not only the points on the edges of the Voronoi cells. A number of EVT algorithms calculate the Voronoi diagram as an intermediate step.

By using the EVT information during the FMM based path planning the planned path is automatically pushed away from obstacles and smoothed at the same time because the path does not follow the obstacle edges as it may happen with funnel algorithm.

In (Garrido, Moreno and Blanco 2009) the authors suggest to include the distance to the obstacles as a velocity parameter when calculating the wavefront propagation. This makes analogy to propagation of light ray through the medium with non-homogeneous refractive index. The light in such a medium is not refracted but bent smoothly. For

the path planning the distance is not used directly, but is transformed analogously to electric potential by a logarithmic function, e.g. (Garrido, Moreno and Blanco 2009):

$$F_{i,j} = c_1 \log(R_{i,j}) + c_2 \quad (9)$$

where $R_{i,j}$ stands for the distance of point (i, j) to the nearest obstacle. For the purpose of the presented study this function was further modified to

$$F_{i,j} = \begin{cases} c_1 \log(\frac{R_{i,j}}{\Delta x} + c_3) + c_2, & \log(\frac{R_{i,j}}{\Delta x} + c_3) > 0 \\ 0, & \log(\frac{R_{i,j}}{\Delta x} + c_3) \leq 0 \end{cases} \quad (10)$$

Equality $\Delta x = \Delta y$ is assumed, and the proposed function enables balanced results with various grid sizes. Weights c_i define the shape of the velocity field and consequently influence the shape of the derived path. In particular, c_1 is related to the velocity of an object moving along the path, c_2 influences the curvature of the path, and c_3 can be used to define a safety margin around obstacles. At the same time the function is consistent with the approach used by FMM, where the interior of obstacles is indicated by adjoining the corresponding grid elements by zero velocity of wavefront propagation.

Additional distance terms $F_{i,j}^k$ may be added, e.g. measuring the distance to some predefined environment borders. These are calculated in a similar manner as (10). The velocity field given by EVT is then calculated as a sum of partial distance functions. The corresponding weights c_i^k may be used as additional design parameters.

With the described calculation of EVT the proposed path planning method can be summarized in the following steps:

- The QT or CDT segmentation of the environment with known obstacle positions is calculated.
- The environment is covered by a grid of points and the EVT transformation is calculated interpreting every point of the grid as a pixel of the image.
- Start and target points are chosen and a corresponding path is planned by an A* type search algorithm. As a result, a path and a corresponding set of segmentation cells are obtained, the cells defining a corridor surrounding the calculated path.
- The information about the corridor boundaries is used to adjust EVT transformation by setting the calculated values of the EVT to zero for any point outside the corridor.
- The newly obtained EVT is used to parametrize the FMM based path search algorithm to obtain a smooth suboptimal path within the corridor.

If new path has to be calculated for another set of points, it suffices to repeat only the last three steps of the proposed procedure.

An example of the path planning results for the CDT-A*-FMM based method is shown in Fig. 7. Configuration of the obstacles as well as starting and target points are the same as in Fig. 2 and Fig. 3. Three paths obtained by varying weight c_2 are shown, which demonstrates how

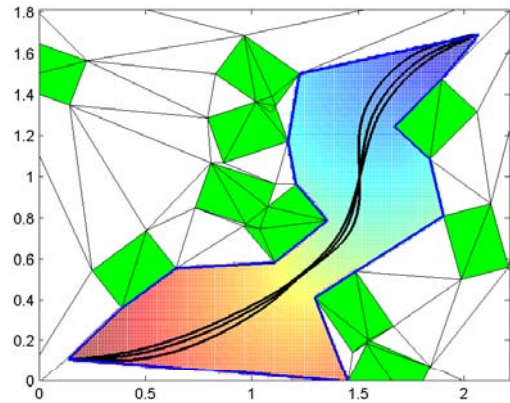


Figure 7: Example of the path smoothing in triangular space segmentation – arrival time function within the corridor and resulting path obtained by the FMM algorithm.

parameters of (10) can be used to trim the shape of the calculated trajectory.

Some preliminary results regarding computational complexity of the proposed method are shown in Tables 3 and 4. The first table shows the results obtained by the combination of quadtrees and FMM method while the second table shows the results obtained by the combination of constrained Delaunay triangulation and FMM method.

The number of obstacles N_{obst} is shown, the min/max size of segmentation cells, raster size used by FMM method, computing time of EVT transformation, and the remaining columns show the computation time needed for adjusting EVT to the given corridor (t_{EC}), for calculation of arrival time function within the corridor by FMM method (t_{FMM}), and for calculation of the path by following the largest gradient by Heun's method (t_H). By the last part, the integration step size was chosen according to the raster size: $step = 0.1/raster$. Similarly as before, the last three computation times refer to the average of eight consequent runs of the path planning algorithm with different start and target points.

Table 3: Combined method computation times - quadtrees

Environment size 10 x 10 m, maximal obstacle size 1 x 1 m						
N_{obst}	MinDim [m]	raster [m]	t_{EVT} [s]	t_{EC} [s]	t_{FMM} [s]	t_H [ms]
100	10/128	0.02	612	9.4	1.0	21.5

Table 4: Combined method computation - triangulation

Environment size 10 x 10 m, maximal obstacle size 1 x 1 m						
N_{obst}	MaxDim [m]	raster [m]	t_{EVT} [s]	t_{EC} [s]	t_{FMM} [s]	t_H [ms]
100	1	0.05	96	2.5	0.42	4.1
		0.02	612	12.1	2.6	22.5

4.1. Experimental results

The proposed approach operation was also validated experimentally. The experiments were performed on the small,

two-wheeled, differentially driven mobile robot in the environment 2.2×1.8 m where obstacles are implemented by A4 paper sheets which are randomly placed in the environment as shown in Fig. 8.

The robot measures $7.5 \times 7.5 \times 7.5$ cm and weighs 0.6 kg. It contains a C167 microcontroller running at a 20-MHz clock, a 12-V battery supply, two powerful DC motors equipped with incremental encoders (512 pulses per revolution), and a gear reduction head.

The reference path for the robot is calculated using the presented algorithm with A* path optimization on QT space representation (Fig. 1), corridor determination and funnel algorithm or FMM inside the corridor. The path planning part of the algorithm is implemented in Matlab environment while the path-tracking control part is implemented in C++ environment. robot current pose is obtained by the overhead camera with 33 ms sampling rate.

The determined corridor and the shortest path within the corridor obtained by the funnel algorithm are given in Fig. 9. The path is piecewise linear, which enables good tracking within the segments but could result in tracking errors at the junctions of the linear segments.

The overall tracking performance can be improved by the proposed combined planning approach. Therefore the path within the corridor is determined by FMM taking into account distance to the obstacles as well as distance to the borders of the corridor.

The feasibility of the planned path is analysed by calculation of the linear and angular velocities as well as tangential and radial accelerations. For the used mobile robot the maximal velocity is constrained at 1 m/s while the analysis in (Lepetic, *et al.* 2003) shows that the path can be tracked when the maximal tangential acceleration remains below $2m/s^2$ and maximal radial acceleration is below $2m/s^2$.

The velocity profile is calculated directly from the points obtained by Heun's method. The obtained points are interpreted as points on the path sampled at equidistant sample times. The linear and angular velocities are then calculated by backward difference approximation of

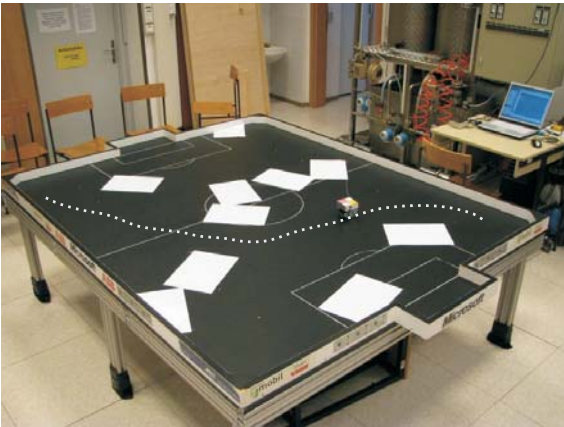


Figure 8: Experiment setup. The robot needs to travel between the obstacles (A4 paper sheets) from right corner to the left corner.

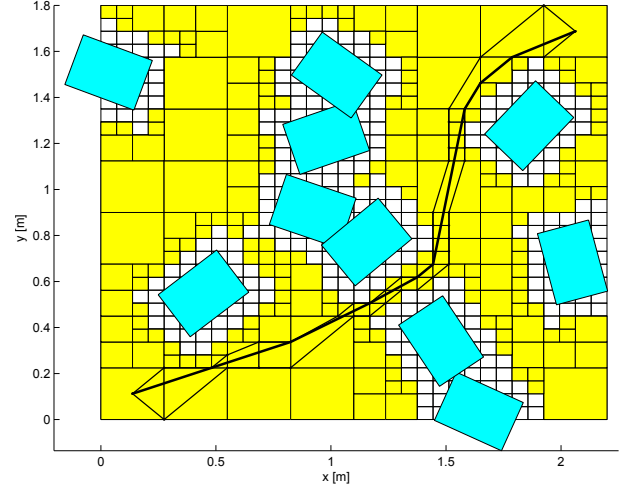


Figure 9: Example of the optimal path (thick line) determination inside the corridor (thin lines) obtained by A* path finding algorithm and QT space representation.

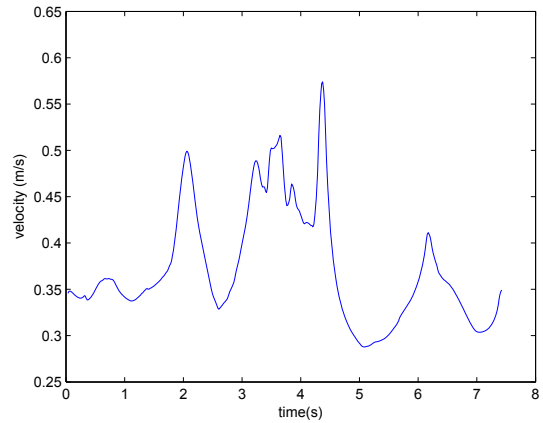


Figure 10: Velocity profile of the mobile robot for the planned trajectory.

expressions

$$v_{ff}(t) = \sqrt{\dot{x}_r^2(t) + \dot{y}_r^2(t)} \quad (11)$$

$$\omega_{ff}(t) = \frac{\dot{x}_r(t)\ddot{y}_r(t) - \dot{y}_r(t)\ddot{x}_r(t)}{\dot{x}_r^2(t) + \dot{y}_r^2(t)} \quad (12)$$

The velocity profile obtained this way is shown in Fig. 10. The velocity remains below the limit of 1 m/s. Similarly, the tangential and radial acceleration are calculated by approximation of

$$a_t(t) = \frac{dv_{ff}(t)}{dt} \quad (13)$$

$$a_r(t) = v_{ff}(t)\omega_{ff}(t) \quad (14)$$

Fig. 11 shows that the calculated accelerations remain within the limits given by (Lepetic, *et al.* 2003), which proves the feasibility of the planned path.

To further validate the feasibility of the approach the tracking of the planned paths was also tested experimentally. To drive the robot with differential kinematic on the

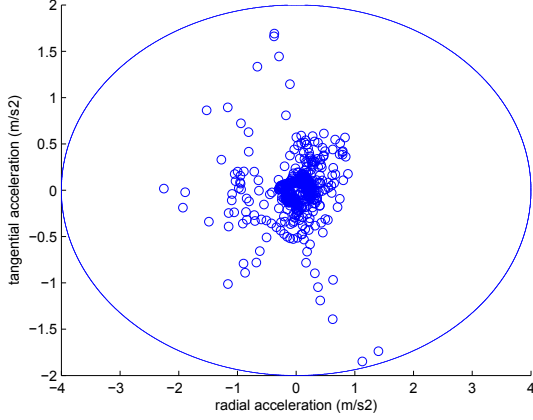


Figure 11: Radial and tangential acceleration of the mobile robot for the planned trajectory.

planned path the control law (Oriolo, Luca and Vandittelli 2002) is as follows

$$\begin{aligned} v(t) &= v_{ff}(t) \cos e_\theta(t) + k_x e_x(t) \\ \omega(t) &= \omega_{ff}(t) + k_y v_{ff}(t) \frac{\sin e_\theta(t)}{e_\theta(t)} e_y(t) + k_\theta e_\theta(t) \end{aligned} \quad (15)$$

where k_y is a positive constant, while $k_x(\cdot)$ and $k_\theta(\cdot)$ are continuous positive bounded functions. The trajectory tracking errors $\mathbf{e}(t) = [e_x(t), e_y(t), e_\theta(t)] = \mathbf{R}(\mathbf{q}_r(t) - \mathbf{q}(t))$ are expressed in robot local coordinates (fixed to the robot center and x axis is in the robot forward direction) where \mathbf{R} is the rotation matrix between global and robot coordinate frame, $\mathbf{q}_r(t)$ is the reference robot pose on the planned path and $\mathbf{q}(t)$ is the current robot pose. The feedforward inputs are calculated from the planned reference trajectory by (11) and (12).

In Fig. 12 the robot was controlled to follow the optimal planned (sequentially linear) path obtained using the funnel algorithm inside the corridor. Robot starts in the upper right corner ($x = 2.06m$, $y = 1.69m$) and ends in the lower left corner ($x = 0.14m$, $y = 0.11m$) which coincides with situation in Fig. 1. Robot feedforward inputs were selected as $v_{ff}(t) = 0.4$ m/s and $\omega_{ff} = 0$ 1/s and gains in Eq. 15 are $k_x = 4$, $k_y = 30$ and $k_\theta = 4$. It can be seen that robot has bigger tracking error when the path changes discontinuously. At higher tracking errors it becomes possible for the robot to leave the safe corridor and can even hit some obstacle.

Better tracking performance is obtained in Fig. 13 where the smooth reference path is obtained using FMM inside the corridor. The feedforward inputs $v_{ff}(t)$ and $\omega_{ff}(t)$ are calculated from the smooth reference path using equations (11) and (12). The robot can follow the reference path with much smaller tracking error therefore the possibility that the robot escapes the safety corridor and hit some obstacle becomes much smaller.

In the presented experiments the environment is partitioned in 813 cells where the smallest cell is approximately the size of the robot. On 2.4 GHz PC the QT algorithm takes approximately 0.4 s, optimization with A* algorithm takes around 15 ms, funnel algorithm inside the corridor

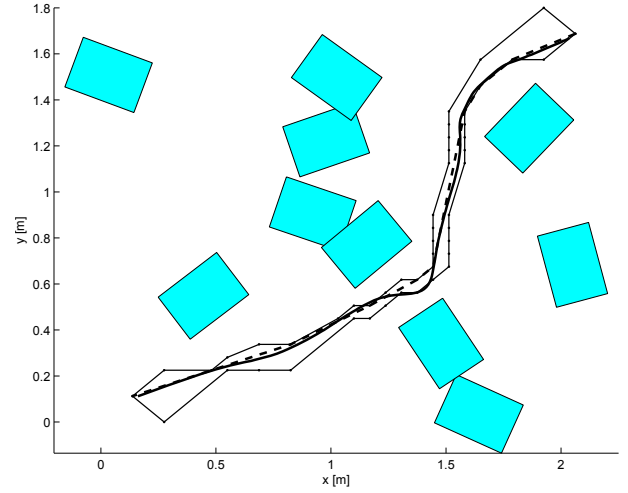


Figure 12: Mobile robot trajectory tracking experiment (thick line). The reference path (thick dotted line) is obtained by Funnel algorithm inside the determined corridor (thin lines).

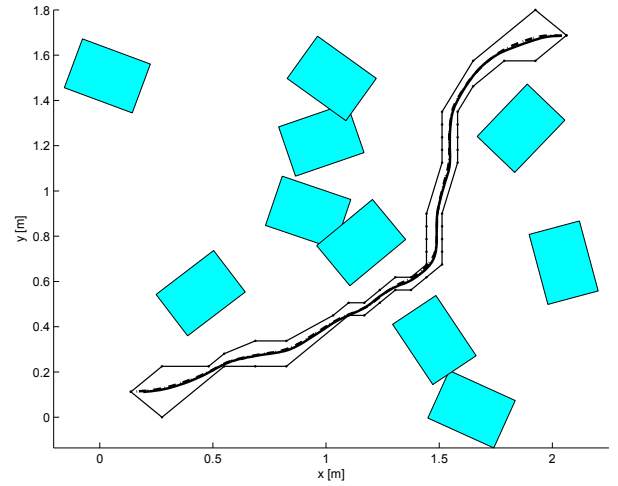


Figure 13: Mobile robot trajectory tracking experiment (thick line). The reference path (thick dotted line) is obtained by FMM inside the determined corridor (thin lines).

takes 7 ms while the FMM (with raster $\Delta x = \Delta y = 1$ cm) takes some 3.5 s for the whole image and approximately 350 ms for the calculation inside the corridor only.

5. CONCLUSIONS AND FUTURE WORKS

The results show the feasibility of the proposed path planning method and confirm the significant reduction in computation times of FMM in the corridor compared by the FMM for the whole search space. Due to rather small size of the minimal cells of the QT a small FMM raster size must be chosen in this case to avoid numeric problems when following the gradient in the narrow parts of the corridor. This indicates that smaller granulation of QT is not necessary advantageous. Besides, the larger granulation yields more room in the corridor for FMM to smooth down the planned path. On the other hand, larger granulation prevents the path planning algorithm to draw the

planned path over the narrow passages among obstacles. The resulting paths may therefore not be optimal in terms of their lengths. Presented path planning approach is evaluated by experiments on differential mobile robot where the proposed corridor constrained FMM approach results in smooth planned paths which enable good trajectory tracking results. Future work will try to increase the robustness of the algorithm in exceptional cases, and further investigate the parametrization of the velocity field with respect to kinematic constraints. The study how to upgrade the proposed approach for dynamic environments where A^* is used only for initial planning and D^* for each path replanning will also be performed.

REFERENCES

- Baerentzen, J.A., 2001. On the Implementation of Fast marching Methods for 3D Lattices, *Technical Report IMM-REP-2001-13*, Technical University of Denmark.
- Botea, A., Müller, M., Schaeffer, J., 2004. Near optimal hierarchical path-finding, *Journal of Game Development*, 1, 7-28.
- Davis, I., 2000. Warp speed: Path planning for star trek: Armada, *AAAI Spring Symposium on AI and Interactive Entertainment*, AAAI Press, Menlo Park, CA.
- Farag, A.A., and M.S. Hassouna, 2005. Theoretical Foundations of Tracking Monotonically Advancing Fronts Using Fast Marching Level Set Method, *Technical report*, University of Louisville.
- Garrido, S., L. Moreno and D. Blanco, 2009. Exploration of 2D and 3D Environments using Voronoi Transform and Fast Marching Method, *Journal of Intelligent and Robotic Systems*, 55, 55–80.
- Ferguson, D. and A. Stentz, 2006. Using Interpolation to Improve Path Planning: The Field D^* Algorithm, *Journal of Field Robotic Systems*, 23(2), 79–101.
- Hongyang, Y, H. Wang, Y. Chen, D. Dai, 2008. Path Planning Based on Constrained Delaunay Triangulation, *Proceedings of the 7th World Congress on Intelligent Control and Automation*, June 25 - 27, Chongqing, China.
- Kallmann, M., 2005. Path Planning in Triangulations, *Proceedings of the Workshop on Reasoning, Representation, and Learning in Computer Games, International Joint Conference on Artificial Intelligence (IJCAI)*, Edinburgh, Scotland, July 31, pp. 49-54.
- Koenig, S., M. Likhachev and D. Furcy, 2004. Lifelong Planning A^* , *Artificial Intelligence*, 155(1-2), 93–146.
- Sethian, J.A., 1999. Fast Marching Methods, *SIAM Review*, 41(2), 199–235.
- LaValle, S.M., 2006. *Planning Algorithms*, Cambridge University Press.
- G. Oriolo, A. Luca and M. Vandittelli, 2002. WMR Control Via Dynamic Feed-back Linearization: Design, Implementation, and Experimental Validation, *IEEE Transactions on Control Systems Technology*, 10(6), 835–852.
- M. Lepetic, G. Klančar, I. Skrjanc, D. Matko, B. Potocnik, 2003. Time optimal path planning considering acceleration limits, *Robotics and Autonomous Systems*, 45, 199–210.

AUTHOR BIOGRAPHY

GREGOR KLANČAR received B.Sc. and Ph.D. degrees in electrical engineering from the University of Ljubljana, Slovenia in 1999, and 2004, respectively. His research interests are in the area of fault diagnosis methods, multiple vehicle coordinated control and robot soccer related problems. His Web page can be found at <http://msc.fe.uni-lj.si/Staff.asp>.

GAŠPER MUŠIČ received B.Sc., M.Sc. and Ph.D. degrees in electrical engineering from the University of Ljubljana, Slovenia in 1992, 1995, and 1998, respectively. He is Associate Professor at the Faculty of Electrical Engineering, University of Ljubljana. His research interests are in discrete event and hybrid dynamical systems, supervisory control, planning, scheduling, and industrial production control. His Web page can be found at <http://msc.fe.uni-lj.si/Staff.asp>.

MODELING AND SIMULATION ARCHITECTURE FOR CLOUD COMPUTING AND INTERNET OF THINGS (IOT) BASED DISTRIBUTED CYBER-PHYSICAL SYSTEMS (DCPS)

Xie Lulu^(a), Wang Zhongjie^(b)

^(a) College of Electronics and Information Engineering, Tongji University, Shanghai, China, 201804

^(b) College of Electronics and Information Engineering, Tongji University, Shanghai, China, 201804

^(a) lilymaomao@gmail.com, ^(b) wang_zhongjie@tongji.edu.cn

ABSTRACT

Distributed Cyber-Physical Systems (DCPS) consists of many spatio-temporal heterogeneous CPS subsystems and components, making the modeling, management and control of resources complicated. In this paper, firstly, objectives and research backgrounds are described; opportunities and challenges in modeling and simulation of coordinated and efficient DCPS are discussed. Secondly, abstraction and deployment of networked DCPS is presented and analyzed: 1) an Internet of Things (IoT) framework for ubiquitous and self-managed environment of DCPS has been proposed. Which can bridging autonomous DCPS nodes with the future internet; 2) an optimized Cloud-based open and reusable modeling and simulation architecture for the management, scheduling and control of large-scaled dynamic and heterogeneous resources and services is designed. Thirdly, a multi-model based hierarchical architecture for the modeling and simulation of an open and reusable DCPS platform is then proposed, with an experimental technology framework towards cooperative and robust DCPS being constructed and discussed in the last.

Keywords: distributed cyber-physical systems, internet of things, cloud-based resource management, model based architecture

1. INTRODUCTION

Cyber-Physical System (CPS) integrates computing, communication and storage capabilities with monitoring and control of entities in the physical world dependably, safely, securely, efficiently and in real-time, (Cardenas, Amin and Sastry 2008). It connects the virtual information world with the physical world through the integration and interactions of Cyber and Physical components, Lee (2008). It requires close interaction between the cyber and physical worlds both in time and space, and the interaction among components are autonomous managed.

Some basic theories of CPS are derived from the integration of Distributed Real-time Embedded Systems; Wireless Sensor Networks; Networked Control Systems and Decision Support Systems, Bujorianu (2009).

Comparing with these technologies, CPS is much more complicated and with stricter system properties. CPS focuses on its real-time and resource-saved performances, being interactive, accurate, coordinate, intelligent, secure, robust, efficient and autonomous. Thus makes some existing technologies in traditional embedded systems; intelligent systems and hybrid systems can be optimized and employed into the modeling and simulation of DCPS.

Being an auto-controlled real-time intelligent system with excellent performances, CPS technologies are intended to improve the quality of human life, promoting a harmonious living environment. The idea of CPS can be widely applied to various research areas, such as: smart vehicles in the area of transportation; smart grids in the area of infrastructures and energy; various smart devices and robots in medical treatment, biology, industries, agricultures. Thus constructed a distributed heterogeneous DCPS Environment, consists of either same kinds or different kinds of CPS subsystems and resources, with physical entities being monitored, coordinated, controlled and integrated by the computing and communication core, and the coupling among system components being manifested from the nano-world to distributed wide area systems, at multiple spatio-temporal scales.

2. BACKGROUNDS

2.1. United States

Concepts of CPS are first proposed in USA, 2002. With attractive performances and gradually emerged importance, this technology has gained a lot of attention from both the governments and research institutions around the world since 2005. Many basic researches and trial applications for the single CPS have been aroused, mainly based on the architecture of embedded autonomous systems and hybrid systems. There are some novel research intentions and meaningful applications.

Campbell and Garnett (2006), proposed the ideas of CPS Environments and CPSs Sensor Grids, intending to sense the physical environment in different granularities, and with dynamic and various disturbance elements, for

the evaluation and simulation of systems' overall abilities and performances.

Edward (2009), combining the distributed embedded software with dynamic physical models. An efficient programmable temporal semantics abstraction based synergy language has been proposed, for hard-real time control of DCPS.

(Ilic and Xie 2008; Zhang and Ilic 2009), study the modeling and control methods of Distributed Smart Grids. Support Vector Machine (SVM) and Markov-State based control and prediction model have been proposed for the modeling and control of Smart Grids, with the optimized scheduling and dispatching among distributed electric power stations.

Thiagarajan and Ravindranath (2009), aim to providing a Trust-based intelligent vehicle navigation system: VTrack, which also is a typical application of DCPS. They have done a lot of work in the coordination and stability control between cars and between car and the traffic signals.

Correll and Bolger (2010), take part in the research of Distributed Robot Garden. With which a team of robots can take care of the tomato plants in the garden autonomously, through distributed sensing (each plant is equipped with a sensor node monitoring its status), navigation, manipulation; wireless networking and coordination.

Asynchronous mixed-signal modeling and verification methods and tools have been suggested by Thacker and Myers (2010), to study the battery-based DCPS, as for the optimized coordination, scheduling and simulation between DCPS nodes.

Chen and Ding (2010), use Grammatical reasoning models, combing with distributed Multi-Agents simulation algorithm and other basic models and symbols, to abstract and achieve synthesis and effective control and learning strategies between DCPS.

To encourage more researchers to take part in the researches related to DCPS, the research of self-interacted, coupled, collaborative and integrated platform of DCPS has been listed into the NSF's important research agenda in 2011. The platform is expected to realize the intelligent coordination and interaction among different DCPS components, and to improve the overall performance of DCPS,

2.2. Europe

European scholars focus primarily on the structure and theoretical foundations for the design, modeling, and implementation, performance and applications of DCPS. Intelligent modeling algorithms have been proposed for the control and optimization of DCPS, such as the ant colony, immune and hormonal algorithms integrated methods, Rammig (2008). International project called "RoboEarth" attempts to let the robots to share information with each other and to store and update their knowledge in a self-managed manners, (Zweigle, Andrea and Haussermann 2009). Besides, Europeans have done a lot of work in intelligent electronic systems; SCADA systems; integration of multiple components;

and modeling and control of complex systems, which are beneficial to the research of DCPS.

2.3. Japan, South Korea, Australia

In Japan and South Korea, DCPS started to get concerned around 2008, Easwaran and Insup (2008), applications and software frameworks of DCPS have been studied, such as the automated integration of embedded objects and computing equipments under hybrid communication networks. Modeling and control experiments of intelligent robots with CPS properties have been studied. Researchers in Australia have also launched many interesting researches in Smart Grids, and Smart Cars, Lyster (2010).

2.4. China

In China, the research of DCPS has been proposed and selected as one of the major development directions by the High-tech Research and Development Program of China and NSFC since 2009. There are some exploratory works.

Xia and Ma (2008), have made some progress in the QoS and real-time high performance control of medical DCPS, based on feedback control between medical CPS nodes.

Zhang (2010), try to design DCPS with networked clouds, high confidence middleware and information exchange technologies.

Zhao (2010), announced a cut based on geometric topology control algorithm for the energy balance in DCPS, using network topology control algorithm to improve the energy efficiency and the robustness of the system, with the optimizing of MAC layer protocol to achieve lower energy consumption, reduce transmission delay and optimize network performance.

Xiao and Yu (2010), proposed a series of control methods to improve the reliability of DCPS, using a Petri-nets based Case modeling.

(He 2010; Ma 2010), proposed a "sensing-control network" concept, for CPS, focuses on the theories research of a single CPS, involving mathematical modeling, analysis, verification methods and theory-based research of CPS, to address key issues encountered, such as real-time, cross-layer, composability, predictability, dynamic evolution.

2.5. Opportunities and challenges

Taking into account the existing studies, there are many meaningful researches in the modeling and control of DCPS applications. While the technologies for the overall management and control of DCPS are currently scarce; most of the existing models and algorithms are applied only to the fixed applications; most of them just focused on some parts of a specific CPS, lacking the analysis and modeling, control or scheduling strategies of the overall DCPS environment, with a large amount of physical equipments, computing devices and communication resources are inactivated most of the time, with low utilization rates.

As for the complexity of the environment and resource heterogeneity, the coordination and collaboration of the CPS components will greatly affect the real-time property and the overall performance of DCPS. Compared with the networked control technology, embedded technology and the internet of things, DCPS environment has better coordination and collaboration mechanisms, and is capable of achieving a much efficient and real-time performance. Its ultimate goal is to perceive the environment and resources accurately, monitor and coordinate different components, and make real-time decisions based on the feedback of DCPS performance to implement appropriate behavior and actions without any manual supervision.

There is a large amount of researchers like Cardenas and Sastry (2009), who have found that although there are already many useful researches on the communication and computing security of the networked DCPS, researches of robustness are somewhat limited. Especially that DCPS environment hold series of dynamic and heterogeneous coupled complex CPS subsystems and applications, it is difficult to discover the unexpected events, and it will be hard to guarantee a safe and stable environment for the DCPS. It is important to design and construct a stable and efficient architecture to have the abilities to avoid the cascading failures and malicious attacks, and be prepared even in an uncertain and unmanned environment.

There are already some useful abstractions and architectures, such as the multi-model based real-time architecture proposed by Lee (2008), and the feedback control based architecture by Xia and Ma (2008), the spatio-temporal event based architecture by Tan and Mehmet (2009), ect.

We plan to integrate the features of these architectures using the multi-model based layered management and control architecture.

There are several research problems that should be especially considered in the management and control of DCPS:

- How to deal with the huge information and resources in the large and space-time heterogeneous DCPS environment?
- How to build the self-adapted and intelligent learning or understanding of the optimized modeling methods and algorithms for the modeling agents and unified model interfaces or services under dynamic environment with various requirements and restrictions?
- How to improve the efficacy and robustness of the schedule and control methods assuring both real-time and low energy consumption?
- How to locate and schedule system resources intelligently and robustly to achieve a stable and sustained DCPS environment under restricted ability constraints?

- How to realize the ubiquitous hard real-time management and control of different DCPS applications over wired and wireless communication environments covering wide areas and different networks?
- How to detect and predict potential threats through the verification of the models and simulated results?
- How to quantize and estimate the variety of the environment, to reduce the errors among the lab simulation results and the actual control and action results under undertrained and dynamic environment?

We suggest an universal management and control architecture for the existed CPSs, with the abilities to support the building and testing of newly constructed CPS models and even new applications. That the related CPSs' knowledge, models and hardware resources among various application areas can be collected, shared, analyzed and reused, with indexes built according to different properties and events concerned. And all of the DCPS resources can be remotely located and managed, and can be customized dynamically integrated, optimized and updated for the possible future uses.

Architecture for DCPS has been proposed and designed in this paper. Based on the modeling and interaction of each technological layer, the architecture aims to construct a cooperative and robust modeling and simulation framework for the sustained management and schedule of DCPS components and resources based on the cloud computing and cloud simulation concept (Li, Chai and Hou 2009). Using Complex Network based topology abstraction and behavior prediction to support the research of stability and robustness of CPS under perturbations or uncertain environment, and consider the autonomous interaction and self-coordination among CPS components and (or) CPS subsystems through Multi-Agents and Petri-net, with the intelligently optimized simulation and verification results, system models can be updated and to be self-adapted, so that the efficiency of resource utility and system performance can be improved.

3. ABSTRACTION OF NETWORKED DCPS ENVIRONMENT

Based on the concept, structure and composition of DCPS, announced by Lee (2008), Rajhans (2009), Chun (2010). DCPS environment can be considered as a set of distributed decision support systems, which combine real-time embedded methods with the networked coordination and control technologies. Networked DCPS environment (as shown in "Figure 1") bridges and associates the cyber world of computing, communication, and control with the physical world, Vincenzo (2007). It contains both the physical and computational components, with interaction between physical layer and information layer under the network communication environment to make decisions.

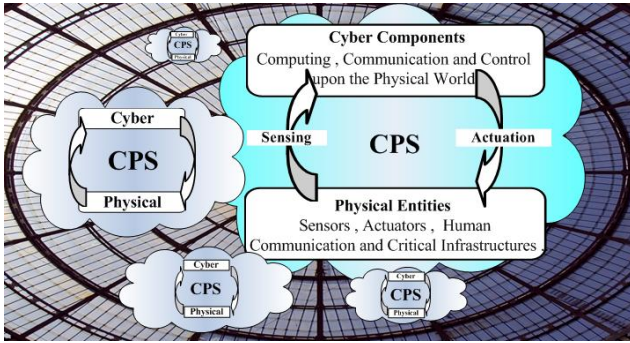


Figure 1: Abstraction of Networked DCPS Environment

DCPS environment is a self-managed ubiquitous networked environment; components in DCPS environment can be self-organized, self-adaptive, self-optimized and self-configurable, self-protected and self-healed to make an autonomous controlled DCPS environment. In the DCPS environment, one needs to consider not only the micro-level interaction and integration between the cyber and physical components, but also the macro-level interaction and collaboration between the different CPS subsystems.

Taking into account the complexity of the environment and resource heterogeneity, effective coordination and collaboration of the CPS components will promote the real-time response and even the overall performance of DCPS. Compared with the networked control technology, embedded technology and the internet of things, DCPS environment has better coordination and collaboration mechanisms, and is capable of achieving a much efficient and real-time performance. Its ultimate goals are to perceive the environment and resources accurately, monitor and coordinate different components, and make real-time decisions based on the feedback of DCPS performance to implement appropriate behavior and actions without any manual supervision.

4. RELATED WORKS

4.1. Internet of Things (IoT)

Internet of Things (IoT) consists of smart devices which are ubiquitous and will be constantly connected to the public Internet. It can be definite as “Things having identities and virtual personalities operating in smart spaces using intelligent interfaces to connect and communicate within social, environmental, and user contexts”. “Things” refers not only to the uniformed devices, but also the heterogeneous devices with different application areas, functions, editions and core technologies, (Dillon and Zhuge, 2011). All these devices can be included into a common community belongs to the same communication environment to identify each other with seamless integrations.

Comparing IoT with DCPS, there are some major differences:

1) DCPS is more complicated in modeling and control. Devices in IoT just include sensors, actuators,

enabled objects and RFID tags, living things are not included, and nodes in IoT aren't auto-controlled; they can't control each other; and without dynamical self-adaption.

2) Applications of DCPS require bi-directional, close-looped, real-time processing between the cyber world and the physical world with end-to-end QoS determinism and predictability, Dillon and Zhuge (2011). While IoT is not required to be hard real-time restricted.

3) Furthermore, existing solutions do not address the scalability requirements for a future IoT, they provide inappropriate models of governance and fundamentally neglect privacy and security in their design.

IoT aims at realizing the seamless integration of heterogeneous IoT technologies into a coherent architecture, which is strongly related to the development of the future internet environment required by the characters of DCPS. With the unified and standard communication protocols and frameworks of IoT, it will be easier to provide a unified CPS application development environment to support and promote a much more rapid and cost-effective CPS application development.

4.2. Cloud computing

Events and demands in DCPS are analyzed and described as semantic information and service oriented structures. The amount of the information is always huge, real-time allocation, management and control of spatio-temporal heterogeneous resources can't be realized by the traditional scheduling algorithms.

Cloud computing embraces cyber-infrastructure, and builds upon decades of research in virtualization, distributed computing, utility computing, and more recently networking, web and software services, Vouk (2008). It delivers infrastructure, platform, and software (applications) as services, which are made available as subscription-based services in a pay-as-you-go model to consumers. These services in industry are respectively referred to as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), Calheiros and Ranjian (2009). Five essential characteristics of Cloud computing are: broad network access; measured services; on-demand self-services; rapid elasticity; resource pooling,

Cloud computing implies a service-oriented architecture (SOA), reduced information technology overhead for the end-user, great flexibility, reduced total cost of ownership, on-demand services and many other things. Therefore, it can be used in the management and optimization of resources.

4.3. Combination of IoT and Cloud computing

There are many design challenges in the research of IoT: such as limited bandwidth; low memory; low transmission power; low bit rate; low computational power; low throughput; low lifetime and low utility ratio. Most of these problems result from constrains of

limited computation and storage abilities and resources of the IoT nodes. Therefore, we consider combining the IoT technology with Cloud computing to promote the utility of the IoT resources.

Because Cloud computing and IoT are both service-oriented dynamical technologies, IoT can be transferred into virtual resources, (Gyu and Crespi, 2010). Therefore, combination of Cloud computing and IoT is technically feasible. Besides, seamless integration of the ubiquitous IoT communication environment will make the information access in the clouds more convenient, cloud services can be visited by more people, accelerating the use ratio and the spread scale. Then useful services can be more attractive and valuable, making a higher profit; and useless services can be removed or updated, to improve the systems' performances.

5. INTERNET OF THINGS (IOT) AND CLOUD COMPUTING BASED DCPS ENVIRONMENT

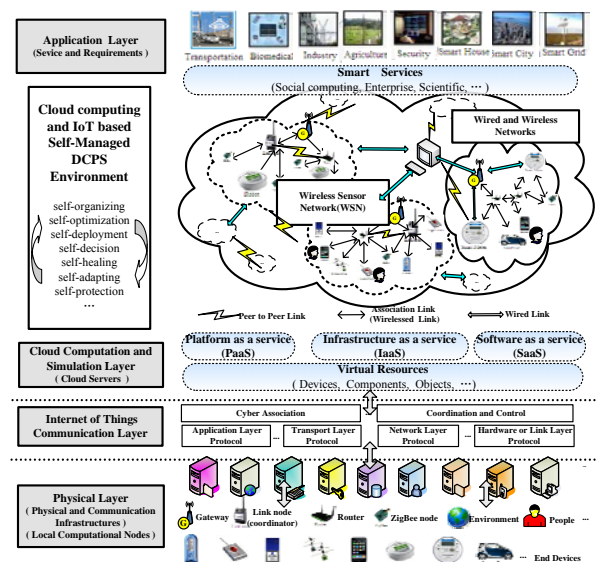


Figure 2: Structure of IoT and Cloud Computing Based Self-managed DCPS Environment.

With the rapid developments of Internet of Things (IoT), cloud computing and semantic services, all of the physical entities can be ubiquitously connected, and can be intelligent and remotely identified, located, tracked, monitored and managed under the future internet. In the DCPS world, entities can see and sense the environment, can do some not very complicated computing, and will be auto-controlled, even with the intelligent to thinking, and can be prepared and self-adapted for instant events.

The DCPS environment can be considered as a set of distributed decision support systems, which combine real-time embedded methods with the networked coordination and control technologies. Therefore, technologies of IoT combing with cloud computing, will construct a perfect ubiquitous computing and communication networked physical environment for the research and development of DCPS.

According to the characteristics and description of IoT and Cloud computing, a hierarchical structure of self-managed DCPS environment based on the IoT communication framework and Cloud Services is presented in “Figure 2”. Detailed deployment of this structure is discussed below.

5.1. Physical Layer

Physical layer usually contains a variety of different physical entities; these entities are cloud resources at the system level of IaaS. Resources and can be abstracted into different types of nodes: 1) Physical nodes: Simple physical nodes, such as sensors, actuators, sensor+actuators, infrastructures, and end devices. 2) Computation or communication nodes: Modules or methods with limited computing, communication or storage features. 3) Computation-Physical nodes: Software and hardware middlewares and cloud based services available for cyber-physical interactions. 4) Cyber-Physical nodes: Nodes are autonomously controlled and self-managed. They are intelligent nodes with both the ability of sensing and actuating, including humans. Because of the integration of information components and physical components, these nodes have a certain amount of computing, storage, and reasoning ability. Such as smart cars, smart meters, smart vehicles, robots etc. With different coupling strength and combination methods, these nodes can either be considered as a physical node with computing and decision-making capacity or be some computing nodes which can manage physical entities through communication and control.

These nodes are heterogeneous in their distributions. Being mobile and adaptive, they have the ability of perception, memory, reasoning and learning, and have characteristics of living things. All kinds of nodes and resources in DCPS environment can be abstracted; simulated into graphs and network topologies, and be simplified with semantic services supported models of events detecting and behavior perception. Therefore, technology of intelligent agents can be used to model these DCPS nodes.

5.2. Internet of Things Communication Layer

Communication layer consists mainly of network middleware, access equipments, standards, various communication protocols and routing algorithms. This layer communicates with the User-Level Middleware (SaaS) development in the Cloud programming over future internet, and being associated with the reasoning and calculation models in the computation layer, to realize the transmission of information and the management of the associated DCPS network nodes and ability restricted resources with spatio-temporal synchronization clocks.

Communication of DCPS can be either wired or wireless. For wired communication, communication layer may contain high-performance server farms and complex industrial equipments. For wireless communication, communication layer may involve a

large number of wireless sensor nodes with constraints of size, cost and energy power. Therefore, it is hard for traditional methods to compute network delay and packet loss rate, and even harder to detect and handle cascade failures or malicious attacks with a timely response. In this paper, a complex network analysis method has been proposed to solve the above problems, in which the DCPS stability is guaranteed through flow based route optimization and key nodes analysis.

Due to the diversity of communication, the locations of DCPS resources are distributed. And for the pervasive and flexible management or control of the DCPS environment, we proposed an open and pervasive cloud services based DCPS communication and control architecture, as shown in “Figure 3”.

5.3. Cloud Computation and Simulation Layer

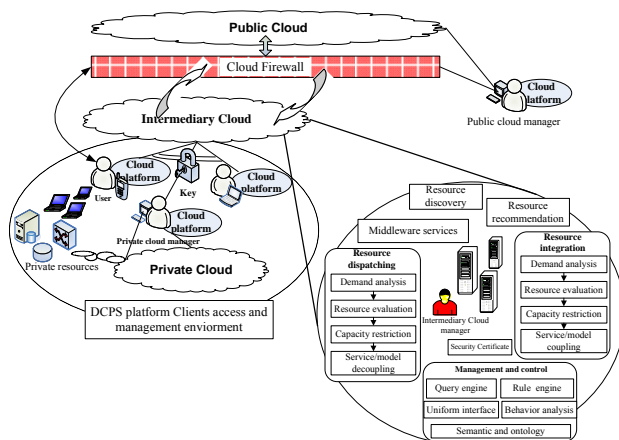


Figure 3: An Efficiency Optimized Cloud Computing and Simulation Architecture for DCPS Resources

Computation layer of DCPS environment contains clouds of virtual resources, such as virtual data server, virtual communications server, virtual log server, virtual high-performance computing servers, virtual firewalls and security equipments. Judged by the functions, there are storage clouds, computing clouds and simulation clouds.

Actual physical data and information of the devices and objects and people; are being virtualized by the Virtual Machine through its management and deployment. These components store historical data, related models and methods, and monitor the operation of the system in a real-time feedback control loop.

Associated with the IoT communication environment, this layer will also improve the autonomic of the systems/components, and is associated with all kinds of social and environmental information and services in the future dynamic networks. Interactions of these two methods at this layer take part in the PaaS stage, covering a series of core middleware services: QoS Negotiation, Admission Control, Pricing, SLA Management, Monitoring, Execution Management, Metering, Accounting, Billing, and Virtual Machine Management and Deployment.

With these services of adaptive resource storage and allocation; resources indexing and updating; virtual machine management and deployment, remote cloud-based computing, simulation and control may be realized.

This architecture has four layers, three basic layers are Private Clouds, Public Clouds, and Public Cloud Firewall; and the additional layer is a coordinator layer, it aims to optimize the computing and simulation efficiency of DCPS resources.

Private clouds contain all of the user's private resources; public clouds share and storage authorized and common resources; the cloud firewalls between the user terminals integrated safety standards for different systems and resources, make sure that the users can locate and use remote resources safely.

Huge amounts of heterogeneous dynamical requirements and services are difficult to proceed, and are always time consumed. To conquer this problem, we proposed an idea of “Intermediary Cloud”. Intermediary clouds are responsible for the discovering, computing, analyzing, managing and deployment of domain specific services, resources or protocols of both private clouds and public clouds. Advanced computing intelligence algorithms can be used to accelerate the speed of information searching and decision making. Social elements will be considered and evaluated, such as the price of the services; behavior of customer; policies differences; ability of suppliers. Intermediary clouds can be considered as the social coordinators of DCPS. With intelligent coordination and management strategies, accuracy and efficiency of the resources allocation can be improve; scheduling and dispatching can be more efficient; along with the safety and robustness can be considered, and the overall performances of DCPS are guaranteed.

5.4. Application Layer

Application layer interacted with user interfaces, and supplies all kind of smart services here. Standard cloud applications services including: Social computing, Enterprise, ISV, Scientific, CDNs, etc. Customers can design and order the services they are interested in.

Designing of this layer should consider two main aspects: 1) Construction of industrial equipments and embedded components will cost enormous human and financial resources, so it is an important issue to make an effective and efficient reuse of current software and hardware resources. 2) A real physical environment often covers a number of different applications, so the bottleneck of DCPS application is how to abstract a variety of network topologies from the existing CPS subsystems to satisfy different requirements, and to realize resource sharing and reuse among CPS subsystems.

Swarm algorithms and community discovery methods can be employed to analyze the above two problems. Based on the interaction mechanisms of multi-agents, coupling and decomposition methods can be used to improve the system performance. As a self-

managed DCPS environment should have the abilities of autonomy, coordination, real-time feedback, low energy consumption, and with high-performances. There are still many factors required be considered, such as heterogeneous composition of the DCPS components, uneven distribution of resources, dynamic or uncertain behavior of the components and systems, and environmental complexities.

6. MODELING AND SIMULATION ARCHITECTURE FOR AN OPEN AND REUSABLE DCPS PLATFORM

As discussed above, it is necessary to build a generalized DCPS architecture that can support cooperation and coordination among heterogeneous CPS components and resources. Based on the structure of IoT and Cloud computing based DCPS computing, communication, and control environment, considering the key research problems in the research of DCPS, a multi-model based hierarchical architecture for the modeling and simulation of an open and reusable DCPS platform has been proposed and discussed in “Figure 4”.

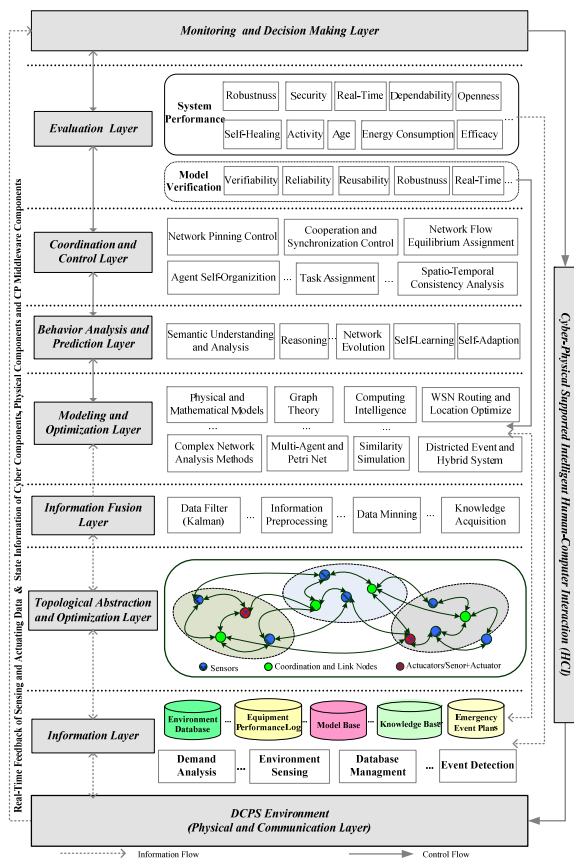


Figure 4: Multi-model Based Modeling and Simulation Architecture for An Open and Reusable DCPS Platform

Information layer consists of many databases, model bases, log bases and plan bases. This layer acquires, pre-processes and storage all kind of resource data and information sensed and collected from the DCPS environment. With all of these databases, there are huge information pools of different resources.

Topological abstraction and optimization layer is used to abstract the CPS into a topological network. For this procedure, complex network analysis expert experiences will be employed. Graph theory can be used to optimize the topological network.

Information fusion layer focuses on spatio-temporal information analysis and data mining. With demand analysis, environment perception and detection of the events, this layer achieves pre-process of information flow between different levels. Kalman Filter can be used to remove the redundant data. The information similarity extraction method based on semantic or some other computational intelligence models are used to process the data, and translate them into understandable knowledge.

Modeling and optimization layer can be used to discover the events; to establish the models required in each layers; and to build services. For the openness and reusability of the platform, modeling should be smart and efficient. Models, methods and components in different layers should be associated and optimized. Model bases for DCPS should be constructed, and may include physical and mathematical models, graphs, computation intelligence and many models of hybrid systems. As for social elements related problems, models of complex networks, context-aware cognitive, agents and similarity simulation methods, wireless sensing network routing and relocating methods, may all be helpful. Optimization is based both on the modeling, simulation and evaluation.

Behavior analysis and prediction layer aims at realizing a self-learning and self-adaptive intelligent DCPS environment. This layer is novel and special modular, considering for the dynamic and uncertain social elements in DCPS. The behaviors of CPS nodes and network evolution are predicted. Execution of these events will be monitored, recorded and analyzed. Analysis methods include semantic understanding methods, reasoning methods, network evolution analysis and key nodes discovery methods and flow equilibrium methods.

Coordination and control layer receives behaviors of the CPS nodes from the behavior analysis layer, and makes a decentralized control to coordinate different nodes. With this layer, CPS is able to achieve better utilization of resource and performances. The coordination and control methods may include spatio-temporal synchronization, intermittent feedback, intelligent perception, perturbation control, pinning control and adaptive control, which will contribute to the rational and efficient management, schedule, dispatching and utilization of limited resources, fits the research of cloud services, internet of things and green computing.

Evaluation layer mainly involves in system performance analysis and model verification. CPS performance includes efficiency, surviving period, accuracy (prediction, classification, sorting, etc.), correlation, security, robustness and vulnerability.

Model verification refers to verifiability, reliability, security, robustness, and reusability etc.

Monitoring and decision making layer consists of both centralized control and decentralized control. Centralized control mode is mainly for the cooperation and coordination of the nodes and resources. Decentralized control mode mean that every node is autonomous controlled, and so centre node caused sudden failures and cascading failures can be avoided, which is especially useful for emergencies. Therefore, elements such as the energy consumption and life-cycle of DCPS nodes; computation complexity or task importance; economic profit of services; government policies and other factors should be considered to gain a balance or creating a better control mode.

7. TECHNOLOGY FRAMEWORK

This technology framework in “Figure 5” is an experimental architecture particularly designed for a DCPS taking “cooperative” and “robust” as the most important performances. This design is based on the modeling and simulation architecture of the unified DCPS platform. Methods of complex network and multi-agents are used to build and analyze the DCPS’s topology network. Topology abstraction of DCPS is constructed with three main elements: Nodes, Links (relationship between nodes), and Flow (information streams, Cheng and Wang (2010), tasks, and event flow through the nodes and edges).

Description of the technique steps are as follows:

(1) The environment state is firstly perceived. After the state noise being removed by the data fusion methods such as the Kalman filter, the data will be normalized. Then the data ontology and semantic presentations can be built.

(2) For different CPS applications, analyze and quantify the system requirements, detect the possible events, acquire and store the knowledge for the control and modeling.

(3) According to the quantified requirements and complex network construction rules, the similarity analysis method can be used to define the nodes and links, and abstract the DCPS environment into a topological space.

(4) In order to increase the robustness of DCPS, the “Hub” node in the topology network, Ulieru (2007), will be detected and analyzed with the statistic characteristics of complex network centrality, such as degree distribution, power-law, closeness, betweenness centrality, random walk betweenness centrality.

(5) Analyze the components’ clusters and the community composition in DCPS environment; reconstruct the CPS subsystems if necessary.

(6) With dynamical collaboration mechanism and flow evolution control methods, analyze and control the network flow and make a macro-view over the whole DCPS environment.

(7) Analyze and control the link flow and make a micro-view over the resource distribution and the node utilization. Several methods can be combined or

decoupled to achieve network and flow equilibrium assignment, for example, the equilibrium methods, the ordinary differential equation (ODE) models, recursive algorithms, and the complex network coordination and synchronization control methods.

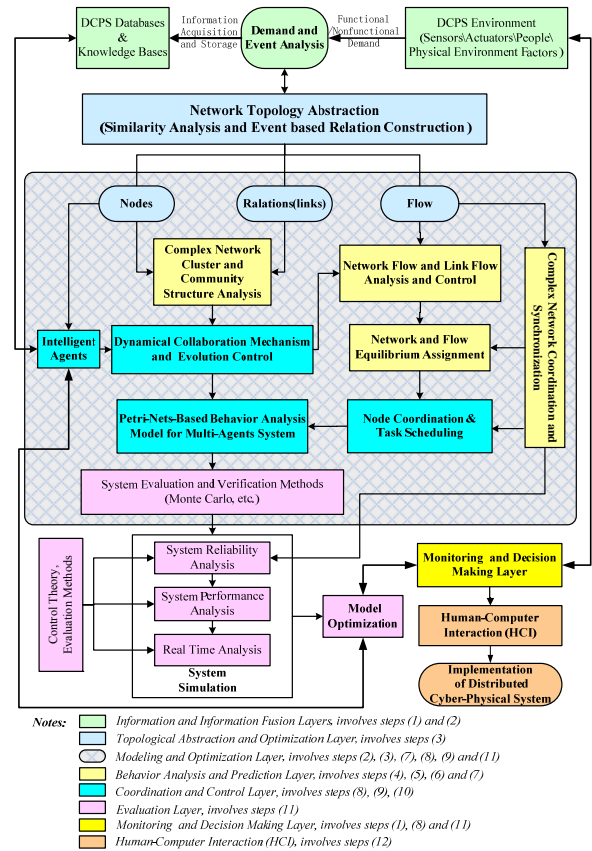


Figure 5: An Experimental Technology Framework

(8) Design and modeling the agents for DCPS with different functions and attributes. The Agents will be autonomous controlled, and can be designed based on a BDI (Belief, Desire, Intention) framework, capable of storing, reasoning, computing and communicating, and should have open and reusable services, middlewares and device interfaces.

(9) Build a Petri-nets-based behavior analysis model for multi-agents system by combining and extending Petri-nets. Analyze the cooperation and interaction mechanism between different agent-groups.

(10) Try to combine the Petri-nets model with the complex network coordination and synchronization control methods to achieve the coordination of the nodes and the efficient scheduling of the tasks.

(11) With the system verification methods, test and analyze the system stability, performance and the operation time. Results can be used to optimize the related models.

(12) Construct a networked DCPS supported Intelligent Human-Computer Interaction (HCI), to test the DCPS platform, and to design customizable services.

8. FUTURE WORK

Smart Grids is a typical application of DCPS. The proposed architecture will be deployed into the analysis, management, schedule and optimize of the distributed smart power grids in China. The accordingly modeling and simulation architectures and the open management and control platform will be designed, verified and performed. The task is to optimize the deployment of electric power in the advanced distributed power grids. With the consuming and saving behavior of the electric power be tracked, modeled, simulated and predicted; and the relationships between the distributed energy generation equipments, transmission controllers and stations are achieved, coordinated and optimized.

9. CONCLUSIONS

Cyber-Physical Systems (CPS) is an emerging technology with excellent performances, which can be applied into many different areas, thus constructed a large-scale heterogeneous environment of Distributed Cyber-Physical Systems (DCPS). This paper aims to manage and control DCPS resources efficiently and smartly. An Internet of Things (IoT) and Cloud computing based DCPS management and deployment structure is designed and proposed to improve the performances in computing and communication. And a multi-model based hierarchical modeling and simulation architecture for an open and reusable DCPS management and control platform over the IoT and Cloud-based self-managed environment is then presented, with an experimental technology framework towards a cooperative and robust DCPS being constructed and described. Future researches will focus on the simulation, realization and optimization of this modeling and simulation architecture in the application area of Smart Grids under the future internet.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grant No. 71071116), the Shanghai Key Program for Basic Research of China (Grant No. 10JC1415300), and the National High-Tech. R&D Program of China.

REFERENCES

Bujorianu, M.C., and Bujorianu, M.L., 2009. A Unifying Specification Logic for Cyber-Physical Systems. *Proceedings of the 17th Mediterranean Conference on Control and Automation*, pp 1166-1171. June 24-26, Tessaaloniki, Greece.

Calheiros, R.N., Ranjan, R. and others., 2009. CloudSim: A Novel Framework for Modeling and Simulation of Cloud Computing Infrastructures and Services, *Echnical Report, GRIDS-TR-2009-1, Grid Computing and Distributed Systems Laboratory*, 9 pages. March 13, The University of Melbourne, Australia.

Campbell, R.H., Garnett, G. and McGrath, R.E., CPS Environments. *NSF Workshop on Cyber-Physical Systems*, October 16-17 Austin, TX.

Cardenas, A.A., Amin, S., and Sastry, S., 2008. Secure Control: Towards Survivable Cyber-Physical Systems. *Proceedings of ICDCS Workshops*, pp. 495-500. June 17-20, Beijing, China.

Cardenas, A., Sastry, S. and others., 2009. Challenges for Securing Cyber Physical Systems. *Workshop on Future Directions in Cyber-physical Systems Security*, July 23, DHS.

Carney, D., Cetintemel, U. and others., 2002. Monitoring streams: a new class of data management applications. *Proceedings of the 28th international conference on Very Large Data Bases*, pp 215-226, August 20-23, China.

Chen, Y., Ding, X.C., Stefanescu, A. and Belta, C., 2010. A Formal Approach to Deployment of Robotic Teams in an Urban-Like Environment. *Proceedings of 10th International Symposium on DARS*, pp 14 . Nov 1-3, Lausanne, Switzerland.

Cheng, L., Wang, Z.J., 2010. A Stream-based Communication Framework for Network Control System, *Proceedings of the CISP2010-BMEI2010*, Vol 1, pp 2828-2833, Oct 16-18, Yantai.China.

Correll, N., Bolger, A., Bollini, M. and others., 2010. Building a Distributed Robot Garden. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp 219-232 . Oct 11-15, St. Louis.USA.

Dillon, T., Zhuge, H., Wu, C. and Singh, J., 2011. Web of things framework for cyber-physical systems. *Concurrency and Computation: Practice and Experience*, 23 (9) : 905–923.

Easwaran, A., Insup, L., 2008. Compositional schedulability analysis for cyber-physical systems. *SIGBED Review*, 5 (1): 11 -12.

Gyu, L., Crespi, N., 2010. Shaping Future Service Environments with the Cloud and Internet of Things: Networking Challenges and Service Evolution, *Leveraging Applications of Formal Methods, Verification, and Validation, Lecture Notes in Computer Science*, 6415: 399-410, Heidelberg: Springer Berlin.

He, J.F., 2010. Cyber-physical Systems. *China Computer Federation Communication*, 6 (1): 25-29. (in Chinese version)

Ilic, M.D., Xie, L. and Khan, A.U., 2008. Modeling Future Cyber-Physical Energy Systems. *Proceedings of IEEE Power Engineering Society General Meeting*, pp 1 - 9, July 20-24, Pittsburgh.

Ingeol, C., Jeongmin, P. and others., 2010. Autonomic Computing Technologies for Cyber-Physical Systems. *Proceedings of International Conference on Advanced Communication Technology, ICACT*, pp1009-1014. February 7-10, South Korea.

Lee, E.A., 2008. Cyber physical systems: Design challenges. *Proceedings of 11th IEEE Symposium on Object Component Service-Oriented Real-Time*

- Distributed Computing*, pp363-369. May 5-7, Orlando.
- Li, B., Chai, X. and Hou, B., 2009. Networked Modeling & Simulation Platform Based on Concept of Cloud Computing—Cloud Simulation Platform. *Journal of System Simulation*, 21 (17): 5292-5299.
- Liberatore, V., 2007. Networked Cyber-Physical Systems: An Introduction. *Proceedings of 14th Annual IEEE International Conference and Workshops on the Engineering of Computer-Based Systems*, pp 1-2. January 13, Tucson.
- Lyster, R., 2010. Smart Grids: Opportunities for Climate Change Mitigation and Adaptation. *Monash University Law Review*, 36 (1):137-191
- Ma, W.F., 2010. CPS: a sense of control from sensor networks to the network. *China Computer Newspaper*, 7: 25-27. (in Chinese version).
- Rajhans, A., Cheng, S.W. and others., 2009. An Architectural Approach to the Design and Analysis of Cyber-Physical Systems.
- Rammig, F.J., 2008. Cyber Biosphere for Future Embedded Systems. *Proceedings of the 6th International Workshop on Software Technologies for Embedded and Ubiquitous Systems*, Vol 5287, pp.245-255. Oct 1-3, Anacardi. Italy.
- Tan, Y., Vuran, C. M., and others., 2009. Spatio-Temporal Event Model for Cyber-Physical Systems. *Proceedings of 29th IEEE International Conference on Distributed Computing Systems Workshops*, pp 44-50. June 22-26, Montreal, Quebec, Canada.
- Thacker, R.A., Myers, C.J. and Zheng, H., 2010. Automatic abstraction for verification of cyber-physical systems. *Proceedings of the 1st ACM/IEEE International Conference on Cyber-Physical Systems ACM*, pp 12-21. April 13-15, New York, USA.
- Thiagarajan, A., Ravindranath, L., LaCurts, K. and others., 2009. VTrack: Accurate, Energy-Aware Road Traffic Delay Estimation Using Mobile Phones. *Proceedings of 14th ACM SenSys*, pp 85-98. November 30, Berkeley, CA.
- Ulieru, M., 2007. e-networks in an increasingly volatile world: Design for resilience of networked critical infrastructures, *Proceedings of The Inaugural IEEE International Conference on Digital Ecosystems and Technologies - IEEE-DEST*, pp 540-545. Feb 21-23, Cairns, Australia.
- Vouk, M.A., 2008. Cloud computing — Issues, research and implementations. *Journal of Computing and Information Technology, Cloud Computing – Issues, Research and Implementations. – CIT*, 16 (4):235–246.
- Xia, F., Ma, L.H. and others., 2008. Network QoS Management in Cyber-Physical Systems. *Proceedings of International Conference on Embedded Software and Systems*, pp 302-307. July 29-31, Chengdu.
- Xiao, X., Yu, H.Q. and Fan, G.S., 2010. A Petri Net-Based Approach to Aspect-Oriented Use Case Modeling. *East China University of Science and Technology: Natural Science*, 35 (2): 248-254. (in Chinese version)
- Zhang, L., Leung, Henry., and others. 2008. Information Fusion Based Smart Home Control System and Its Application, *IEEE Transactions on Consumer Electronics* 54:3, 1157-1165.
- Zhang, Y., Ilic, M.D., Tonguz, O.F., 2007. Application of support vector machine classification to enhanced protection relay logic in electric power grids. *Proceedings of Large Engineering Systems Conference on Power Engineering*, pp.30-37. October 10-12, Montreal. Canada.
- Zhang, H.G., Yan, F. and others., 2010. Research on theory and key technology of trusted computing platform security testing and evaluation. *SCIENCE CHINA Information Sciences*, 53 (3): 434-453.
- Zhao, W., 2010. *WSN information in the physical system topology control and MAC Protocol*. Thesis (PhD). Central South University. (in Chinese version)
- Zweigle, O., Andrea, R., Haussermann, K., 2009. RoboEarth: connecting robots worldwide. *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*, pp184-191, Nov.24-26, Seoul.

AUTHORS BIOGRAPHY

WANG Zhong-Jie is a Professor and Ph.D. supervisor in the Department of Control Science and Engineering, College of Electronics and Information Engineering, Tongji University, Shanghai, China. Her research interests focus on the modeling, simulation and scheduling of complex environments and hybrid systems, including Dynamic Programming and Optimal Control; Cyber-Physical Systems; Cloud Manufacture; Smart Grids and Social Networks.

XIE Lu-Lu is a Ph. D. candidate in the Department of Control Science and Engineering, Tongji University, Shanghai, China. Her research interests are in the field of intelligent modeling, control and optimization techniques of Distributed Cyber-Physical Systems.

TRANSPORT NETWORK OPTIMIZATION: SELF-ORGANIZATION BY GENETIC PROGRAMMING

J. Göbel^(a), A.E. Krzesinski^(b), B. Page^(a)

^(a) Department of Informatics, University of Hamburg, 22527 Hamburg, Germany

^(b) Department of Mathematical Sciences, University of Stellenbosch, 7600 Stellenbosch, South Africa

^(a) goebel.page@informatik.uni-hamburg.de, ^(b) ae1@cs.sun.ac.za

ABSTRACT

This goal of the paper is transport network optimization. *Transport networks* are defined as network topologies where entities are forwarded from node to node constrained by capacity restrictions both on nodes and links. Examples include urban traffic (vehicles/signalized intersections) and IP networks (packets/routers). Optimization of such networks particularly has to provide the logic the nodes use to determine which entity to process next. Such logic can be imposed by a central authority based on global knowledge of the network state. In contrast, a *self-organizing* network solely relies on local decision rules to prioritize entities. At the cost of a potential loss in performance, such a decentralized network control is scalable and robust. This paper proposes *genetic programming* to evolve local node rules. Results indicate that the performance is similar to centrally (near-optimally) controlled systems.

Keywords: transport network optimization, genetic programming, discrete event simulation, simulation framework, Java

1. INTRODUCTION

The target of the work is the optimization of a general class of networks, namely *transport networks*. These are defined as graph topologies formed by nodes and links; continuously, entities “appear” at originating nodes (all nodes or only a subset may qualify as originating nodes). Such entities are queued for being “processed” by their origin nodes. After processing, the entities traverse links, thus queuing for processing at the next nodes on their routes until eventually reaching their destination nodes. Examples of such networks include urban traffic (vehicles advancing from one intersection to the next), conveyor-based manufacturing systems (items processed successively by different workstations) and telecommunication networks (e.g. IP packets being forwarded from one router to the next).

Optimizing such a transport network typically may involve minimizing waiting or travel times or maximizing throughput. Apart from discarding entities or adjusting their routes (which may or may not be feasible, depending on the network type) and long-term

improvements to the network topology itself (e.g. increasing nodes’ capacities, establishing additional links), the only degree of freedom for achieving such targets is the logic the nodes use to determine which entity (e.g. vehicle, item, IP packet) to process next.

This logic for entity prioritization can be set up by a central authority, which is typically provided with global knowledge of the network state. For example, based on global (estimated) traffic density data, signals can be coordinated such that platoons of vehicles are able to traverse the network without stopping (“green waves”). However, such attempts to centrally optimize such networks typically imply exponential computational complexity (Holland 1995), yielding bad scaling behaviour. Further assuming the requirement to adaptively adjust to dynamic changes in traffic patterns, such approaches depend on the availability of the central server and the communication to this authority.

This motivates applying decentralized optimization: Without being dependent on a central authority, each node (router, workstation, traffic light) independently decides which entity is processed next. This decision is based on information that is locally available (e.g. queue lengths, local flow estimations) only, enabling nodes to act autonomously if assuming means of obtaining these data (e.g. induction loops and cameras and image processing capabilities at an urban intersection).

Literature – e.g. in Bazzan 2005, Cools 2007, Gershenson 2005, Helbing 2008, Lämmer 2007 attempting decentralized urban traffic optimization – already provides local node control logic performing almost as well as or better than centrally controlled systems for some special network topologies, e.g. for Manhattan grids with one-way traffic (i.e. north to south and west to east traffic only, Gershenson 2005) or intersection inflows from different directions assumed to be mutually exclusive (Lämmer 2007).

The remainder of this paper is organized as follows: Section 2 provides further details about decentralized transport network optimization. Section 3 proposes genetic programming (GP) as potential solution, evaluated in Section 4 by experiments in a simulation environment. The paper concludes with a summary and outlook about further work in Section 5.

2. DECENTRALIZED TRANSPORT NETWORK OPTIMIZATION

Transport network optimization can be conducted by a central authority to which all relevant information is made available; examples include Diakaki 2003, Pohlmann 2010 or commercial systems like SCOOT (see e.g. United Kingdom Department for Transport 1995). However, apart from the dependence on the communication of each node to this central authority, the run-time performance of such approaches scales badly with the network size, compare Section 1. Furthermore, as optimization is typically conducted in cycles of 15-60 minutes, reaction to patterns of traffic shifting or the failure of an adjacent node is delayed.

This motivates applying decentralized optimization, analogously to the concept of *self-organization* in thermodynamic and other natural sciences: A *self-organizing* system autonomously acquires and maintains order despite external influence subject to perturbations, which is typically achieved by a set of microscopic (local) decision rules independently used by each component (Wolf 2005). However, the development of such rules if not known in advance is difficult; designing a self-organizing system can be interpreted as reverse-engineering such rules from the desired macroscopic behaviour of the system: This behaviour (e.g. efficient transportation) is an emergent property of such rules, for which research thus far does not offer an agreed-upon and universally applicable means of obtaining (Zambonelli 2004).

Nonetheless, the development of such local rules facilitating decentralized optimization of a transport network can be approached from the input side, i.e. the information locally available at the nodes: From Bazzan 2005, Cools 2007, Gershenson 2005, Helbing 2008, Lämmer 2007 and other approaches to solve special cases of the network optimization problem, a set of criteria can be obtained, upon which the decision as to which entities to prioritize at a given instant is based.

Using from now on urban traffic terminology for example, these criteria apply either to a specific lane (a queue for vehicles arriving at a node on a certain link, waiting for processing and departure on one or more other links), to an intersection as whole (e.g. maximum queue length, total estimated arrival rate), or even to the overall network (e.g. switching penalty, during which all lanes are “red” for safety reasons). See Table 1 for further examples for such criteria.

Lämmer 2007 has also shown that any node logic can be expressed as function which he refers to as the *priority index*: Based on a subset of these criteria, one can determine the priority index of each lane; serve the vehicles from the lane with the highest priority, unless network stability would be violated if a lane did not receive any service for some maximum waiting period. Table 2 shows an extension of this mechanism using a set of lanes instead of single lanes, taking into account intersections serving more than one lane simultaneously (e.g. opposing traffic from north and south proceeding straight ahead).

Flow (lane)
Queue length (F, LQL)
Longest waiting time (F, LWT)
Est. arrival rate overall (F, LAO)
Est. arrival rate current period (F, LAP)
Current phase since (F, LPS)
No green since (F, LGS)
Utilization of incoming link (F, LIU)
Maximum utilization of all outgoing links (F, LOU)
Est. arrivals next period (F, LEA)
Est. duration until next platoon arrival (F, LPA)
Est. feasible flow/absolute (F, LFA)
Est. feasible flow/relative (F, LFR)
Node (intersection)
Max. queue length (F, NQL)
Longest waiting time (F, NWT)
Est. arrival rate/overall (F, NAO)
Est. arrival rate/current period (F, NAP)
Duration of current phase (F, NPL)
Idle (B, NID)
blocked (B, NBL)
Global (network)
Average acceleration (F, GAA)
Switching penalty (F, GSP)
Passing possible in queues (B, GPP)

Table 1: Examples of criteria for entity (vehicle) prioritization in urban traffic; the suffix states the data types, either floating point numbers (F) or Boolean (B), and an abbreviation.

Lane set selection can for example be conducted on *greedy* basis: The lane with the highest priority is set to green. Repeatedly, from all lanes not mutually exclusive to a lane already set to green, the lane with the highest priority is set to green until no further lane exists that is neither set to green nor mutually exclusive to a lane already set to green.

Determine priority index for each lane
Unblock set of lanes with highest accumulated priority
Repeatedly re-evaluate priority indices and switch to a different set of lane once their accumulated priority exceeds the priority of the current flows by a positive Δ
Ensure minimum green and maximum waiting periods are not violated

Table 2: Link choice based on priority indices, extended from Lämmer 2007

With the *input* (the above-mentioned criteria) and the *result* (a priority index function) being specified, the open problem is determining how to obtain the latter from the former. For special network topologies, this

problem is already solved (compare literature quoted above), but not for the general case.

Part of a more general solution might be setting local rules such that certain desirable patterns – like green waves – are facilitated, see Göbel 2009. However, for network configurations, in which no such desirable patterns are known or where patterns available do not suffice for fully specifying a priority index function, we propose a different approach based on *genetic programming* (GP) in Section 3.

3. NODE LOGIC EVOLUTION BASED ON GENETIC PROGRAMMING

Our target is to obtain priority index functions for decentralized transport network control based on the input criteria from Table 1. In the absence of any restrictions of which of these criteria to use and how to algebraically and logically determine a priority index (PI) from their respective values, genetic programming, see e.g. Koza 1992 or Poli 2008, has been chosen because is a flexible and robust search technique not relying on any preconditions which would limit its applicability to special cases of transport networks: Mimicking nature, programs to determine priority indices are evolved by successively improving existing programs. Note that GP can be interpreted as a generalization of genetic algorithms (GA): While the search space of a typical GA is a static chromosome on which parameter values to optimize are stored, this structure the chromosome in GP itself is subject to evolution.

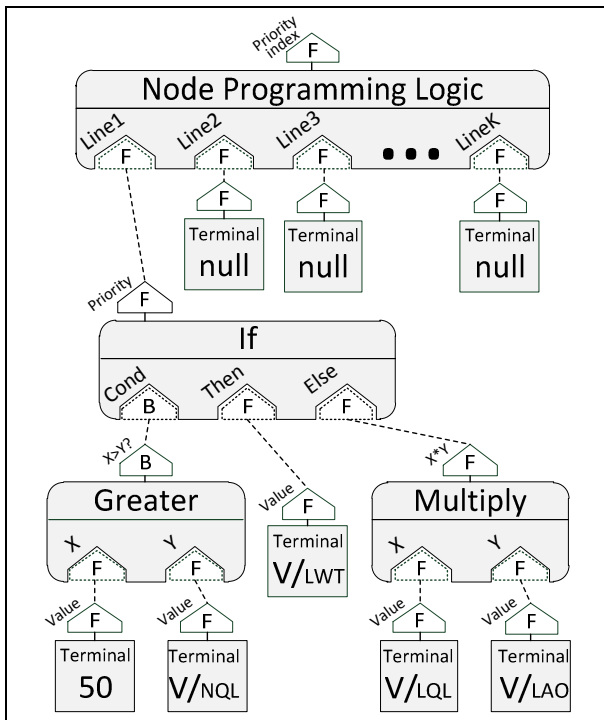


Figure 1: Example of a GP node priority index function

To facilitate the evolution of genetic programs of arbitrary complexity, genetic programs typically have a tree-like structure consisting of nodes (“building

blocks”) representing functions that can be flexibly combined, with the only restriction that arguments and result data types of adjacent node functions have to match.

An example priority index function is shown in Figure 1, composed of functions like for instance “Multiply” (bottom right), representing a function accepting two floating point numbers (F) as arguments and returning the product as floating point number as result. The overall result of the GP program is the return type of the top-level function (“Node programming logic”), which itself requires arguments determined by functions on the second level and so forth. Note that the closure of genetic programs assumes the availability of functions not requiring any argument, so-called terminals, e.g. numerical or Boolean constants or values read from variables (notation V/... to distinguish from constant terminals), like the node’s maximum queue length (V/NQL), the longest waiting time of an entity at a lane (V/LWT), a lane’s queue length (V/LQL) or a lane’s estimated overall arrival rate (V/LAO), compare Table 1.

A typical genetic programming cycle is summarized by Figure 2: Based on an initial population of programs generated at random, so-called genetic operators are applied to simulate biological evolution: The fitter (i.e. better network control performance) a program, the higher its probability of being selected for offspring composition by means of recombination; Figure 3 shows examples of recombination operators, particularly chromosome transfer ❶, aggregation ❷, projection ❸, and swapping ❹.

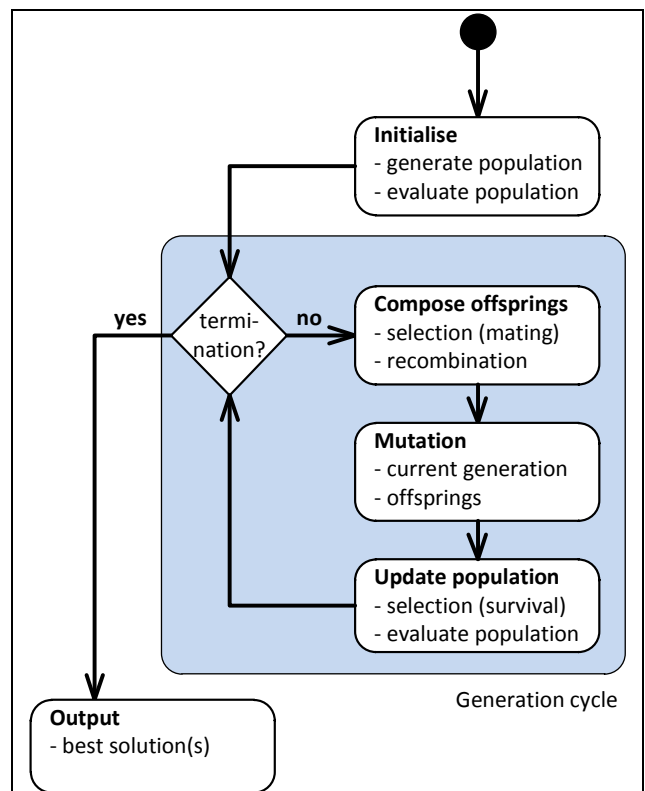


Figure 2: GP evolution, based on Koza 1992

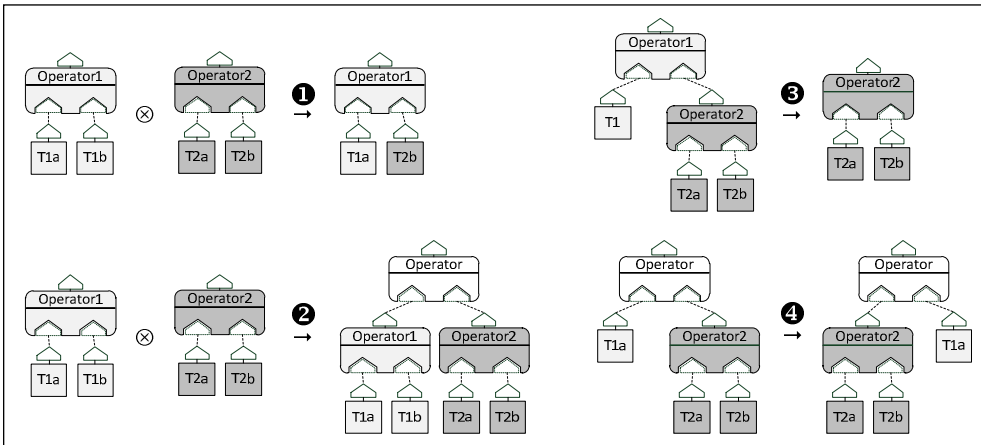


Figure 3: GP recombination operators

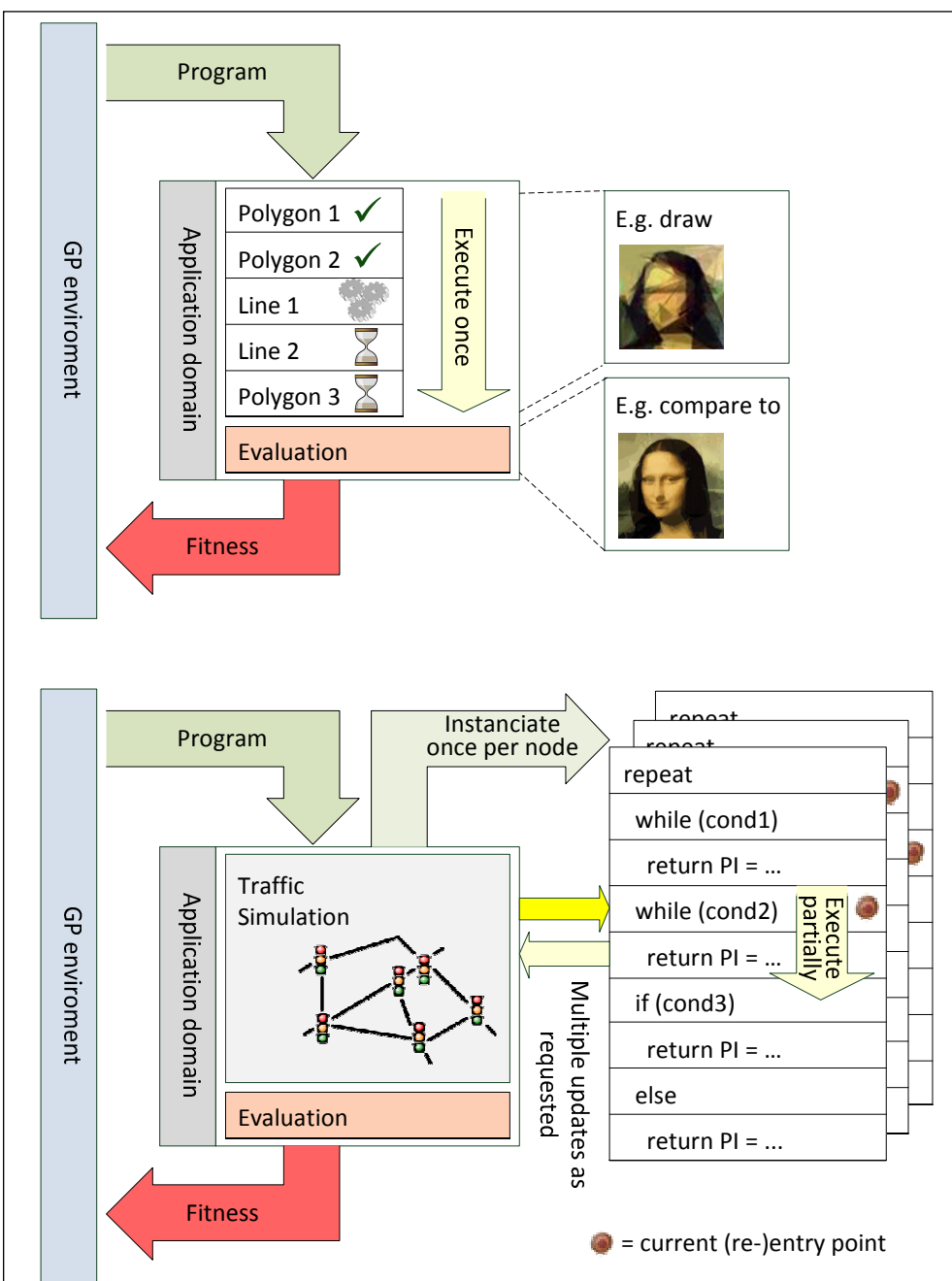


Figure 4: Alternative means of evaluating the fitness of a program

Both existing programs and new offsprings have a small chance of undergoing a random mutation, e.g. functions replaced by other functions with the same input and result parameter types, including terminals potentially being replaced by other terminals of the same type. For the resulting set of programs, a fixed number is selected to “survive” and form the next generation.

The set of functions typically used to GP-based mathematical functions includes numerical (plus, minus, multiply, divide, power, root, exp, log, abs) and logical operators (and, or, xor, not, greater, smaller) as well as statements to define piecewise functions (if/else, switch).

Inversion of control and state-dependent priority indices

The standard GP approach of determining the fitness of a program is shown in Figure 4, upper half: The full program – e.g. serving the purpose of creating a Mona Lisa forgery by drawing lines and filled polygons on a canvas (courtesy of Meffert 2011) – is executed once, followed by evaluating the fitness by comparison to the “real” Mona Lisa.

For two reasons, this paradigm is not appropriate for determining node priority index (PI) functions:

First, the evaluation of a PI function is driven by a simulation experiment calling the PI function whenever desired.

Secondly, local PI functions as proposed by Gershenson 2005 or Lämmer 2007 are dependent on previous calls to the PI functions since they use partial conditionals (e.g. “if” without “else”, leaving PI calculation to the next line if the condition does not apply) and loops. Allowing such program components potentially yields programs evaluated only partially, compare Figure 4, bottom half: An equation determining the PI inside the “while”-function for instance may be evaluated repeatedly until the “while”-condition is no longer fulfilled.

Our implementation, based on JGAP (Java Genetic Algorithms Package, see Meffert 2011) reflects this control flow. Particularly, we propose a special function referred to as “Node Programming Logic” (see also Figure 1), which keeps a reference to the current (re-)entry point, i.e. the “line” to use at the moment to determine the priority of a lane, thus emulating program execution. Note that multiple nodes in a network require multiple program instances with a different (re-)entry point token each to reflect nodes potentially being in different states. The return type of “Node Programming Logic” is a PI floating point value, which allows for hierarchically nesting partial conditionals and loops using multiple “Node Programming Logic” functions. After all lines have been used for determining a set of lanes to set to green, the program execution resumes at the first line.

Wrapping up, the program from Figure 1 will assign the highest priority to the lane of the vehicle with longest waiting time, which yields a FIFO (First in, first out) service, see the “then” branch of the “if” clause, as long as congestion is moderate (less than 50 entities queued in total). Otherwise, lane priority is the product of queue length and arrival rate: As congestion increases, the function tends to prefer main roads and to serve multiple vehicles before switching to other links.

Section 4 will investigate the performance of this GP approach of decentralized network optimization for three example networks (one intersection, two intersections, a small city area).

4. EVALUATION

We have built a discrete event simulation environment which is sufficiently parametrizable to represent different kinds of transport networks like urban traffic and IP packet routing; see Göbel 2009 for details about this environment. The mesoscopic logic of entity movement is derived from the traffic queuing model proposed by Nagel 2003. This simulation is based on DESMO-J, a framework for discrete event modelling and simulation in Java (see Page 2005 and the web page at <http://www.desmoj.de>), developed at the University of Hamburg.

To evaluate the performance of the described GP approach of transport network optimization, we have investigated three network scenarios S1, S2, S3:

- S1 consists of a single isolated intersection with incoming traffic from two directions with

identical conditions (same speed limit of 50 km/h, single lanes).

- In S2, two intersections are located 100 meters apart. Symmetrical traffic flow is restricted to W→N and E→S, yielding mutual exclusiveness at both intersections (see Figure 5, assuming right-hand traffic).
- S3 is a network consisting of 11 intersections from southern Hanover/Germany subject to various flows from almost any entry to almost any exit, see Pohlmann 2010.

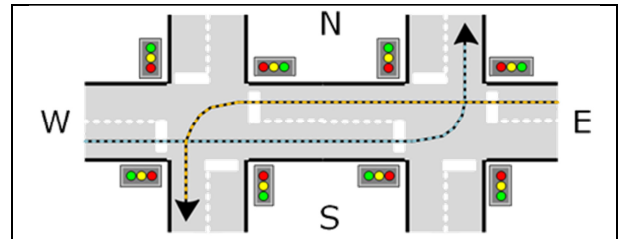


Figure 5: Scenario 2 (Two intersections)

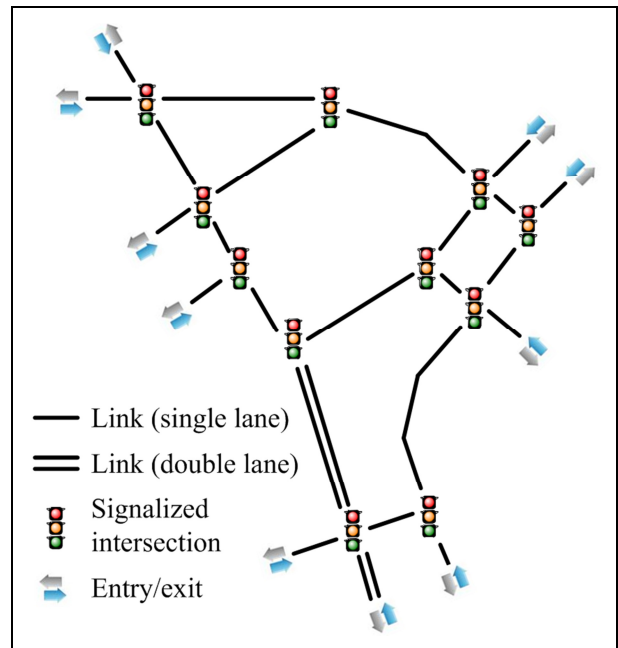


Figure 6: Scenario 3 (Hanover)

The optimization target is to minimize the vehicles’ average overall waiting times. For S1, a near-optimal strategy is alternately (“round robin”) serving each flow until the queue is empty: Switching earlier is not optimal as intersection capacity would be given away due to the switching penalty incurred in terms of a two second safety period in which all traffic lights have to be red; switching later most likely wastes capacity as no vehicle is served (unless the next arrival of a vehicle not yet queued is very close at hand).

Since the space between the intersections is limited in S2, the optimization problem particularly involves synchronizing their traffic lights such that both intersections never waste capacity by being unable to server either of the flows. This undesired situation

occurs if the link towards the other intersection is fully congested (thus no further incoming vehicles from W at the western intersection and from E at the eastern intersection can be served) while at the same time no vehicles bound for N/S already served by the other intersection are waiting for service. A near-optimal centralized solution for S2 is “fill and clear”, exploiting the symmetry of the traffic flows offered: At both intersections the incoming flows from W at the western intersection and from E at the eastern intersection receive green until the link between the intersections is filled or until both queues incoming are empty. Synchronously, the vehicles bound for N/S now receive green at both intersections until the link between the intersections is cleared. Assuming enough “supply” of incoming vehicles, neither of the intersections ever wastes capacity apart from symmetry deficits caused by stochastic noise, e.g. one intersection clearing its vehicle queue on the link between the intersections faster than the other.

For S3, a heuristic is used serving the longest queues (typically four protected flows on a four-way intersection, e.g. W→S, E, N and N→W in right-hand traffic) until the queues for different combinations of flows are least 25 vehicles longer, yet at least for 5 seconds.

Network	Control	Avg. Wait	Throughput
S1 (low load)	Round robin	5.7	3167
	GP	5.4	3161
S1 (high load)	Round robin	17.4	6243
	GP	19.0	6085
S1 (overload)	Round robin	141.6	6376
	GP	135.5	6459
S2	Fill and clear	13.6	5170
	GP	13.0	5201
S3	Longest queue	77.3	18663
	GP	52.9	23377

Table 3: Experiment performance results

Network	Program
S1/S2	repeat if (V/NID) return PI = (V/LIU) else return PI = exp(V/LFA)
S3	repeat if (V/NID) return PI = V/LAO else return PI = exp(V/LPF) * V/LWT while (N/NQL > 68.4445) return PI = V/GAA*exp(exp(V/LQL))

Table 4: GP evaluation results

Table 3 compares these means of intersection control to the best GP solution found in 100 generations of size 100 for S1/S2 (combined) and for S3; the table

states average waiting time and vehicle throughput during 5 hours (average of 10 runs). S1 has been evaluated with three different load levels (approx. 3200, 6400 and 9200 vehicles offered). The GP fitness function was the average waiting duration (the lower, the better), subject to a penalty proportional to the node-count of the genetic program, thus implicitly bounding the complexity the programs evolved. Table 4 shows the “fittest” programs found in each of the runs S1/S2 and S3.

Comparing the results of GP to the (near-)optimal solutions in the case of S1 and S2 or to the heuristic in S3 yields a GP performance similar or better (with the exception of the second S1 load level): Although not applying centralized control, e.g. explicit traffic light synchronization in S2, the GP solutions perform approximately equally well or in some cases even slightly better by exploiting the marginal remaining optimization potential, e.g. asymmetrical link clearance in S2 used to advance the traffic light switch at the relevant intersection which is advantageous in terms of overall waiting durations if the flow set to green is slower than its counterpart receiving green later.

5. SUMMARY AND OUTLOOK

This paper has presents a GP-based approach to decentralized, transport network optimization, providing local rules in terms of priority index functions. To the standard paradigm of GP evolution (Koza 1992), adjustments were necessary to cover inversion of control in fitness evaluation (simulation calling the program to be evaluated, not vice versa) and state-dependently only partially executing the program to be evaluated; these adjustment were implemented extending JGAP (Java Genetic Algorithms Package, see Meffert 2011). Experiment results indicate that the performance is similar to centrally (near-optimally) controlled systems while at the same node control is scalable and not dependent on a central authority. “Performance” of course is not restricted to minimizing waiting times as conducted in the experiments in Section 4; the GP-based transport network is sufficiently flexible to use any fitness function, e.g. a weighted combination of waiting times and fuel consumption/emission production.

Further work will address the run-time performance of the GP evolution of node PI functions: As the fitness evaluation of a single program is relatively expensive due to the discrete event simulation runs to be executed (approx. 1 day for scenarios S1/S2, approx. 4 days for scenario S3 on a single machine), recognizing and not evaluating inferior programs may provide large improvements in run-time performance. Examples of such inferior programs are all programs containing branches that are never executed (e.g. all lines after a while(true) {...} statement) or programs not containing a single lane-specific criterion (compare Table 1) as they yield the same priority for all lanes at an intersection.

The convergence of the GP can also be improved by providing “higher level” criteria, e.g. including the

optimal priority for an isolated intersection subject to uniform flows (no stochastic noise) as determined by Lämmer 2007, thus relieving the GP evolution from producing such terms.

Another means of facilitating GP convergence is removing the need for co-evolution by allowing a sub-tree referenced more than once: If the results determined by a certain sub-tree, the current GP approach would be required to create this repeating pattern more than once (Figure 7, left). Allowing multiple references (Figure 7, right) has to ensure infinite recursion is avoided, yet provides smaller programs without need to multiple branches undergoing the same evolution.

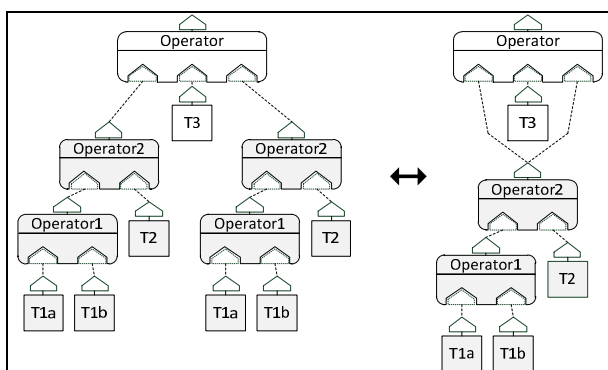


Figure 7: Multiple references to a sub-tree

REFERENCES

- Bazzan, A., Oliveira, D. de, and Lesser, V., 2005. Using Cooperative Mediation to Coordinate Traffic Lights: A Case Study. *Proceedings of Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 463-469, July 25-29, Utrecht (The Netherlands).
- Cools, S.-B., Gershenson, C., and D'Hooghe, B., 2007. Self-organizing traffic lights: A realistic simulation. In M. Prokopenko (ed): *Self-Organization: Applied Multi-Agent Systems*, pp. 41-49. London (UK): Springer.
- Diakaki, C., Dinopoulou, V., Aboudolas, K., Papageorgiou, M., Ben-Shabat, E., Seider, E., and Leibov, A., 2003. Extensions and new applications of the traffic signal control strategy TUC. *Transportation Research Board*, 1856:202-211.
- Gershenson, C., 2005. Self-Organizing Traffic Lights. *Complex Systems* 16(1):29-53.
- Göbel, J., 2009. On Self-Organizing Transport Networks – an Outline. *Proceedings of the 6th Vienna International Conference on Mathematical Modelling (MATHMOD) 2009*, p. 82. Feb 11-13, Vienna (Austria).
- Helbing, D., and Lämmer, S., 2008. Self-Control of Traffic Lights and Vehicle Flows in Urban Road Networks. *Journal of Statistical Mechanics: Theory and Experiment* 4(P04019):1-33
- Holland, J. H., 1995. *Hidden Order – How Adaption builds complexity*. New York (New York, USA): Basic Books.
- Koza, J. R., 1992. *On the programming of computers by means of natural selection*. Cambridge (Massachusetts, USA): MIT Press.
- Lämmer, S., 2007. *Reglerentwurf zur dezentralen Online-Steuerung von Lichtsignalanlagen in Straßennetzwerken*. PhD Thesis, Technical University of Dresden (Germany).
- Meffert, K. et al., 2011: *JGAP – Java Genetic Algorithms and Genetic Programming Package*. URL: <http://jgap.sf.net>
- Nagel, K., 2003. Traffic networks. In S. Bornholdt, H. G. Schuster (eds): *Handbook on networks*. New York (NY, USA): Wiley.
- Page, B. and Kreuzer, W., 2005. *The Java Simulation Handbook – Simulating Discrete Event Systems with UML and Java*. Aachen (Germany): Shaker.
- Pohlmann, T., 2010. *New Approaches for Online Control of Urban Traffic Signal Systems*. PhD Thesis, Technical University of Braunschweig (Germany).
- Poli, R., Langdon, W. B. and McPhee, N.F., 2008. *A Field Guide to Genetic Programming*, Raleigh (North Carolina, USA): Lulu.com
- United Kingdom Department for Transport, 1995. “SCOOT” Urban Traffic Control System. *United Kingdom Department for Transport Traffic Advisory Leaflet 04/1995*.
- Wolf, T. de and Holvoet, T., 2005. Emergence Versus Self-Organisation. In S. A. Brueckner, et al. (eds): *Engineering Self-Organising Systems*, pp. 1-15. Berlin (Germany): Springer.
- Zambonelli, F., Gleizes, M.-P., Mamei, M., and Tolksdorf, R., 2004. Spray Computers: Frontiers of Self-Organization. *Proceedings of the First IEEE International Conference on Autonomic Computing (ICAC'04)*, pp. 268-269, May, Miami (Florida, USA).

AUTHORS BIOGRAPHY

J. Göbel holds a diploma in Information Systems from the University of Hamburg, Germany. He is scientific assistant and PhD candidate at the Center of Architecture and Design of IT-Systems at the University of Hamburg; his research interests focus on discrete event simulation and network optimization.

A. E. Krzesinski obtained the MSc from the University of Cape Town and the PhD from Cambridge University, England. He is a Professor of Computer Science at the University of Stellenbosch, South Africa. His research interests centre on the performance evaluation of communication networks.

B. Page holds degrees in Applied Computer Science from the Technical University of Berlin, Germany, and from Stanford University, USA. As professor for Applied Computer Science at the University of Hamburg he researches and teaches in the field of Modelling and Simulation as well as in Environmental Informatics.

3D Physics Based Modeling and Simulation of Intrinsic Stress in SiGe for Nano PMOSFETs

Dr. A. El Boukili

Al Akhawayn University, Ifrane 53000, Morocco

Email: a.elboukili@aui.ma

Abstract

We are proposing a new analytical model, in three dimensions, to calculate intrinsic stress that builds during deposition of Silicon Germanium pockets in source and drain of strained nano PMOSFETs. This model has the advantage of accurately incorporating the effects of the Germanium mole fraction and the crystal orientation. This intrinsic stress is used to calculate the extrinsic stress distribution in the channel after deposition. Simulation results of channel stress based on this model will be presented and discussed for Intel technology based nano PMOS transistors.

Keywords: 3D Modeling, Intrinsic Stress, Silicon Germanium, Nano PMOSFETs

1 Introduction

The originality of this paper is the development of new analytical model, in three dimensions (3D), to calculate accurately the intrinsic stress in Silicon Germanium ($Si_{1-x}Ge_x$) due to lattice mismatch between $Si_{1-x}Ge_x$ and Silicon where x represents the Germanium mole fraction. This intrinsic stress is generated during deposition of $Si_{1-x}Ge_x$ pockets in source and drain of Intel nano PMOSFETs (Ghani 2003). In the literature, there are only few papers and only in two dimensions (2D) about the modeling of intrinsic stress in SiGe (Rieger and Vogl 1993; Van de Walle and Martin 1986; Fischetti and Laux 1996; Brash, Dewey, Doczy, and Doyle 2004). These papers were developed in the context of device simulations for mobility modeling under the effects of stress. On the other hand, in most advanced commercial or noncommercial process simulators as FLOOPS, Sentaurus, or Athena,

the intrinsic stress is a user defined input. And, it is not calculated internally by the simulator.

In this paper, we are attempting to extend the 2D models found in the literature to 3D in the context of process simulation. We are following the 2D model of Van de Walle (Van de Walle and Martin 1986).

Most of nano semiconductor device manufacturers as Intel, IBM and TSMC are intentionally using this intrinsic stress to produce uniaxial extrinsic stress in the Silicon channel. And, it is now admitted that the channel stress enhances carrier mobilities for both nano PMOS and NMOS transistors by up to 30% (Krivokapic 2003).

The need of the hour is the development of accurate physics based models and the use of TCAD simulation tools to understand the physics of intrinsic and extrinsic stress and how to attain the desired stress in the channel.

This paper is organized as follows. Section 2 outlines the different sources of intrinsic stress in $Si_{1-x}Ge_x$ pockets generated during deposition. Section 3 describes the proposed 3D model to calculate accurately the intrinsic stress due to lattice mismatch between $Si_{1-x}Ge_x$ and Silicon. After deposition process, intrinsic stress produces an extrinsic stress distribution in the whole device. This section, also outlines the 3D elastic model for the extrinsic stress and how it is related to the intrinsic stress. Section 4 presents 3D simulation results and analysis of extrinsic stress distribution in 45nm Intel strained PMOSFETs (Ghani 2003) using the proposed intrinsic model. This section will also present 3D numerical results showing the effects of Germanium (Ge) mole fractions and crystal orientations on the intrinsic stress. For qualitative and quantitative validations, the channel extrinsic stress profiles will be calculated using the proposed 3D intrinsic stress model and will

be compared with channel stress profiles found in the literature.

At this point, we could not find any experimental values of intrinsic stress in 3D. Therefore, we could not provide any comparisons with experiments.

2 Sources of intrinsic stress in SiGe

The deposition process plays a key role in determining the intrinsic stress in $Si_{1-x}Ge_x$ films. At first, we should note that the deposition takes place at elevated temperatures. When the temperature is decreased, the volumes of the grains of $Si_{1-x}Ge_x$ film shrink and the stresses in the material increase. The stress gradient and the average stress in the $Si_{1-x}Ge_x$ film depend mainly on the Silicon-Germanium ratio, the substrate temperature and orientation, and the deposition technique which is usually LPCVD (low pressure chemical vapor deposition) or PECVD (plasma enhanced chemical vapor deposition). It was observed that the average stress becomes more compressive, if the Ge concentration decreases (Hollauer 2007). Thus, it is expected that a film with higher Ge concentration has a higher degree of crystallinity and larger grains, which leads to higher film density and to higher intrinsic stress. The intrinsic stress observed in thin films has generally the following main sources.

2.1 Intrinsic stress due to lattice mismatch

During deposition, thin films are either stretched or compressed to fit the substrate on which they are deposited. After deposition, the film wants to be smaller if it was stretched earlier, thus creating tensile intrinsic stress. And similarly, it creates a compressive intrinsic stress if it was compressed during deposition. In this paper, we are focusing on developing an analytical model in 3D for this type of intrinsic stress.

2.2 Intrinsic stress due to thermal mismatch

Thermal mismatch stress occurs when two materials with different coefficients of thermal expansion are heated and expand/contract at different rates. During thermal processing, thin film materials like $Si_{1-x}Ge_x$, Polysilicon, Silicon Dioxide, or Silicon Nitride expand and contract at different rates compared to the Silicon

substrate according to their thermal expansion coefficients. This creates an intrinsic strain and stress in the film and also in the substrate. The thermal expansion coefficient is defined as the rate of change of strain with temperature.

2.3 Intrinsic stress due to dopant

Boron doping in p-channel source/drain regions introduces a local tensile strain in the substrate due to its size mismatch with Silicon. Boron (B) atom is smaller in size than Silicon atom and when it occupies a substitutional lattice site, a local lattice contraction occurs because the bond length for Si-B is shorter than for Si-Si (Randell 2005, Horn 1955). We will deal with the 3D modeling of intrinsic stress due thermal mismatch and doping in future work.

3 Proposed 3D analytical model

In this section, we are going to describe the proposed analytical model in 3D to calculate the three normal components ($\sigma_0^{xx}, \sigma_0^{yy}, \sigma_0^{zz}$) of the intrinsic stress in $Si_{1-x}Ge_x$ due to lattice mismatch at the interfaces between Silicon and Silicon Germanium.

We did follow the same strategy used in the 2D model of Van de Walle (Van de Walle and Martin 1986). We first calculate the strained lattice constants parallel and perpendicular to the interfaces in x , y and z directions. Then, from these lattice constants, we calculate the strain parallel and perpendicular to the interfaces in x , y and z directions. And, finally, we get the 3 normal stress components in 3D from the calculated strain using a modified Hookes's law. The restriction of the proposed 3D model to 2D gives exactly the 2D model of Van de Walle. And, this is a great advantage for validations and even comparison issues.

In 3D PMOSFET with SiGe source and drain as shown in the Figure 2, there are two interfaces between Si and SiGe: a vertical interface and a horizontal interface. In the Figure 2, the vertical interface is defined in the yz -plane and the horizontal interface is defined in the xz -plane. Let's assume that $Si_{1-x}Ge_x$ pocket grown in source or drain area has thickness D_{SiGe}^h and D_{SiGe}^v at horizontal and vertical interface respectively. Let D_{Si}^h and D_{Si}^v be the thickness of the Silicon substrate at horizontal and vertical interface respectively. Strains will be generated due to lattice mismatch of

the lattice constants. Let A_{Si} and A_{SiGe} be the lattice constants of unstrained Silicon and $Si_{1-x}Ge_x$. Let $A_{\parallel,h,x}^i$, $A_{\parallel,h,z}^i$ and $A_{\parallel,v,y}^i$ be the strained lattice constants parallel to the horizontal interface in x and z directions and parallel to vertical interface in y direction. The index i represents the Silicon or Silicon Germanium materials. Let $A_{\perp,h,x}^i$, $A_{\perp,h,z}^i$, and $A_{\perp,v,y}^i$, be the strained lattice constants perpendicular to the horizontal interfaces in x and z directions and perpendicular to the vertical interface in y direction. Please see Figure 1 to have an idea about the lattice constants parallel and perpendicular to a given interface between Silicon and Silicon Germanium.

In 2D, Van de Walle assumed that $A_{\parallel,h,x}^{Si} = A_{\parallel,h,x}^{SiGe}$ and $A_{\parallel,v,y}^{Si} = A_{\parallel,v,y}^{SiGe}$. In 3D, we are also assuming that $A_{\parallel,h,z}^{Si} = A_{\parallel,h,z}^{SiGe}$. For simplicity, let's assume that $A_{\parallel,x} = A_{\parallel,h,x}^i$, $A_{\parallel,z} = A_{\parallel,h,z}^i$, and $A_{\parallel,y} = A_{\parallel,v,y}^i$. The 2D model of Van de Walle gave expressions to calculate the strained lattice constants parallel and perpendicular to the interfaces in x , y directions. For the interface in z direction, we use the same expression given by:

$$A_{\parallel,z} = \frac{(A_{Si}G_{Si}^z D_{Si}^h + A_{SiGe}G_{SiGe}^z D_{SiGe}^h)/(G_{Si}^z D_{Si}^h + G_{SiGe}^z D_{SiGe}^h)}{1}$$

$$A_{\perp,h,z}^i = A_i[1 - D_i^z(\frac{A_{\parallel,z}}{A_i} - 1)]$$

The shear modulus G_i^z for Silicon and SiGe depend on the elastic constants of the material i and depend on the the orientation of interface in z direction. It is given by:

$$G_i^z = 2(C_{11}^i + 2C_{12}^i)(1 - \frac{D_i^z}{2})$$

In this paper, the constant D_i^z depend on the elastic constants C_{11}^i , C_{12}^i , C_{44}^i of each material i . And, they also depend on the interfaces's orientations that are (001), (110), or (111). In this work, the elastic constants C_{11} , C_{12} , C_{44} for $Si_{1-x}Ge_x$ depend on the Germanium mole fraction x and on the elastic constants C_{11} , C_{12} , C_{44} of Silicon and Germanium that we get from Van Der Walles Table I. We are going to use a nonlinear extrapolation method of (Rieger and Vogl 1993) to calculate C_{11} , C_{12} , and C_{44} for $Si_{1-x}Ge_x$. We calculate the constant D_i^z that depends on the orientation of the interface in z direction by following the 2D model of (Fischetti and Laux 1996; Van de

Walle and Martin 1986):

$$\begin{aligned} D_{i,(001)}^z &= \frac{2C_{12}^i}{C_{11}^i} \\ D_{i,(110)}^z &= \frac{C_{11}^i + 3C_{12}^i - 2C_{44}^i}{C_{11}^i + C_{12}^i + 2C_{44}^i} \\ D_{i,(111)}^z &= \frac{C_{11}^i + 3C_{12}^i - 2C_{44}^i}{C_{11}^i + C_{12}^i + C_{44}^i} \end{aligned} \quad (1)$$

The ratio of strained lattice constants $A_{\parallel,x}$, $A_{\parallel,y}$, $A_{\parallel,z}$ and $A_{\perp,h,x}^i$, $A_{\perp,v,y}^i$, and $A_{\perp,h,z}^i$ to unstrained lattice constants A_i determines the intrinsic strain parallel and perpendicular to the interfaces in x , y , and z directions: $(\epsilon_{\parallel,h,x}, \epsilon_{\perp,h,x}, \epsilon_{\parallel,v,y}, \epsilon_{\perp,v,y}, \epsilon_{\parallel,h,z}, \epsilon_{\perp,h,z})$. At horizontal interfaces in x and z directions, and at vertical interface in y direction, we have:

$$\epsilon_{\parallel,h,x} = (\frac{A_{\parallel,x}}{A_{SiGe}} - 1), \quad \epsilon_{\perp,h,x} = (\frac{A_{\perp,h,x}^i}{A_{SiGe}} - 1)$$

$$\epsilon_{\parallel,h,z} = (\frac{A_{\parallel,z}}{A_{SiGe}} - 1), \quad \epsilon_{\perp,h,z} = (\frac{A_{\perp,h,z}^i}{A_{SiGe}} - 1)$$

$$\epsilon_{\parallel,v,y} = (\frac{A_{\parallel,y}}{A_{SiGe}} - 1), \quad \epsilon_{\perp,v,y} = (\frac{A_{\perp,v,y}^i}{A_{SiGe}} - 1)$$

Let $\sigma_0^{xx,h,x}$, $\sigma_0^{zz,h,z}$, $\sigma_0^{xx,v,y}$ and $\sigma_0^{yy,v,y}$ be the intrinsic stress at the horizontal interface in x and z directions and at the vertical interface in y direction respectively. We use a modified Hookes's law to get these intrinsic stress components from the intrinsic strains as follows:

$$\sigma_0^{xx,h,x} = (C_{11} + C_{12})\epsilon_{\parallel,h,x} + C_{12}(\epsilon_{\perp,h,x} + \epsilon_{\parallel,h,z})$$

$$\sigma_0^{zz,h,z} = E\epsilon_{\perp,h,z}$$

$$\sigma_0^{yy,v,y} = (C_{11} + C_{12})\epsilon_{\parallel,v,y} + C_{12}(\epsilon_{\perp,v,y} + \epsilon_{\parallel,h,z})$$

$$\sigma_0^{xx,v,y} = E\epsilon_{\perp,v,y}$$

The elastic constants C_{11} , C_{12} and the Young's modulus E are those of $Si_{1-x}Ge_x$. Finally, the normal intrinsic stress components σ_0^{xx} , σ_0^{yy} , and σ_0^{zz} in $Si_{1-x}Ge_x$ in the proposed 3D model are calculated as follows:

$$\sigma_0^{xx} = \sigma_0^{xx,h,x} + \sigma_0^{xx,v,y}$$

$$\sigma_0^{yy} = \sigma_0^{yy,v,y} \quad (2)$$

$$\sigma_0^{zz} = \sigma_0^{zz,h,z}$$

We should note that the proposed 3D model given by the equations (2) reduces to 2D model of Van de Walle if we take $\sigma_0^{zz,h,z} = 0$ and $\epsilon_{\parallel,h,z} = 0$. The 3 shear intrinsic stress

components σ_0^{xy} , σ_0^{yz} , and σ_0^{zx} are taken to be zero. Then, the 6 components of the intrinsic stress tensor σ_0 in $Si_{1-x}Ge_x$ are given by: $\sigma_0 = (\sigma_0^{xx}, \sigma_0^{yy}, \sigma_0^{zz}, 0, 0, 0)$.

The intrinsic stress tensor σ_0 is used as a source term to calculate, in the whole 3D nano MOSFET structure, the extrinsic stress tensor $\sigma = (\sigma^{xx}, \sigma^{yy}, \sigma^{zz}, \sigma^{xy}, \sigma^{yz}, \sigma^{zx})$. We note that σ^{xx} , σ^{yy} , and σ^{zz} represent the extrinsic stress along the channel, vertical to the channel, and across the channel. We assume that Silicon and Silicon Germanium are elastic materials. And, to calculate the stress tensor σ , we use the elastic stress model based on Newton's second law of motion, and the following Hookes law relating stress to strain: $\sigma = D(\epsilon - \epsilon_0) + \sigma_0$. Here D is the tensor of elastic constants C_{11}, C_{12}, C_{44} , $\epsilon_0 = 0$ is the intrinsic strain and σ_0 is the intrinsic stress given by the equations (2). A detailed description of this elastic model is given in (El Boukili 2011).

Table 1: Ge Mole Fraction Effects on Stress

Ge%	σ_0^{xx}	σ_0^{yy}	σ_0^{zz}
17	$-1.432e^{10}$	$3.269e^9$	$1.752e^9$
20	$-1.674e^{10}$	$3.821e^9$	$2.047e^9$
30	$-2.454e^{10}$	$5.607e^9$	$3.000e^9$
40	$-3.197e^{10}$	$7.312e^9$	$3.914e^9$
50	$-3.900e^{10}$	$9.321e^9$	$4.780e^9$

Table 2: Substrate Orientation Effects on Stress

Orientation	σ_0^{xx}	σ_0^{yy}	σ_0^{zz}
(100)	$-8.859e^9$	$1.186e^{10}$	$6.361e^9$
(110)	$-1.432e^{10}$	$3.269e^9$	$1.752e^9$
(111)	$-1.556e^{10}$	$1.327e^9$	$7.116e^8$

4 3D Numerical Results and Analysis

The proposed 3D model for intrinsic stress given by the equations (2) is used to simulate numerically the 3D extrinsic stress in the channel of an Intel 45nm gate length PMOSFET shown in Figure 2. For the following numerical results we used (001) for the substrate orientation and 17% as the Germanium mole fraction. In the future, we will do more investigations using different gate lengths (32nm, 22nm and below), different substrate orientations and Germanium mole fractions. The Table 1 shows the effects of Germanium mole fraction on the intrinsic stress

in $Si_{1-x}Ge_x$. From Table 1, we observe that the intrinsic stress along channel becomes more compressive, if the Ge concentration decreases. This is in great agreement with what was reported in (Hollauer 2007). Table 2 show the effects of substrate orientations on the intrinsic stress. The results in Figures 3 and 4 show that the stress components σ_{xx} and σ_{zz} along the channel, and across the channel respectively are all significant. A similar stress distribution has been reported in (Victor 2004). The values of the calculated 3D extrinsic stress are also qualitatively and quantitatively in good agreement with those calculated in (Victor 2004).

Figure 5 shows that the distribution of x stress component is compressive along channel as expected. Figure 6 shows that the distribution of the z stress component is really nonuniform in the channel. A similar result was reported in (Victor 2004). These numerical results confirm that our implementation of intrinsic and extrinsic stress models in 3D provide valid and correct results. We also believe that these results are of great interest to the semiconductor community including industrials and academia.

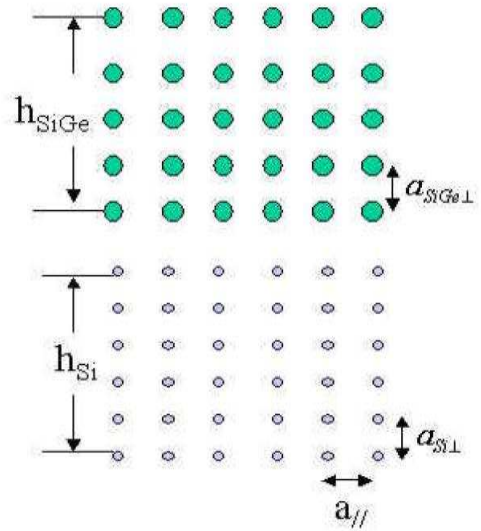


Figure 1: Parallel and perpendicular strained lattice constants.

References

- [1] Brash, B., Dewey, G., Doczy, M., Doyle, B., 2004. Mobility enhancement in compressively strained SiGe surface channel PMOS transistors with HfO₂/TiN gate stack. *Elec-*

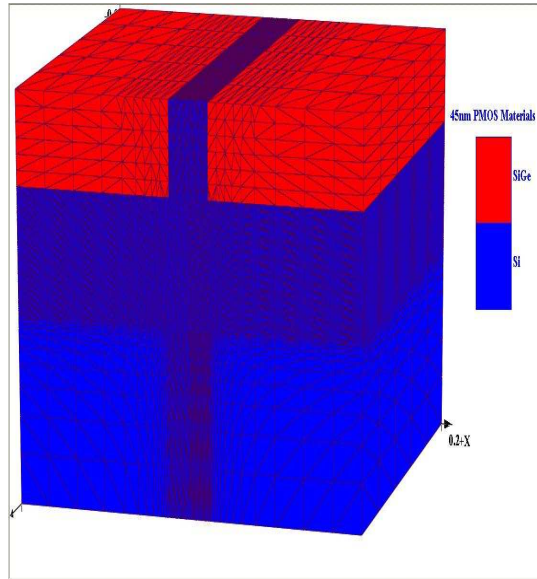


Figure 2: Materials and mesh of the simulated structure.

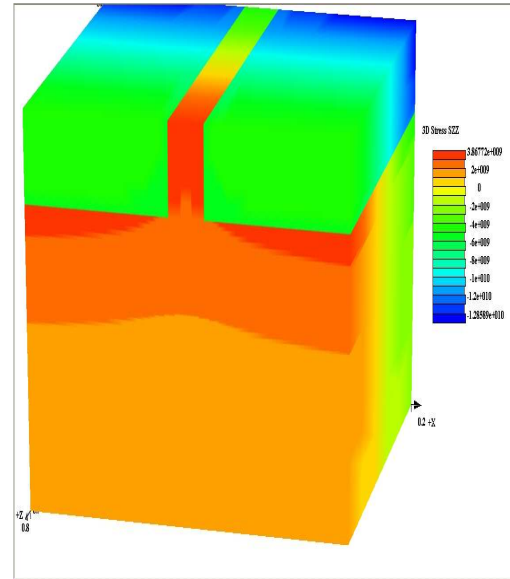


Figure 4: 3D distribution of z stress component across channel.

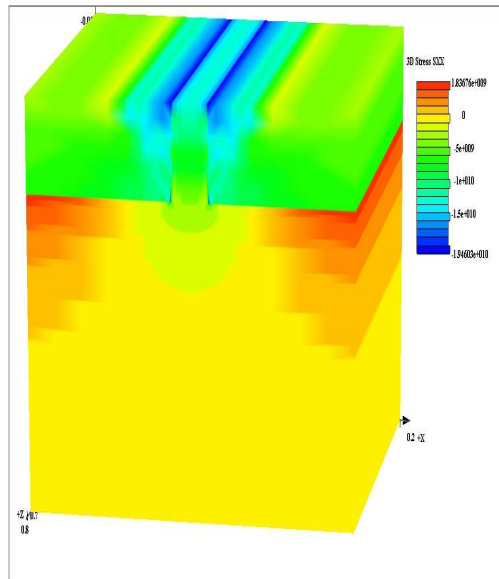


Figure 3: 3D distribution of x stress component along channel.

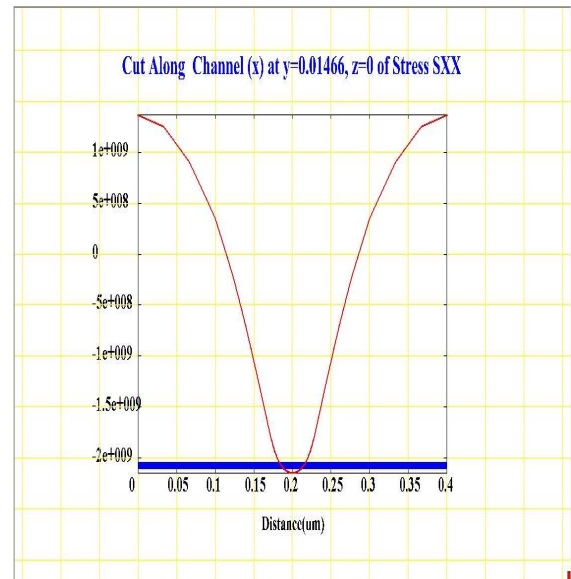


Figure 5: Cut along channel of x stress component.

trochemical society proceedings,12-30, 2004, Volume 07, San Antonio, California, USA.

[2] EL Boukili, A., 2011, 3D Stress Simulations of Nano Transistors. To appear in *Proceedings of the 16th European Conference on Mathematics for Industry*. July 26-30, 2010. Wuppertal, Germany.

[3] Fischetti, M., Laux, S., 1996. Band Structures, Deformation Potentials, and Carrier Mobility in Strained Si, Ge, and SiGe Alloys. *J. Appl. Phys.* Vol. 80: 2234-2240148

[4] Ghani T. et al., 2003. A 90nm High Volume Manufacturing Logic Technology Featuring Novel 45nm Gate Length Strained Silicon CMOS Transistors. *Proceedings of IEDM Technical Digest*, 978-980, December 2003. Washington, DC, USA.

[5] Hollauer, C., 2007. *Modeling of thermal oxidation and stress effects*. Thesis (PhD). Technical University of Wien.

[6] Horn, F., 1955. Densitometric and Electrical Investigation of Boron in Silicon. *Physi-*

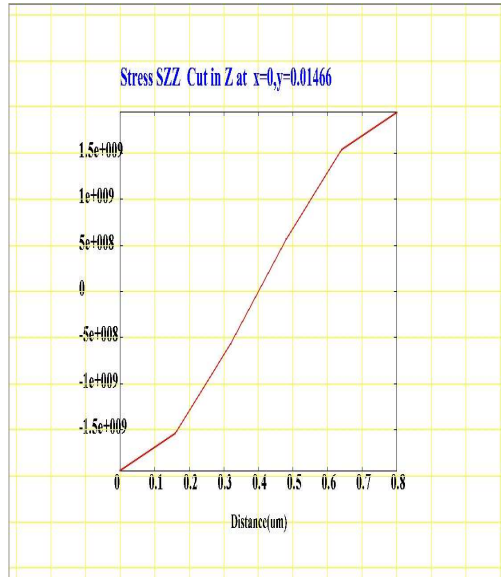


Figure 6: Cut in z-direction of z stress component.

cal Review Vol. 97: 1521-1525.

- [7] Krivokapic, Z. et al., 2003. Locally strained ultra-thin channel 25nm Narrow FDSOI Devices with Metal Gate and Mesa Isolation. *Proceedings of IEDM, IEEE International*, 445-448, 2003. Washington, DC, USA.
- [8] Randell, H., 2005. *Applications Of Stress From Boron Doping And Other Challenges In Silicon Technology*. Thesis (Master). University of Florida.
- [9] Rieger, M., Vogl P., 1993. Electronic-band parameters in strained Si(1-x)Ge(x) alloys on Si(1-y)Ge(y) substrate. *Phy. Rev. B.*, Vol. 48 No 19: 14276-14287.
- [10] Rim, K., et al., 2000. Fabrication and Analysis of Deep Submicron Strained-Si N-MOSFETs. *IEEE Transactions on Electron Devices*, Vol. 47, No.7: 1406-1415.
- [11] Takagi, S. et al., 2003. Channel Structure Design, Fabrication and Carrier Transport Properties of Strained-Si/SiGe-On-Insulator (Strained-SOI) MOSFETs. *IEDM Technical Digest*, 57-60, 10 December, Washington, DC, USA.
- [12] Van de Walle, C., Martin, R., 1986. Lattice constants of unstrained bulk Si(1-x)Ge(x). *Phy. Rev. B.*, Vol. 34: 5621-5630.

AUTHOR'S BIOGRAPHY

Abderrazzak El Boukili received both the PhD degree in Applied Mathematics in 1995,

and the MSc degree in Numerical Analysis, Scientific Computing and Nonlinear Analysis in 1991 at Pierre et Marie Curie University in Paris-France. He received the BSc degree in Applied Mathematics and Computer Science at Picardie University in Amiens-France. In 1996 he had an industrial Post-Doctoral position at Thomson-LCR company in Orsay-France where he worked as software engineer on Drift-Diffusion model to simulate heterojunction bipolar transistors for radar applications. In 1997, he had European Post-Doctoral position at University of Pavia-Italy where he worked as research engineer on software development for simulation and modeling of quantum effects in heterojunction bipolar transistors for mobile phones and high frequency applications. In 2000, he was Assistant Professor and Research Engineer at the University of Ottawa-Canada. Through 2001-2002 he was working at Silvaco Software Inc. in Santa Clara, California-USA as Senior Software Developer on mathematical modeling and simulations of vertical cavity surface emitting lasers. Between 2002-2008, he was working at Crosslight Software Inc. in Vancouver-Canada as Senior Software Developer on 3D Process simulation and Modeling. Since Fall 2008, he is working as Assistant Professor of Applied Mathematics at Al Akhawayn University in Ifrane-Morocco. His main research interests are in industrial TCAD software development for simulations and modeling of opto-electronic devices and processes. <http://www.aui.ma/perosnal/A.Elboukili>.

Simulation model for the calcination process of cement.

Idalia Flores^(a)Guillermo Perea^(b)

(a) Facultad de Ingeniería, UNAM

(b) Facultad de Ingeniería, UNAM

(a) idalia@unam.mx, (b) guillermo.perea@live.com.mx

Abstract

Simulation is an important tool when a phenomenon or input-output relationships of a system makes its operation or testing impossible, expensive, dangerous or impractical. This paper develops a simulation model for the burning process of Portland cement. The methodology used is the one used in simulation, which establishes the definition of the problem, analysis of the variables to be modeled, executes a basic model, a detailed model development, validation, reports and conclusions.

Keywords: *Simulation, calcinations, clinker, Cruz Azul cement, Arena, Simio.*

1. Introduction

Cement is one of the main inputs in the construction industry in Mexico; domestic production was 42 million tons in 2010. The calcinations of cement unit is a system consisting of a Preheater, Kiln and Cooler (PHE), which raises the temperature of the limestone powder to 1.450° C, causing physicochemical changes and the formation of silicates in a granular mixture called clinker. Simulating the system PHE will allow us to analyze the formation of clinker through a mass-energy balance.

To optimize this process we require a model that allows us to manipulate the different variables of the system. The aim of this paper is to build a simulation model of the calcination process in cement production, assessing the behavior of the input, distribution of the process, and output variables.

Figure 1 shows some components of the cement.



Figure 1 Cement components

2. Manufacturing process.

2.1 Obtaining raw materials.

The cement manufacturing process begins with the extraction of raw materials that are found in deposits, usually in open quarries. The quarries are operated by controlled blasting in the case of hard materials such as limestone and slates, while excavators are used to dig out the soft materials (clays). Once the material is extracted and classified, it is then crushed to a particle size suitable for the mill product and is transported by conveyer belt or truck to the factory for storage in the prehomogenization pile.

2.2 Homogenization and grinding of raw.

In the prehomogenization pile, the crushed material is stored in top layers to be selected later in a controlled manner. The blending bed can prepare the proper dosage of components by reducing variability.

Subsequently, these materials are ground in ball or vertical mills to make them smaller and thus make it easier to fire

them in the kiln. In the vertical mill, the material is crushed by the pressure of its roller on a turntable. From there, the raw material (powder or rawmix) is stored in a silo to increase the uniformity of the mixture.

2.3 Preheater, kiln and cooler (PHE).

The kiln is powered by means of the cyclone preheater that heats the feedstock to facilitate firing. Ground material or rawmix is inserted through the top of the tower and drops through it. Meanwhile, the gases from the kiln, which are at a high temperature, rise against the current, thus the rawmix is preheated before entering the kiln.

As the rawmix progresses in the kiln while it rotates, the temperature increases to reach 1.500 ° C. At this temperature complex chemical reactions occur that result in the clinker.

To achieve the temperatures required for firing the raw materials and the production of clinker, the kiln has a main flame that burns at 2, 000 ° C. In some cases there is also a secondary flame located in the combustion chamber in the preheater tower.

Once the clinker leaves the kiln, a cooler is introduced in inject cold air to lower the temperature from 1.400 ° C to 100 ° C. The hot air generated in this device is returned to the kiln to support combustion, thereby improving the energy efficiency of the process.

2.4 Grinding of the clinker.

Once the clinker is obtained, it is mixed in a cement mill with gypsum and additives, in the right proportions. Inside, the materials are ground, mixed and homogenized.

The mills can consist of (horizontal and vertical) rollers or balls. The later consists of a large rotating tube with steel balls inside. Thanks to the rotation of the mill,

the balls collide, crushing the clinker and additives to a fine homogeneous rawmix: **cement**.

2.5 Distribution.

Finally, the cement is stored in silos, separated according to its various classes before being bagged or loaded onto a truck for transport by road or rail.

3. The simulation model for the cement.

3.1. Calcination process analysis.

The reactions that occur in the calcination process are:

- Evaporation of water from the mixture.
- Elimination of combined water in the clay.
- Dissociation of magnesium carbonate.
- Dissociation of calcium carbonate.
- Reaction in the kiln, mixing the lime and clay.

The kiln (heat exchanger-cooler) is the equipment that determines the production, being the most important part of the process.

The clinker is produced by heating the properly dosed rawmix at high temperatures in an oxidizing atmosphere generally. The reactions of clinker produced essentially four main elements: CaO, SiO₂, Al₂O₃, Fe₂O₃ to form silicates with hydraulic properties. In overall, the clinker formation process can be divided into four parts:

Temperature (°C)	Reactions		ferroaluminato
20-100	$\text{CaCO}_3 \cdot \text{MgCO}_3 \cdot \text{Al}_2\text{O}_3 \cdot \text{SiO}_2 \cdot \text{Fe}_2\text{O}_3 \cdot \text{H}_2\text{O}$ $\gg \text{CaCO}_3 \cdot \text{MgCO}_3 \cdot \text{Al}_2\text{O}_3 \cdot \text{SiO}_2 \cdot \text{Fe}_2\text{O}_3 + \text{H}_2\text{O}_{(v)}$ Dehydration of the mixture (evaporation of free water)	1260-1450	$3\text{CaO} + \text{SiO}_2 \gg 3\text{CaO} \cdot \text{SiO}_2$ Formation of tricalcium silicate (C3S) from C2S and free lime $\text{CaO} + 2\text{CaO} \cdot \text{SiO}_2 \gg 3\text{CaO} \cdot \text{SiO}_2$
Table 3 Sintering and clinker			
100-400	$\text{Al}_2\text{O}_3 \cdot 2\text{SiO}_2 \cdot \text{H}_2\text{O} \gg \text{Al}_2\text{O}_3 + \text{SiO}_2 + 2\text{H}_2\text{O}_{(v)}$ It expels water of crystallization (water removal combined with clay)	Temperature (°C)	Reactions
400-900	Chemical water is released.	1450	Belita Formation of and Alita
		1300-1240	Crystallization of aluminates and ferrites

Table 1 Drying

Table 4 Cooled

Temperature (°C)	Reactions
500-900	$\text{CaCO}_3 \gg \text{CaO} + \text{CO}_2$ Decarbonation $\text{MgCO}_3 \gg \text{Mg} + \text{CO}_2$ Dissociation of magnesium carbonate CO_2 is expelled
Debajo de 800	$\text{CaO} + \text{Al}_2\text{O}_3 \gg \text{CaO} \cdot \text{Al}_2\text{O}_3$ Formation of calcium aluminate $\text{CaO} + \text{Fe}_2\text{O}_3 \gg \text{CaO} \cdot \text{Fe}_2\text{O}_3$ Formation of ferrous oxide
800-900	$\text{CaO} + \text{SiO}_2 \gg \text{CaO} \cdot \text{SiO}_2$ Formation of calcium silicate
900-950	$5\text{CaO} + 3\text{Al}_2\text{O}_3 \gg 5\text{CaO} \cdot 3\text{Al}_2\text{O}_3$ Formation of calcium trialuminato
950-1200	$2\text{CaO} + \text{SiO}_2 \gg 2\text{CaO} \cdot \text{SiO}_2$ Formation of dicalcium silicate (C2S)

Table 2 Calcination

Temperature (°C)	Reactions
1200-1300	$3\text{CaO} + \text{Al}_2\text{O}_3 \gg 3\text{CaO} \cdot \text{Al}_2\text{O}_3$ Formation of tricalcium aluminate (C3A)
1260	$4\text{CaO} + \text{Al}_2\text{O}_3 + \text{Fe}_2\text{O}_3 \gg 4\text{CaO} \cdot \text{Al}_2\text{O}_3 \cdot \text{Fe}_2\text{O}_3$ Formation of Tetracalcium

3.2 System mass balance-preheater kiln-cooler (PHE).

PHE Process is developed in the following steps:

3.2.1. Precalcination of the raw mixture. The preheater has a preheater, which heats the raw mixture of 60-70 ° C to 800-850 ° C, usually fueled by natural gas as fuel and use the waste gases from the kiln.

3.2.2. Formation of clinker.

The formation of the clinker takes place in the rotary kiln, which is fed with the raw mixture from the preheater and, in turn, introduces hot air (secondary) cooler.

3.2.3. Cooling of clinker.

The cooler consists of fans with variable flow through variable speed drives.

3.2.4 Cooling gases.

The waste gases are cooled in a cooling tower, which is constituted by a system of nozzles and decanting to separate the oil carried by the gases. However, decanting is not enough, so an electrostatic filter is also used.

3.2.5. Separation of dust from waste gases.

An electrostatic precipitator, consisting of plate-rapping systems and electric fields, is used to separate or precipitate dust from raw waste gases.

3.2.6. Separation of dust from the cooler.

It uses an electrostatic filter that separates the particles of dust from clinker cooler air

Mass balance:

The aforementioned integrated process is summarized in the following block diagram:

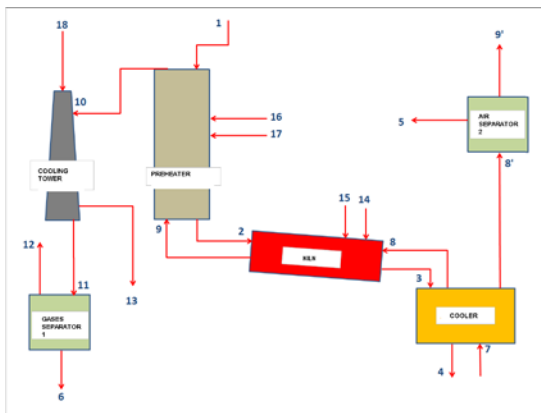


Figure 2 Mass balance

To calculate the mass flow rates, the incidence matrix for mass was developed, according to the mass flows that get in and out of the equipment:

No.	EQUIPO	INCIDENCE MATRIX OF MASS																						
		FLOW MATERIAL BETWEEN EQUIPMENTS										LEVEL VARIATIONS												
		2	3	8	9	10	11	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
1	PREHEATER	-1	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
2	KILN	1	-1	1	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
3	COOLER	0	1	-1	-1	0	0	0	0	-1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
4	COOLER TOWER	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0
5	SEPARATOR 2 (AIR)	0	0	0	1	0	0	0	0	-1	0	0	-1	0	0	0	0	0	0	0	0	0	0	0
6	SEPARATOR 4 (AIR)	0	0	0	0	0	1	0	0	0	-1	0	0	-1	0	0	0	0	0	0	0	0	0	0
0	ENVIRONMENT	0	0	0	0	0	0	0	-1	1	1	-1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	ADD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 5 Incidence matrix for mass

3.3 Simulation model for calcination unit No. 9 of Cruz Azul cement plant, Hidalgo, Mexico.

Cooperativa La Cruz Azul S.C., a homegrown company from the Mexican state of Hidalgo, currently ranks third in national cement production after Cemex and Apasco.

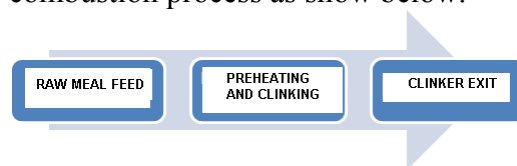
Due to the complexity and automation of most processes in the Cooperativa La Cruz Azul SCL, it is somewhat difficult to find areas of opportunity for improvement. At the present time, Cruz Azul has project engineering areas and an optimization department whose job is to constantly search for innovative technologies or technical information that would maximize existing resources, evaluate the replacement of equipment, performance and energy fuels and, if necessary, to supervise the construction of an entirely new factory. Usually these areas (or project optimization) work separately and there is a significant difference between the scope of each of them.

However, in both the Project area and the area of optimization, the firing is the key process in cement manufacturing. From the roasting process is designed the size and capacity of the kiln, which in turn determines the capacity of the preheater building itself and, in consequence, the various skills of the teams that will take part in the design of a production line or a complete plant.

The following questions arise:

- Why not analyze the input-output flows under a controlled environment of mass-energy and time?
- Why not gradually change the way people work in the process engineering department, using a scientific methodology provided by simulation?

The system to be analyzed consists of a combustion process as show below.



3.3.1 Collecting the information.

Solving the system of mass balance equations with real data of the calcination unit No.9, we have:

COOPERATIVA LA CRUZ AZUL S.C.L. MASS FLOW FOR UNIT 9			
Crude at the entrance of the preheater	m1	265.00	Ton/hr
Crude output of the preheater	m2	238.50	Ton/hr
Clinker from the oven that enters the cooler	m3	184.18	Ton/hr
Clinker cooler exit	m4	157.60	Ton/hr
Recovered clinker dust separator 2	m5	4.84	Ton/hr
Dust recovered oil separator 1	m6	29.16	Ton/hr
Inlet air cooler	m7	452.27	Ton/hr
Secondary air cooler and comes out the kiln	m8	296.60	Ton/hr
Air leaving the cooler and into the separator 2	m8'	182.24	Ton/hr
Kiln waste gases entering the preheater	m9	377.31	Ton/hr
Air "dust" coming from the separator 2	m9'	177.40	Ton/hr
Gases exiting the preheater and enter to the tower	m10	430.45	Ton/hr
Waste gases leaving the tower and enter the separator 1	m11	438.94	Ton/hr
Dust-free waste gases leaving the separator 1	m12	409.78	Ton/hr
Recovered oil coming out of the tower	m13	29.16	Ton/hr
Primary air enters the kiln	m14	18.88	Ton/hr
Fuel entering the kiln	m15	7.50	Ton/hr
Fuel entering the preheater	m16	8.80	Ton/hr
Air entering the preheater (cool)	m17	17.85	Ton/hr
Water enters the cooling tower	m18	37.64	Ton/hr

Table 6 Mass flow of calcination unit 9

3.3.2 Basic model.

The basic model was developed in Arena, which is an initial flowchart where the flow of rawmix fed into the Preheater-Kiln-Cooler system (PHE), and the chaotic movement is undergoes in the cyclone preheater (allocation probability) determines, in a linear fashion, both the consumption of fuel and electrical power.

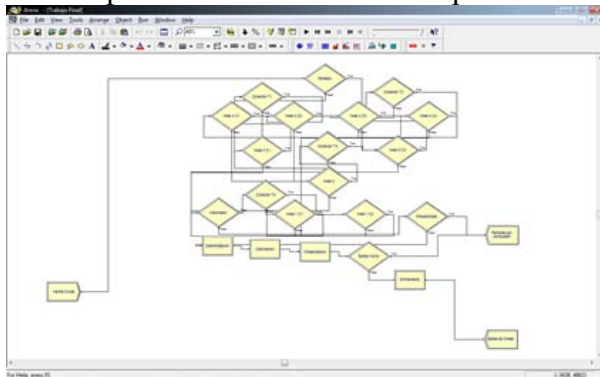


Figure 3 Flow chart for the PHE

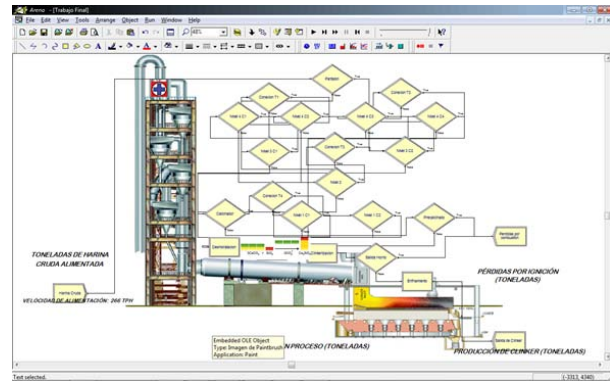


Figure 4 Basic model with Arena

3.3.3. Complete Model.

From the calculated mass balance, we developed a new model from the inputs and outputs, by considering mass and the stoichiometric analysis of the raw materials and the heat capacity of the fuel. APE software was used.

3.3.4 Flowcharts

According to the block diagram in which inputs and outputs represent the mass, we proceeded to develop flow charts for the calcination process.

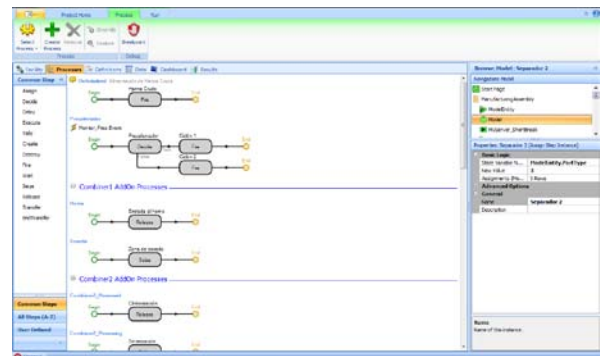


Figure 5 Flowchart for the calcination process

3.3.5 Display

We subsequently assigned the variables representing the masses (m_1, m_2, m_3, \dots) to form the input-output system mass. The arrival of a continuous entity called rawmix is determined to simulate a power of 265 Ton / hr within the system.

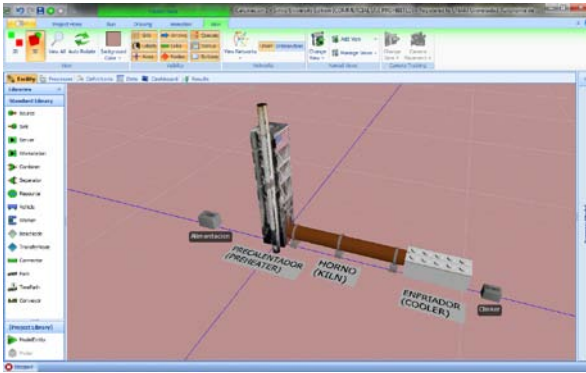


Figure 6 Simulations with SIMIO

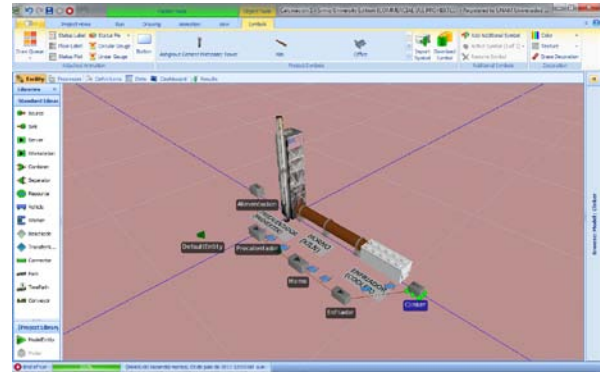


Figure 7 Simulation experiments

As a result, we obtained the following data:

We observe in the Arena model that there is a chaotic movement of rawmix particles in the cyclone, so that we determine that they are probability fluxes (approximately 65% -35%). The temperature is monitored at the inlet to the Preheater, and at the inlet, center and outlet from the kiln, at the inlet and outlet of the clinker cooler. Fuel consumption is based on the flow of rawmix, which has already been decarbonated in order to achieve more efficient calcination. It simulates the consumption of coke (petroleum), whose consumption is approximately 7.500 kg / hr in the preheater (preheating) and 8.800 kg / hr in the main burner (kiln). The flow in the supply of rawmix is between 64.7 - 65 Ton / hr, so we used the linear function L (64.7-65) to simulate the behavior.

3.3.6 Experiments

In order to have more results, 10 experiments were performed, as shown below:

COOPERATIVA LA CRUZ AZUL S.C.L.		
MASS FLOW UNIT CALCINATION No.9		
Crude at the entrance of the preheater	m1	265,00 Ton/hr
Crude output of the preheater	m2	238,50 Ton/hr
Clinker from the oven that enters the cooler	m3	184,18 Ton/hr
Clinker cooler exit	m4	157,60 Ton/hr
Recovered clinker dust separator 2	m5	4,84 Ton/hr
Dust recovered oil separator 1	m6	29,16 Ton/hr
Inlet air cooler	m7	452,27 Ton/hr
Secondary air cooler and comes out the kiln	m8	296,60 Ton/hr
Air leaving the cooler and into the separator 2	m8'	182,24 Ton/hr
Kiln waste gases entering the preheater	m9	377,31 Ton/hr
Air "dust" coming from the separator 2	m9'	177,40 Ton/hr
Gases exiting the preheater enter to the tower	m10	430,45 Ton/hr
Waste gases leaving the tower and enter the separator 1	m11	438,94 Ton/hr
Dust-free waste gases leaving the separator 1	m12	409,78 Ton/hr
Recovered oil coming out of the tower	m13	29,16 Ton/hr
Primary air enters the kiln	m14	18,88 Ton/hr
Fuel entering the kiln	m15	7,50 Ton/hr
Fuel entering the preheater	m16	8,80 Ton/hr
Air entering the preheater (cool)	m17	17,85 Ton/hr
Water enters the cooling tower	m18	37,64 Ton/hr

Experiment No.										
1	2	3	4	5	6	7	8	9	10	
265,00	265,00	265,00	265,00	265,00	265,00	265,00	265,00	265,00	265,00	265,00 Ton/hr
238,50	238,50	238,50	238,26	238,50	238,50	238,50	238,50	238,50	238,50	238,50 Ton/hr
184,18	184,18	184,18	183,81	184,18	184,18	184,18	184,18	184,18	184,18	184,18 Ton/hr
157,60	157,60	157,60	157,13	157,60	157,60	157,60	157,60	157,60	157,60	157,60 Ton/hr
4,84	4,84	4,84	4,82	4,84	4,84	4,84	4,84	4,84	4,84	4,84 Ton/hr
29,16	29,16	29,16	29,01	29,16	29,16	29,16	29,16	29,16	29,16	29,16 Ton/hr
452,27	452,27	452,27	449,56	452,27	452,27	452,27	452,27	452,27	452,27	452,27 Ton/hr
296,60	296,60	296,60	294,53	296,60	296,60	296,60	296,60	296,60	296,60	296,60 Ton/hr
182,24	182,24	182,24	180,79	182,24	182,24	182,24	182,24	182,24	182,24	182,24 Ton/hr
377,31	377,31	377,31	373,92	377,31	377,31	377,31	377,31	377,31	377,31	377,31 Ton/hr
177,40	177,40	177,40	175,64	177,40	177,40	177,40	177,40	177,40	177,40	177,40 Ton/hr
430,45	430,45	430,45	425,74	430,45	430,45	430,45	430,45	430,45	430,45	430,45 Ton/hr
438,94	438,94	438,94	433,70	438,94	438,94	438,94	438,94	438,94	438,94	438,94 Ton/hr
409,78	409,78	409,78	404,49	409,78	409,78	409,78	409,78	409,78	409,78	409,78 Ton/hr
29,16	29,16	29,16	28,75	29,16	29,16	29,16	29,16	29,16	29,16	29,16 Ton/hr
18,88	18,88	18,88	18,60	18,88	18,88	18,88	18,88	18,88	18,88	18,88 Ton/hr
7,50	7,50	7,50	7,38	7,50	7,50	7,50	7,50	7,50	7,50	7,50 Ton/hr
8,80	8,80	8,80	8,65	8,80	8,80	8,80	8,80	8,80	8,80	8,80 Ton/hr
17,85	17,85	17,85	17,53	17,85	17,85	17,85	17,85	17,85	17,85	17,85 Ton/hr
37,64	37,64	37,64	36,93	37,64	37,64	37,64	37,64	37,64	37,64	37,64 Ton/hr

Table 7 Simulation experiments

3.3.7 Model Validation

In order to validate the model, we consider the nominal production of the calcination unit No.9 Cruz Azul, according to the information provided by the area of new projects, nominal output

is 157.65 tons / hr of clinker. Considering this fact, along with the 10 experiments and the linear nature of a controlled process, we can validate the statistical behavior of the production of clinker simulated by using the Student t test as shown in the following table:



COOPERATIVA LA CRUZ AZUL S.C.L.
MASS FLOW UNIT CALCINATION No.9

Crude at the entrance of the preheater	m1	265,00	Ton/hr	m1
Crude output of the preheater	m2	238,50	Ton/hr	m2
Clinker from the oven that enters the cooler	m3	184,18	Ton/hr	m3
Clinker cooler exit	m4	157,60	Ton/hr	m4
Recovered clinker dust separator 2	m5	4,84	Ton/hr	m5
Dust recovered oil separator 1	m6	29,16	Ton/hr	m6
Inlet air cooler	m7	452,27	Ton/hr	m7
Secondary air cooler and comes out the kiln	m8	296,60	Ton/hr	m8
Air leaving the cooler and into the separator 2	m8'	182,24	Ton/hr	m9
Kiln waste gases entering the preheater	m9	377,31	Ton/hr	m10
Air "dust" coming from the separator 2	m9'	177,40	Ton/hr	m11
Gases exiting the preheater enter to the tower	m10	430,45	Ton/hr	m12
Waste gases leaving the tower and enter the separator 1	m11	438,94	Ton/hr	m13
Dust-free waste gases leaving the separator 1	m12	409,78	Ton/hr	m14
Recovered oil coming out of the tower	m13	29,16	Ton/hr	m15
Primary air enters the kiln	m14	18,88	Ton/hr	m16
Fuel entering the kiln	m15	7,50	Ton/hr	m17
Fuel entering the preheater	m16	8,80	Ton/hr	m18
Air entering the preheater (cool)	m17	17,85	Ton/hr	m19
Water enters the cooling tower	m18	37,64	Ton/hr	m20

clinker production per hour (m4)

EXPERIMENTS										
1	2	3	4	5	6	7	8	9	10	
265,00	265,00	265,00	265,00	265,00	265,00	265,00	265,00	265,00	265,00	265,00
237,55	238,50	238,50	238,50	238,50	238,50	238,50	238,50	238,50	238,50	238,50
182,70	184,18	184,18	184,18	184,18	184,18	184,18	184,18	184,18	184,18	184,18
155,72	157,60	157,60	157,60	157,60	157,60	157,60	157,60	157,60	157,60	157,60
4,76	4,84	4,84	4,84	4,84	4,84	4,84	4,84	4,84	4,84	4,84
28,58	29,16	29,16	29,16	29,16	29,16	29,16	29,16	29,16	29,16	29,16
441,52	452,27	452,27	452,27	452,27	452,27	452,27	452,27	452,27	452,27	452,27
288,40	296,60	296,60	296,60	296,60	296,60	296,60	296,60	296,60	296,60	296,60
176,49	182,24	182,24	182,24	182,24	182,24	182,24	182,24	182,24	182,24	182,24
363,94	377,31	377,31	377,31	377,31	377,31	377,31	377,31	377,31	377,31	377,31
170,43	177,40	177,40	177,40	177,40	177,40	177,40	177,40	177,40	177,40	177,40
411,89	430,45	430,45	430,45	430,45	430,45	430,45	430,45	430,45	430,45	430,45
418,33	438,94	438,94	438,94	438,94	438,94	438,94	438,94	438,94	438,94	438,94
388,98	409,78	409,78	409,78	409,78	409,78	409,78	409,78	409,78	409,78	409,78
27,57	29,16	29,16	29,16	29,16	29,16	29,16	29,16	29,16	29,16	29,16
17,78	18,88	18,88	18,88	18,88	18,88	18,88	18,88	18,88	18,88	18,88
7,03	7,50	7,50	7,50	7,50	7,50	7,50	7,50	7,50	7,50	7,50
8,22	8,80	8,80	8,80	8,80	8,80	8,80	8,80	8,80	8,80	8,80
16,60	17,85	17,85	17,85	17,85	17,85	17,85	17,85	17,85	17,85	17,85
34,88	37,64	37,64	37,64	37,64	37,64	37,64	37,64	37,64	37,64	37,64
155,72	157,60	157,60	157,60	157,60	157,60	157,60	157,60	157,60	157,60	157,60

$$H0: \mu = 158 \text{ Ton/hr}$$

$$H1: \mu \neq 158 \text{ Ton/hr}$$

Using the t-test, the mean and the standard deviation is calculated with a sample of the 10 experiments:

$$\mu = 157.41$$

$$\sigma = 0.596$$

$$n = 10$$

4. Conclusions

The linearity of an industrial process such as cement calcination represents very slight changes in control variables, because of the importance and criticality

of this operation on the final quality of the cement. It has been demonstrated to simulate a controlled continuous production process (24 hours a day 360 days a year) finally yields results that are very close to reality, regardless of the number of variables involved. The simulation that we developed was a process "in constant motion".

There are several international companies that develop the engineering and construction for cement plants, using very complex mass and energy balances to determine the specific capacity of each piece of equipment to be installed, though, of course, the heart of the system is the installed kiln capacity and overall clinker production rate of the calcination unit. Simulation can give different scenarios for the future and allows the company to change or modify important parameters in the production of the cement.

References

Aguilar Barona Byrthzee Rubén. Tesis de Maestría: **“La Simulación de Sistemas en la mejora de procesos de manufactura y servicios”**. Facultad de Ingeniería UNAM. Año 2004.

Barceló Jaime. **“Simulación de Sistemas Discretos”**. Primera Edición. Año 1996. Editorial Isdefe. Madrid España.

García Dunna Eduardo, García Reyes Heriberto y Cárdenas Barrón Leopoldo. **“Simulación y análisis de sistemas con ProModel”**. Primera Edición. Editorial Prentice Hall, México 2006.

Kelton David, Sadowski Randall y Sturrock David. **“Simulación con Software Arena”**. Cuarta edición. Editorial Mc Graw Hill. Año 2008.

Rivett Patrick. **“Construcción de modelos para análisis de decisiones”**. Primera

Edición. Año 1983. Editorial Limusa.
México D.F.

Deolalkar S P. **“Handbook for Designing
Cement Plants”**. Año 2009. BS
Publications.

Ackoff Russell L., Sasieni Maurice W.
**“Fundamentos de Investigación de
Operaciones”** Sexta Edición. Año 1984.
Editorial Limusa. México D.F.

Web sites

<http://canacem.org.mx/canacem.htm>

<http://www.cruzazul.com.mx>

<http://www.holcim.com.mx>

<http://www.ieca.es>

<http://minerals.usgs.gov/minerals/pubs/mcs/2011/mcs2011.pdf>

<http://www.investigacion-operaciones.com/Historia.htm>

<http://www.vaticgroup.com/unlimitpages.asp?id=81&pid=-1>

<http://www.flsmidth.com>

(UPIICSA), currently studying a Masters in Systems Engineering with specialization in Operations Research at the National Autonomous University of Mexico (UNAM). He has worked in the cement industry in the planning and control of projects specifically in Cementos Cruz Azul. Within the concrete industry in the same group has developed databases and applied statistical tools that have increased the efficiency of the information in the technical area He is currently coordinator of new projects.

AUTHORS BIOGRAPHY

Dra. Idalia Flores de la Mota...

Dr. Idalia Flores de la Mota is mathematics at the Faculty of Sciences of the UNAM and studied the Masters and Ph.D. in Operations Research at the Faculty of Engineering of the UNAM. She has published notes, chapters in books, booklets and articles disclosed in international journals. He has been referee of the journals, Simulation, the Journal of Accounting and Administration, Computing Reviews online and Revista Iberoamericana de Automática e Informática Industrial. It belongs to the Institute for Operations Research and the Management Science and is Director of the Center for Simulation McLeod in Mexico.

Ing. Guillermo Perea Rivera.

Industrial engineer, graduated from the National Polytechnic Institute

JOB SATISFACTION MODELLING IN AGENT-BASED SIMULATIONS

Alexander Tarvid

University of Latvia

atarvid@inbox.lv

ABSTRACT

Theoretical labour market models that incorporate social networks have largely focused on the steady-state of the system, ignoring their short- and medium-term dynamic effects. In many agent-based models of job search, the unemployed were either static, taking any vacancy proposed to them, or chose among vacancies based on either proposed wage or whether there were any of their friends employed in the firm. Thus, job satisfaction, an important multi-faceted concept in the labour market literature, has been overlooked. We propose a way to measure job satisfaction and illustrate how it can be incorporated in an agent-based model of the labour market. We use a simulation to study the dynamics of this model.

Keywords: job satisfaction, agent based modelling, labour market, social network

1. INTRODUCTION

Empirical studies show that social networks are important in the labour market. Bewley (1999) reports that 96 out of 161 (or 60 per cent) US businesses interviewed use personal contact networks to find job candidates, where in most cases, this meant employee referrals. Based on a survey of 6066 employers in Latvia, Hazans (2011) found that networking is the most popular recruitment method used by enterprises (depending on language used in enterprises, 30% to 50% of them hire by referral), but the intensity of systematic use of social networks decreases with firm size. Latvia is not an exception—indeed, Kuddo (2009) notes that in all Eastern European countries, a usual way of finding and hiring for vacancies is through informal channels (relatives, friends, acquaintances), especially in the small and medium enterprise sector. Employees also use their social networks in the process of job search. Montgomery (1991) cites several studies reporting that around 50 per cent of employees in the US found their jobs through friends and relatives. In Estonia, using Estonian Labour Force Survey data, we find that every year during 2001-2009, 30 per cent of respondents reported asking relatives and friends as their most important step taken to find a job (30% mentioned watching job ads, 15%—directly contacting employers, and 15% found it most helpful to seek through the state employment office).

Granovetter (2005) mentions two reasons why social networks are so much used in the hiring process. Firstly, they help mitigate the problem of bilateral asymmetric information, when both prospective employers and employees do not know the other side's quality. In these settings, they search for more information about one another from personal sources they can trust. Secondly, the cost of searching for a new employee in existing social networks, which are maintained mainly for non-economic reasons, is far lower than using the formal channels. One could argue that existing employees may inflate the real qualifications of the friend they recommend, but this would contradict their long-term interest in the company. Therefore, using referral hiring is a theoretically clean way of reducing costs.

At this point, we would like to stress that we do not touch upon normative theories of human resource management concerning whether appointing one's friend to a position inside the company is a right way of doing management. Rather, we adhere to the literature on labour economics and observe what actually happens in real-world labour markets.

Realizing the importance of social networking for labour markets, researchers started investigating the interplay between social networks and the economic situation of workers and firms. Several theoretical results (see, e.g., Bramoullé and Saint-Paul 2010, Calvó-Armengol and Jackson 2007, Krauth 2004) have been obtained for the steady-state of Markov processes describing employment and social network dynamics. To analyse dynamic non-equilibrium short- and medium-term effects, agent-based models were built. However, there still are restrictive assumptions under many of them.

In some models, the unemployed were static—they were simply taking any vacancy the labour market proposed them. Abdou and Gilbert (2009) focus on the level of homophily driving the probability of both changing the social network and changing the employment status in a particular firm. They assume, however, that only social networking and homophily are the main determinants of labour status. Gemkow and Neugart (2011) use the experience-weighted attraction algorithm to guide agents in their network formation decisions. Nevertheless, by assuming that workers apply to all available vacancies, they do not

model choice between them. The probability of being employed in their model depends only on whether an applicant has friends in the firm hosting the vacancy. Tassier and Menczer (2008) assume that agents learn about open vacancies after a formal search and based on information from their friends. Nevertheless, all vacancies are identical.

Other models introduced heterogeneous vacancies. For instance, Tassier and Menczer (2001) present an evolutionary model where vacancies differ by the associated wage rate, and the person chooses the vacancy with the highest proposed wage. However, the social network plays only a role of informing its member on the vacancies available.

In reality, individuals' decisions on which vacancy to choose or whether to leave the current job depend on a combination of monetary and social rewards, rather than on each of these in isolation. In particular, it is quite well-known that job satisfaction (JS) is an important predictor of the decision to quit (Acker 2004, Manger and Eikeland 1990, Parry 2008). Carless and Arnup (2011) found that JS increases statistically significantly after a job change, which means that workers take into account expected job satisfaction when choosing among several job proposals.

Kalleberg (1977, p. 126) defines JS as “an overall affective orientation on the part of individuals toward work roles which they are presently occupying” and views it as the result of an interplay between the values workers attach to job characteristics and the extent to which these values are satisfied. He proposes six dimensions of values: intrinsic (associated with the task itself), convenience, financial, relationships with co-workers (satisfaction of social needs), career, and resource adequacy. While he did not find that relationships with co-workers significantly affect JS, this does not mean that co-workers are irrelevant to it. Indeed, in his definition of resources he extensively mentions help, authority, information, supervision, and competency of co-workers, and these resources are found to significantly influence JS. Harris, Winkowski, and Engdahl (2007) arrive at a similar conclusion with a more recent dataset. Empirically, the level of social support from co-workers was found to be significant in many occupations (Alexander, Lichtenstein, Oh, and Ullman 1998; Brough and Frame 2004; Cortese, Colombo, and Ghislieri 2010; Ducharme and Martin 2000; Roxburgh 1999). This importance of the social support resource may come from it being a buffer against high job demands to prevent job strain and because it affects motivation and productivity, according to the Job Demands-Resources model (Bakker and Demerouti 2007).

This paper proposes a way to incorporate JS in an agent-based model of labour market, explicitly making several factors affect individuals' decision-making. The paper is structured as follows. Section 2 introduces our method of modelling JS. In Section 3, we put our JS model in the context of an artificial labour market. Section 4 provides the description of simulation setup

that implements the labour market model and the analysis of its dynamics. The last section concludes.

2. MODELLING JOB SATISFACTION

To formally model job satisfaction (JS), we propose to separate it into two components: expected JS and current JS. The difference between the two is that the former can be measured for any job, as it depends only on the current situation in the firm hosting the job. The latter depends on the former but, in addition, incorporates a stochastic component tracking the experience of the person on the job.

In this paper, we assume that expected JS, s_j^e , is a function of the ratio of the wage of agent i to his reservation wage, w_{ij}/w_i^r , and the ratio of the number of friends he has in the firm hosting the job (also referred to as “the number of local friends”) to the maximum number of friends he can have, n_i^f/\bar{n}_i . In other words, a person expects a certain level of monetary compensation and social support on the job. (Note that with the number of local friends we approximate a broad notion of social support rather than dividing it into support from management and from colleagues. More sophisticated frameworks should account for these two facets separately.) This is quite realistic, as normally, this is the only information individuals have before actually starting working in the firm. We also assume the following properties of expected JS:

- Its partial derivatives with respect to both parameters are decreasing functions of the absolute values of the respective parameters. Thus, any next friend working in the same firm would add less to job satisfaction. The same would go for any next dollar of wage change;
- Its range is bounded: $s_j^e \in [\underline{s}, \bar{s}]$. Firstly, the level of satisfaction cannot be arbitrarily large or arbitrarily low. Secondly, this requirement is consistent with empirical data from surveys, where satisfaction is normally measured on a Likert scale, which would help in validating the model;
- The same level of job satisfaction can be gained by different combinations of the relative wage and the number of friends.

On the contrary, when the person starts working, many other parameters start influencing his current JS—working conditions, job demands, role clarity, and other facets. As already noted above, we assume that these factors are pure noise captured by a random disturbance $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. The change in current JS in period t , therefore, is the sum of this random disturbance and the change in expected JS (due to changes in wage and the number of local friends):

$$s_{ijt}^c - s_{ijt-1}^c = (s_{ijt}^e - s_{ijt-1}^e) + \varepsilon_t \quad (1)$$

Note that both s_{ijt}^c and s_{ijt}^e should remain in $[\underline{s}, \bar{s}]$.

3. SIMULATION CONTEXT

To illustrate how the proposed job satisfaction model could be used in a real simulation, we incorporate it in the following artificial labour market.

3.1. General Characteristics

The timing is discrete, one period representing one month, and 12 months constituting a year. Throughout the paper, we will use subscript t to refer to the monthly periods and τ to the yearly ones. Most actions in the labour market, such as changes in job satisfaction and in the workforce of a firm, are made on a monthly basis. Changes in the population and in wages are made once a year.

There are two types of agents in the economy: persons and firms. Initially, the economy is populated with N_0 persons. Each year $\tau \geq 1$, N_τ new persons are added to the population, with the number of new entrants growing at a fixed rate of g , $N_\tau = gN_{\tau-1}$. This can be regarded as an inflow of new secondary school graduates in the labour market. Persons are born with zero age and zero experience (including those in the initial population) and start seeking a job. They retire at the age of \bar{a} years, at which moment they are removed from the simulation. Firms, on the contrary, are assumed to live forever, the number of firms being fixed at M .

3.2. Job Search

There is a unique vacancy list in the economy that is available to everyone for free. To find new labour, firms post vacancies on the vacancy list. Persons use the list to find new jobs.

A vacancy is a three-tuple (f, x, w) , where

- f is the firm hosting the vacancy;
- $x \in \mathbb{Z}, 0 \leq x \leq \bar{x}$ is the required working experience measured in years, \bar{x} being the sufficient experience, which is common for all vacancies;
- $w \in \mathbb{Z}, w \geq w_m$ is the proposed wage rate at the required experience x , w_m being the minimum wage, which is the same for the whole economy.

The proposed wage rate at experience $x_i \leq \bar{x}$ is

$$w_i = w + q(x_i - x), \quad (2)$$

where q is a constant equal for all vacancies. The proposed wage rate at experience $x_i > \bar{x}$ is given by the same equation with x_i taken equal to \bar{x} .

A person in search for a job browses through the vacancy list and creates a sub-list of vacancies that require a working experience not higher than his experience and, for his experience (q is known by everyone), propose a wage rate not lower than his reservation wage. He then sends applications to k vacancies from the sub-list with the highest expected job satisfaction, where k is constant for all persons.

A firm screens the applicant list for each of its vacancies and, if it finds its employees' friends, it randomly chooses among them; in the other case, it chooses randomly from all applicants. Successful

candidates receive acknowledgements. If a person receives acknowledgements for several applications, he chooses the one with the highest expected job satisfaction. He then sends an acknowledgement in reply to the chosen vacancy and starts working immediately.

If the vacancy failed to attract a new employee, the hosting firm re-posts this vacancy in the next period, raising the proposed wage rate w by the factor of h , which is the same for all firms. Required experience does not change, because the firm needs qualified personnel for its vacancies. Instead, the firm realises that the reason of the failure of the vacancy is a lack of motivation, that is, expected job satisfaction, which the firm can improve only by raising the proposed wage rate. If the re-posted vacancy also fails, it is completely removed from the vacancy list.

For a working person, reservation wage is his current wage. For a person with no working experience, it is given by the minimum wage. For an unemployed, it is a decreasing function of his last wage and the length (in months) of the current unemployment period t_i^u ,

$$w_i^r = \varphi^{t_i^u - 1} w_i, \quad w_i^r \geq w_m \quad (3)$$

The parameter $\varphi, 0 < \varphi < 1$ is the same for all persons. Thus, the longer a person is unemployed, the lower wage rate he is ready to accept. Reservation wage, however, cannot fall below the minimum wage.

We also model on-the-job search. If current JS falls below the minimal level, which is the same for everyone, the person starts seeking job as if he was unemployed. However, in this case, he only considers the appropriate vacancies with expected JS not lower than his current JS. If he is selected to fill a vacancy, he quits his current job and then starts working on the new position (it could be hosted by the same firm where he worked before).

3.3. Firm Inside Dynamics

All firms start with no workforce. In the first month, each firm publishes N_0/M vacancies. Thus, at the beginning, all firms try to be of equal size.

Each month, firms randomly change the size of their workforce by δ_{ft} persons, which is distributed uniformly in $[-\underline{\delta}, \bar{\delta}]$. If $\delta_{ft} < 0$, the firm contracts its workforce by randomly firing $|\delta_{ft}|$ employees. If the change is positive, it publishes δ_{ft} vacancies.

Each vacancy for a new position is created with the required experience x uniformly chosen from $[0, \bar{x}]$. Given that value of x , the corresponding proposed wage w is set to the average wage currently received by the firm's employees having experience x . If no such persons are currently employed, the firm considers wages earned by the relevant employees who were working in the firm in the nearest month during the last year. If no such persons worked in the firm during the last year, it makes an interpolation from the point $(0, w_m)$ using Eq. (2).

A firm can also publish a vacancy that would substitute employee i who just quit the firm, either

because of reaching the retirement age or due to a low job satisfaction (fired workers are not substituted). In this case, the vacancy is published with the required experience being two years smaller than that employee's experience, checking that the resulting experience is inside $[0, \bar{x}]$. The proposed wage corresponding to the required experience is then set so that an applicant with experience equal to that employee's experience had the same proposed wage as that employee, correcting it if it falls below the minimum wage. This is summarized by the following equations:

$$x = x_i - 2, \quad 0 \leq x \leq \bar{x} \quad (4)$$

$$w = w_i - q(x_i - x), \quad w \geq w_m \quad (5)$$

At the start of each year $\tau \geq 1$, each firm posts $\theta_{ft} N_\tau$ vacancies characterised by the tuple $(f, 0, w_m)$, where θ_{ft} is the firm's current labour market share. Thus, firms try to hire a share of fresh graduates that is consistent with their current labour market share, providing these graduates with vacancies with the lowest experience requirements, but also proposing them the minimum wage. While we do not explicitly model production and selling, by placing a cap on the share of graduates that can be hired, we are preventing the situation when a small firm hires an arbitrarily large number of new workers, for which it may not have enough resources.

For a firm's employee, wage can change only once a year—thus, we model stickiness of wages. Wages change as in the trinomial option pricing model, i.e., by a factor taken from $\{w_u, 1, w_d\}$, where $w_u > 1$ and $w_d < 1$ with the corresponding probabilities $\{p_w^u, p_w^n, p_w^d\}$, $p_w^u + p_w^n + p_w^d = 1$; all these parameters are fixed for all firms. In the beginning of the year, firms choose one of these factors and throughout the year, they change wages of all workers with expiring yearly contracts by this factor.

3.4. Social Network Dynamics

According to Granovetter (2005, p. 34), “people have cognitive, emotional, spatial and temporal limits on how many social ties they can sustain.” Thus, maintaining a particular number of friends has an inherent cost for a person. We do not model such costs explicitly. Rather, we assume that each person has a maximal number of friends, \bar{n}_i , which depends on the importance of friends in his life, $\lambda_i \in [0, 1]$, which, in turn, is generated by $\log N(\mu_\lambda, \sigma_\lambda^2)$. Lognormal distribution was chosen because it approximates the degree distribution in networks of friends quite well (see, e.g., Toivonen et al 2009). The functional form of $\bar{n}_i(\lambda_i)$ can then be chosen so that maximal number of friends is distributed log-normally, too.

The person starts his life with a random number of friends from his generation that does not exceed the maximal number of friends he is ready to make. These can be regarded as his school-friends. Coming to a new workplace, he tries to make new $\Delta n_i = \lceil \bar{n}_i / 10 \rceil$ friends working in the firm hosting this workplace. He succeeds in creating a friendship tie with a random firm's employee with probability 1/2, as the employee can

refuse the proposed friendship. If, due to additional ties created in the workplace, the number of friends exceeds the allowable ceiling, the person removes these extra friends. He first removes currently unemployed friends, starting with those with the longest period of unemployment. If this is not sufficient, he randomly removes friends who do not work with him in one firm. Finally, after there are no more such friends, he randomly removes his colleagues from his friendship circle.

4. SIMULATION RESULTS

4.1. Simulation Setup

We implemented a simulation of the model described in the two previous sections in Repast Symphony. Expected JS was represented as a sum of two logistic functions, which correspond to the three desirable properties stated in Section 2:

$$s_{ijt}^e = (1 - \lambda_i) P\left(6 \left[\frac{w_i}{w_i^f} - 1\right]\right) + 2\lambda_i \left[P\left(\frac{6n_i^f}{\bar{n}_i}\right) - \frac{1}{2}\right] \quad (6)$$

In this formula, $P(\cdot)$ is the logistic function. In the definition of both summands, we took into account that $P(6) \approx 1$ and $P(-6) \approx 0$. Thus, when $w_i \ll w_i^f$, the first summand approaches zero, while it approaches one when $w_i = 2w_i^f$. The second summand is zero when the person has no local friends and approaches one when he has all his possible friends working with him. We also take into account the importance of friends, λ_i , so that the more important friends are for a person, the higher is the weight of the second summand. Since $\lambda_i \in [0, 1]$, it follows that $s_{ijt}^e \in [0, 1]$.

The maximum number of friends is determined according to the following linear relationship:

$$\bar{n}_i = \lceil 100\lambda_i \rceil \quad (7)$$

As $0 < \lambda_i \leq 1$, Eq. (7) guarantees that maximum number of friends is distributed log-normally in the interval $[1, 100]$.

Table 1 reports the values of the simulation's parameters. Annual population growth rate was taken approximately equal to the one characterising the situation in Europe in the last decade, as reported by Eurostat. Critical job satisfaction level is the level at which the person starts on-the-job search. Parameters of the friend importance distribution, μ_λ and σ_λ , were chosen so that the median importance is 0.07, leading to the median maximum number of friends equal to seven. The standard deviation of job satisfaction noise, σ_ε , is not reported in Table 1, since we will compare model dynamics depending on the values of this parameter.

4.2. Analysis

As the retirement age was set to 20 years and initially, everyone is of age zero, the size of the population grows rapidly until year $\tau = 20$, at which time the persons born in the first periods start retiring and the labour force size grows much less rapidly. Thus, we analyse only the last 20 years of the simulation, when the artificial labour market should have already stabilised.

Table 1: Simulation Parameter Values

Parameter	Value
— Simulation length (in months)	480
M Number of firms	20
N_0 Initial population size	200
g Annual population growth rate	1.005
\bar{a} Retirement age (in years)	20
w_m Minimum wage	100
h Wage change factor for failed vacancy	1.1
w_u Wage increase factor	1.05
p_w^u Probability of wage increase	0.6
w_d Wage decrease factor	0.95
p_w^d Probability of wage decrease	0.1
μ_λ Mean of friend importance	-2
σ_λ Std. dev. of friend importance	0.8
$\underline{\delta}, \bar{\delta}$ Workforce change boundary	5
φ Reservation wage modifying factor	0.9
\bar{x} Sufficient experience (in years)	10
k Number of simultaneous applications	5
q Wage-experience multiplier	1.1
— Critical job satisfaction level (%)	20

We compare four setups of the model, which differ, firstly, on the values of current JS noise: medium ($\sigma_\varepsilon = 0.1$) vs. low ($\sigma_\varepsilon = 0.05$), and secondly, on whether JS incorporates the local friend component (the second summand in Eq. (6)) or it consists of the wage component only. In models where friends are irrelevant to JS, persons still make friends and firms continue to hire by referral. Persons simply do not take the number of local friends into account when choosing among vacancies or considering starting on-the-job search.

Figure 1 compares the distribution of friend importance (which is identical to the distribution of the maximum number of friends) with the actual number of friends in the last period of the model. We can observe that the friend distribution is more peaked and has a thinner tail than the friend importance distribution.

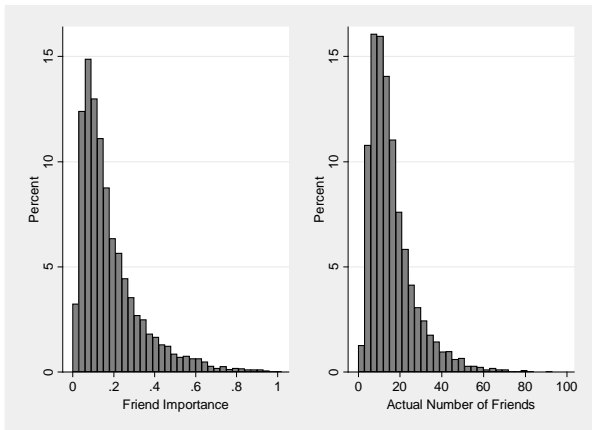


Figure 1: Distributions of friend importance and actual number of friends for the last period of the simulation.

This means that persons with a high maximum number of friends generally fail to make that many friendship ties, which should lead to their second

component of expected JS being less than that of persons with lower friend importance—other things equal, the former are less happy than the latter because they do not meet their needs for social interaction. Consequently, we expect persons with higher friend importance to change jobs more frequently. Logistic regressions of leaving the job because of a low current JS (see Table 2) confirm this: in all four models, the largest effects in absolute terms are from friend importance and squared experience, and friend importance effects are positive. Note that the effect of friend importance on the probability of quitting the job increases if JS is based solely on the wage component.

Table 2: Marginal Effects after Logit Regression of Leaving the Job

Model	Med noise	Med noise, wage only	Low noise	Low noise, wage only
Friend imp.	.010***	.013***	.007***	.012***
# local friends	.000***	-.001***	.000**	-.001***
Wage	.000***	.000***	.000***	.000***
Age	.001***	.000***	.001***	.000*
Age ² /100	.004***	.002***	.003***	.002***
Experience	-.003***	-.004***	-.003***	-.004***
Exper. ² /100	-.009***	-.004***	-.008***	-.002**
Pseudo-R ²	.0848	.1232	.0806	.1210

*** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$

The table shows median values for marginal effects and median pseudo-R²s for runs of each model. Standard errors allow for intra-group correlations, where a group is defined as all observations belonging to one person.

To see whether friend importance systematically differs by the labour status of a person, we ran Kolmogorov-Smirnov two-sample tests on the runs of each of the four setups. The tests show that in the models where JS contained information on friends, the unemployed generally have a lower friend importance than the employed. For the two models where JS is based solely on wages, however, the situation is reversed.

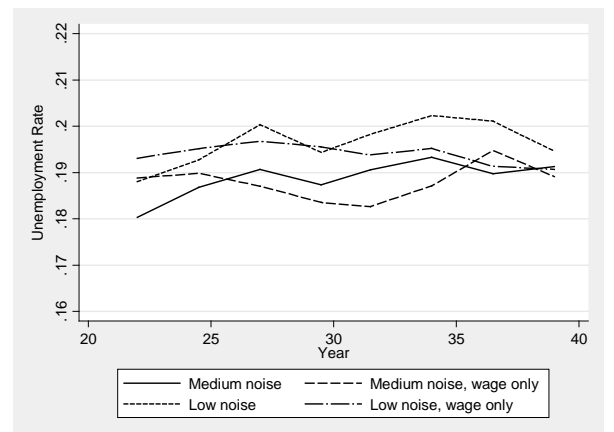


Figure 2: Median annual unemployment rates for the runs of the four setups.

Annual unemployment rates, however, do not differ much between the groups with differing JS

functions—in all four cases shown in Figure 2, median unemployment rates are in the narrow interval (0.18,0.20).

To check whether the difference in friend importance between the employed and the unemployed is significant in real terms, we compare the average of this characteristic for the two groups in each of the four model setups (see Table 3). The table shows that the differences between the groups are minor.

Table 3: Mean Friend Importance by Labour Status

Model	Friend Importance	
	Employed	Unemployed
Medium noise	0.18	0.17
Medium noise, wage only	0.18	0.19
Low noise	0.18	0.17
Low noise, wage only	0.18	0.19

Our final check for the link between friend importance and unemployment concerns the length of the longest period of unemployment experienced by the person. Results (see Table 4) show that, firstly, higher friend importance tends to reduce time to find the job, and secondly, that this reduction is much larger when JS is based on wages only. Note also that for wage-only JS models, regression fit is considerably lower than for the other two models. Another result is that the lower is JS noise, the more pronounced is the friend importance effect.

Table 4: Regression of the length of the longest period of unemployment experienced by the person

Model	Med noise	Med noise, wage only	Low noise	Low noise, wage only
Friend imp.	-1.356*	-3.691***	-1.448**	-4.082***
Age	4.025***	2.821***	4.338***	3.061***
Age ² /100	.826***	-5.098***	-1.792***	-5.474***
Experience	-5.214***	-3.173***	-5.914***	-3.296***
Exper. ² /100	5.527***	8.374***	8.166***	8.759***
Constant	21.117***	17.846***	21.570***	18.004***
R ²	.2887	.1870	.2736	.196

*** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$

The table shows median values for regression coefficients and median R²s for runs of each model.

Next, we check how the situation with person's friends relates to his own situation in the labour market. We find that the correlation coefficient between a person's wage and his friends' average wage is positive and quite high in all four model specifications (see Table 5), meaning that there is a certain degree of income homophily among persons. The coefficient is generally higher when friends are taken into account in JS. The coefficient value increases slightly with JS noise variance; moreover, for a lower noise, the difference in the correlation coefficient between JS specifications becomes larger.

Table 5: Correlation coefficient between person's wage and his friends' average wage

Model	Med noise	Med noise, wage only	Low noise	Low noise, wage only
Corr. coeff.	0.819***	0.816***	0.809***	0.780***

*** $p < 0.01$

Finally, we analyse situation within firms. We divide them into three groups by the average number of employees in the last month of each of the last 20 years of the simulation (see Table 6). While all firms were of the same size initially, several large (>500 employees) and medium-sized (50-500 employees) firms have evolved in the artificial economy. Note that with a lower JS noise, more large and medium firms evolve in the standard JS function setup. That can be explained by current JS changing less and, thus, decisions to quit taken less often—as a result, persons work longer at the same jobs, and firms can better accumulate workforce.

Large firms have a higher average friend importance than medium and small firms in the models where JS incorporates friends, while the situation is reversed when JS depends on wage only, and the magnitude of JS noise does not affect the results. The same holds for average total number of friends of firms' employees. Average number of local friends, on the contrary, is always greater, the larger the company, which was also expected. Note, however, that once friends become a component of current JS, the number of local friends in large and medium-sized companies increases 1.5 to 2 times, at the expense of a minor decrease of this characteristic for small companies. In other words, friendship networks are more clustered within companies than in models where JS depends only on wages.

Table 6: Firm-Level Statistics

Model	Firm Size	Number of Firms	Average Friend Importance	Avg. Number of Friends	Avg. No. of Local Friends
Medium noise	Large ^a	2	.18	15.47	10.27
	Medium ^b	2	.15	12.93	3.76
	Small ^c	16	.15	12.03	1.68
Medium noise, wage only	Large ^a	2	.17	14.38	6.10
	Medium ^b	4	.22	18.93	2.62
	Small ^c	14	.22	18.50	1.82
Low noise	Large ^a	3	.19	15.84	10.75
	Medium ^b	3	.15	12.83	4.42
	Small ^c	14	.15	11.93	1.67
Low noise, wage only	Large ^a	2	.17	14.28	5.67
	Medium ^b	4	.21	17.99	2.60
	Small ^c	14	.21	18.26	1.80

^a Number of employees > 500.

^b Number of employees in [50,500].

^c Number of employees < 50.

The table shows average results for runs of each model.

5. CONCLUSION

In the present paper, we proposed a way to model job satisfaction and to incorporate it into an agent-based model of labour market. In our model, job satisfaction depends on two components: monetary benefits and social support, which were found to be empirically important factors; the influence of other factors is gauged by a random disturbance term. We created an artificial labour market simulation with heavy usage of social networking—during referral hiring, in choosing among vacancies, and in considering whether to start on-the-job search. The two latter choices are actually made based on job satisfaction.

We found that friend importance is an important determinant of the probability to quit the job and of the length of the longest unemployment period; both effects increase in absolute terms when job satisfaction does not depend on social support.

We also found evidence of social clustering. Firstly, labour market dynamics resulted in the emergence of a small number of large and medium-sized firms. Secondly, there is a substantial positive correlation between friends' wages, indicating income homophily of social groups. While, naturally, the average number of a person's friends working in the same firm increases with firm size, this number is nearly two times higher when job satisfaction depends both on wages and on social support than when it depends on wages only.

The model presented here has several limitations. Firstly, it does not distinguish among several components of social support, modelling it as a single factor. Moreover, it assumes that job satisfaction determinants other than monetary compensation and social support are pure noise, while they might be firm/job-specific and show serial correlation. Secondly, it portrays the behaviour of small and large firms analogically, while in a large firm, overall social support may be less important than social support in the department where the person works—thus, it may be useful to model the effects of social ties between workers of the same department and of different departments differently. In addition, empirical findings show that large firms rely on referral hiring to a lower extent than the small ones. Thirdly, it needs to be verified whether the same results hold when the production-consumption decisions are incorporated in the model. Further research should aim to overcome these limitations.

ACKNOWLEDGMENTS

The research was supported by the European Social Fund Project No. 2009/0138/1DP/1.1.2.1.2/09/IPIA/VIAA/004 (“Support for Doctoral Studies at University of Latvia”). The author also thanks two anonymous reviewers for their helpful comments.

REFERENCES

- Abdou, M., Gilbert, N., 2009. Modelling the Emergence and Dynamic of Social and Workplace Segregation. *Mind & Society*, 8(2), 173-191.
- Acker, G.M., 2004. The Effect of Organizational Conditions (Role Conflict, Role Ambiguity, Opportunities for Professional Development, and Social Support) on Job Satisfaction and Intention to Leave Among Social Workers in Mental Health Care. *Community Mental Health Journal*, 40(1), 65-73.
- Alexander, J.A., Lichtenstein, R., Oh, H.J., Ullman, E., 1998. A Causal Model of Voluntary Turnover Among Nursing Personnel in Long-Term Psychiatric Settings. *Research in Nursing & Health*, 21(5), 415-427.
- Bakker, A.B., Demerouti, E., 2007. The Job Demands-Resources Model: State of the Art. *Journal of Managerial Psychology*, 22(3), 209-238.
- Bewley, T.F., 1999. *Why Wages Don't Fall During a Recession*. Cambridge, MA: Harvard University Press.
- Bramoullé, Y., Saint-Paul, G., 2010. Social networks and labor market transitions. *Labour Economics*, 17(1), 188-195.
- Brough, P., Frame, R., 2004. Predicting Police Job Satisfaction and Turnover Intentions: The role of social support and police organisational variables. *New Zealand Journal of Psychology*, 33(1), 8-16.
- Calvo-Armengol, A., Jackson, M.O., 2007. Networks in labor markets: Wage and employment dynamics and inequality. *Journal of Economic Theory*, 132(1), 27-46.
- Carless, S.A., Arnup, J.L., 2011. A longitudinal study of the determinants and outcomes of career change. *Journal of Vocational Behavior*, 78(1), 80-91.
- Cortese, C.G., Colombo, L., Ghislieri, C., 2010. Determinants of nurses' job satisfaction: the role of work-family conflict, job demand, emotional charge and social support. *Journal of Nursing Management*, 18(1), 35-43.
- Ducharme, L.J., Martin, J.K., 2000. Unrewarding Work, Coworker Support, and Job Satisfaction: A Test of the Buffering Hypothesis. *Work and Occupations*, 27(2), 223-243.
- Gemkow, S., Neugart, M., 2011. Referral hiring, endogenous social networks, and inequality: an agent-based analysis. *Journal of Evolutionary Economics*, 1-17.
- Granovetter, M., 2005. The impact of social structure on economic outcomes. *Journal of Economic Perspectives*, 19(1), 33-50.
- Harris, J.I., Winkowski, A., Engdahl, B.E., 2007. Types of workplace social support in the prediction of job satisfaction. *The Career Development Quarterly*, 56(2), 150-156.
- Hazans, M., 2011. Labor market integration of ethnic minorities in Latvia. In M. Kahanec, K.F. Zimmerman, ed. *Ethnic diversity in European*

labor markets: Challenges and solutions.
Cheltenham, UK: Edward Elgar.

- Kalleberg, A.L., 1977. Work Values and Job Rewards: A Theory of Job Satisfaction. *American Sociological Review*, 42(1), 124-143.
- Krauth, B.V., 2004. A dynamic model of job networking and social influences on employment. *Journal of Economic Dynamics and Control*, 28(6), 1185-1204.
- Kuddo, A., 2009. *Employment Services and Active Labor Market Programs in Eastern European and Central Asian Countries.* Washington DC: World Bank.
- Manger, T., Eikeland, O.-J., 1990. Factors predicting staff's intentions to leave the university. *Higher Education*, 19(3), 281-291.
- Montgomery, J.D., 1991. Social networks and labor-market outcomes: Toward an economic analysis. *American Economic Review*, 81(5), 1408-1418.
- Parry, J., 2008. Intention to leave the profession: antecedents and role in nurse turnover. *Journal of Advanced Nursing*, 64(2), 157-167.
- Roxburgh, S., 1999. Exploring the Work and Family Relationship: Gender Differences in the Influence of Parenthood and Social Support on Job Satisfaction. *Journal of Family Issues*, 20(6), 771-788.
- Tassier, T., Menczer, F., 2001. Emerging small-world referral networks in evolutionary labor markets. *IEEE Transactions on Evolutionary Computation*, 5(5), 482-492.
- Tassier, T., Menczer, F., 2008. Social network structure, segregation, and equality in a labor market with referral hiring. *Journal of Economic Behavior & Organization*, 66(3-4), 514-528.
- Toivonen, R., Kovanen, L., Kivelä, M., Onnela, J.-P., Saramäki, J., Kaski, K., 2009. A comparative study of social network models: Network evolution models and nodal attribute models. *Social Networks*, 31(4), 240-254.

AUTHOR'S BIOGRAPHY

Alexander Tarvid is a PhD student at the University of Latvia. His research focuses on higher education policy, choice of field of higher education and its impact on the individual's position in the labour market, in particular, on unemployment, shadow employment, and over-education. He also does research on the application of agent-based simulations to the modelling of dynamics in the labour market, including various effects of social networks. In 2010-11, he participated in the World Bank project *Multi-Country Policy Study of Unregistered Employment and the Shadow Economy* as research assistant.

SIMULTANEOUS SCHEDULING OF MACHINES AND OPERATORS IN A MULTI-RESOURCE COINSTRAINED JOB-SHOP SCENARIO

Lorenzo Tiacci^(a), Stefano Saetta^(b)

^(a)^(b)Dipartimento di Ingegneria Industriale – Università degli Studi di Perugia
Via Duranti, 67 – 06125 Perugia - Italy

^(a)lorenzo.tiacci@unipg.it, ^(b)stefano.saetta@unipg.it

ABSTRACT

In the paper the simultaneous scheduling of different types of resources is considered. The scenario is constrained by machines and human resources, and its complexity is increased by the presence of two types of human resources, namely the equipper, that performs only an initial action of each task (the ‘setup’), and the normal operator, that loads and unloads each piece from machines. A conceptual model of the shop is build in order to simultaneously handle priority rules for each one of the three types of resources considered (machine, equipper and operator). A simulation model has been implemented and a simulation experiment performed in order to explore the effect on mean flow factor reduction of different combination of priority rules.

Keywords: dual-resource constraints, multi-resource constraints, priority rule scheduling, job shop control.

1. INTRODUCTION

Job shop scheduling has attracted researchers for many decades, and still now is one of the most studied subjects in literature related to industrial problems. However, multi or dual resource constrained scheduling problems are significantly less analyzed, although being more realistic (Scholz-Reiter, Heger and Hildebrandt 2009).

ElMaraghy, Patel and Abdallah (2000) defined the machine/worker/job scheduling problem as: “Given process plans for each part, a shop capacity constrained by machines and workers, where the number of workers is less than the number of machines in the system, and workers are capable of operating more than one machine, the objective is to find a feasible schedule for a set of job orders such that a given performance criteria is optimized”.

Optimal solution are difficult to find also for the single resource scheduling problem, so that many heuristics approaches have been used in literature to find good but non optimal solutions for the machine constrained problem. These approaches include: Simulating annealing (Laarhoven, Aarts and Lenstra 1992); Genetic algorithms (Zhou, Cheung and Leung

2009; Manikas and Chang 2008); Tabu search (Zhang, Li, Guan and Rao 2007)).

Complexity increases in dual resource constrained problems, and extending these often quite complex heuristics to more realistic scenarios is usually not straightforward. Dautère-Pères, Roux and Lasserre (1998) developed a disjunctive graph representation of the multi-resource problem and proposed a connected neighborhood structure, which can be used to apply a local search algorithm such as tabu search. Matie and Xie (2008) developed a greedy heuristic guided by a genetic algorithm for the multi-resource constrained problem.

However, in most real-world environments, scheduling is an ongoing reactive process where the presence of a variety of unexpected disruptions is usually inevitable and continually forces reconsideration and/or revision of pre-established schedules (Ouelhadj and Petrovic 2009). Most of the above-mentioned approaches have been developed to solve the problem of static scheduling and are often impractical in real-world environments, because the near-optimal schedules with respect to the estimated data may become obsolete when they are released to the shop floor. As a result, Cowling and Johansson (2002) addressed an important gap between scheduling theory and practice, and stated that scheduling models and algorithms are unable to make use of real-time information.

A quick, intuitive, and easy to be implemented method for dynamic scheduling is utilizing priority (or dispatching) rules. The application of priority rules gives raise to a completely reactive scheduling, where no firm schedule is generated in advance and decisions are made locally in real-time. A priority rule is used to select the next job with highest priority to be assigned to a resource. This is done each time the resource gets idle and there are jobs waiting. The priority of a job is determined based on job, machine or in general resources attributes.

Priority-scheduling rules have been developed and analyzed for many years (Haupt 1989, Blackstone, Philips and Hogg 1982, Rajendran and Holthaus 1999, Geiger, Uzsoy and Aytu 2006, Geiger and Uzsoy, 2006). Although priority rules have also been applied to

dual-resource constrained problems (Scholz-Reiter, Heger and Hildebrandt, 2009), there are no studies in literature that deal with the presence of different types of human resources, each one competent to perform a specific action of the job cycle. In fact, resources heterogeneity is usually considered just in terms of different work efficiency of resources on different tasks.

In this work we analyze a multi-resource constrained job-shop scenario in which scheduling is constrained by machines and by two types of human resources, namely ‘equippers’ and ‘operators’. Equippers and operators do not perform the same action with different efficiency, but are assigned to completely different and non-overlapping actions related to the job cycle. A conceptual model of the company’s shops is built in order to simultaneously handle priority rules for each one of the three types of resources considered (machine, equipper and operator). A simulation model has been built and a simulation experiment performed in order to explore the efficacy on flow factor reduction of different combination of priority rules.

The paper is organized as follows. The job shop scenario is described in section 2. In section 3 the conceptual model of the shops is illustrated. Section 4 deals with the implementation of the simulation model, while in section 5 the simulation experiment is described and results are discussed. In section 6 conclusions are drawn.

2. THE JOB-SHOP SCENARIO

The scenario is representative of a real case study of a manufacturing company in the field of precision metal and mechanical processing. The company is specialized in the production of very complex components for industrial, aeronautical and aerospace applications. In the aerospace and aeronautical fields, the company produces 1/A class components such as, for example, actuators, stabilizers, worm gears, landing devices, turbine’s bearing rings and axle rotors. In the industrial sector, the company produces high quality components for machine tools and laser cutting.

2.1. Areas

The company is organized in different areas, in which there are homogeneous machines. Every job assigned to a certain area can be processed indifferently in one of the machine belonging to that area. There are 5 areas: the cutting area, area 1 (turning), area 2 (milling), area 3 (drilling), and a control area. The cutting area and the control area are not critic for the scheduling problem, because resources assigned to these areas do not constrain the solution. However, they have been considered in our model in order to get a realistic representation of the flow time of each job.

2.2. Jobs

Each job is represented by (see Fig. 1):

- a quantity of pieces that have to be processed (lot size);

- a set of tasks that have to be performed on each piece of the job, and the associated area;
- the sequence of tasks that have to be performed;
- the processing times of actions connected to each task.

JOB 1 (lot size: 9)			
Task sequence	Task	AREA	
1	Turning	AREA 1	
		ACTION	TIME (min)
		Set-up	0.85
		Load	2.5
		Run	5
		Unload	2.5
		Inspection	5
2	Turning	AREA 1	
		ACTION	TIME (min)
		Set-up	131.45
		Load	2.5
		Run	3
		Unload	2.5
		Inspection	5
3	Milling	AREA 2	
		ACTION	TIME (min)
		Set-up	203.92
		Load	2.5
		Run	45
		Unload	2.5
		Inspection	5
4	Milling	AREA 2	
		ACTION	TIME (min)
		Set-up	202.93
		Load	2.5
		Run	45
		Unload	2.5
		Inspection	5
5	Control	CONTROL AREA	
		ACTION	TIME (min)
		Control	15

Figure 1: Example of data representing a job.

2.3. Machines

In each area of interest (areas 1,2 and 3) there are computer numerical control (CNC) machines. Each machine can be equipped with a variable set of tools that allow completing the run with no interruptions. Every time a job is changed, the set of tools have to be changed depending on the new task requirements, and the controlling software has to be appropriately programmed. Then each piece belonging to the job has to be loaded, processed (run), and unloaded. After the first piece of a job has finished its run and has been unloaded, it must be inspected before that the remaining pieces of the lot can start being processed (see Figure 2).

2.4. Equippers

The machine set-up is performed by the equipper operator at the beginning of each task, before processing the first piece of the lot. The inspection action on the first piece is also performed by the equipper, that controls if the run has been properly executed. If everything is ok, the other pieces of the lot can start to be processed, and load and unload operations are then carried out by the normal operator,

without the need of the participation of the equipper. The equippers assigned to a certain area are able to equip all the machines inside that area.

2.5. Operators

The normal operator performs loading and unloading of the pieces of a job. Processing starts directly after the machines are loaded. Unloading begins after processing, but if there is no operator available for unloading, the machine stays idle. The operators are not needed during processing and can work on other machines in that time period. The operators assigned to a certain area are able to load/unload all the machines inside that area.

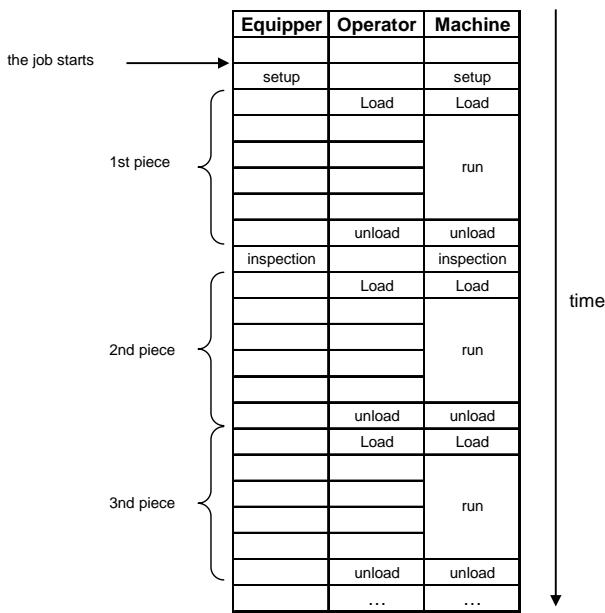


Figure 2: Actions involving the different resource types.

2.6. Shifts

An important feature of the scenario is that shifts are different between Equippers and Operators. The operators work is organized in three shifts per day, each shift during 8 hours. Equippers work on a single shift per day, from 8.00 to 17.00, with an interval of one hour between 13.00 and 14.00. Thus, while operators (as machines) are available during all the 24 hours, equippers are available only in the central part of the day. The number of resources per shift in each area is reported in Table 1.

Table 1: Resources

	Machines	Equippers	Operators
Availability	24h/day	8.00-13.00 14.00-17.00	24h/day
Area 1	4	2	3
Area 2	4	1	1
Area 3	3	2	2

2.7. The company’s scheduling process

Production scheduling is performed through a commercial software that considers different types of resources, such as machines, operators, equipments, transporters etc. However, the scheduling is done primarily only considering the machines as limiting resource. Then, the resulting schedule is verified, checking if the capacity constraints related to the other resources are respected. In case of capacity shortages, the solution can be manually modified, for example by introducing overtime, or re-calculated, by modifying jobs due dates.

This approach however does not allow a real simultaneous scheduling of machines and human resources, and the ‘trial and error’ nature of the procedure make it quite rigid, and unsuited to reacting to uncertainties of the environment. Furthermore, in the specific case study, the equipper (the most specialized human resource) is not always available during the day (see Table 1) and this makes it the primary candidate for being the limiting resource in many situations.

In the next paragraphs, we describe an alternative way to approach the scheduling problem, based on dispatching rules. The approach developed is based on the fact that, besides the rules considered to assign jobs to machines, it is necessary to simultaneously consider the rules to assign operators and equippers to jobs.

3. THE CONCEPTUAL MODEL

The systems has been modeled through a series of queues, some of which are ordered following different possible priority rules, through which decide the pick-up order of the elements. The logical flow of entities in the systems is depicted in Fig. 3 (where AREA 2 is considered). There are 4 types of queues in each area, namely: PQ1, PQ2, VQ1 and VQ2.

When a new job arrives, it tries to enter the area corresponding to the task that has to be performed. If all the machines in that area are busy, the job has to wait in a queue of type PQ1. When one of the machines in the area is free, the job is assigned to that machine, and the lot is divided into a number of pieces equal to the lot size. Pieces are then allocated in the PQ2 queue of the assigned machine (input buffer).

Pieces in PQ2 queues may claim an equipper or an operator, depending on the action needed to complete their current task. If they claim an equipper, they also enter the virtual queue VQ1, which is served from the equippers of the area; if they claim an operator, they also enter the VQ2, which is served from the operators of the area. When an equipper or an operator is available, pieces are removed from VQ1 or VQ2, and the required action on the pieces are performed.

When a task of a job has been completed, i.e. the last piece of the lot has been unloaded from the machine, the machine is released, and the job tries to enter the area corresponding to the next task of its processing sequence.

Considering each area, we classify the queues into physical and virtual queues.

3.1. Physical Queues (PQ)

- PQ1. The first physical queue is related to jobs that are waiting for entering an area. The queue is physical because we can associate a job to the lot that is waiting (in a trolley for example) in a certain part of the shop. Jobs are picked up from this queue as soon as one of the machines of the area is available to work. Each area has one queue of type PQ1.
- PQ2. The second physical queue is related to pieces of jobs that are waiting to be processed by a machine, i.e. they belong to a job that has already been assigned to a machine, and are waiting in the input buffer of the machine. Each machine has one queue of type PQ2.

3.2. Virtual queues (VQ)

- VQ1. The first virtual queue is related to pieces of jobs that are waiting for an equipper, i.e., the first pieces of a job that have been already assigned to an available machine and are waiting or for the setup action on the machine, or for the inspection action. Each area has one queue of type VQ1.
- VQ2. The second virtual queue is related to pieces of jobs that are waiting for an operator, i.e., pieces belonging to jobs already initiated, and waiting for loading or unloading actions. Each area has one queue of type VQ2.

It is noteworthy that VQ1 and VQ2 are not physical queues. In particular, elements waiting in VQ1 can be physically in the input buffer of a machine (the first piece of a job waiting for the setup action) or in the machine itself (the first piece of a job waiting for the inspection action). Similarly, elements waiting in VQ2 can be physically in the input buffer of a machine (waiting for load action) or inside the machines (waiting for unload action).

3.3. Priority rules

Priority rules are defined to select elements waiting in PQ1, VQ1 and VQ2 queues. These queues are the ones to which priority rules are applied, because elements of these queues claim a resource: PQ1 - machines, VQ1 - equippers and VQ2 - operators.

The priority rules for PQ1 (machines) are:

1. FIFO (First-IN, First-OUT);
2. LIFO (Last-IN, First-OUT);
3. SPT (Shortest Processing Time);
4. LPT (Longest Processing Time);
5. RANDOM;

In addition to these 5 rules, we considered two additional decision rules for VQ1 (equippers) and VQ2 (operators):

6. LMQL (Longest Machine Queue Length);
7. SMQL (Shortest Machine Queue Length).

Note that while rules 1 to 5 are related to a local characteristic of the queue, rules LMQL and SMQL are taken on the basis of the length of PQ2 queues.

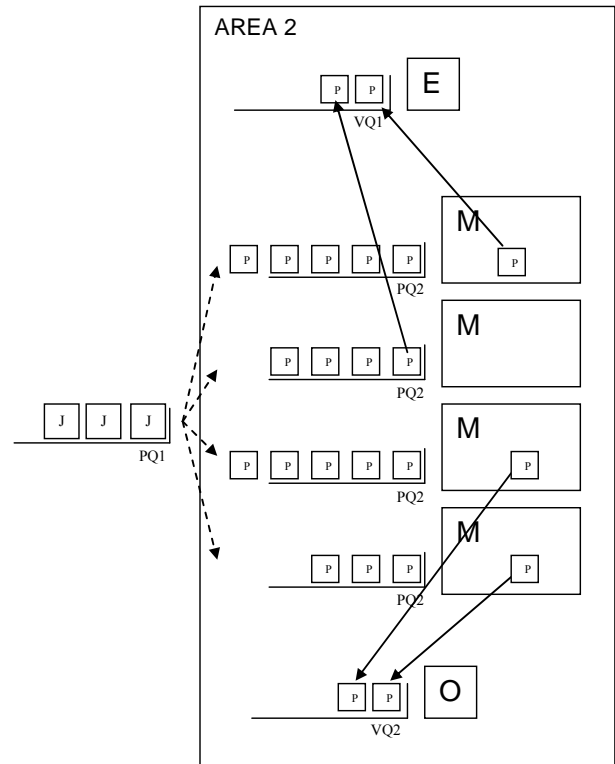


Figure 3: Queues and logical flow (J = job; P = piece)

4. THE SIMULATION MODEL

The simulation model has been built with Arena 11, using basic process, advanced process, advanced transfer, calendar schedules, and animation tools. In the next section, the main part of the model implementation is described.

4.1. The simulation model implementation

The basic entity is one piece of a job. Pieces are batched, and batches move through the system when they move between different areas. A batch is separated into pieces when the job enters an area and pieces have to be processed in the machines.

Machines, Operators and Equippers are modeled as Resources. The number of operators and equippers in each shift (see Table 1) is modeled using the Calendar Schedule utility of Arena, through which the detailed time patterns and the capacity of each resource can be set up.

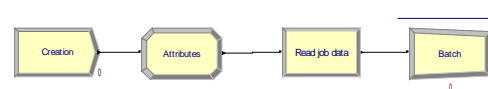


Figure 4: Pieces and job creation

A piece is created and a series of attributes are assigned to it (e.g. the job code, the current time). Then other attributes related to the job (lot size, tasks, actions times, tasks sequence etc.) are read from file. Pieces are batched according to the lot size of the job, and the batch proceeds to the area of destination, according to its task sequence. The Route and Station modules of the Advanced Transfer tools of Arena have been utilized to perform entities movements throughout the model.

The system is modeled through different sets of *Station*: Area Stations, Machine Stations, Equipper and Operator Stations, Actions Stations.

4.1.1. Area Stations

An Area Station is associated to each area of the shop floor. When a batch arrives at the Area Station (e.g. Area 3 in Figure 5), it enters PQ1 queue of the area. The order of the queue can be set according to one of the priority rules (1-5) defined in section 3.3. A batch can leave the queue according to a ‘scan for condition’ rule. The condition is verified when at least one of the PQ2 queues of the machines in the area is empty. The decide module directs the batch to the Station related to the available machine.

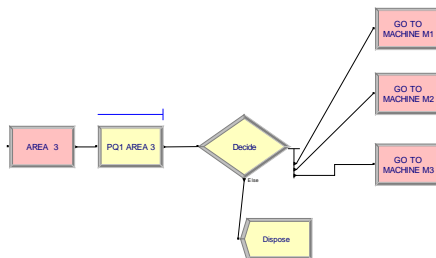


Figure 5: Area Stations.

4.1.2. Machine Stations

A Machine Station is associated to each machine of an area. When the batch arrives to the machine (e.g. machine M1 in Figure 6), it is separated into pieces. A Counter module and a Decide module allow to identify the first piece of the job (the related attribute ‘First Piece’ is associated to the entity). An attribute (named ‘Current Machine’), related to the machine to which the piece has been assigned, is also stored. Then the entity moves to the Seize module corresponding to the PQ2 queue of the machine. The seize module is associated to a virtual resource of fixed capacity equal to 1, that is seized when the entity exits the queue, and is released by the same entity when all the actions associated to its task on the machine have been performed (see later in section 4.1.4). The introduction of this virtual resource is necessary because when the entity exits the queue and seizes the resource, this means that one machine is available to work, but not that the machine will be immediately seized: in fact, the piece could have to be waiting for an equipper or an operator to be loaded in or unloaded from the machine. So, the seize module cannot be associated to the machine resource if one want to accurately evaluate the actual machine utilization.

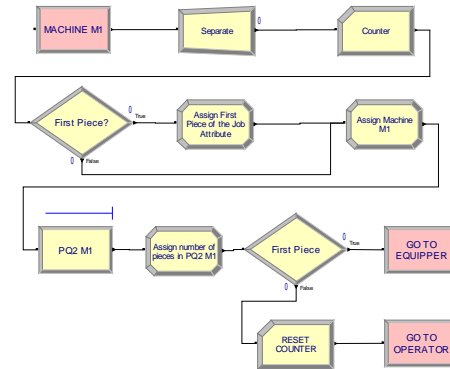


Figure 6: Machine Stations.

When a piece exits the PQ2 queue, an attribute describing the length of that queue is assigned. This will allow ordering the succeeding VQ1 or VQ2 queues on the basis of this attribute, in order to implement rules 6 and 7. After that, another Decide module sends the entity to the Equipper or the Operator Station modules, depending on the ‘First Piece’ attribute.

4.1.3. Equipper and Operator Stations

There are one Equipper Station and one Operator Station for each area of the shop floor. When a piece arrives to the Equipper Station of an area (Figure 7) it enters the *Seize* module corresponding to the VQ1 queue of the area. The order of the queue can be set according to one of the priority rules 1-7. The Seize module is associated to the Equipper resource of the area, whose capacity is set up through the *Calendar Schedule* utility. If at least one equipper is available, the piece exits the queue and seizes the equipper, which will be then released by the same entity when the action performed by the equipper (set up or inspection) has been performed. The Decide module directs the entity to the Station corresponding to the next action to be performed on the machine. This is possible thanks to the ‘current machine’ attribute (previously assigned to the entity), and another entity attribute, which is updated during the model execution, that describes the next action that the entity has to perform.

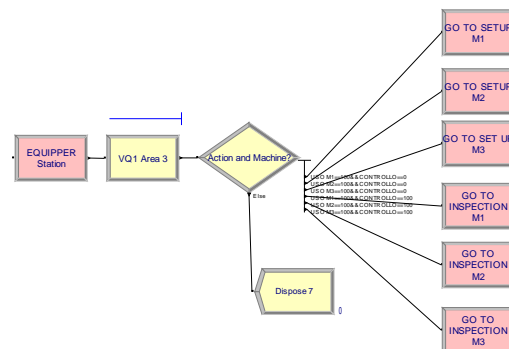


Figure 7: Equipper Stations.

If the piece is directed to the Operator Station (Figure 8), it follows a very similar path. Here the queue is the VQ2 queue, which can be ordered according to

priority rules 1-7. The associated resource that will be seized is one operator of the area.

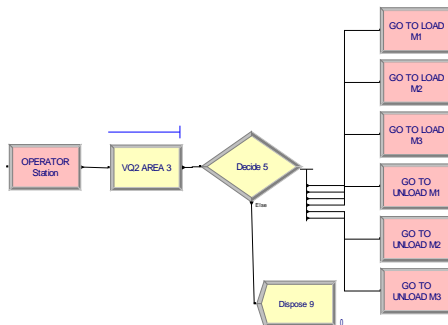


Figure 8: Operator Stations.

4.1.4. Action Stations

There are four Action Stations for each machine. Each station is related to a determined action performed by equippers or operators on the machine, namely: *set up*, *load*, *unload* and *inspection* (see Figure 9).

A piece that arrives at the *setup* Station is necessarily coming from the *equipper* Station (Figure 7), where it had seized one *equipper*. The entity now seizes the machine, and is delayed by a time equal to the *setup* time. Then it releases the *equipper* (but not the machine) and is routed toward the *Operator Station*.

A piece that arrives at the *load* Station it is necessarily coming from the *Operator Station* (Figure 8), where it had seized one *operator*. If the piece is the first one of the job, the machine must not be seized, because it has already been seized during the *setup*. The entity is here delayed for an amount of time equal to the *load* action, after which releases the *operator* (but not the machine). The succeeding delay module corresponds to the *run* action performed by the machine. An attribute specifying the next action to be performed (*unload*) is stored before the entity is routed again towards the *Operator Station*.

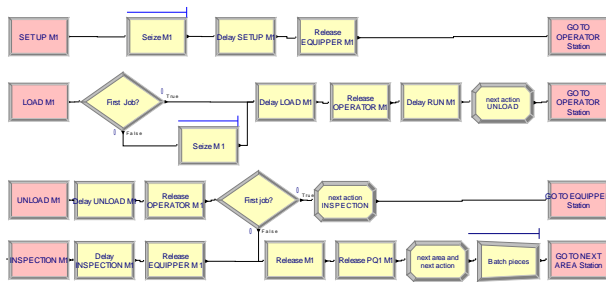


Figure 9: Actions Stations

When a piece arrives to the *unload* Station (coming from an *Operator Station*, where it had seized an *operator*), it is delayed for unloading, and then it releases the *operator*. If the piece is the first of the job, it has to be inspected by an *equipper*: the opportune next action attribute (*inspection*) is stored, and the entity is routed towards the *Equipper Station*. Otherwise the piece has finished his task in the machine: it releases the machine and then releases the virtual resource

associated to the *PQ1* queue of the machine (to allow a new piece to be loaded by an *equipper* or an *operator*, see Figure 6). The opportune ‘next action’ and ‘next area’ attributes are stored, and finally the piece enters the *batch* module.

A similar path is followed by the first piece of a job that arrives to the *inspection* Station. The only difference is that the entity is delayed for a time equal to the *inspection* time, and then releases the *equipper*.

When all the pieces of the job have been processed, the batch is ready to be routed to the next area of destination.

4.2. Verification and validation

During the time for the simulation model realization, many meetings with company’s managers have been organized. For the valid modelisation of the human resources (*operators* and *equippers*) and the possible scheduling logic that could be implemented, the continuous confrontation with company’s staff during the model development has been very profitable. In this way, the essential aspects of the scheduling and the production processes have been outlined by those which operate in the day by day operations activities in the company. This confrontation also brought to renounce adopting complicated approach that are often studied by a theoretical point of view, but that are scarcely applicable to real cases. This allowed also to gain the company’s management accreditation for the use of simulation for the specific purpose of searching for alternative scheduling techniques with the aim to reduce the jobs mean flow factor.

The conceptual model has been validated by the operational experts of the company: they confirmed that the assumptions underlying the proposed conceptual model were correct and that the proposed simulation design elements and structure (simulation’s functions, their interactions, and outputs) would have lead to results realistic enough to meet the requirements of the application. After the implementation, the same experts, comparing the responses of the simulation with expected behaviours of the system, confirmed that those responses were sufficiently accurate for the range of intended uses of the simulation.

We also verified our model through two widely adopted techniques (see Law and Kelton, 2000). The first one consists in computing exactly, when it is possible and for some combination of the input parameters, some measures of outputs, and using it for comparison. The second one, that is an extension of the first one, is to run the model under simplifying assumptions for which its true characteristics are known and, again, can easily be computed. Furthermore, in order to check the correct implementation of dispatching rules logic, the animation capability of *Arena* has also been exploited (see Figure 10).

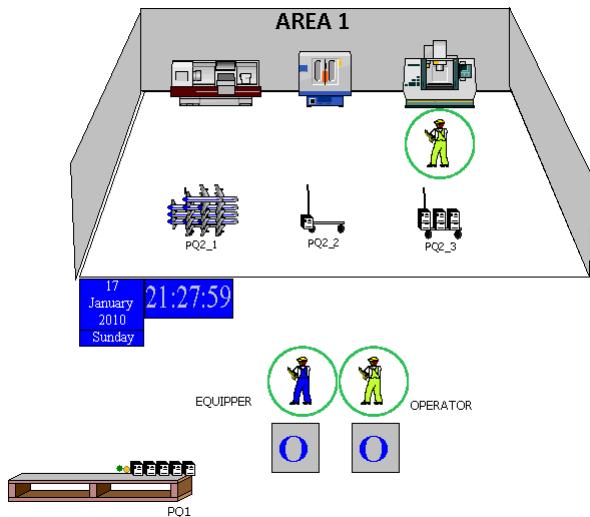


Figure 10: A screenshot of the animation.

5. THE SIMULATION EXPERIMENT

The simulation experiment is conducted using real data provided by the company, and refers to orders arrived during 4 months, for a total number of different jobs equal to 24. Processing times performed by equippers and operators have been modeled as normal distributed. Standard deviations data were available for some of the considered jobs, those ones that had already been manufactured by the company and for which work-sampling activities had already been performed. A coefficient of variation equal to 0.03 has been assumed for new jobs, accordingly to historical data related to similar jobs.

The simulation experiment is divided into two parts. In the first one it is assumed that the same decision rule is assigned both to equippers and to operators. So different scenarios have been evaluated considering all the $5 \times 7 = 35$ combinations of the 5 decisions rules for machines (queue PQ1) and the 7 decision rules for operators and equippers (queues VQ1 and VQ2). The aim is to find the best rule for machines, and then to perform the second experiment maintaining the selected machines rule fixed, and exploring all the 7×7 combinations of rules for equippers and operators. This is performed in the second part of the experiment. Each scenario has been replicated 20 times.

5.1. Performance Measures

Traditionally, the focus of performance in this type of scheduling problems has been on the Flow Time, which is defined as the amount of time that a given job spends in the system. If the i -th job arrives at time $r(i)$, has Processing Time $p(i)$ (that is known at the time of its arrival), and a Completion Time $C(i)$, its flow time will be $C(i) - r(i)$. However, Flow Time measures the time that a job is in the system regardless of the service it requests. Relying on the intuition that a job that requires a long service time must be prepared to wait longer than jobs that require small service times, practitioners and researchers have used the 'Flow Factor' (Scholtz-Reiter,

Heger and Hildebrandkt, 2009) or 'Stretch' (Bender, Muthukrishnan and Rajaraman, 2004) to measure the effect of scheduling on an individual job. The Flow Factor (or Stretch) of a job is the ratio of its Flow Time to its Processing Time: $[C(i) - r(i)]/p(i)$. Flow factor is particularly suited in this case, where multiple jobs with different processing times are considered. The mean flow factor of all the jobs has been indicated by the expert personnel of the company as the measure through which compare different scheduling combinations of dispatching rules. In particular, each combination can be compared also with the scheduling decided by the company in the same period, that obtained a mean flow factor equal to $\mu_c = 27.33$.

5.2. Results

Figure 11 shows the simulation experiment results related to the first part (same equippers and operators rules).

Table 2: The simulation experiment results (first part).

machine rule (PQ1)	equippers and operators rule (VQ1 and VQ2)	MEAN FLOW FACTOR	P-Value
FIFO	FIFO	30.14	1.000
FIFO	LIFO	29.54	1.000
FIFO	LMQL	32.42	1.000
FIFO	LPT	32.01	1.000
FIFO	RND	34.49	1.000
FIFO	SMQL	28.38	1.000
FIFO	SPT	28.92	1.000
LIFO	FIFO	32.39	1.000
LIFO	LIFO	32.84	1.000
LIFO	LMQL	38.04	1.000
LIFO	LPT	31.88	1.000
LIFO	RND	32.76	1.000
LIFO	SMQL	28.94	1.000
LIFO	SPT	29.55	1.000
LPT	FIFO	37.40	1.000
LPT	LIFO	35.63	1.000
LPT	LMQL	41.05	1.000
LPT	LPT	40.09	1.000
LPT	RND	37.65	1.000
LPT	SMQL	35.41	1.000
LPT	SPT	36.67	1.000
RND	FIFO	26.27	0.369
RND	LIFO	25.74	0.001
RND	LMQL	29.43	1.000
RND	LPT	28.42	1.000
RND	RND	26.22	0.278
RND	SMQL	28.21	1.000
RND	SPT	25.12	0.000
SPT	FIFO	24.39	0.000
SPT	LIFO	24.17	0.000
SPT	LMQL	26.16	0.170
SPT	LPT	25.94	0.018
SPT	RND	25.85	0.006
SPT	SMQL	22.77	0.000
SPT	SPT	22.00	0.000

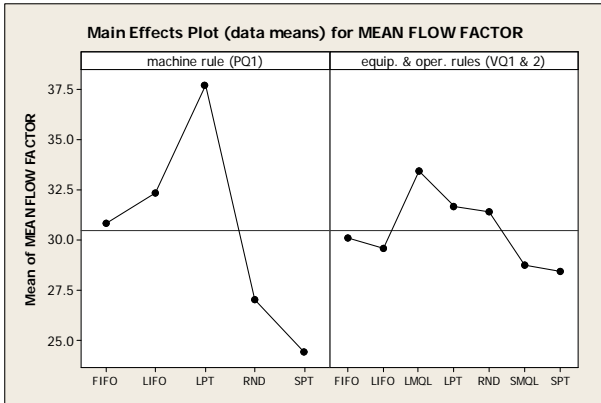


Figure 11. Main effects for Mean Flow Factor, first experiment.

The table reports the average value μ (over 20 replications) of the Mean Flow Factor for each scenario. The P-Value of the t-test in the last column indicates the smallest level of significance at which the null hypothesis ($H_0: \mu = \mu_C$) would be rejected in favor of the alternative hypothesis ($H_1: \mu < \mu_C$). The lowest value of mean flow factor is obtained when the Short Processing Time (SPT) rule is adopted for all the resources of the system (machines, operators and equippers). It is noteworthy that in the most part of scenarios obtaining a mean flow factor significantly lower than the one obtained by the company, the SPT rule is adopted for machines. Figure 11 reports the main effects plots for Mean Flow Factor, in which is also shown that SPT is the machine rule that performs better in combination with all the other equippers and operator rules.

In the second experiment the machines rule was fixed (SPT), while all the combination of equippers and operators rules have been evaluated. Results are reported in Table 3. It is easy to see that the number of scenarios with $\mu < \mu_C$ is significantly higher. The best combination is obtained when also equippers follow the SPT rule, while operators follow the SMQL rule. Main effects plots (reported in Figure 12) confirm that these rules are the ones that on the average perform better when combined with all the other rules.

Some considerations about the validity of these rules for equippers and operators can be drawn. Equippers are available only in the central part of the day (see Table 1), while machines and operators are available during all the 24 hours. An undesirable situation would be that a machine has to be set up when the equippers are not available (eg. during the night). This would cause in fact an idle time both for the machine and potentially for operators, that cannot proceed with load and unload actions on the pieces of the job. A good situation would be that the equipper set up the machine during its shift in such a way that machines and operators can continue processing the job during the night, without the need of a set-up. By giving precedence to jobs with the shortest processing time during its shift, the equipper tends to serve the longest

processing time jobs at the end of the shift. In this way there is a higher probability that jobs starting at the end of the shift will last a reasonable amount of time, and that the successive set-up will not occur just a short time after the end of the equipper shift.

Table 3. Main effects for Mean Flow Factor.

equippers rule (VQ1)	operators rule (VQ2)	MEAN FLOW FACTOR	P-Value
FIFO	FIFO	24.39	0.000
FIFO	LIFO	25.38	0.000
FIFO	LMQL	27.82	1.000
FIFO	LPT	25.19	0.000
FIFO	RND	24.47	0.000
FIFO	SMQL	23.47	0.000
FIFO	SPT	23.72	0.000
LIFO	FIFO	23.64	0.000
LIFO	LIFO	24.17	0.000
LIFO	LMQL	25.87	0.008
LIFO	LPT	25.85	0.006
LIFO	RND	26.08	0.085
LIFO	SMQL	21.78	0.000
LIFO	SPT	22.06	0.000
LMQL	FIFO	25.38	0.000
LMQL	LIFO	24.28	0.000
LMQL	LMQL	26.16	0.170
LMQL	LPT	24.59	0.000
LMQL	RND	24.91	0.000
LMQL	SMQL	23.28	0.000
LMQL	SPT	23.73	0.000
LPT	FIFO	25.79	0.003
LPT	LIFO	24.01	0.000
LPT	LMQL	27.34	1.000
LPT	LPT	25.94	0.018
LPT	RND	23.14	0.000
LPT	SMQL	24.79	0.000
LPT	SPT	25.31	0.000
RND	FIFO	25.58	0.000
RND	LIFO	23.67	0.000
RND	LMQL	28.44	1.000
RND	LPT	24.20	0.000
RND	RND	23.56	0.000
RND	SMQL	21.96	0.000
RND	SPT	22.77	0.000
SMQL	FIFO	24.64	0.000
SMQL	LIFO	25.33	0.000
SMQL	LMQL	24.65	0.000
SMQL	LPT	24.25	0.000
SMQL	RND	22.72	0.000
SMQL	SMQL	22.77	0.000
SMQL	SPT	22.59	0.000
SPT	FIFO	22.91	0.000
SPT	LIFO	23.79	0.000
SPT	LMQL	25.15	0.000
SPT	LPT	25.05	0.000
SPT	RND	23.30	0.000
SPT	SMQL	21.51	0.000
SPT	SPT	22.00	0.000

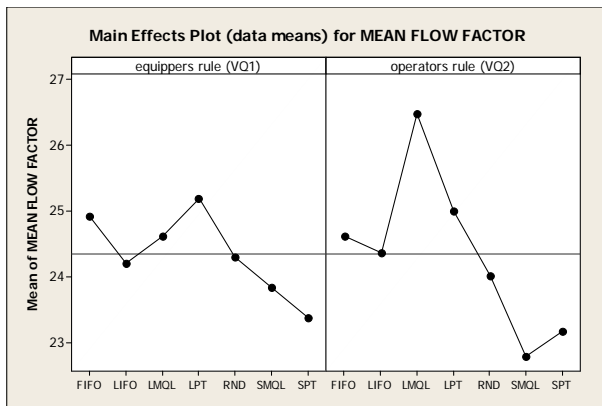


Figure 12. Main effects for Mean Flow Factor, second experiment.

As far as the SMQL rule for the operators is concerned, it is noteworthy the importance to implement a rule that is not based on a job attribute, but on a machine attribute (PQ2 queue length). On the basis of the two preceding choices dictated by machine and equipper rules, the operator has only to choose among jobs that have already been assigned to a machine and set-up. In this case, serving the machine with the lowest number of pieces in the queue means to speed up the machine release, and to favor the entering of a new job.

Possible improvements in the application of dispatching rules to this job shop scenario could be reached by allowing the selection of dispatching rules depending from time. For example, it would be possible to assign the SPT rule for the equipper in the first part of its shift and the LPT in the final part, in order to increase the probability that the successive set-up will not be needed just little time after the finish of the equipper shift. Analogously, dispatching rules for machines and operators could be differentiated in the case the equippers are present (the central shift of the day) with respect to when they are not.

6. SUMMARY

In the paper a job-shop scheduling scenario is considered, derived from a case study of a manufacturing company that works for the aeronautical industry. A conceptual model of the shops has been built in order to implement a priority rules approach for the simultaneous scheduling of machines and two types of human resources: equippers and operators. The modelisation through virtual and physical queues allowed to define rules that are related both to jobs attributes and to machines attributes, and facilitated the implementation of a simulation model. Different combination of priority rules for machines, equippers and operators have been simulated and results have been compared on the basis of the mean flow factors of the considered jobs.

Results have been compared to the mean flow factor obtained by the company for the same input data, and allow identifying the best rules combination that

over performs the company's scheduling. The approach allows to gain insights into priority rule performance, and to individuate a simple and implementable scheduling logic that provides a completely reactive scheduling.

Aknolwedgments

Authors thanks Dr. Cristiano Antinori for his supporting activity during the model implementation.

REFERENCES

- Bender, M.A., Muthukrishnan, S., Rajaraman, R., 2004. Approximation algorithms for average stretch scheduling. *Journal of Scheduling*, 7, 195-222.
- Blackstone, J.H., Philips, D.T., Hogg, G.L., 1982. A state-of-the-art survey of dispatching rules for manufacturing job shop operations. *International Journal of Production Research*, 20(1), 27-45.
- Cowling, P. I., Johansson, M. (2002). Using real-time information for effective dynamic scheduling. *European Journal of Operational Research*, 139(2), 230-244.
- Dauzère-Pérès, S., Roux, W., Lasserre, J.B., 1998. Multi-resource shop scheduling with resource flexibility. *European Journal of Operational Research*, 107(2), 289-305.
- ElMaraghy, H., Patel, V., Abdallah, I.B., 2000. Scheduling of manufacturing systems under dual-resource constraints using genetic algorithms. *Journal of Manufacturing Systems*, 19(3), 186-201.
- Geiger, C., Uzsoy, R., Aytu, H., 2006. Rapid modeling and discovery of priority dispatching rules: an autonomous learning approach. *Journal of Scheduling*, 9, 7-34.
- Geiger, D., Uzsoy, R., 2006. Learning effective dispatching rules for batch processor scheduling. *International Journal of Production Research*, 46(6), 1413-1454.
- Haupt, R., 1989. A survey of priority rule-based scheduling. *OR Spectrum*, 11, 3-16.
- Laarhoven, P.J.M.v., Aarts, E.H.L., Lenstra, J.K., 1992. Job shop scheduling by simulated annealing. *Operations Research*, 40(1), 113-125.
- Manikas, A., Chang, Y., 2009. Multi-criteria sequence-dependent job shop scheduling using genetic algorithms. *Computers & Industrial Engineering*, 59(1), 179-185.
- Matie, Y., Xie, X., 2008. A genetic-search-guided greedy algorithm for multi-resource shop scheduling with resource flexibility. *IIE Transactions*, 40(12), 1228-1240.
- Ouelhadj, D., Petrovic, S., 2009. A survey of dynamic scheduling in manufacturing systems. *Journal of Scheduling*, 12, 417-431.
- Rajendran, C., Holthaus, O., 1999. A comparative study of dispatching rules in dynamic flowshops and

- jobshops. *European Journal of Operational research*, 116, 156-170.
- Scholz-Reiter, B., Heger, J., Hildebrandt, T., 2009. Analysis and comparison of dispatching rule-based scheduling in dual-resource constrained shop-floor scenarios. *Proceedings of the World Congress on Engineering and Computer Science*, pp. 921-927. October 20-22, San Francisco (California, USA).
- Zhang, C., Li, P., Guan, Z., Rao, Y., 2007. A tabu search algorithm with a new neighborhood structure for the job shop scheduling problem. *Computers & Operations Research*, 34(11), 3229-3242.
- Zhou, H., Cheung, W., Leung, L.C., 2009. Minimizing weighted tardiness of job-shop scheduling using a hybrid genetic algorithm. *European Journal of Operational Research*, 194(3), 637-649.

AUTHORS BIOGRAPHY

Lorenzo Tiacci. Laurea Degree in Mechanical Engineering, doctoral Degree in Industrial Engineering, he is Assistant Professor at the Department of Industrial Engineering of the University of Perugia. He is currently teaching courses of Facilities Planning & Design, Production Planning and Control, and Project Management at the University of Perugia. His research activity covers modeling and simulation of logistic and productive processes, plants design, production planning and inventory control, supply chain management, transportation problems.

Stefano Saetta. Stefano SAETTA is Associate Professor at the Engineering Faculty of the University of Perugia. His research fields covers essentially the following subjects: modelling and simulation of logistic and productive processes, methods for the management of life cycle assessment, discrete event simulation, supporting decision methods, lean production. He was involved in several national and international research projects He is the organising committee and in the scientific committee of many international conferences.

EFFECT OF REJECT OPTION ON CLASSIFIER PERFORMANCE

S. Dreiseitl^(a), M. Osl^(b)

^(a)Upper Austria University of Applied Sciences at Hagenberg, Austria

^(b)Division of Biomedical Informatics, UCSD, USA

^(a)stephan.dreiseitl@fh-hagenberg.at, ^(b)mosl@ucsd.edu

ABSTRACT

Binary classifier systems that provide class membership probabilities as outputs may be augmented by a reject option to refuse classification for cases that either appear to be outliers, or for which the output probability is around 0.5. We investigated the effect of these two reject options (called “distance reject” and “ambiguity reject”, respectively) on the calibration and discriminatory power of logistic regression models. Outliers were found using one-class support vector machines. Discriminatory power was measured by the area under the ROC curve, and calibration by the Hosmer-Lemeshow goodness-of-fit test. Using an artificial data set and a real-world data set for diagnosing myocardial infarction, we found that ambiguity reject increased discriminatory power, while distance reject decreased it. We did not observe any influence of either reject option on the calibration of the logistic regression models.

Keywords: classifier systems, reject option, performance evaluation

1. INTRODUCTION

Decision support systems in biomedicine can augment a physician’s diagnostic capabilities by providing an automated second opinion. There are a number of approaches to building such systems, ranging from capturing an expert’s domain knowledge in explicit form to using machine learning methods that learn a model from given data without additional human intervention.

Here, we consider only systems of the second kind, and further restrict our attention to models that distinguish between two classes (e.g., classifying cases as either healthy or diseased). Some of these systems, such as logistic regression or neural network models, provide explicit class membership probabilities, i.e., their output is a measure to which degree a case is healthy or diseased. Other machine learning models, such as support vector machines, must be explicitly augmented to provide probability outputs.

The advantages of probability outputs are numerous: Besides facilitating accurate assessments of the system’s discriminatory power and calibration via ROC analysis

(Bradley 1997; Fawcett 2006) and goodness-of-fit tests (Hosmer et al. 1997; Pigeon and Heyse 1999), probability estimates can also be used for implementing a *reject option*. Such an option allows the system to refrain from making a decision if the predicted membership probability for both classes is around 50%, i.e., if the system cannot make a decision with a reasonable level of certainty.

In addition to rejecting uncertain cases, it may also be desirable for a decision support system to make recommendations only for cases that are similar to the ones that were used for building it. This goal is more difficult to achieve, because it involves estimating how similar a new case is to a set of previously known cases.

In the literature (see Section 2.), the first reject option (around probabilities of 50%) is known as *ambiguity reject option*, and the second (for outlier cases) as *distance reject option*.

In this work, we investigate to which extent the ambiguity and distance reject options have an influence on the quality of a classifier’s performance, as measured by its discriminatory power and calibration. For the ambiguity reject option, we use a logistic regression model and do not classify cases for which the model output is close to 0.5. For the distance reject option, we additionally use a one-class SVM to estimate the regions in input space where most of the cases lie. The points outside these regions are then considered to be outliers and rejected from classification.

2. PREVIOUS WORK

The idea of not classifying cases in regions of substantial class overlap, and thus class membership probabilities of around 50%, was proposed by Chow (1970), who was also the first to conduct an investigation into the benefits of the reject rule from a theoretical point of view. Dubuisson and Masson (1993) were the first to consider distance rejection, using nearest neighbor distances to decide whether a point is too far from the remainder of a data set. The particular case of a reject option for nearest neighbor classifiers had been studied earlier by Hellman (1970). Muzzolini et al. (1998) noted that rejection thresholds have to be adjusted to the covariance structure of mixture models to be unbiased, and proposed a method for performing this adjustment. The work of Landgrebe

et al. (2004, 2006) focused on the distance reject option when the classification task is ill-defined in the sense that one clearly defined target class is to be distinguished from another poorly defined class in the presence of an unknown third outlier class. Tax and Duin (2008) proposed a novel method for performing classification with a distance reject option by combining multiple one-class models, one for each of the individual classes.

In the statistical literature, there is some theoretical research on the effects of reject options when rejection costs are different from misclassification costs. The framework of empirical risk minimization provides a theoretical background for the works of Herbei and Wegkamp (2006), Yuan and Wegkamp (2010), and Bartlett and Wegkamp (2008), who derived an SVM classifier with a reject option.

3. METHODS

In this section, we first describe the algorithms we used for building machine-learning models, and then the methods for evaluating the performance of these algorithms.

3.1. Machine learning algorithms

We consider dichotomous classification problems as specified by an n -element data set of m -dimensional input vectors x_1, \dots, x_n and corresponding class labels $y_1, \dots, y_n \in \{-1, 1\}$. For logistic regression, we assume there is an additional constant 1 at the first position of the x_i in order to simplify the notation below; these augmented data points are thus $(m + 1)$ -dimensional.

In a logistic regression model, the optimal values for the $(m + 1)$ -dimensional parameter vector β are determined by minimizing a negative log-likelihood function. We additionally consider L_2 -regularization of logistic regression models by calculating the maximum likelihood estimate β_{ML} as

$$\beta_{\text{ML}} = \arg \min_{\beta} \sum_{i=1}^n \log(1 + e^{-y_i \beta^T \cdot x_i}) + \lambda \beta^T \beta.$$

The regularization parameter λ is usually chosen by cross-validation.

The model predictions for new cases x are then given as class-membership probabilities

$$P(y = +1 | x, \beta_{\text{ML}}) = \frac{1}{1 + e^{-\beta_{\text{ML}}^T \cdot x}}.$$

The model outputs are thus logistic transformations of $\beta_{\text{ML}}^T \cdot x$, i.e., values proportional to the distance of x from the hyperplane parameterized by β_{ML} . Ambiguity rejection can therefore be seen to refuse classification for those cases that are within a certain distance from the separating hyperplane.

For distance rejection, we need to estimate the regions of input space in which data are more dense than in others. Standard parametric and non-parametric density estimation algorithms are susceptible to the curse of dimensionality, and therefore not easily applicable in high

dimensions. A recent addition to the machine learning arsenal allows us to address this problem without regard to data dimensionality (Schölkopf et al. 2001; Schölkopf and Smola 2002). *One-class support vector machines* extend standard support vector machine (SVM) methodology to the case of estimating a given fraction $(1 - \nu)$ of the support of a data set; the remaining fraction ν are considered outliers.

As with other support vector methods, the one-class SVM algorithm projects the data into a different feature space F using a nonlinear mapping $\Phi: \mathbb{R}^n \rightarrow F$. Without a second class, the aim is then to separate the projected data from the origin by as wide a margin ρ as possible. The use of kernel functions k to replace projections and dot product operations is similar to other SVM algorithms. We used a Gaussian kernel

$$k(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

with inverse variance parameter γ . Values of γ were chosen in such a way that the proportion of outliers was close to ν .

One-class SVMs estimate the data distribution by solving the constrained optimization problem

$$\begin{aligned} \min_{w \in F, \xi_i \in \mathbb{R}^m, \rho \in \mathbb{R}} \quad & \frac{1}{2} \|w\|^2 + \frac{1}{n\nu} \sum_{i=1}^n \xi_i - \rho \\ \text{subject to} \quad & w \cdot \Phi(x_i) \geq \rho - \xi_i, \\ & \xi_i \geq 0 \quad \forall i = 1, \dots, n \end{aligned}$$

where w is the parametrization of the separating hyperplane in F , and the ξ_i are slack variables. The dual problem is

$$\begin{aligned} \min_{\alpha_i \in \mathbb{R}^n} \quad & \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \frac{1}{n\nu} \quad \forall i = 1, \dots, n \\ & \sum_{i=1}^n \alpha_i = 1. \end{aligned}$$

Support vectors are those data points x_i for which the corresponding α_i satisfies $0 < \alpha_i < \frac{1}{m\nu}$. Outliers are those points for which the decision function

$$f(x) = \sum_{i=1}^n \alpha_i k(x_i, x) - \rho$$

is negative.

3.2. Evaluation metrics

We used the area under the ROC curve (AUC) as measure of a classifier's discriminatory power, and computed an estimator $\hat{\theta}$ of the AUC via its equivalence to a Mann-Whitney U-statistic as

$$\hat{\theta} = \frac{1}{n_1 \cdot n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(\mathbb{1}(p_i^- < p_j^+) + \frac{1}{2} \mathbb{1}(p_i^- = p_j^+) \right).$$

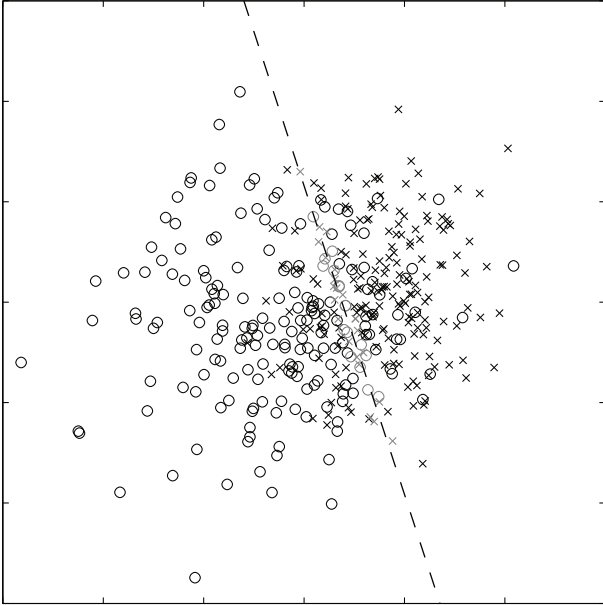


Figure 1: Sample of the artificial data set showing two normally distributed classes with the logistic regression discriminatory line. The points for which no classification is made based on ambiguity rejection are located close to the discriminatory line, and shown in light grey.

Here, p_i^- and p_j^+ are the classifier outputs for cases from classes -1 and $+1$, respectively, and $\mathbb{1}$ is the Boolean indicator function.

The calibration of a classifier is usually assessed with the Hosmer-Lemeshow C-test (Hosmer and Lemeshow 1980). Although often criticised for a number of drawbacks (Bertolini et al. 2000), it is nevertheless the de-facto standard for determining the goodness-of-fit of a model. As a Pearson chi-squared test, it computes the test statistic

$$C = \sum_{i=1}^G \frac{(O_i - E_i)^2}{E_i(1 - \frac{E_i}{n_i})},$$

as the sum of standardized squared differences between the number of observed cases O_i and expected cases E_i for a grouping of classifier outputs into G groups, each with n_i cases. By definition, $G = 10$ and the data is grouped by sorted classifier outputs. Hosmer and Lemeshow (1980) observed that C has an approximate chi-squared distribution with $G - 2$ degrees of freedom.

4. EXPERIMENTS

Our experiments on the effect of the reject option on classifier performance utilized two data sets, one simple artificial toy problem, and one real-world data set from the domain of predicting acute myocardial infarction.

4.1. Data sets

For the artificial data set, we generated 500 data points each from two multivariate normal distributions with diagonal covariance matrices, each representing one of the

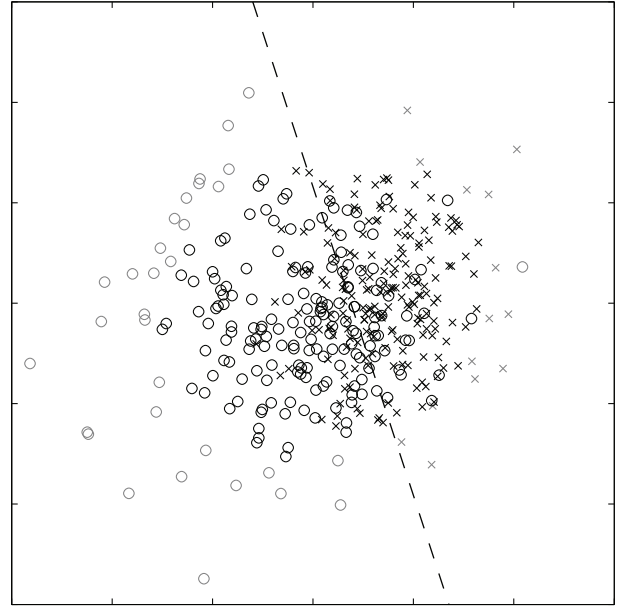


Figure 2: Sample of the artificial data set showing two normally distributed classes with the logistic regression discriminatory line. The points for which no classification is made based on distance rejection are located at the outer edge of the data set, and shown in light grey.

two classes. The parameters were chosen to achieve an AUC of about 0.9. A sample of the data with the separating line as determined by logistic regression, along with points in the ambiguity reject region, is shown in Figure 1. The same data, now with points rejected based on distance highlighted, is presented in Figure 2.

The myocardial infarction data set consists of information collected from 1253 patients presenting at the emergency department of the Edinburgh Royal Infirmary in Scotland with symptoms of acute myocardial infarction (AMI). A total of 39 features were recorded, comprising patient data (smoker, diabetes, ...), clinical information (location of pain, sensation of pain, hypertension, ...), and results of an ECG test (LBB, abnormal T wave, ...). To increase diagnostic difficulty, we removed data about ECG measurements, retaining a total of 33 features. The gold standard diagnosis was made by expert physicians based on a combination of blood serum tests with clinical and ECG data. Of the 1253 patients, 274 were diagnosed with AMI, and 979 patients were either declared healthy or to be suffering from other ailments.

4.2. Results

Our experiments were carried out using MATLAB (MathWorks, Natick, MA), with our own implementation of logistic regression models and the libsvm implementation of one-class SVMs (Chang and Lin 2001). For both the artificial as well as the myocardial infarction data set, we trained the logistic regression models using 60% of the data, with the remaining 40% reserved for testing. All data features were normalized to zero mean and unit variance. The experiments were performed 50 times, each

Table 1: Discrimination and classification of logistic regression models on the artificial data set. The results for ambiguity and distance reject options are listed by varying fractions of rejected cases. HL denotes the Hosmer-Lemeshow test statistic; the critical value for $\alpha = 0.05$ is 15.51.

	AUC		HL	
	mean	std	mean	std
logistic regression (baseline)	0.880	0.01	15.07	7.16
ambiguity reject				
$\tau = 0.1$	0.897	0.01	15.22	8.58
$\tau = 0.2$	0.913	0.01	13.93	9.35
$\tau = 0.3$	0.925	0.01	14.73	7.82
$\tau = 0.4$	0.936	0.01	14.82	9.61
distance reject				
$\nu = 0.05$	0.871	0.01	13.93	6.49
$\nu = 0.1$	0.862	0.02	13.10	5.99
$\nu = 0.2$	0.851	0.02	9.65	3.95

with different random allocations of data to the training and test sets. All results are reported as averages and standard deviations on the test set over these 50 runs.

The parameters of our experiments were ν , the fraction of outliers for the distance reject option, and τ , the fraction of cases for the ambiguity reject option. The ambiguity reject cases were the proportion τ of cases for which the model output (class membership probability) was closest to 0.5. A kernel parameter of $\gamma = 0.001$ gave a number of outliers within 10% of the desired value, as specified by ν . The number of support vectors was slightly larger. This is in concordance with theory, which states that ν is an upper bound on the fraction of outliers, but a lower bound on the fraction of support vectors (Schölkopf et al. 2001).

The results of our experiments on the artificial data set are summarized in Table 1. One can observe that ambiguity rejection had a positive effect on AUC. This is to be expected, because ambiguity rejection removes those cases for which most misclassification errors occur. Furthermore, it is also reasonable that the increase in AUC is not as pronounced for larger values of τ , because fewer and fewer ambiguous cases get removed.

On the other hand, there seemed to be no effect of τ on the value of the Hosmer-Lemeshow (HL) test statistic. As is known from the literature (Bertolini et al. 2000), this test depends strongly on the particular grouping of data points, and showed high volatility in our experiments as well (as indicated by the large standard deviations).

As for distance rejection, the AUC value decreased with increasing numbers of rejected cases, while the HL test statistic showed better model fit. A possible explanation for the first phenomenon is the fact that the rejected points were almost all correctly classified by the model

Table 2: Discrimination and classification of logistic regression models on the myocardial infarction data set. The results for ambiguity and distance reject options are listed by varying fractions of rejected cases. HL denotes the Hosmer-Lemeshow test statistic; the critical value for $\alpha = 0.05$ is 15.51.

	AUC		HL	
	mean	std	mean	std
logistic regression (baseline)	0.842	0.12	14.51	9.03
ambiguity reject				
$\tau = 0.1$	0.851	0.12	14.10	11.47
$\tau = 0.2$	0.860	0.13	14.29	10.98
$\tau = 0.3$	0.874	0.13	13.14	10.10
$\tau = 0.4$	0.886	0.13	12.19	8.76
distance reject				
$\nu = 0.05$	0.841	0.12	19.38	20.39
$\nu = 0.1$	0.838	0.12	22.61	19.29
$\nu = 0.2$	0.834	0.12	34.81	21.23

(because they were on the correct side of the discrimination line). Removing them therefore had a detrimental effect on the AUC. The second phenomenon may just be a fluke observation, as evidenced by the high standard deviations and the fact that it was not present in the real-world data.

Table 2 provides the same information for the myocardial infarction data set. Again, we find that ambiguity rejection increased AUC without a clear effect on the HL test statistic (which exhibits even higher standard deviations than on the artificial data set). And again, we observed that distance rejection had a negative effect on AUC. The effect on the HL test statistic was in the opposite direction of the effect it had on the artificial data set.

Distance rejection is also not beneficial if it is performed *before* training, as suggested by Landgrebe et al. (2006). In this case, one rejects data from the training set, and not from the test set. The reasoning for this is that the model may be a better representation of the underlying data generator when outliers are removed prior to model building. Table 3 shows that this is not the case: There was no difference in AUC for the artificial data set, and an even larger negative effect for the real-world data set. There were no discernible effects on the HL test statistics.

5. CONCLUSION

We investigated the effect of the ambiguity and distance reject options on performance of a logistic regression model on an artificial and a real-world data set. We observed ambiguity rejection to increase AUC, and distance rejection to decrease it. Both reject options did not have an effect on classifier calibration.

Table 3: Discrimination and classification of logistic regression models on the artificial and on the myocardial infarction data set, when a fraction ν of cases are removed by distance rejection from the training set prior to model building. HL denotes the Hosmer-Lemeshow test statistic; the critical value for $\alpha = 0.05$ is 15.51.

	AUC		HL	
	mean	std	mean	std
artificial data				
logistic regression (baseline)	0.880	0.01	15.07	7.16
$\nu = 0.1$	0.880	0.01	14.48	6.92
$\nu = 0.2$	0.880	0.01	15.61	7.73
myocard. inf. data				
logistic regression (baseline)	0.842	0.12	14.51	9.03
$\nu = 0.1$	0.828	0.12	15.20	11.75
$\nu = 0.2$	0.818	0.12	14.28	10.40

ACKNOWLEDGEMENTS

This work was funded in part by the Austrian Genome Program (GEN-AU), project Bioinformatics Integration Network (BIN) and the National Library of Medicine (R01LM009520).

REFERENCES

- Bartlett, P. and Wegkamp, M. (2008). Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(Aug):1823–1840.
- Bertolini, G., D’Amico, R., Nardi, D., Tinazzi, A., and Apolone, G. (2000). One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. *Journal of Epidemiology and Biostatistics*, 5(4):251–253.
- Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159.
- Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Chow, C. (1970). On optimum error and reject tradeoff. *IEEE Transactions on Information Theory*, IT-16(1):41–46.
- Dubuisson, B. and Masson, M. (1993). A statistical decision rule with incomplete knowledge about classes. *Pattern Recognition*, 26(1):155–165.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.
- Hellman, M. (1970). The nearest neighbor classification rule with a reject option. *IEEE Transactions on Systems Science and Cybernetics*, SSC-6(3):179–185.
- Herbei, R. and Wegkamp, M. (2006). Classification with reject option. *Canadian Journal of Statistics*, 4(4):709–721.

- Hosmer, D. and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, A10:1043–1069.
- Hosmer, D. W., Hosmer, T., le Cessie, S., and Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16:965–980.
- Landgrebe, T., Tax, D., Paclík, P., and Duin, R. (2006). The interaction between classification and reject performance for distance-based reject-option classifiers. *Pattern Recognition Letters*, 27(8):908–917.
- Landgrebe, T., Tax, D., Paclík, P., Duin, R., and Andrew, C. (2004). A combining strategy for ill-defined problems. In *Proceedings of the 15th Annual Symposium of the Pattern Recognition Association of South Africa*, pages 57–62.
- Muzzolini, R., Yang, Y.-H., and Pierson, R. (1998). Classifier design with incomplete knowledge. *Pattern Recognition*, 31(4):345–369.
- Pigeon, J. G. and Heyse, J. F. (1999). An improved goodness of fit statistic for probability prediction models. *Biometrical Journal*, 41(1):71–82.
- Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471.
- Schölkopf, B. and Smola, A. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.
- Tax, D. and Duin, R. (2008). Growing a multi-class classifier with a reject option. *Pattern Recognition Letters*, 29(10):1565–1570.
- Yuan, M. and Wegkamp, M. (2010). Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11(Jan):111–130.

AUTHOR BIOGRAPHIES



STEPHAN DREISEITL received his MSc and PhD degrees from the University of Linz, Austria, in 1993 and 1997, respectively. He worked as a visiting researcher at the Decision Systems Group/Harvard Medical School before accepting a post as professor at the Upper Austria University of Applied Sciences in Hagenberg, Austria, in 2000. He is also an adjunct professor at the University of Health Sciences, Medical Informatics and Technology in Hall, Austria. His research interests lie in the development of machine learning models and their application as decision support tools in biomedicine.



MELANIE OSL received her Ph.D. degree from the University of Medical Informatics and Technology (UMIT) in Hall, Austria in 2007. She is currently a postdoc fellow at the Division of Biomedical Informatics at the University of California, San Diego. Her research interests include knowledge discovery and data mining in biomedicine, clinical bioinformatics and machine learning.

NEW DISCRETE TOPOLOGY OPTIMIZATION METHOD FOR INDUSTRIAL TASKS

Sierk Fiebig^(a)

^(a)Volkswagen Braunschweig, Braunschweig, Germany

^(a)sierk.fiebig@volkswagen.de

ABSTRACT

Nowadays the development of mechanical components is driven by ambitious targets. Engineers have to fulfill technical requirements under the restrictions of reducing costs and weights simultaneously. Therefore in the last years optimization methods have been integrated in the development process of industrial companies. Today, especially topology optimization methods, have gained in importance and are standard for developing casting parts. Stress or strain-energy information is used for sensitivities in all topology optimization methods. The method SIMP, today's standard in industry, uses continuous material modeling and gradient algorithms. ESO/BESO use discrete modeling and specific algorithms depending on the individual approaches. The new Topology Optimization method uses a discrete modeling, too. The number of modified elements is controlled by the progress of the constraint.

For solving tasks in the industrial development process, a topology optimization method must enable an easy and fast usage and must support manufacturing restrictions.

Keywords: topology optimization, mechanical components, discrete modeling of material

1. INTRODUCTION

Today several approaches exist for topology optimization. The starting point of FEA based topology optimization was at the end of the eighties [Roz01]. Bendsøe introduced first his homogenization method [Ben89]. Parallel to the homogenization method, Bendsøe presented the SIMP approach (Solid Isotropic Microstructure with Penalization) [BenSig03]. This method has become popular, because other researchers use it [Roz92]. Today the SIMP approach is one of the standard methods for topology optimization. For example, the commercial tool Tosca[®] from FE-Design is based on SIMP. SIMP uses continuous design variables. Here the density is used as design variable. The coupled Young-Modulus transfers the modifications of the optimization to the structure results. At the end of each topology optimization, a clear discrete distribution for interpreting the results is needed. Due to this, the SIMP approach penalizes intermediate density values using a penalization factor as a power. In this way, low stiffness values are

assigned to intermediate density values [Edw07]. SIMP is combined to a gradient algorithm, e.g. the method of moving asymptotes [Svan87].

Since 1992 another important approach has been developed. The evolutionary structural optimization (ESO) is focused to remove unnecessary material from too conservative designed parts [Que00]. To ESO, it is only possible to remove material and uses a discrete element modeling in comparison to SIMP [HuaXie10]. To enable the opposite, Querin introduces the additive evolutionary structural optimization method, called AESO [Que00b]. AESO adds material to the highest stressed points in order to become an optimal structure. The combination of ESO and AESO is the bidirectional evolutionary structural optimization [BESO] method [Que00] [HuaXie10]. The main idea behind ESO, AESO and BESO is to remove under stressed elements and to add material to higher stressed areas. To designate these elements, two reference levels are defined. During the optimization these levels are adapted to the optimization progress.

All elements under a reference level are removed and all elements above a second level are added. BESO uses here - depending on the individual approach - direct, gradient or interpolated information about material properties to change the structure [HuaXie10].

For industrial usage the SIMP method in combination with gradient algorithm has a large distribution. One main reason for the success of the approach is the integration of manufacturing restrictions. Without these restrictions, it isn't possible, in most cases to get a feasible design for real life problems. At the moment no proposals for the integration of manufacturing restriction to BESO are offered.

2. THE NEW APPROACH FOR TOPOLOGY OPTIMIZATION

The motivation for the new approach is based on three reasons. The main focus is the usage of the method in industry. The overall interest of industry is to recognize parts with lower weight and cost compared to the older reference structure. In contrast to optimization from a mathematical or theoretical view, the task of optimization isn't to find the absolute optimum. In the opinion of engineering and praxis, optimization means the improvement of the result.

This mean, that better optimization results are the first motivation for the new method.

To achieve this and to improve the universal usage, linear and nonlinear FEA analysis should be possible with the new Topology Optimization method. Nonlinear effects are for example plastic behavior of material, nonlinear behavior in bushing and contact problems. Finally the last point, manufacturing requirements should be fulfilled.

2.1. Basic functionalities similar to ESO/BESO

Using stress or strain-energy information for sensitivities are the basic ideas in all topology optimization methods, see [BenSig03] [HuaXie10] [Mat94]. Depending on this main idea, the new approach uses the stress-values for reducing or adding discrete material in the design space. Another important similarity is the discrete modeling of material.

Following ESO and BESO, the lowest stressed elements are removed from the structure. This is a simple but effective method. This mechanism has also an analogy in nature, during the development and growth of plants [Mat94] and is rooted from experience for solving problems in engineering also.

Due to the fact of discrete material, new elements can only be added to the borders of an existing structure. Without interpolation information the new material is placed to the areas with the highest stress levels, see the AESO method of Querin [Que00].

Depending on the discrete modeling, both methods are possible to handle linear and nonlinear effects in the FEA analysis. The only difference to a regular FEA simulation lies in the surface of the FEA model. Up to now the models from topology optimizations with discrete modeling is not as smooth as a model from a regular simulation.

2.2. Main differences to ESO/BESO

The new Topology Optimization has beside some similarities clear differences to the ESO/BESO methods. The main idea of ESO/BESO is a full stressed design, means all elements receive the same stress level. For this method the compliance-volume product can be assumed as an objective function [Edw07]. Opposite to this, the new Topology Optimization method uses only the volume as target or object function.

For the optimization the new method needs constraints. Remembering the motivation, the new Topology Optimization allows several constraints, e.g. displacement or reaction force. Also the combination of all constraints is possible. Normally a min-max formulation is used. But also other mathematical operators are possible to use, e.g. weighed or distance formulations. In the original approach of BESO, no constraints are used. Only the stress levels are important.

With the main focus to a normalized stress level, BESO adds material by comparing each element stress level to a reference level. Comparing this to the new Topology Optimization method more elements are added in each

iteration. The reason is that the new Topology Optimization method adds only at the highest stressed elements(often called hotspots) material.

Starting optimization with infeasible solutions forces the optimization method to add material first to the structure. The BESO method finds the same solution in this case as an optimization run starting from full design space [Que00b]. The new approach offers in this case different solutions, because the process is controlled by the constraint limit. For industrial purposes this behavior is more powerful in later development phases, e.g. when load conditions must be changed to new requirements, the engineer wants to find a new feasible and as light as possible design but with a minimum of changes in the part.

2.3. Main process of the new Topology Optimization method

The flow chart in figure 1 illustrates the main steps of the new Topology Optimization method. The step size controller calculates first a basic rate. Depending on this basic rate, the number of removing and adding elements is defined. After the controller the necessary elements are inserted. In this way, hotspot areas are corrected. After this correction process, the lowest stress elements according to the reduction rate are removed. After adding and removing elements, it is important to check if the structure is connected. All force transmission points must be connected to the supports. If this check fails, the controller modifies the correction and reduction rate in order to produce a feasible structure. In the heuristic steps, non connecting elements are removed from the structure.

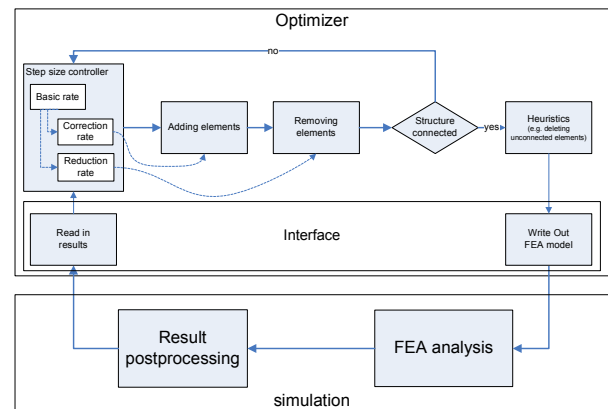


Figure 1: Flow chart of New Topology Optimization Method

The necessary interfaces to the FEA solver are integrated in the optimizer. After finishing all changes and checks, the optimizer writes the element input decks. After the FEA analysis, the result postprocessing evaluates all target functions and constraints. The read in process transfers this information back to the controller.

2.4. Integration into the industrial development process

Several steps are needed for the procedure of a topology optimization. Normally a topology optimization is based on FEA analysis. Due to this, the topology optimization must be coupled with a FEA solver.

One basic idea of this new approach is the integration in the standard development process, especially simulation process. Through this, an external FEA solver should be used. This demands interfaces to read and write the special formatted input decks of the solver. The optimizer supports two FEA solvers: Abaqus from Simulia® and Nastran.ND® from MSC®. Other FEA solvers can be integrated. Only the necessary interfaces have to be programmed in C++.

To minimize the complexity of the development, the new Topology Optimization doesn't manage the process of the topology optimization. The workflow is controlled through an external program, such as Optimus® from Noesis®.

2.4.1. Preprocessing

The preprocessing can be divided into two parts. First the normal FEA preprocessing has to be done. In figure 2, the two steps: meshing the part and define loads and structure supports are illustrated.

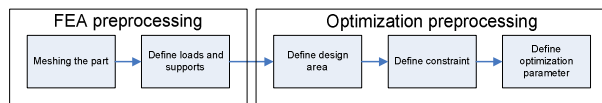


Figure 2: Preprocessing

After this, for the optimization preprocessing, different files in ASCII® format are used. The design areas, the constraint and the optimization parameter are chosen. All is flexible and can be adapted to the specific problem.

2.4.2. Interface between the optimizer and the FEA solver

The optimizer works internal with a data grid, see figure 3. The information from the internal data grid, called "matrix", can be transferred to the FEA model. On the initial run of the optimizer the elements are mapped to the matrix. This mapping is fixed over the whole optimization. The optimizer changes the status of the matrix. Status 0 means no material, Status 1 means material. After finishing the optimization steps, this information is mapped to the FEA model.

Only the elements with status 1 are written to the FEA data file. No other elements are available for the FEA solver.

Therefore, the results of the FEA analysis, apart from the aliasing effects at the border, have the same result quality as a normal analysis.

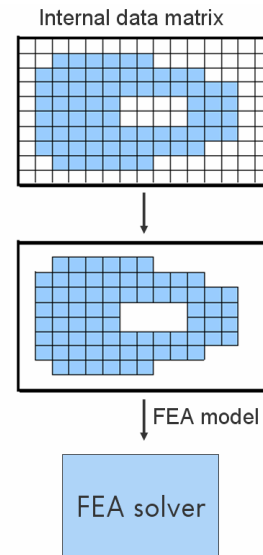


Figure 3: Interface between optimizer and FEA solver

2.4.3. Optimization Workflow

For the workflow Optimus® is used. The optimizer is integrated as User Algorithm. The optimizer writes the FEA input decks on his own. To Optimus a reference to this file is transferred. During the process, Optimus transfers files, starts the FEA solver and the postprocessing scripts to evaluate the stress values and the constraints. The process is shown in figure 4.

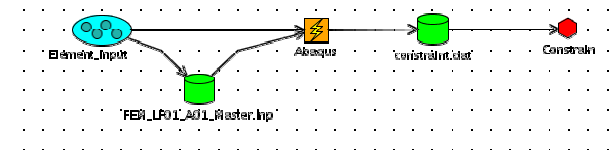


Figure 4: Preprocessing

2.4.4. Interface between the FEA solver and the optimizer

The figure 5 illustrates the process between FEA postprocessing and the interface of the optimizer.

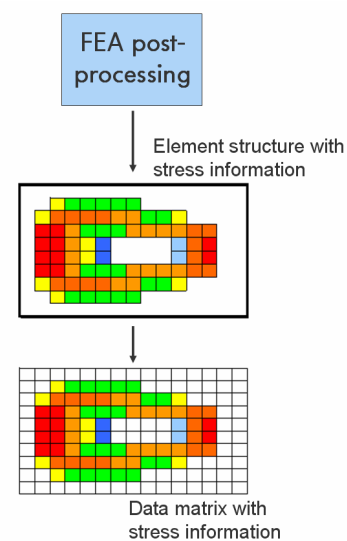


Figure 5: Postprocessing and interface to the optimizer

The result of the postprocessing is a list of all elements. Each line of this list represents one element. The line begins with the element id. Second entry is the stress value of the element.

The first step of the interface maps the stress value of each element to the internal element list.

After this, the stress values of the elements are mapped to the internal data matrix.

2.5. Integration of manufacturing restriction

For a feasible industrial casting part design, numerous manufacturing restrictions have to be fulfilled. Besides minimum and maximum material strength, normally a forming direction has to be taken into account. Special production processes - especially forging - need closed structures, at best, without any holes.

Additionally it is sometimes necessary to design symmetric parts, maybe for using the same part on the left and right side of a car. As well as a minimum strength restriction, casting directions, forging and symmetry restrictions are implemented.

2.5.1. Casting direction

Due to a casting direction, no material inside the structure can be deleted. In this way, no undercuts exists.

The figure 4 illustrates the differences between a part with active and non active casting restrictions. Without casting restriction, all elements can be removed from the structure. The optimization starts with the lowest stressed elements.

With casting restriction only the visible elements can be removed from the structure. After removing one element, the next element becomes visible. In each step, the current lowest element is deleted.

The red line in figure 6 shows the maximum element number in one row and which is possible to remove each iteration. Due to this, the algorithm can repair too large cuts in the next iteration.

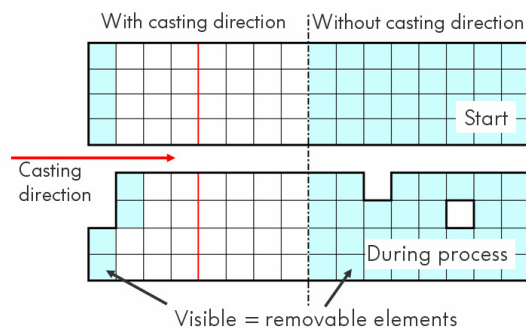


Figure 6: Casting direction

2.5.2. Forging

This restriction avoids parts with holes in the structure. To implement the mechanism, the last elements in one row are blocked for adding them to the visible group. Without getting visible, these elements can't be removed from the structure, see figure 7.

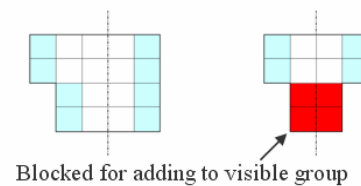


Figure 7: Forging

2.5.3. Symmetry

At the moment, only a plane symmetry is implemented. But other symmetries, like point symmetry, are easy to add. The mechanism for the plane symmetry can be directly transferred to them.

For a plane symmetry, all elements are divided into two groups. The first group allows modifications. After adding and removing elements in this group, the changes are mapped to the second group, illustrated in figure 8.

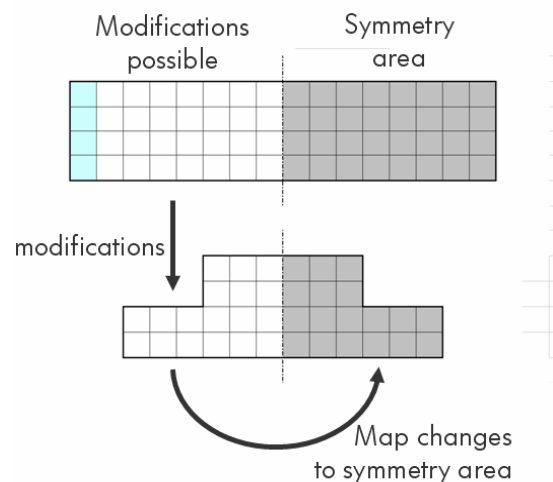


Figure 8: Plane symmetry

2.5.4. Minimum Strength

To avoid too small structures, which aren't possible to manufacture, normally filters are used. To this reason, discrete element modeling and using the half length of the minimum material strength for the elements length, the structure has a natural material strength. If the strength of the structure is smaller, the risk of collapsing in the FEA analysis is too high. This is demonstrated for example in figure 9.

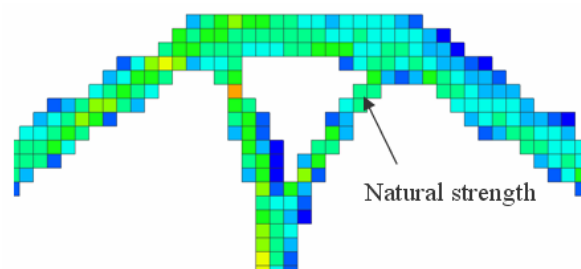


Figure 9: Minimum material strength

2.6. Example: Cantilever example with plastic material

One classical problem for testing topology optimization is the cantilever problem. In this case, it should directly demonstrate, how the new Topology Optimization method works using a nonlinear FEA analysis with plastic material. The material characteristic in this example has the specification of steel. On the left, two fixed supporting elements form the boundary as indicated in figure 10. In the middle of the plate on the right an enforced displacement of 20 mm at an angle of 90° to the main describes the load. In FEA simulation nonlinear geometry is activated and a real flow curve is used. As constraint function the reaction force at the node where the enforced displacement is applied, is used. The part should be optimized to a level of 10 kN. The target function is the minimization of weight, measured in elements.

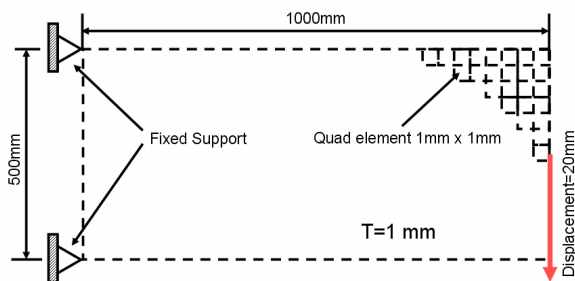


Figure 10: Cantilever problem with plastic material behavior

The optimization starts with a full design space with 125000 elements and an inertial basic reduction rate of 0.1. The general optimization process can be divided into three phases. The first phase is described by large reductions of elements up to the moment, where the constraint function rises strongly. In this second phase the constraint rises to the point where the constraint limit is reached. At this point two following iterations violate the limit. The optimization makes a cutback caused by the control mechanism and adds nearly 20% of elements to the structure. As a result the constraint offers the possibility to reduce the elements once again. Up to iteration 36 the optimization run reaches the point, where the cutback was made. Now the optimization control function is under the constraint limit. So the optimization progress enters the third phase. Characterized by slow step sizes, the optimization run offers improvements in detail. Through the oscillation around the constrain limit the last unnecessary elements are removed.

In figure 11 and 12 the optimization process is described through the number of elements and a normalized constraint. The value is the reaction force through the constraint limit of 10 kN. At ~25000 elements the cutback level is reached. After the second phase the optimization minimizes the weight to 20.35% of the starting value. In iteration 100 the optimization ends with a final value of 19.75%.

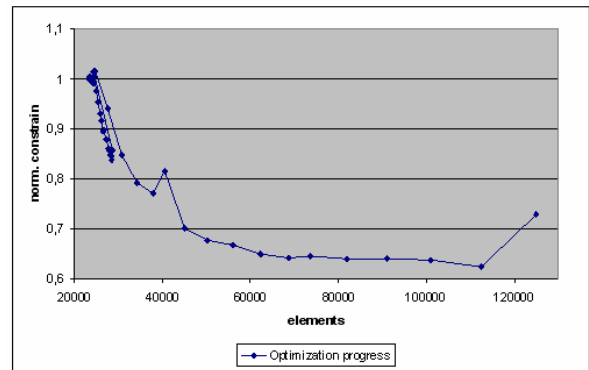


Figure 11: Optimization process of a cantilever problem with plastic material behavior described in Figure 10

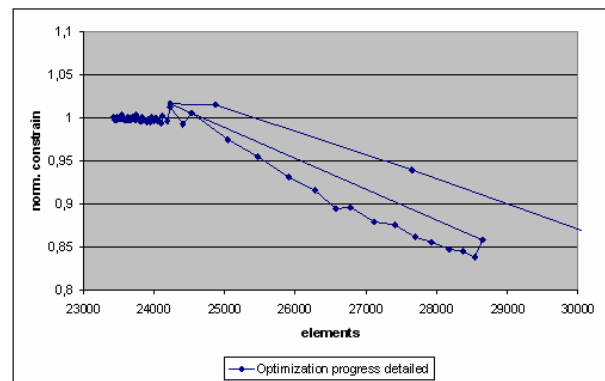


Figure 12: Detailed optimization process of a cantilever problem with plastic material behavior from Figure 10

The changes of the structure and the stress plots during the optimizer are illustrated in figure 13.

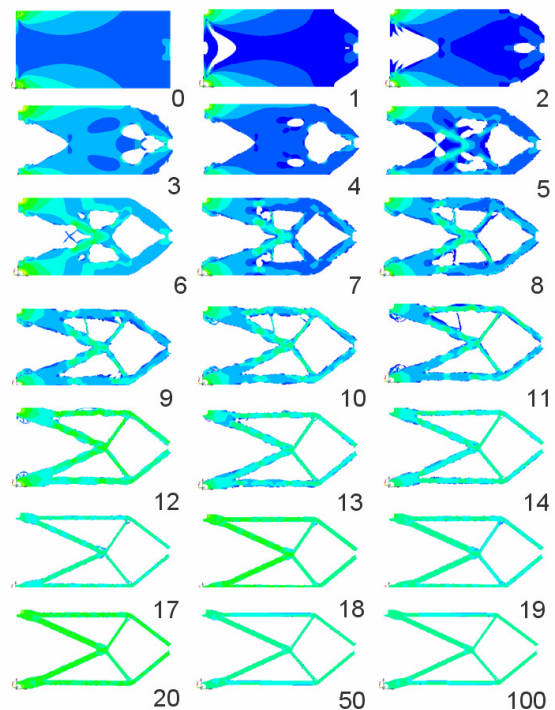


Figure 13: changes in optimization process of a cantilever problem with plastic material behavior from Figure 10

Figure 14 demonstrates the result quality of the new approach. Using nonlinear FEA analysis during the optimization run, it is possible to dimension all areas in the structure correctly. Due to this, the result shows a very good utilization of material in nearly all areas. Another positive aspect is the simplicity of the structure. No ramifications are proposed. This structure is easier to be constructed and manufactured.

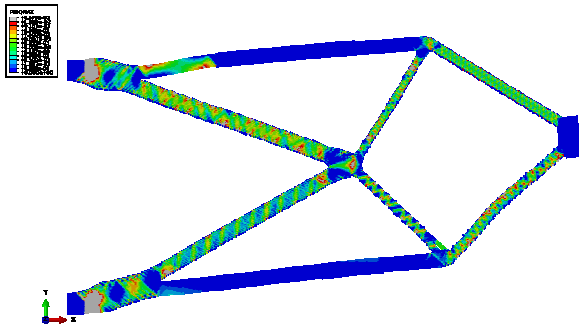


Figure 14: Nonlinear FEA analysis of optimization results from the example in Figure 10

The figure 15 shows the reaction force of the final iteration. In the figure is the progress of the reaction over the displacement illustrated. The final value at 20mm displacement is 10008,8 N. This demonstrates, that the optimizer has the ability to deliver a result exactly to the necessary constraint limit.

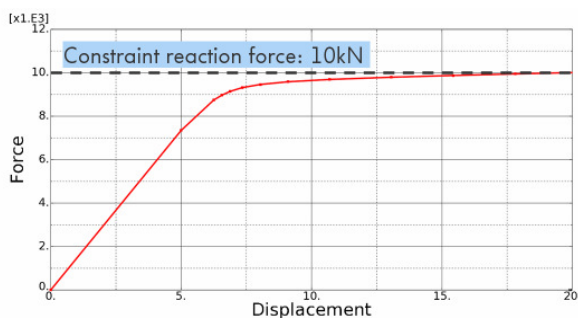


Figure 15: Reaction force of final iteration from the example in Figure 10

3. 6. CONCLUSION

In this paper an approach for a new Topology Optimization method is proposed. The method is developed based on requirements from the automotive industry. The main focus is the combination of finding a minimum weight and the best material distribution in one optimization run. To fulfill this task a discrete approach to include all nonlinear effects, e.g. plastic material behavior was chosen.

The example demonstrates the quality of the new optimization method. For the cantilever problem the new approach shows an advantage of 30% compared with a conventional industrial gradient based topology optimization methods. The new developed method shows the usability for real life development problems. The quality of the results is significantly increased

compared with conventional gradient based topology optimizations, especially in cases with nonlinear effects, e.g. plastic material behavior, in the FEA simulation.

Finding a satisfying solution in topology optimization reduces the necessary development time in a development department. The first designs in CAD based on the optimization runs indicate very competitive weight and fulfill immediately the technical requirements and manufacturing restrictions. So development loops and development costs can be saved. New target and constraint functions will increase the usage and more problems can be solved in less time.

REFERENCES

- G.I.N. Rozvany:** Aims, scope, methods, history and unified terminology of computer-aided topology optimization in structural mechanics, *Struct Multidisc Optim* 21, 90–108, Springer-Verlag, Berlin, 2001
- M.P. Bendsøe:** Optimal shape design as a material distribution problem, *Struct. Optim.* 1, 193–202, 1989
- M.P. Bendsoe and O. Sigmund:** Topology Optimization: Theory, Methods and Applications, Springer-Verlag, Berlin, 2003.
- G.I.N. Rozvany, M. Zhou, T. Birker:** Generalized shape optimization without homogenization. *Struct. Optim.* 4, 250–254, 1992
- C. S. Edwards, H. A. Kim, C. J. Budd:** An evaluative study on ESO and SIMP for optimising a cantilever tie-beam, *Struct Multidisc Optim* (2007) 34:403–414, Springer-Verlag, Berlin, 2007
- K. Svanberg:** The method of moving asymptotes-a new method for structural optimization, *Int J Numer Methods Eng* 24:359–373, 1987
- Q.M. Querin, G.P. Steven, Y.M. Xie:** Evolutionary structural optimisation using an additive algorithm, *Finite Elements in Analysis and Design* 34 (2000) 291-308, ELSEVIER, 2000
- X. Huang, Y.M.Xie:** Evolutionary Topology Optimization of Continuum Structures, Wiley, Chichester, 2010
- Q.M. Querin, V. Young, G.P. Steven, Y.M. Xie. :** Computational efficiency and validation of bi-directional evolutionary structural optimization, *Comput. Methods Appl. Mech. Engrg.* 189 (2000) 559-573, ELSEVIER, 2000
- C. Mattheck:** Design in der Natur, Rombach Verlag, Freiburg im Breisgau, 1997

AUTHORS BIOGRAPHY

Sierk Fiebig studied mechanical engineering at the TU Braunschweig. After his studies, he started research work for his Ph.D.-thesis at Volkswagen in optimization methods for the development of mechanical optimization methods. Since joining the company he has worked as a simulation engineer in the chassis development at Volkswagen Braunschweig.

3G MOBILE NETWORK PLANNING BASED ON A TRAFFIC SIMULATION MODEL AND A COST-BENEFIT MODEL TO SERVICE LOS CABOS INTERNATIONAL AIRPORT

Aida Huerta Barrientos^(a), Mayra Elizondo Cortés^(b).

^(a) National Autonomous University of Mexico (UNAM)

^(b) National Autonomous University of Mexico (UNAM)

^(a) aida.huerta@comunidad.unam.mx, ^(b) mayra.elizondo@hotmail.com.

ABSTRACT

We propose an application approach to planning a third-generation mobile network based on a traffic simulation model and a cost-benefit model to service the interior of an airport. This proposal represents an alternative to the mobile network planning traditional process. We developed a network traffic simulation model in terms of service transmission rates of applications such as voice over IP, video phone, FTP file transfer and high definition video-phone. The simulations are executed using ARENA software. From the results of the traffic simulation model, we obtained the network capacity in terms of the cell radius. Based on the cost-benefit model and on the network capacity, we got the cell radius that maximizes the net profit percentage of the network and satisfies the user's requirements. Under this approach is not taken into consideration some of technical aspects, but rather it is to highlight the economic aspect of the network planning process.

Keywords: 3G mobile network, traffic simulation, network planning, economic model

1. INTRODUCTION

The telecommunications industry has been expanded substantially since the past decade. Technology advancement along with the liberation of once closed markets and privatization of government-held monopolies changed the nature of the industry in the 1990s and continues to shake up the industry every now and then. In early 2000, the industry scaled new highs with respect capitalization. Both business and technology disruptions have introduced significant expansion and innovation. The global mobile cellular subscription was closed to 5.3 billion by the end of 2010. That is equivalent to 77 percent of the world population. So 90 percent of the world now lives in a place with access to a mobile network. For people living in rural communities this is lower at 80 percent, according to the estimation of the International Telecommunication Union.

The mobile cellular technology, as happens in others technology fields, has had an innovation process in which has been defined and implemented successive generations. Each of these generations has responded to

specific service demands from network users and operators. The first generation cellular mobile (1G) technology enabled the human communication via voice. While the second generation cellular mobile (2G) technology enabled the human communication via voice and text messages, and the third generation (3G) technology enabled the human communication via voice, text messages, data and video.

The same forces that fed the development of new services and the entrance of new players also saw margins grow slimmer for most services as well as significant customer churn as competitors offered alternative choices. So for network operators has been important to give mobile cellular service indoor and outdoor along cities and rural communities. To give the service, the network operators need, by first time, to plan the network. The network planning is a process on which operators take in consideration aspects such as demand of services and applications from users, service area based on the potential user locations, service rates, quality of service, environmental morphology and return of investment, for example.

By one hand, in today's extremely challenging business environment, many telecommunications operators and carriers are measuring their success by the size and growth of their profit margins. As a result, operators are under intense pressure to reduce or eliminate the major threats to these slim margins including revenue leakages and frauds, churn, inefficient network usage, and least-cost routing plans. These competitive and market pressures are also making the telecommunications industry reassess its business model and redefining the path that will return it to competitiveness and profitability (Pareek 2007).

By another hand, under competitive conditions, the customer becomes the central focus of the carrier's activities. Customer requirements not only determine service offerings, but also shape the network. In this sense, we propose a network design to service Los Cabos International Airport, which is located in Mexico country, based on user's requirements and using 3G mobile cellular technology specifications. This proposal is based on a traffic simulation model and a cost-benefit model and the main objective is to get the cell radius that maximizes the net profit percentage of the network.

So we obtain the cell number and their corresponding configuration to cover the airport total physical area. The study is done considering the operator's view and we hope it can help to support the decision making in the telecommunication industry, specifically in the network design area.

2. MOBILE CELLULAR TECHNOLOGY

In order to have access in the mobile cellular telecommunication networks, users must have a terminal device based on a specific technology and operators must implement a network based on a specific technology too. Mobile cellular technology has evolved over generations. Thus, we have the so-called first generation (1G) technology, characterized by analog devices, while the so-called second generation (2G) technology, characterized by digital devices and based on standards as GSM (Global System Mobile) and CDMA (Code Division Multiple Access). And the so-called third generation (3G) technology based on standards such as UMTS (Universal Mobile Telecommunications System) (see Fig. 1).

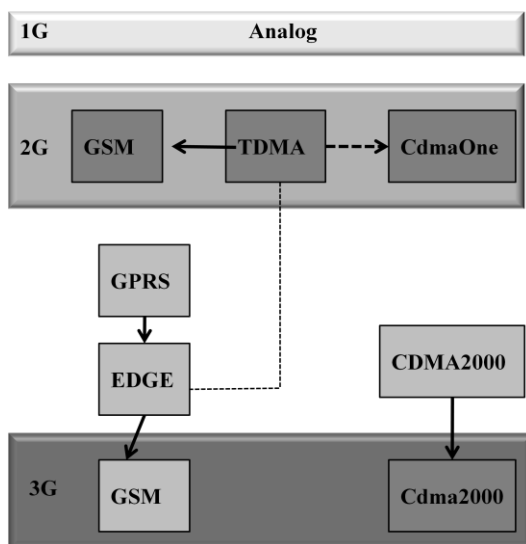


Figure 1: Mobile cellular technology evolution

In accordance with the UMTS specifications (3GPP 2002), using this technology makes possible to increase network bandwidth in order to get a major and better data transmission. In fact, this situation has encouraged to operators to offer a wider range of mobile services and applications, through introducing a new technology platform over their actual network (located physically between the user terminal device and the network controllers). So, using UMTS technology, the network must be planned in accordance with criteria evaluated by operators in order to satisfy customer requirements but customer requirements depend of customer communications needs, in certain places and in certain time. The operators must know the customer communications needs in order to be a very good option in a competitive market as is nowadays the mobile cellular telecommunication market.

3. MOBILE CELLULAR NETWORK PLANNING

3.1. The process

The network planning process consists of ten specific activities (Mishra 2007) which are carried out by different technical teams (see Fig. 2).

It starts with the network design, which may or not be based on field measurements of an existing network. The next activity is the locations acquisition. The locations acquisition means that networks operators rent a physical space where sites will be built. After the acquisitions and building, the equipment is implemented through installation and commissioning. The commissioning is the activity on which the engineer field sets the correct value of the equipment parameters. This activity includes the integration of the new site to the total network according to the operator criteria. Once the commissioning is without technical and operational problems, the site is put into service. It means that phone calls can be processed by the site through the network. The activity which closed the process is the optimization.

In order to optimize a new site and its integration to the telecommunication network, we need to collect measurements about the level and the quality of the signal and then to eliminate signal interferences in order to improve indicators such as the dropped call rate for example.

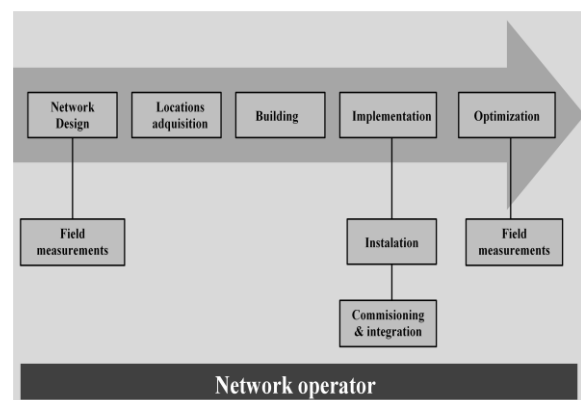


Figure 2: The network planning process

3.2. Different type of cells

The network design, the first activity in the network planning process, is carried out based on certain criteria which must be specified by the network operator in accordance with specific customer requirements. Normally, the network operator must adapt these requirements to his business model. As noted, the network design is an activity which may or may not be based on field measurements. To get field measurements is not always possible because for that, it is necessary that exists a mobile telecommunication network, does not matter if the existing mobile telecommunication network works with a different technology, the most important thing is to get the field measurements. If does not exist a network, the design is

made using simulation models based on theoretical assumptions. At the end of the network design, we can get the different types of cells, which will be implemented, as well as their capacity, both in terms of the number of transmission channels. The different types of cells depend on the physical area to which the cell needs to give the communication service.

Thus, pico cell can be designed to provide service within buildings and are characterized because their dimensions are small and their transmission power is low. Technically, a one location correspond one kind of this cell. While micro cells are designed to provide service to shopping centers, open parks and business centers, for example, and are characterized as well as the pico cell by a low transmission power.

We can also implement macro cells, which reach 2000 meters as radio service area with a transmission rate of 144 kb/s. If the applications require a better transmission rate, the network can be adapted to a major one, but the radio service area must decrease until to reach almost 500 meters (ETSI 1997). This type of cells is characterized by a big transmission power (see Fig. 3).

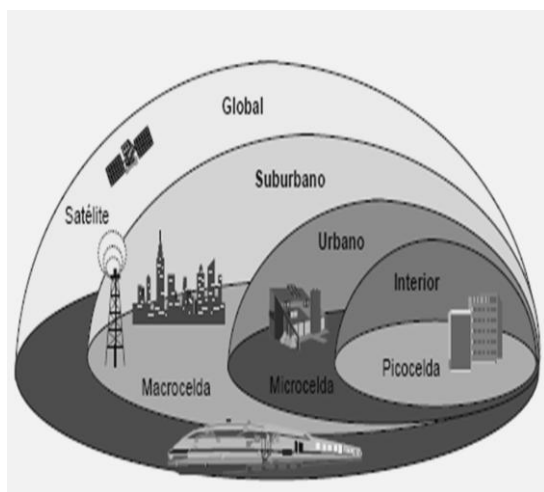


Figure 3: Different types of cells

In order to be most competitive, in some cases the operators design a telecommunications network based on the service area criteria. This criteria works well in the cases of macro cells, it means, to cover a big area such as a community, because it helps operators to attract new customers. But in the case of a network designed to service inside of a building with pico cells, is very important to consider the cost-benefit approach and the specific user requirements because the customer will move around all the building and need to get the service in any point inside the building and the service operator need to be provided in all point inside the building in order to get customer satisfaction which will be converted in revenue for operators. So network operators must implement a network that first satisfies customer service needs and then, guarantees revenue for him.

We do not say that the cost-benefit criteria is not taken in account when a network operator design his own network based on macro cells, instead, we say that in many cases is more convenient for operators, in terms of trade, to base the network design on the service area criteria for macro cells and for the network design based on pico cells is more convenient for operators to be ensure the availability of the network in terms of services, applications and transmission channels.

4. ANOTHER APPROACHES

There are many approaches to design a mobile telecommunications networks and it depends on the priorities of the network operator. Once operators have defined their design criteria, it is necessary to select the technical design tools to be used. One of the most used tools in these cases is the simulation. At the present time, there are many kinds of simulators for UMTS telecommunications networks design which are configured in accordance with the parameters that we need to modeling and analyzing (Alonso y Lopez 2005).

Simulators permit to analyze the network at different levels. The advantages for working with network simulators are the increased productivity in the network development, less time to market so total cost is decreased. Some examples of simulators used in the network planning are the follows:

- Network simulator (System level), is possible to analyze the traffic, QoS, handover, admission control,
- Link simulator (Link level), controls the transmission errors,
- Physical layer simulator, is used to evaluate coverage area, power transmission, cells and interferences,
- Protocols simulator, verifies, analyzes and optimizes protocols,
- Integrated simulators, this kind of simulators has many functions integrated as the name indicates.

Many studies are development each year in accordance with the specific criteria of the network operators. The criteria could be technical, economic or both of them.

5. THE PROPOSAL

In this paper, we propose an application approach to design a third-generation mobile network based on a traffic simulation model and a cost-benefit model, to provide 3G mobile telecommunication service at the interior of an airport. This approach consists of five general steps (see Fig. 4).

1. First, we develop a network traffic simulation model in terms of transmission services and third-generation applications,
2. Then, we perform the simulation experiments using the software ARENATM,

3. From the results of the experiments simulation, we obtain the network capacity measured in Mbits/s depending on the cell radius,
4. Based on the cost-benefit model and the network capacity, both as functions of the radio cell, we get the cell radius that maximizes the net profit percentage of the network,
5. So we obtain the network configuration, number of cells and radius cells, which satisfy the operator and users requirements.

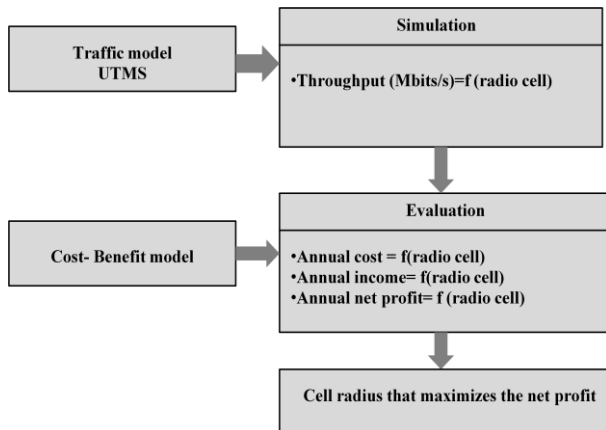


Figure 4: An approach for design a 3G mobile network based on cost-benefit

This approach represents an alternative to the mobile network design traditional process and we propose it to be used by network operators because this approach take account the users requirements and the operator cost-benefit, which are operation basic criteria in a competitive industry as actually is the telecommunication industry.

5.1. The scenario

The scenario considered is the international airport of Los Cabos, Baja California Sur, Mexico. Currently, this airport is ranked as the seventh most important in Mexico. In 2005, the airport received a total of 2,466,733 passengers, of whom 431,724 were domestic and 2,035,009 were international, in accordance with GAP (Grupo Aeroportuario del Pacifico). The physical dimension of the terminal is 8440 meters square and it receives, on average, 730 passengers per hour every day (see Fig. 5).

5.2. The 3G mobile services and applications

According to the ITU (International Telecommunications Union), the network services and applications based on technology 3G are grouped as follows:

- Interactive (Conversation, messages and information download).
- Distribution (emission and cyclic).



Figure 5: The international airport of Los Cabos, BCS

By one hand, in accordance with Ferreira and Velez (2005), an application is defined as a work which requires the communication between one or more information flow, between two or more parts geographically distributed. The applications are characterized by the service attributes, the communication and the traffic characteristics. One essential data to the analyses purposes is the utilization rate for each application. By another hand, the services and applications distribution requirements can be over real time and over no real time as follows:

- Real time, in this case, the applications need the information distributed for immediate use,
- No real time, in this case, the information is stored in specific reception points to be used later.

5.3. The network traffic model

The most important traffic parameters included in a network design are transmission rate and average duration. The transmission rate is an average number of bits which are transferred between two devices, in each unit time. A possible measurement unit is kb/s. While the average duration represents the duration which each user uses a specific application, for own purposes.

In accordance with Ferreira, Gomez and Velez (2003), a traffic generation model can be used to measure and to describe the traffic over the network only if it is based on the population density and on the service insight, so we can be able to know the call rate for each service. The potential services are grouped as follows:

- Sound,
- Multimedia,
- Narrow band,
- Wide band.

For each service, we select one application in order to get a traffic model (see Table 1). The utilization figures are approximations proposed in this study and that can be tailored for particular cases.

Table 1: Services and applications

Service	Application	Utilization
Sound	Voice over IP	50%
Multimedia	Video-telephone	22%
Narrow band	File Transfer Protocol	16%
Wide band	High Definition Video-Telephone	12%

The transmission rate and the duration statistic distribution, which characterize the applications selected, are included in Table 2 (Antoniou, Vassiliou and Jacovides 2003).

Table 2: Technical specifications for applications

Application	Duration statistic distribution	Transmission rate
Voice over IP	EXP/3 min	12 kb/s
Video-telephone	EXP/3 min	128 kb/s
File Transfer Protocol	EXP/0.1 min	384 kb/s
High Definition Video-Telephone	EXP/30 min	1920 kb/s

5.4. The simulation study

Now, we need to simulate the traffic model proposed in order to be able to analyze the traffic as function of the network capacity and the applications. Since the early days of simulation, people have constantly looked for new and better ways to model a system, as well as novel ways to use existing computer hardware and software in simulation (Law and Kelton 2000). For the purposes of this work, we use discrete-event simulation to model and analyze the network traffic. The steps that will compose a simulation study are showed in Figure 6.

On the next sub-sections, we conduct the simulation study according to the steps in Figure 6.

5.4.1. Data collected

The data which characterized the scenario were described in section 5.1 and the data which characterized the traffic were described in section 5.3. We propose the call arrival based on the statistics from the user arrival peak times at the airport. We considered that the periods on which more call are processed are when the users arrive to the airport in peak times, turn on the cellular phone and make a call.

The periods considered in one day are:

- 8:00 hrs. – 12:00 hrs.,
- 14:00 hrs. – 16:00 hrs.,
- 17:00 hrs-18:00 hrs.

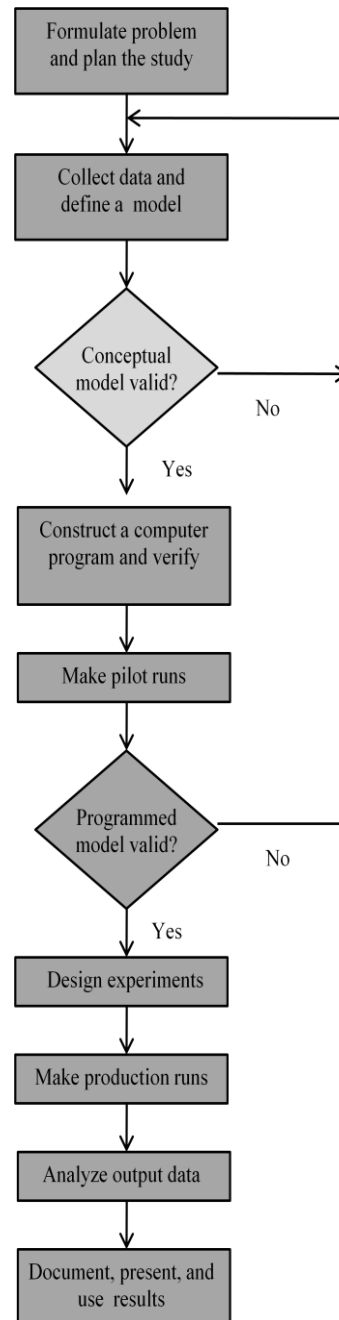


Figure 6: The steps for a simulation study

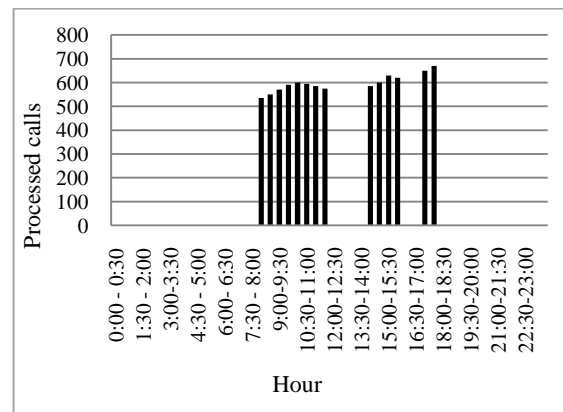


Figure 7: Peak processed calls

In this study, we consider six configurations which take into account a different network capacity. The network capacity is based on the theoretical cell radius, we considered omnidirectional cells. Also, we considered that each cell is composed by seven traffic channels. For a radius cell of 30 meters corresponds a total capacity of 21 traffic channels. While for a radius cell of 21 meters corresponds a total capacity of 42 traffic channels, and so on until reaching a cell whose radius is less than 14 meters, which account for a total capacity exceeding the 98 traffic channels.

5.4.2. The conceptual model defined

The conceptual model about the traffic generated in a network is described in Figure 8.

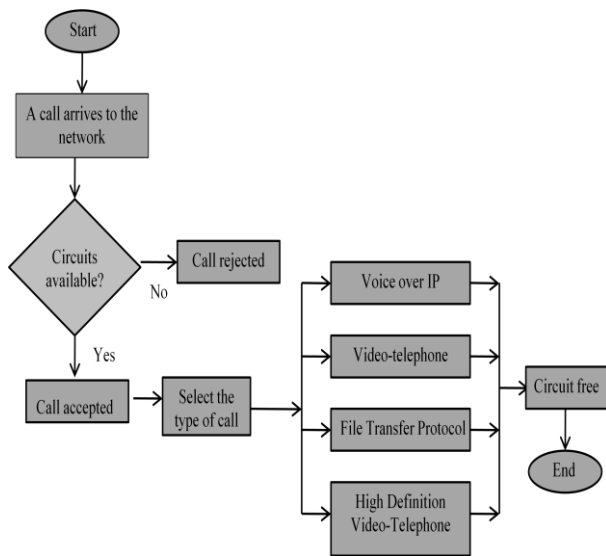


Figure 8: Conceptual model

Once a call starts, the network verifies if a circuit is available. For case negative, the call is rejected and the counter for this kind of calls is incremented. For case positive, the call is accepted and then is classified according to the type of call. There are four types of calls, each type corresponds a one application considered in the traffic model:

1. Voice over IP,
2. Video-telephone,
3. File Transfer Protocol,
4. High Definition Video-Telephone.

Once a call is classified, it is processed in accordance with their duration statistic distribution specified in Table 2. Then the counter is incremented and when the call is finished, the circuit is free to be used by another call.

5.4.3. Software selection

Elizondo and Flores de la Mota (2006) suggest a process which is based on some questions, in order to select a software simulation. The process is divided in two phases. By one side, the first phase is related with

references, documentation and compatibility software and, by another side the second phase considers the problem characteristics as follows. We select ARENA software to this simulation study, so some of the questions are answered about this software.

Phase one

- There is a user manual? Yes,
- The language code is compatible with actual computers? Yes,
- The software has enough documentation and error diagnosis? Yes,
- The language is known and easy to be learned? Yes,
- The software is compatible with another kind of software? Yes.

Phase two

- What kind of real problems can be analyzed by the software? Process simulation, business simulation, supply chain simulation, logistic simulation.
- Is easy to store and modify the system data? Is easy through modules.
- Is easy to include user subroutines? Yes, users can create theirs owns modules and add to the software in order to create new systems.

5.4.4. Computer program

Based on the conceptual model and on the ARENA software modules, we develop the computer program (see Figure 9).

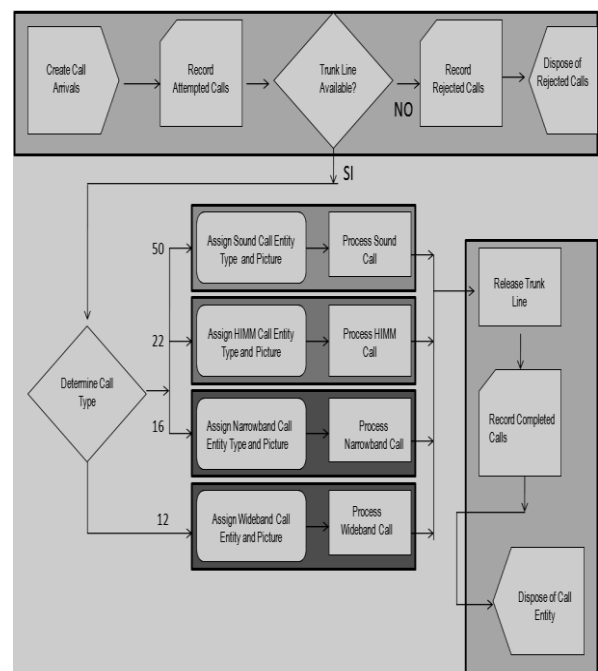


Figure 9: The computer program using ARENA software

5.4.5. Verification & validation

Once the model has been implemented, it is necessary to verify its performance. Based on the technique number 4 suggested by Elizondo and Flores de la Mota (2006), we verified the computer program. This technique consisted in the simulation execution considering many different scenarios so it was necessary to make many changes in the program parameters. After the executions, we checked the consistency of the results in accordance with the expert's opinion.

As results, we found that when the network capacity was increased, the number of rejected calls decreased and the number of call processed increased. This result corresponded with a real situation over a telecommunications network because a major network capacity correspond a major number of call processed and a minor number of calls rejected.

5.4.6. Design of simulation experiments

We carried out 17 simulation experiments, in each experiment we increased the number of traffic days. So, in the first experiment we simulated 5 traffic days, in the second experiment we simulated 10 traffic days, and so on, until in the experiment 17 we simulated 80 traffic days.

5.4.7. Analysis of output data

1652 total calls were processed for the network configuration with capacity of 21 traffic channels, 3155 total calls were processed for the network configuration of 42 traffic channels, 3505 total calls were processed for the network configuration with capacity of 48 traffic channels, 4146 total calls were processed for setting network of 70 traffic channels, 4173 total calls were processed for the network configuration with capacity of 84 traffic channels, and finally, 4171 total calls were processed for the network configuration with a capacity greater than 98 traffic channels (see fig. 10).

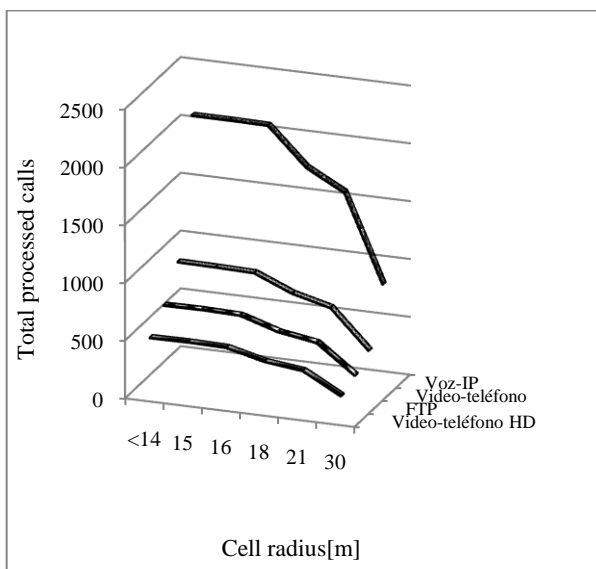


Figure 10: Total calls processed

5.4.8. Use of the results

At this point, we use the total number of calls processed in the cost-benefit model, in order to obtain the annual net profit as a function of cell radius for the different network configurations considered.

5.5. The cost-benefit model

As stated in Gavish and Sridhar (1995), the economic aspect for a telecommunication network can be analyzed by four different approaches: user's approach, service provider's approach, regulator's approach and manufacturer's approach.

By one side, for users, the most important aspects about the network are the service quality and cost. By another side, for service providers represented by networks operators the most important aspect is to find the network configuration which maximizes the expected revenues. While for the regulator, the most important aspect is the social welfare, to promote competition between network operators and to manage the frequency spectrum. Then, for manufacturers, the most important aspect about a telecommunication network is the equipment cost. So, each approach has their necessities to satisfy.

The approach used in this contribution is the service provider. Because of the service provider business model, in this study we will take in account the impact of the other tree approaches. Cabral *et al.* (2005) suggest that the total annual cost of a radio network is determined by a fixed cost and a cost proportional to the number of cells required to service the area required by the operator. For this particular case, the fixed cost includes the cost of an operating license that the operator must ask the Mexican government. While the cost of each cell (node) is determined by the cost of equipment, the cost of installation and cost of operation and the maintenance as in (1).

$$Annual\ cost[\$]=Fixed\ cost+(Cell\ annual\ cost)*(Cell) \quad (1)$$

On another side, the net income diary is obtained as a function of traffic carried across the network as in (2), i.e. the total number of calls processed by all cells in the network. It means taking into account the traffic flow at peak hours the network, as in the rest of the hours of operation of the network, calls are processed very few reaching sometimes be invalidated.

$$Annual\ net\ income[\$]=Net\ income\ diary*Traffic\ Days \quad (2)$$

The total annual net profit is obtained for different network configurations as in (3).

$$Total\ annual\ net\ profit\ [\$]=Annual\ net\ income - annual\ cost \quad (3)$$

For the purposes of this study, the trade data was taken from the dominant mobile phone operator in Mexico. So, the annual net profit (%) obtained for different network configurations, according to the

traffic simulation model and the cost-benefit model is showed in Figure 11.

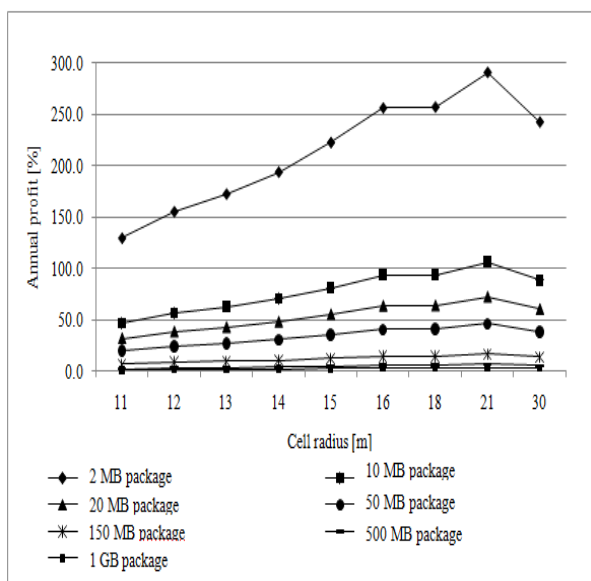


Figure 11: Annual net profit as a function of cell radius of different network configurations

5.6. Results and conclusions

Thus, to attend the use demand of voice over IP (50%) video phone (22%), FTP file transfer (16%), high definition video-phone (12%) at the international airport of Los Cabos in Baja California Sur Mexico, it requires a UMTS network with a total capacity of 42 channels, consisting of 6 cells, each one with a radius of 21 meters, so that the network operator maximizes the percentage of annual net profit network up to 300%.

This proposal represents an alternative to the traditional process of mobile communications network planning. To conceptualize a network as a discrete event model, we can get an approximation of the network configuration that meets the demands of users and supports the decision of the operator, using the simulation. It is clear that under this approach is not taken into consideration technical aspects of a communications network as the propagation loss for example, but rather it is to highlight the importance of the economics of network planning.

REFERENCES

- 3rd Generation Partnership Project, 2002. Technical specification group services and systems aspects: Network architecture V3.5.0. *3GPP Technical Standards*. <http://www.3gpp.org>
- Alonso, M., Lopez, C., 2005. *Simuladores UMTS*. Sevilla: Universidad de Sevilla.
- Antoniou, J., Vassiliou, V., and Jacovides, N., 2003. A Simulation Environment for enhanced UMTS Performance Evaluation. *Proceedings of the austyalian Telecommunications Networks and Applications Conference*. December, Melbourne (Melbourne, Australia).

- Cabral, O., Velez, F.J., Hadjipollas, G., Stylianou, M., Antoniou, J., Vassiliou, V., and Pitsillides, A., 2005. *Proyecto de requerimiento científico REEQ/1201/EEL/2005*. Lisboa (Lisboa, Portugal).
- Elizondo, M., Flores de la Mota, I., 2006. *Apuntes de simulación*. México: DEPEFI, UNAM.
- European Telecommunications Standards Institute, 1997. Universal Mobile Telecommunications System (UMTS); UMTS Terrestrial Radio Access (UTRA); Concept evaluation. *UMTS 30.06, ETSI TR 101 146*.
- Ferreira, J., Velez F. J., 2005. E- UMTS Services and Applications Characterisation. *Teletronikk-strategies in Telecommunications*, 101(1), 113-131.
- Ferreira, J., Gomes, A., and Velez, F. J., 2003. Enhanced UMTS Deployment and Mobility Scenarios. *Proceeding of the 12th IST Mobile & Wireless Communications Summit*. Aveiro (Aveiro, Portugal).
- Gavish, B., Sridhar, S., 1995. Economic aspects of configuring cellular networks. *Wireless Networks*, 1, 115-128.
- Law, A. M., Kelton, W. D., 2000. *Simulation Modeling and Analysis*. 3rd ed. Singapore: McGraw-Hill.
- Mishra, A., 2007. *Advanced cellular network planning and optimization*. Chichester: John Wiley & Sons Ltd.
- Pareek, D., 2007. *Business intelligence for telecommunications*. New York: Auerbach Publications.

AUTHORS BIOGRAPHY

AIDA HUERTA BARRIENTOS currently is an Operational Research PhD student in the Department of Systems at the National Autonomous University of Mexico, UNAM. She received in 2001 her B.S. degree in telecommunications engineering and in 2010 her M.S. degree in systems engineering from UNAM. Her professional experience includes planning, operation, optimization and maintenance of 2G and 3G telecommunications networks from network operators as Vodafone in Spain and Telefonica in Mexico. Also, has managed telecommunications projects in Nokia Mexico. Her doctoral research is in aspects of simulation modeling and analysis of complex systems.

MAYRA ELIZONDO CORTÉS is a Definitive Professor of Operations Research at the Department of Systems of the National Autonomous University of Mexico, UNAM. Her main research activities are in the Simulation and Optimization areas applied to Logistics and Supply Chain. She had published papers internationally and had dictated international courses.

PROVIDING SEMANTIC INTEROPERABILITY FOR INTEGRATED HEALTHCARE USING A MODEL TRANSFORMATION APPROACH

Barbara Franz^(a), Herwig Mayr^(b)

^(a) Upper Austria University of Applied Sciences, Research Center Hagenberg, Austria

^(b) Upper Austria University of Applied Sciences,
Department of Medical Informatics and Bioinformatics, Hagenberg, Austria

^(a)Barbara.Franz@fh-hagenberg.at, ^(b)Herwig.Mayr@fh-hagenberg.at

ABSTRACT

Integrated care and the achievement of high quality healthcare over institutional borders require technical and semantic interoperability between different healthcare providers. Thus, a meta model was developed, based on the application of a system conformant to Integrating the Healthcare Enterprise (IHE) and the use of Clinical Document Architecture (CDA) as document format. This meta model combines the properties of several health information systems and different health service domains (HSD).

For each HSD, a model can be derived, which describes used IHE profiles, available document types to the point of coding systems used in exchanged documents. Thus, it provides information about the domain, workflows, patient etc., which are used for the transformation into another domain. Using this model transformation, HSD models may be compared, checked and completed, where applicable. The results can then in turn be applied to the CDA documents. Thus, it helps to improve the communication among healthcare institutions.

The evaluation of developed models and transformations is conducted using genuine healthcare data, provided by e-Care, an IHE-conformant system for exchange of healthcare data between several healthcare providers.

Keywords: model transformation, semantic interoperability, Clinical Document Architecture, Integrating the Healthcare Enterprise

1. THE CHALLENGE OF HEALTHCARE DATA EXCHANGE

Efficient data exchange between healthcare providers is the cornerstone of integrated healthcare (Arrow et al. 2009). To achieve a high quality of healthcare over institutional borders, it is important to regard which activities and examinations have already taken place and in which quality they were conducted. Therefore it is necessary for healthcare providers to document in such a way that access to documented information as well as a comparison and an integration of the data into other systems is also possible for other providers.

This is a big challenge as the project e-Care, where several healthcare providers were connected, has shown (Franz et al. 2009a). The most demanding factors are:

- various electronic health information systems,
- different domain languages and understanding,
- special processes and workflows in various organizations,
- varying documentation with different goals,
- varying structures in documentation,
- different designations and
- little to no use of coding systems.

Thus, the challenge is the successful combination of technologies and information models in integrated healthcare. Part of this can be met by relying on Integrating the Healthcare Enterprise (IHE). The worldwide initiative IHE aims to optimize the interoperability of IT-Systems in medicine and healthcare, using international standards, cf., e.g., (HL7 2010a). IHE provides several so-called integration profiles which define use cases and suggest technical solutions (IHE 2010a).

One IHE profile suggests the use of Clinical Document Architecture (CDA) as a uniform format for the exchange of clinical documents (IHE 2010c). Therefore, current national and international projects like e-Care, ELGA (ELGA 2010), VHitG (VHitG 2007) and epSOS (epSOS 2010) boost the use of fully structured or at least semi-structured healthcare data in form of HL7 V3 CDA-documents. But as long as there are no strict requirements for system providers to provide their data in highly granular structures, loads of unstructured data will always exist which cannot be interpreted unambiguously (Wozak et al. 2008).

Since uniform semantic standards are only used rudimentarily in these documents (as described above) it is seldom possible to directly compare, check (on plausibility, correctness and completeness) or integrate document content. Such an integration is only possible in consideration of semantics and domain specific differences. Using semantic and technical interoperability in various health service domains, documented information has to be homogenized in a way that an access to structured as well as unstructured information, consistent with a certain context, is possible. Such a knowledge transfer over institutions demands a clear and structured language (Perhab 2007) and particularly interoperability.

2. TECHNICAL AND SEMANTIC INTEROPERABILITY

National as well as international initiatives for information integration in healthcare aim at the increase of interoperability of information systems and at minimizing integration efforts (Norgall 2003; Sunyaev et al. 2008). The term *interoperability* denotes the ability of systems for collaboration. *Technical interoperability* indicates the interaction of technical components and systems (Heitmann und Gobrecht 2009). To achieve interoperability of healthcare systems, IHE and also Continua Health Alliance (CHA 2010a) provide efficient methods and standards.

Semantic interoperability, on the contrary, denotes the interpretation of data, while preserving the intended meaning. It can be achieved using common information models and uniform terminology, thus enabling cross-community interpretation, processing and storage. Besides technical standards, an agreement is essential, how medical and domain specific terminology is used. Nevertheless, maintenance as well as further development and optimization of the terminologies and their structural elements should not be underestimated (Heitmann und Gobrecht 2009).

Domain specific terminology has to be used unambiguously to prohibit misunderstandings and to improve collaboration. Only this way semantic interoperability can be achieved (at least partially) in one domain as well as across several domains.

2.1. IHE Integration Profiles

The use of IHE, conformance to IHE integration profiles and adherence to suggested standards are one way towards smooth integration (IHE 2010b). Four of the most important integration profiles, which enable interoperability, are:

- *Patient Identifier Cross-Referencing (PIX)* supports the unique identification of patients with different patient identities in several domains. For this purpose, patient data are sent from a source to a Patient Identifier Cross-Reference Manager (IHE 2010a).
- *Patient Demographic Query (PDQ)* defines a central patient register, which allows distributed applications to query demographic and case-related patient information (IHE 2010a).
- *Cross Enterprise Document Sharing (XDS)* supports the patient care of several healthcare providers by allowing registration, distribution and access of documents referring to one patient. XDS is content neutral, i.e. it supports different types of documents, independent of their content or format (IHE 2010a), although other profiles force the use of certain standards for documents.
- *Cross Enterprise Document Workflow (XDW)* focuses on the management of cross-enterprise healthcare workflows using a specific workflow document that references all documents

related to a clinical workflow and manages changes in document states (Zalunardo and Cocchiglia 2011).

- *Patient Care Coordination (PCC)* describes the cross-enterprise sharing of medical summaries and personal health records (IHE 2010c).

2.2. Health Level Seven

IT standards, as suggested by IHE integration profiles, are required for data exchange between different information systems. In healthcare, the most important standards, besides Digital Imaging and Communications in Medicine (DICOM) for the exchange of images, are Health Level Seven (HL7) with Version 2 and 3 as well as CDA, which allow the exchange of information like documents and messages.

HL7 standards have been specifically developed for the health sector. They define the exchange of messages, document based communications as well as co-operating services, their implementation and necessary infrastructural services (HL7 2010a). Two important HL7 types for this work are:

- *HL7 Version 2*, which is primarily used in hospitals for message exchange between well-established systems,
- *HL7 Version 3*, which is based on the Reference Information Model (RIM, c.f. Fig.1) and on the Extended Markup Language (XML) and is used for trans-sectorial message exchange in the entire health sector.

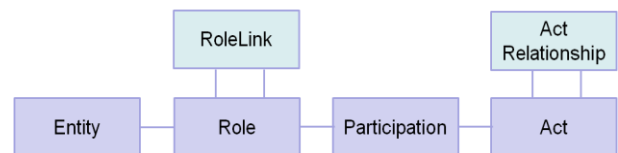


Figure 1: Core classes of the RIM

The formal computational exchange format for HL7 artefacts is defined by the *Model Interchange Format (MIF)*, which is a part of the HL7 Version 3 methodology documentation (McKenzie et al. 2011). To represent the information content required to support a particular domain within HL7 and thus provide interoperability for a specific domain, a *Domain Information Model (DIM)* can be derived from the RIM.

The RIM is a generic healthcare specific information model. Figure 1 shows a part of the HL7 RIM, i.e. the four core classes and the two most important additional classes of the model:

- *Act*: e.g. an observation or treatment,
- *Entity*: e.g. a person,
- *Role*: e.g. author,
- *Participation*: relationship between act and role, e.g. a doctor performs a treatment,
- *ActRelationship*: relationship between acts, e.g. a diagnosis results from an observation,

- *RoleLink*: relationship between roles, e.g. organizational hierarchy.

The goal of HL7 Version 3 is the development of a uniform understanding of objects and processes in the healthcare environment. The use of RIM provides specifications to structure, type, content as well as semantics, used vocabulary and underlying processes necessary for data transfer and interoperability.

Besides the exchange of messages, HL7 Version 3 also provides CDA as specification for the structure, content and exchange of clinical documentation. CDA is based on the RIM and defines the structure and content of medical documents using XML. A CDA document consists of a structured header and a structured or unstructured body. The header contains information about the document, patient, patient encounter, participants and relations to other documents, guidelines or templates. The body contains the actual content of the document.

As already mentioned in section 1, CDA is exerted in several developments and projects, e.g. the European project ePSOS, where it is the specified document format. These projects and developments specify CDA guidelines, which describe the underlying models and therefore the structure and parts of the semantic of the documents and their content. Connections between classes and attributes in the model are shown and possible restrictions on the CDA schema (XML schema) specifications are detailed. Partial structures of the CDA which are defined as CDA templates can be identified over a template ID.

CDA templates are predefined models which simulate the structure of documents or parts thereof and of data elements. Other guidelines are partially based on existing templates, which are referenced. Using the template OID in CDA documents, an automatic technical conformity-check against guidelines is possible. Examples for templates which are often referenced are Continuity of Care Document (CCD) and parts of the IHE profile PCC.

2.3. Use of Code Systems

The unambiguous use of domain specific terminology is a means towards semantic interoperability. In the health sector, different code systems are used, which provide a uniform definition of terminology (Heitmann und Gobrecht 2009). Examples for such code systems are:

- International Statistical Classification of Diseases and Related Health Problems (ICD),
- Logical Observation Identifier Names and Codes (LOINC),
- Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT),
- International Classification of Functioning, Disability and Health (ICF),
- Medical Dictionary for Regulatory Activities (MedDRA),
- Nationality and language codes (ISO tables),

- Unified Code for Units of Measure (UCUM).

Since such code systems are not always used and fully structured information can therefore rarely be provided, unstructured and only partly structured (semi-structured) information have to be taken into account. Therefore, the various types of information have to be connected to provide semantic interoperability, which is necessary for comparing and checking structured as well as semi-structured data.

Current research like (Kilic und Dogac 2009; Bointner und Duftschmid 2008; Rinner und Duftschmid 2009) only consider finely structured CDA data, whereas other work in this research area like (Spat et al. 2007; Faulstich et al. 2008) are limited to fully unstructured free text documents. Also, the comparison of models in one domain has already been successful, as described in (Franz et al. 2009b). However, preserving semantic interoperability is required not only in one domain, but across several domains.

3. A MODEL BASED APPROACH USING IHE AND CDA DATA

Our approach presented in this paper can be used for structured data as well as free text and semi-structured information in form of CDA documents which are exchanged across several domains.

Based on the application of an IHE conformant system and the use of CDA as document format a meta model has been developed which combines the properties of several health information systems and different health service domains (HSD). The HSD meta model contains information about IHE profiles, for example which meta data are available about documents and patients etc. HSD models are derived from this meta model.

A HSD model describes which metadata is actually available in the domain, e.g. which IHE profiles are used, which document types are available, how CDA documents are structured to the point of which coding systems are used in the exchanged documents. Thus, we have information about the domains, workflows, patient etc., which are used for the transformation into another domain.

Derived domain models build the foundation for a transformation between different models and for the cross-domain integration of healthcare data. They can be compared, checked and completed, where applicable. The results can then in turn be applied to the CDA documents.

3.1. Analysis

In the analysis process the systems and healthcare service environments which are considered were identified and it was defined how to use the relations between these systems and domains. Furthermore, information for detailing the domain models like coding systems and similarities in the structure of CDA documents is recorded, to set the semantics into context using a domain specific description.

To achieve this, an analysis of domain knowledge, for example about the processes and terminology of a specific healthcare domain, was necessary and forms the basis for adequate methods and techniques for the model transformation. During a hospital stay, for example, various (CDA) documents are produced (diagnoses, medical summaries, care summaries, operation reports,...). Analyzed IHE profiles provide metadata about these documents as well as detailed patient data, metadata about organizations, domains and workflow information.

Health service domains, inductive connections and especially national and international CDA guidelines were reviewed and checked for similarities, for example Patient Summary (epSOS), Austrian nursing care summary (ELGA) and the Continuity of Care Document (CCD). In all analyzed guidelines, CDA bodies are to be structured at least into sections. What stands out is that the CDA body sections as well as their content in CDA body sections are coded differently (for example LOINC-codes vs. national codes), even if containing equal information. Besides the use of HL7 vocabulary in the CDA header for certain data like gender, marital status and religious affiliation, suggested coding systems vary in general.

3.2. Modeling

The modeling process is based on the analysis results: Domain specific properties of the health service environments are represented in form of a meta model (Kühne 2006), which is – in this case – an abstraction of domain specific models. A domain is in this context a health service environment, which comes with a particular semantic, i.e. with particular terminology.

Each domain provides specific information which is taken into account when deriving a specific health service environment model for a domain. Since the HL7 DIM covers only structured data and requires a detailed definition of processes and data in each domain, building or deriving an explicit DIM from the RIM is not sufficient for practical use in this case. Experience shows that use of terminology and coding systems, which would provide detailed and unambiguous definitions, is only rudimentarily implemented in current practice.

Therefore, more information has to be taken into account and an extended meta model is built in which specifications are embedded like terminology-, coding systems, guidelines, templates, standards and IHE profiles. This results in the use of not only detailed but also rough definitions as domain specific constraints for deriving health service environment models (c.f. Fig.2), e.g.

- which IHE profiles are applied,
- which standards are implemented,
- which guidelines and templates systems, documents etc. are in conformance with, and
- which terminology and coding systems are used.

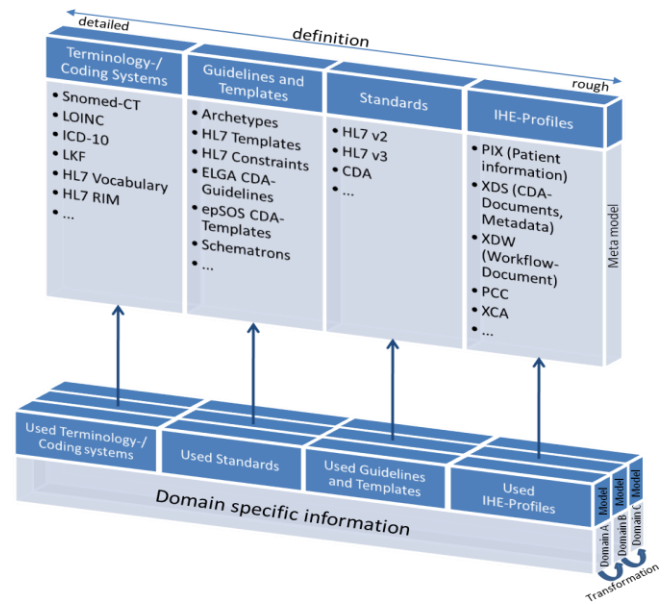


Figure 2: Modeling approach

Using the structure of CDA documents based on the RIM and document metadata provided by a system conformant to IHE recommendations, health service domain models for structured, non-structured and semi-structured data are derived from a meta model.

4. RESULTS AND EVALUATION

The derived HSD models predict content of documents and their relations, based on the particular domain. A specific document of a certain domain describes observed properties. These properties, for example, describe the occurrence of specific terms in a text or in a certain section in a CDA document, which can be deduced by a causal knowledge base.

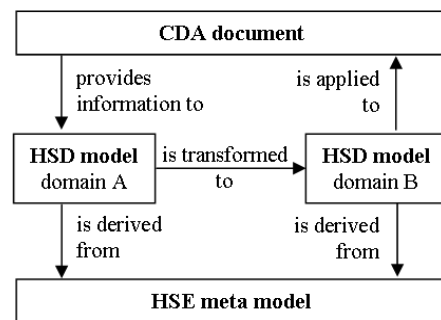


Figure 3: Model transformation

These domain models may be compared, checked and completed, where applicable, by the use of model transformation. The results can then in turn be applied to the CDA documents (c.f. Fig. 3). Thus, it helps to increase the availability of these documents and to improve the communication among healthcare institutions.

Due to known disadvantages of classic knowledge based systems (compare (Giarratano und Riley 2004)), the support of such a system using data driven technologies (Xu und Hou 2009) is focused. Additional technologies to describe the models and their properties, for example MIF, which is more complete than the XML

representation (McKenzie et al. 2011), and an ontology-based approach combined with archetypes as suggested for example by (Duftschmid, Wrba and Rinner 2010), are tested.

The evaluation of the developed models and transformations is currently conducted using our e-Care system. The e-Care system (cf. Fig. 4) is an IHE-conformant platform for the exchange of healthcare data between several healthcare providers (Franz et. al 2009a). It was developed from 2008 to 2010 and has been established since 2010 in an entire region in Upper Austria for the exchange of documents between two clinics, five mobile nursing services and six nursing homes. Therefore, we are able to analyze the practicability of this research in a real-time environment using genuine healthcare data.



Figure 4: e-Care: exchange between hospitals, nursing homes and mobile nursing services

Especially the amount of semantic preservation in an IHE-conformant exchange of healthcare documents is analyzed and how feasible contents can be compared and checked for plausibility.

As plausibility checks cannot be done automatically, the collaboration with users, i.e. healthcare providers, is necessary. We plan to develop further methods and tools to enable health professionals to describe the static and dynamic structure of their health service environment and allow the mapping of one predefined domain into another by themselves.

5. CONCLUSION AND OUTLOOK

This research supports integrated care by enabling comparison, evaluation and completion of documents of various healthcare service environments. Thus, it helps to increase the availability of these documents and to improve the communication among healthcare institutions.

The model transformation approach is not only based on HL7 CDA but also uses metadata provided by an IHE infrastructure, which provides more domain information. Therefore, semantic interoperability may be preserved for highly structured data as well as semi-structured information across domains.

Since the model transformation may not always be automatically achieved, the knowledge of domain experts is necessary. Hence, a specific user interface has to be provided which enables healthcare providers on the one hand to manage the model transformation when

necessary and on the other hand to work efficiently with model transformation results.

Over time, a large amount of documents is accumulated in an IHE-conformant system, especially for multimorbid patients. A next step might be to adapt the models so that they improve and refine themselves in a self-learning way using new documents for a patient. Thus, the interaction with domain experts during the transformation process could be reduced.

ACKNOWLEDGMENTS

This project is supported by the program Regionale Wettbewerbsfähigkeit OÖ 2010-2013, which is financed by the European Regional Development Fund and the Government of Upper Austria.

REFERENCES

- Arrow, K. et al., 2009. Toward a 21st-Century Health Care System: Recommendations for Health Care Reform. *Annals of Internal Medicine*, vol. 150 no. 7 493-495.
- Bointner, K., Duftschmid, G., 2008. Semantische Interoperabilität im elektronischen Gesundheitsdatenaustausch mittels dualer Modellierung – der HL7 Templates Ansatz. In Schreier, G., Hayn, D., Ammenwerth, E., eds.: *Proceedings e-Health2008 & e-Health Benchmarking 2008*, 29-30 May 2008, Vol. 235, Wien, OCG Books.
- Continua Health Alliance, 2010. *About the Alliance*. Available from: <http://www.continuaalliance.org/about-the-alliance.html> [accessed 1 September 2010].
- Duftschmid, G., Wrba, T., Rinner, C., 2010. Extraction of Standardized Archetyped Data from Electronic Health Record Systems based on the Entity-Attribute-Value Model, In: *International Journal of Medical Informatics*; Vol. 79(8); pp. 585 – 597.
- ELGA GmbH, 2010. ELGA. *Elektronische Gesundheitsakte*, Available from: <http://www.arge-elga.at> [accessed 11 August 2010].
- epSOS, 2010. *European Patients Smart Open Services – epSOS*, Available from: <http://www.epsos.eu> [accessed 11 August 2010].
- Faulstich, L. C., Müller, F., Sander, A., Kreuzthaler, M., Kaiser, S., Errath, M., 2008. Semantisches Retrieval medizinischer Freitexte, *Proceedings 53. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie*. Stuttgart, 15-19 September 2008, Düsseldorf, German Medical Science GMS Publishing House
- Franz, B., Lehner, M., Mayr, H., Mayr, H., 2009a. e-Care – IHE-Compliant Patient-Data Exchange in Order to Enable Home & Mobile Care of Elderly People Within an e-Care Affinity Domain. *Proceedings 6th International Conference on Information Technology: New Generations*, 25-27 April 2009, Las Vegas.
- Franz, B., Mayr, H., Mayr, M., Pfeifer, F., Altmann, J., Lehner, M., 2009b. Integrated Care Using a Mod-

- el-Based Patient Record Data Exchange Platform. *Proceedings 21st European Modeling and Simulation Symposium*, 23-25 September 2009, Tenerife, Spain.
- Giarratano, J. C., Riley, G., 2004. *Expert Systems: Principles and Programming*, 4. Ed., Boston, Thomson/PWS Publishing Company.
- Heitmann, K. U., Gobrecht, K., 2009. *HL7 Kommunikationsstandards für das Gesundheitswesen. Ein Überblick*, HL7 Benutzergruppe in Deutschland, Köln, LUP AG Lithographie & Printproduktion.
- Health Level Seven, 2010a. *HL7. Health Level Seven*, Available from: <http://www.hl7.org> [accessed 1 September 2010].
- Health Level Seven, 2010b. *HL7 V3 Clinical Document Architecture, Release 2.0*, ANSI Standard CDA Rel. 2.
- Health Level Seven, 2010c. *HL7 Reference Information Model*, Available from: <http://www.hl7.org/Library/data-model/RIM/C30201/rim.htm> [accessed 3 November 2010].
- IHE International, 2010a. *IHE Information Technology Infrastructure – Technical Framework, Volume 1 Integration Profiles*, Available from: http://www.ihe.net/Technical_Framework/index.cfm#IT [accessed 10 September 2010].
- IHE International, 2010b. *IHE Information Technology Infrastructure – Technical Framework, Volume 2 Transactions*, Available from: http://www.ihe.net/Technical_Framework/index.cfm#IT [accessed 10 September 2010].
- IHE International, 2010c. *Patient Care Coordination. Technical Framework*, Available from: <http://www.ihe.net> [accessed 10 September 2010].
- Kilic, O., Dogac, A., 2009. *Achieving Clinical Statement Interoperability Using R-MIM and Archaic-type-based Semantic Transformations*, IEEE Transactions on Information Technology in Biomedicine, Vol. 13, No. 4, S. 467–477.
- Kühne, T., 2006. *Matters of (Meta-)Modeling*. *Journal on Software and Systems Modeling*, 5(4)369–385, Berlin–Heidelberg, Springer.
- Norgall, T., 2003. Kommunikationsstandard für die Gesundheitstelematik: Status und Ausblick. *Tagungsband e-Health 2003: Telematik im Gesundheitswesen – Vernetzte Versorgung*, Dresden.
- Perhab, F., 2007. Babylonische Sprachverwirrung? Taxonomien, Klassifikationen – gemeinsame Pflegesprache und Wissensmanagement. In: *Österreichische Pflegezeitschrift*. 3/2007, pp. 8–16., Available from: <http://www.oegkv.at> [accessed 11 August 2010].
- Rinner, C., Duftschmid, G., 2009. *Validieren von auf Zweimodell-Ansätzen basierenden, vollstrukturierten EHR-Daten am Beispiel EN/ISO 13606 und HL7 CDA*. In Schreier, G., Hayn, D., Ammenwerth, E., eds.: *eHealth 2009 – Health Informatics meets e-Health: Tagungsband e-Health 2009 & e-Health Benchmarking*, Vol. 250, Wien, OCG Books.
- Spat, S., Cadonna, B., Rakovac, I., Gütl, C., Leitner, H., Stark, G., Beck, P., 2007. *Multi-label Text Classification of German Language Medical Documents*, Health Technol. Inform. 129:1460-1.
- Sunyaev, A., Schweiger, A., Leimeister, J.M., Krcmar, H., 2008. Software-Agenten zur Integration von Informationssystemen im Gesundheitswesen. In Bichler, M., Hess, T., Krcmar, H., Lechner, U., Matthes, F., Picot, A., Speitkamp, B., Wolf, P., eds. *Multikonferenz Wirtschaftsinformatik*, 26-28 February 2008, Berlin, GITO-Verlag.
- VHitG, 2007. *Arztbrief auf Basis der HL7 Clinical Document Architecture Release 2 für das deutsche Gesundheitswesen Version 1.5*, Berlin.
- Wozak, F., Ammenwerth, E., Hörbst, A., Sögner, P., Mair, R., Schabetsberger T., 2008. IHE Based Interoperability – Benefits and Challenges. In Schreier, G., Hayn, D., Ammenwerth, E., eds.: *Proceedings e-Health2008 & e-Health Benchmarking 2008*, 29-30 May 2008, Vol. 235, Wien, OCG Books.
- Xu, J., Hou, Z., 2009. Notes on Data-driven System Approaches. In Tan, T. S., eds. *Acta automatica sinica*, 35(6)668–675, Beijing, The Chinese Association of Automation and the Institute of Automation, Chinese Academy of Sciences.
- Zalunardo, L., Cocchiglia, A., 2011. *IHE IT Infrastructure (ITI) Technical Framework Supplement, Cross-Enterprise Document Workflow (XDW)*. Available from: ftp://ftp.ihe.net/IT_Infrastructure/iheityr9-2011-2012/Technical_Cmte/Profile_Work/XDW/IHE_XDW_v16_23_03_11_draft.doc [accessed 5 April 2011]

AUTHORS' BIOGRAPHY

Barbara Franz MSc, is scientific researcher at the Research Center Hagenberg of the Upper Austria University of Applied Sciences.

Dr. Herwig Mayr is professor for software engineering at the Faculty for Informatics, Communication and Media of the Upper Austria University of Applied Sciences in Hagenberg.

MANAGEMENT OF SUPPLY NETWORKS USING PDES

Carmine De Nicola^(a), Rosanna Manzo^(b), Luigi Rarità^(c)

^{(a), (b), (c)}Dipartimento di Ingegneria Elettronica e Ingegneria Informatica,
University of Salerno, Via Ponte Don Melillo, 84084, Fisciano (SA), Italy

^(a)denicola@diima.unisa.it, ^(b)rmanzo@unisa.it, ^(c)lrarita@unisa.it

ABSTRACT

The paper presents some numerical results for supply networks modeled by a fluid dynamic approach. A mixed continuum-discrete model is examined: the dynamics on each arch is described by a conservation law for the goods density, and an evolution equation for the processing rate. For solving dynamics at nodes, two routing algorithms are considered, maximizing the flux with possible adjustments of the processing rate. Simulations of a supply network have been made assigning a constant input profile with one discontinuity. A functional is defined to locate the discontinuity point in order to maximize the overall production. Parameters dependence is shown solving Riemann Problems with different rules.

Keywords: conservation laws, supply networks, simulation.

1. INTRODUCTION

Control production processes aiming to improve performance in supply networks through the optimal choice of input flow, reduction of dead times and bottlenecks, etc, are of great interest.

Different models have been proposed for supply systems. Most of them are discrete and based on individual parts considerations; others are continuous dealing with ordinary differential equations (see Armbruster et al. 2006a, Armbruster et al. 2006b, Armbruster et al 2004, Daganzo 2003), or/and partial differential equations.

The first paper, that relies on continuous equations, is Armbruster et al. 2006a, where the authors, following a limit procedure on the number of parts and suppliers, have obtained a conservation law, whose flux involves either the goods density or the maximal processing rate.

Due to the difficulties in finding solutions to the general equation proposed in it, other continuous models have been introduced for sequential supply chains (see Bretti et al. 2007, D'Apice et al. 2006, Göttlich et al. 2005), with extensions to networks (Göttlich et al. 2006, D'Apice et al. 2009, Helbing et al. 2004, Helbing et al. 2005).

In this paper, we focus the attention on a discrete-continuous model for supply networks defined in D'Apice et al. 2009. According to it each arch is

modeled by a system of two equations: a conservation law for the goods density, and an evolution equation for the processing rate.

The evolution at a node with one incoming arc and more outgoing ones or with more incoming arcs and one outgoing arc is interpreted thinking of it as a Riemann Problem (RP), a Cauchy Problem with constant initial data on each arc, for the density equation with processing rate data as parameters. RPs are solved using two different "routing" algorithms: the first one allows the redirection of goods to outgoing sub-chains maximizing the flux over incoming sub-chains; the second one is based on the maximization of goods both on incoming and outgoing sub-chains.

Goods flux is maximized for both algorithms also considering two additional rules:

- objects are processed in order to maximize the flux with the minimal value of the processing rate;
- objects are processed in order to maximize the flux: if a solution with only waves in the density exists, then such a solution is taken; otherwise the minimal processing rate wave is produced.

The first rule tends to make adjustments of the processing rate more than the second one, even when it is not necessary for purpose of flux maximization.

Such last rule is more appropriate to reproduce the "Bullwhip effect", see Daganzo 2003: under certain conditions (delays in adaptation of production or delivery rates), the oscillations in delivery and in the resulting inventories (stock level of the products) grow from one producer to the next upstream one, leading to instability with respect to perturbation in the production rate.

The model can be used to study situations characterized by the possibility to reorganize the supply system: in particular, the processing rate can be readapted for some contingent necessity.

Using some ad hoc numerical schemes (see Bretti et al. 2007), based on the classical Godunov method (Godunov, 1959), simulation results have been obtained for a supply network modeling the chips production. In particular, a piecewise constant function with one discontinuity, namely a function of Heavyside type, has been chosen as input profile. In fact, as it happens in

real processes, goods are injected inside supply networks at almost constant levels in different time intervals. The obtained results present some expected features and some unexpected ones: the production, measured by the density on the last arc of the chain, is strongly influenced by the discontinuity point of the input profile; unexpectedly, analysis on the final product flow indicates that final goods start to be produced always at the same temporal instant, independently from the choice of the discontinuity in the input profile. Moreover, discontinuity shifts do not imply simply temporal translations of final product flows, hence indicating the presence of a strong non linearity for the whole system. Finally, a numerical study of the temporal integral of the final product flow (representing the number of produced goods) shows the existence of a time instant at which the discontinuity point of the input profile has to be placed for the maximization of the overall production. This value does not depend on the rules used to solve the dynamics at nodes.

The outline of the paper is the following. Section 2 deals with the mathematical model for supply networks. Riemann Solvers at nodes are described considering different routing algorithms, and finally an example is reported. In Section 3, the numerical results obtained for a supply network are considered and discussed. Finally, the paper ends with conclusions in Section 4.

2. MATHEMATICAL MODEL

A supply network is a finite connected graph consisting of a finite set of arcs (sub-chains) $I = \{I_k : k = 1, \dots, N+1\}$ and a finite set of junctions P .

On each sub-chain I_k (see D'Apice et al. 2009) we consider the system:

$$\begin{cases} (\rho_k)_t + f_\varepsilon^k(\rho_k, \mu_k)_x = 0, \\ (\mu_k)_t - (\mu_k)_x = 0, \end{cases} \quad (1)$$

where $\rho_k(t, x)$ and $\mu_k(t, x)$ are, respectively, the density of the processed objects on I_k and the production rate of I_k , while f_ε^k is the flux, defined as follows:

$$f_\varepsilon^k(\rho_k, \mu_k) = \begin{cases} \rho_k, & 0 \leq \rho_k \leq \mu_k, \\ \mu_k + \varepsilon(\rho_k - \mu_k), & \mu_k \leq \rho_k \leq \rho_k^{\max}, \end{cases} \quad (2)$$

or, alternatively,

$$f_\varepsilon^k(\rho_k, \mu_k) = \begin{cases} \varepsilon\rho_k + (1-\varepsilon)\mu_k, & 0 \leq \mu_k \leq \rho_k, \\ \rho_k, & \rho_k \leq \mu_k \leq \mu_k^{\max}, \end{cases} \quad (3)$$

where ρ_k^{\max} and μ_k^{\max} are, respectively, the maximum density and processing rate. From now on, we assume

that ε is fixed and, for simplicity, we drop the indices, thus indicating the flux by $f(\rho_k, \mu_k)$.

Remark We can consider different fluxes f_ε^k for each sub-chain I_k (also choosing ε dependent on k), or different slopes m_k for each sub-chain I_k :

$$f_\varepsilon^k(\rho_k, \mu_k) = \begin{cases} m_k \rho_k, & 0 \leq \rho_k \leq \frac{\mu_k}{m_k}, \\ \mu_k + \varepsilon(m_k \rho_k - \mu_k), & \frac{\mu_k}{m_k} \leq \rho_k \leq \rho_k^{\max}, \end{cases} \quad (4)$$

where $m_k \geq 0$ represents the velocity of each processor and is given by:

$$m_k = \frac{L_k}{T_k}, \quad (5)$$

with L_k and T_k , respectively, fixed length and processing time of processor I_k .

Sub-chains are connected by junctions P , each one having a finite number of incoming sub-chains and outgoing ones. Hence, we identify P with $((i_1, \dots, i_n), (j_1, \dots, j_m))$ where the first n -tuple indicates the set of incoming sub-chains and the second m -tuple the set of outgoing sub-chains. Each sub-chain can be incoming sub-chain at most for one junction and outgoing at most for one junction.

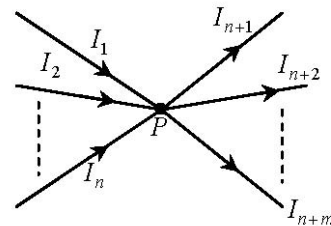


Figure 1: Junction P with n incoming sub-chains and m outgoing ones

The supply network evolution is described on each arc I_k by a finite set of functions (ρ_k, μ_k) defined on $[0, +\infty[\times I_k$. Dynamics at a junction is obtained solving RPs.

Definition A Riemann Solver (RS) for the junction P with n incoming sub-chains and m outgoing ones (of $n \times m$ type) is a map that associates to a Riemann data $(\rho_0, \mu_0) = (\rho_{1,0}, \mu_{1,0}, \dots, \rho_{n+m,0}, \mu_{n+m,0})$ at P a vector $(\hat{\rho}_0, \hat{\mu}_0) = (\hat{\rho}_1, \hat{\mu}_1, \dots, \hat{\rho}_{n+m}, \hat{\mu}_{n+m})$ so that the solution is given by the waves $(\rho_{i,0}, \hat{\rho}_i)$ and $(\mu_{i,0}, \hat{\mu}_i)$ on the sub-

chain $I_i, i=1, \dots, n$ and by the waves $(\hat{\rho}_j, \rho_{j,0})$ on the sub-chain $I_j, j=n+1, \dots, n+m$. We require the consistency condition $RS(RS((\rho_0, \mu_0))) = RS((\rho_0, \mu_0))$.

2.1. Riemann Solvers for suppliers

We discuss RSs for two types of nodes, according to the real case we examine here (for more detail refer to Bretti et al. 2007 and D'Apice et al. 2009):

1. a node with two incoming sub-chains and one outgoing one (2×1);
2. a node with one incoming sub-chain and two outgoing ones (1×2).

For a given arc I_k , (1) is a system of conservation laws in the variables $U = (\rho, \mu)$, namely:

$$U_t + F(U)_x = 0, \quad (6)$$

with flux function

$$F(U) = (f(\rho, \mu), -\mu). \quad (7)$$

Eigenvalues and eigenvectors are:

$$\lambda_1(\rho, \mu) \equiv -1, \quad r_1(\rho, \mu) = \begin{cases} \begin{pmatrix} 0 \\ 1 \end{pmatrix}, & \rho < \mu, \\ \begin{pmatrix} -\frac{1-\varepsilon}{1+\varepsilon} \\ 1 \end{pmatrix}, & \rho > \mu, \end{cases} \quad (8)$$

$$\lambda_2(\rho, \mu) = \begin{cases} 1, & \rho < \mu, \\ \varepsilon, & \rho > \mu, \end{cases} \quad r_2(\rho, \mu) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \quad (9)$$

Hence the Hugoniot curves for the first family are vertical lines above the secant $\rho = \mu$ and lines with slope close to $-1/2$ below the same secant. The Hugoniot curves for the second family are just horizontal lines. Since we consider positive and bounded values for the variables, we fix the invariant region:

$$D = \{(\rho, \mu) : 0 \leq \rho \leq \rho_{\max}, 0 \leq \mu \leq \mu_{\max}, 0 \leq (1+\varepsilon)\rho + (1-\varepsilon)\mu \leq (1+\varepsilon)\rho_{\max} = 2\mu_{\max}\}. \quad (10)$$

Observe that:

$$\rho_{\max} = \mu_{\max} \frac{2}{1+\varepsilon}. \quad (11)$$

We consider a node P of $n \times m$ type and a Riemann initial datum $(\rho_{1,0}, \mu_{1,0}, \dots, \rho_{n+m,0}, \mu_{n+m,0})$. The following Lemma holds:

Lemma *On the incoming sub-chains, only waves of the first family may be produced, while on the outgoing sub-chains only waves of the second family may be produced.*

From such Lemma, given the initial datum, for every RS it follows that:

$$\begin{aligned} \hat{\rho}_i &= \varphi(\hat{\mu}_i), \quad i=1, \dots, n, \\ \hat{\mu}_j &= \mu_{j,0}, \quad j=n+1, \dots, n+m, \end{aligned} \quad (12)$$

where the function $\varphi(\cdot)$ describes the first family curve through $(\rho_{k,0}, \mu_{k,0})$ as function of $\hat{\mu}_k$:

$$\varphi(\hat{\mu}_k) = \begin{cases} \bar{\mu}_k, & \hat{\mu}_k \geq \bar{\mu}_k, \\ \frac{(\varepsilon-1)\hat{\mu}_k + 2\rho_{k,0}}{1+\varepsilon}, & \hat{\mu}_k < \bar{\mu}_k, \rho_{k,0} \leq \mu_{k,0}, \\ \frac{(\varepsilon-1)(\hat{\mu}_k - \mu_{k,0})}{1+\varepsilon} + \rho_{k,0}, & \hat{\mu}_k < \bar{\mu}_k, \rho_{k,0} > \mu_{k,0}, \end{cases} \quad (13)$$

where $\bar{\mu}_k$ is the point at which the first family curve changes:

$$\bar{\mu}_k = \begin{cases} \rho_{k,0}, & \rho_{k,0} \leq \mu_{k,0}, \\ \frac{1+\varepsilon}{2}\rho_{k,0} + \frac{1-\varepsilon}{2}\mu_{k,0}, & \rho_{k,0} > \mu_{k,0}. \end{cases} \quad (14)$$

We define two different RSs at a junction to represent two different routing algorithms:

RA1. *We assume that:*

- (A) *the flow from incoming sub-chains is distributed on outgoing ones according to fixed coefficients;*
- (B) *respecting (A), the processor chooses to process goods in order to maximize fluxes (i.e., the number of goods which are processed) on incoming sub-chains.*

RA2. *We assume that the number of goods through the junction is maximized both over incoming and outgoing sub-chains.*

For both routing algorithms we can maximize the flux of goods considering one of the two additional rules:

SC2. *The objects are processed in order to maximize the flux with the minimal value of the processing rate.*

SC3. *The objects are processed in order to maximize the flux. If a solution with only waves in the density ρ exists, then such solution is taken, otherwise the minimal μ wave is produced.*

To define RPs according to rules RA1 and RA2, we introduce the notation:

$$f_k = f(\rho_k, \mu_k), \quad (15)$$

and define the maximum flux that can be obtained by a wave solution on each production sub-chain:

$$f_k^{\max} = \begin{cases} \bar{\mu}_k, & k = 1, \dots, n, \\ \mu_{k,0} + \varepsilon(\rho_{\max} - \mu_{k,0}) \frac{\rho_{\max} - \mu_{\max}}{\mu_{\max}} - \mu_{k,0}, & k = n+1, \dots, n+m. \end{cases} \quad (16)$$

It is possible to prove that a necessary and sufficient condition for the solvability of RPs at nodes is

$$\sum_{i=1}^n f_i^{\min} \leq \sum_{j=n+1}^{n+m} \left[\mu_{j,0} + \varepsilon(\rho_{\max} - \mu_{j,0}) \frac{\rho_{\max} - \mu_{\max}}{\mu_{\max}} - \mu_{j,0} \right], \quad (17)$$

where

$$f_i^{\min}((\rho_0, \mu_0)) = \begin{cases} \frac{2\varepsilon}{1+\varepsilon} \rho_0, & \rho_0 \leq \mu_0, \\ \varepsilon \rho_0 + \frac{\varepsilon(1-\varepsilon)}{1+\varepsilon} \mu_0, & \rho_0 > \mu_0. \end{cases} \quad (18)$$

2.1.1. One outgoing sub-chain

In this case, algorithms RA1 and RA2 coincide since there is only one outgoing sub-chain.

We fix a node P with 2 incoming arcs (labelled by 1 and 2) and 1 outgoing one (indicated by 3) and a Riemann initial datum given by $(\rho_0, \mu_0) = (\rho_{1,0}, \mu_{1,0}, \rho_{2,0}, \mu_{2,0}, \rho_{3,0}, \mu_{3,0})$. Let us denote with $(\hat{\rho}, \hat{\mu}) = (\hat{\rho}_1, \hat{\mu}_1, \hat{\rho}_2, \hat{\mu}_2, \hat{\rho}_3, \hat{\mu}_3)$ the solution of the RP at P . We introduce a priority parameter $q \in]0, 1[$, that indicates a *level of priority* at the junction of incoming sub-chains. We define:

$$\Gamma = \min\{\Gamma_{inc}, \Gamma_{out}\}, \quad (19)$$

where

$$\Gamma_{inc} = \sum_{i=1}^2 f_i^{\max}, \quad \Gamma_{out} = f_3^{\max}. \quad (20)$$

First, we compute \hat{f}_i , $i = 1, 2, 3$ according to rules (SC2) and (SC3). Introduce the conditions:

$$(A1) \quad q f_3^{\max} < f_1^{\max};$$

$$(A2) \quad (1-q) f_3^{\max} < f_2^{\max}.$$

If $\Gamma = \Gamma_{inc}$, we get that $\hat{f}_i = f_i^{\max}$, $i = 1, 2$, $\hat{f}_3 = f_1^{\max} + f_2^{\max}$.

If $\Gamma < \Gamma_{inc}$, we have that:

- $\hat{f}_1 = q f_3^{\max}$, $\hat{f}_2 = (1-q) f_3^{\max}$, $\hat{f}_3 = f_3^{\max}$ when A1 and A2 are both satisfied;

- $\hat{f}_1 = f_3^{\max} - f_2^{\max}$, $\hat{f}_2 = f_2^{\max}$, $\hat{f}_3 = f_3^{\max}$ when A1 holds and A2 is not satisfied;
- $\hat{f}_1 = f_1^{\max}$, $\hat{f}_2 = f_3^{\max} - f_1^{\max}$, $\hat{f}_3 = f_3^{\max}$ when A1 is not satisfied and A2 holds.

The case of both A1 and A2 false is not possible, since it would be $f_3^{\max} > \Gamma_{inc}$.

Now, we compute $\hat{\rho}_k$ and $\hat{\mu}_k$, $k = 1, 2, 3$. On the incoming sub-chains i , $i = 1, 2$, we have to distinguish two subcases.

If $\hat{f}_i = f_i^{\max}$, according to rules SC2 and SC3, we get:

$$SC2 : \begin{cases} \hat{\rho}_i = \bar{\mu}_i, \\ \hat{\mu}_i = \bar{\mu}_i, \end{cases} \quad SC3 : \begin{cases} \hat{\rho}_i = \bar{\mu}_i, \\ \hat{\mu}_i = \max\{\bar{\mu}_i, \mu_{i,0}\}. \end{cases} \quad (21)$$

If $\hat{f}_i < f_i^{\max}$, for both SC2 and SC3 rules, we get that $\hat{\mu}_i$, $i = 1, 2$, solves the equation:

$$\hat{\mu}_i + \varepsilon(\varphi(\hat{\mu}_i) - \hat{\mu}_i) = \hat{f}_i, \quad (22)$$

while

$$\hat{\rho}_i = \varphi(\hat{\mu}_i), \quad i = 1, 2. \quad (23)$$

On the outgoing sub-chain we have, for both rules SC2 and SC3:

$$\hat{\mu}_3 = \mu_{3,0}, \quad (24)$$

while $\hat{\rho}_3$ is the unique value solving the equation $f_3(\mu_{3,0}, \hat{\rho}_3) = \hat{f}_3$, namely:

$$\hat{\rho}_3 = \begin{cases} \hat{f}_3, & \hat{f}_3 \leq \mu_{3,0}, \\ \frac{\hat{f}_3 - \mu_{3,0}}{\varepsilon} + \mu_{3,0}, & \hat{f}_3 > \mu_{3,0}. \end{cases} \quad (25)$$

2.1.2. One incoming sub-chain

Consider a node P with 1 incoming arc, labelled by 1, and 2 outgoing ones, indicated by 2 and 3 and an initial datum $(\rho_0, \mu_0) = (\rho_{1,0}, \mu_{1,0}, \rho_{2,0}, \mu_{2,0}, \rho_{3,0}, \mu_{3,0})$.

We introduce a distribution parameter $\alpha \in]0, 1[$, that indicates the percentage of goods, which, from the incoming arc 1, is directed to the outgoing arc 2 (obviously, the arc 3 is interested by a percentage of goods equal to $1-\alpha$). We have different solutions for algorithms RA1 and RA2. In what follows, the asymptotic solution is reported only for the RA1 algorithm, since RA2 is solved as for the node with one outgoing sub-chain.

As usual, we first compute the fluxes solutions. Following rules (A) and (B) of the algorithm RA1, we get that:

$$\hat{f}_1 = \min \left\{ f_1^{\max}, \frac{f_2^{\max}}{\alpha}, \frac{f_3^{\max}}{1-\alpha} \right\}, \hat{f}_2 = \alpha \hat{f}_1, \hat{f}_3 = (1-\alpha) \hat{f}_1. \quad (26)$$

Densities and processing rates, $\hat{\rho}_i$, and $\hat{\mu}_i, i=1,2,3$, are obtained as follows.

If $\hat{f}_1 = f_1^{\max}$, we get:

$$SC2 : \begin{cases} \hat{\rho}_1 = \bar{\mu}_1, \\ \hat{\mu}_1 = \bar{\mu}_1, \end{cases} \quad SC3 : \begin{cases} \hat{\rho}_1 = \bar{\mu}_1, \\ \hat{\mu}_1 = \max\{\bar{\mu}_1, \mu_{1,0}\}. \end{cases} \quad (27)$$

If $\hat{f}_1 < f_1^{\max}$, $\hat{\mu}_1$, either for rule SC2 or SC3, satisfies the equation:

$$\hat{\mu}_1 + \varepsilon(\varphi(\hat{\mu}_1) - \hat{\mu}_1) = \hat{f}_1, \quad (28)$$

while:

$$\hat{\rho}_1 = \varphi(\hat{\mu}_1). \quad (29)$$

On the outgoing sub-chain j , $j=2,3$, for both rules SC2 and SC3, we have that:

$$\hat{\mu}_j = \mu_{j,0}, \quad (30)$$

while $\hat{\rho}_j$ solves the equation $f_j(\mu_{j,0}, \hat{\rho}_j) = \hat{f}_j$, namely:

$$\hat{\rho}_j = \begin{cases} \hat{f}_j, & \hat{f}_j \leq \mu_{j,0}, \\ \frac{\hat{f}_j - \mu_{j,0}}{\varepsilon} + \mu_{j,0}, & \hat{f}_j > \mu_{j,0}. \end{cases} \quad (31)$$

Remark For sequential sub-chains (one incoming arc, 1, and one outgoing arc, 2), the fluxes solutions are $\hat{f}_1 = \hat{f}_2 = \min\{f_1^{\max}, f_2^{\max}\}$ while $\hat{\rho}_i$, and $\hat{\mu}_i, i=1,2$, are obtained for rules SC2 and SC3 as before.

2.2. Example

In what follows we report densities and production rates at the instant $t=0$ and after some times (at $t=1$) for different initial data using different routing algorithms.

We consider a node of type 2 x 1, assuming the following data:

$$\begin{aligned} \varepsilon &= 0.25, \mu_i^{\max} = 0.8, i=1,2,3, \\ (\rho_{1,0}, \rho_{2,0}, \rho_{3,0}) &= (0.35, 0.2, 0.6), \\ (\mu_{1,0}, \mu_{2,0}, \mu_{3,0}) &= (0.95, 0.55, 0.3). \end{aligned} \quad (32)$$

As there is only one outgoing sub-chain, algorithms RA1 and RA2 coincide and the choice

$q=0.6$ indicates that 60% of goods flow is directed from arc 1 to the outgoing one. In Table 1, numerical results for asymptotic fluxes, densities and production rates are reported while Figures 1 and 2 show the behaviour of density and production rate waves. For both rules SC2 and SC3, the results are the same, with the exception of values $\hat{\mu}_1$ and $\hat{\mu}_2$ for rule SC3. For sub-chain 3, a shock wave in the density connect the initial and the asymptotic state while, for sub-chain 1 and 2, there is no waves formation (Figure 2). A similar situation happens for production rates (Figure 3): in the case SC2, only sub-chains 1 and 2 are interested by waves formation. For rule SC3, shock formations do not occurs, as all sub-chains have asymptotic states equal to the initial ones. In fact SC2 tends to make adjustments of the processing rate more than SC3.

Table 1: Numerical results for a node of 2 x 1 type

RA1 = RA2		
	SC2	SC3
\hat{f}	(0.35, 0.2, 0.55)	(0.35, 0.2, 0.55)
$\hat{\rho}$	(0.35, 0.2, 1.01)	(0.35, 0.2, 1.01)
$\hat{\mu}$	(0.35, 0.2, 0.3)	(0.95, 0.55, 0.3)

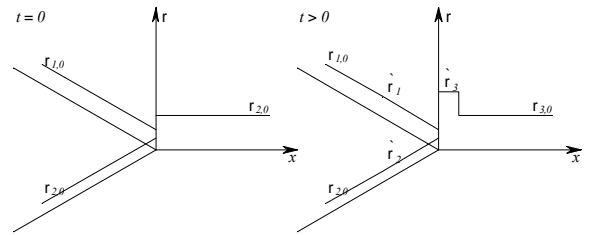


Figure 2: Densities at $t=0$ and $t=1$ on sub-chains for rules SC2 and SC3

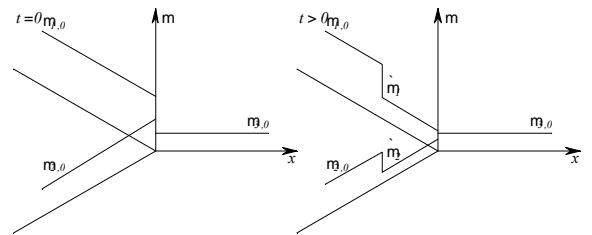


Figure 3: Production rates at $t=0$ and $t=1$ on sub-chains for rule SC2

3. SIMULATIONS

In this section, we present some simulation results to foresee the behaviour of goods fluxes on a supply network. In particular, we study how to choose the injection times of different goods levels to increase the production.

3.1. Numerical methods

We refer to a Godunov method for a 2×2 system (details are in Bretti et al. 2007, Godunov 1959), which

is described as follows. Define a discrete grid in the plane (x, t) , whose points are $(x_j, t^n) = (j\Delta x, n\Delta t)$, $j \in \mathbb{Z}$, $n \in \mathbb{Z}$, and indicate by ${}^k \rho_j^n$ and ${}^k \mu_j^n$, respectively, the approximations of density and production rate of the arc I_k in the point (x_j, t^n) . An approximation scheme for the system (1) reads as:

$$\begin{cases} {}^k \rho_j^{n+1} = {}^k \rho_j^n - \frac{\Delta t}{\Delta x} \left(g({}^k \rho_j^n, {}^k \rho_{j+1}^n) - g({}^k \rho_{j-1}^n, {}^k \rho_j^n) \right), \\ {}^k \mu_j^{n+1} = {}^k \mu_j^n + \frac{\Delta t}{\Delta x} ({}^k \mu_{j+1}^n - {}^k \mu_j^n), \end{cases} \quad (33)$$

where the Godunov numerical flux g is found solving RPs among the states (ρ_-, μ_-) on the left and (ρ_+, μ_+) on the right:

$$g(\rho_-, \mu_-, \rho_+, \mu_+) = \begin{cases} (\rho_-, -\mu_+), & \rho_- < \mu_-, \rho_- \leq \mu_+, \\ \left(\frac{1-\varepsilon}{1+\varepsilon} \mu_+ + \frac{2\varepsilon}{1+\varepsilon} \rho_-, -\mu_+ \right), & \rho_- < \mu_-, \rho_- > \mu_+, \\ \left(\frac{1+\varepsilon}{2} \rho_- + \frac{1-\varepsilon}{2} \mu_-, -\mu_+ \right), & \rho_- \geq \mu_-, \mu_+ > \tilde{\mu}, \\ \left(\frac{1-\varepsilon}{1+\varepsilon} (\mu_+ + \varepsilon \mu_-) + \varepsilon \rho_-, -\mu_+ \right), & \rho_- \geq \mu_-, \mu_+ \leq \tilde{\mu}, \end{cases} \quad (34)$$

with

$$\tilde{\mu} = \mu_- + \frac{1+\varepsilon}{2} (\rho_- - \mu_-). \quad (35)$$

We need to introduce the boundary data value, given by the term ${}^k \rho_{j-1}^n$. For the first arc of the supply network, ${}^k \rho_{j-1}^n$ is defined by an assigned input profile; otherwise, ${}^k \rho_{j-1}^n$ is determined by the solution to RPs at nodes.

Remark. The construction of the Godunov method is based on the exact solution to the RP in the cell $]x_{j-1}, x_j[\times]t^n, t^{n+1}[$. To avoid the interaction of waves in two neighbouring cells before time Δt , we impose a CFL condition like:

$$\frac{\Delta t}{\Delta x} \max \{ |\lambda_0|, |\lambda_1| \} \leq \frac{1}{2}, \quad (36)$$

where λ_0 and λ_1 are the eigenvalues of system (1). Since, in this case, the eigenvalues are such that $|\lambda_0| = 1$, $|\lambda_1| \leq 1$, the CFL condition reads as:

$$\frac{\Delta t}{\Delta x} \leq \frac{1}{2}. \quad (37)$$

3.2. A complex network

We present some simulation results for a supply network, whose topology is in Figure 4.

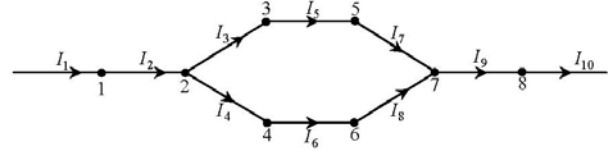


Figure 4: Network with 8 nodes and 10 arcs

Such a network can model the chips production. First, potatoes are washed (arc I_1) and then they are skinned off (arc I_2). Assuming that two different types of fried potatoes are produced (classical and stick, for example), node 2 is a diverging point: a percentage α of potatoes are sent to arc I_3 for stick chips production, and a percentage $1-\alpha$ to arc I_4 for the classical potatoes production. On arcs I_5 and I_6 , potatoes are fried and on arcs I_7 and I_8 they are salted. Node 7 is a merging point: considering a certain priority level q , potatoes are directed to arc I_9 where they are put in envelope; on arc I_{10} , the obtained packets are sealed.

The goods evolution inside the supply network is simulated in a time interval $[0, T]$, with $T = 1000$ min,

using the approximation scheme (33) with $\frac{\Delta t}{\Delta x} = \frac{1}{2}$.

The dynamics at node 2 is solved using the RA1 algorithm. In fact, the redirection of potatoes in order to maximize the production on both incoming and outgoing sub-chains is not possible, since classical and stick potatoes have different shapes. Moreover, at node 2, we use rule SC2 and a distribution coefficient $\alpha = 0.3$ for arc I_3 . At node 7, dynamics is solved using the RA1 algorithm with rule SC3 and priority level $q = 0.4$ for arc I_7 (notice that, for such last node, algorithm RA1 and RA2 coincide).

We assume that, at the beginning of the simulation ($t = 0$), all arcs are empty. Moreover, in Table 2, initial conditions for processing rates, maximal processing rates, lengths and processing times, are reported for each arc I_k , $k = 1, \dots, 10$.

Table 2: Parameters for the supply network

I_k	$\mu_{k,0}$	μ_k^{\max}	L_k	T_k
1	10	15	15	15
2	7	10	30	30
3	7	10	20	20

4	15	20	15	15
5	5	8	20	20
6	5	10	20	20
7	12	12	20	20
8	10	10	25	25
9	15	15	15	15
10	10	10	10	10

Maximal densities on arcs are obtained using equation (11), where we consider $\varepsilon = 0.2$. Boundary data are also needed: for arc 1, it represents the amount of goods, that have to be processed inside the supply network; for arc 8, it is a sort of wished production.

The input profile for arc 1 is chosen as a constant piecewise function with one discontinuity, namely a Heavyside function. In fact, during production processes, goods are injected inside supply networks at almost constant levels in different time intervals:

$$\rho_{1,b}(t,0) = \begin{cases} 30, & 0 \leq t \leq \bar{t}, \\ 5, & \bar{t} < t \leq T, \end{cases} \quad (38)$$

where \bar{t} is the time instant at which the injection levels inside the supply network abruptly change. Notice that levels 30 and 5 of $\rho_{1,b}$ have been chosen according to the following criterion: when $0 \leq t \leq \bar{t}$, the arcs of the supply network process a great amount of goods and often reach the maximal density; when $\bar{t} < t \leq T$, the arcs process goods whose density is always less than the maximal one.

For arc 8, we assume a boundary datum equal to $\rho_{10}^{\max} \square 16.667$, hence we require a possible wished output near to the maximal density processed by arc 10.

The aim is to choose some \bar{t} value, that guarantees maximal production. First we examine the behaviour of $\rho_{10}(t,x)$, for $\bar{t} = 100$ and $\bar{t} = 500$. The overall system is completely influenced by \bar{t} . In Figure 5, we notice one production peak at time, approximately, $t = 400$, but the average level of density is quite low (about 0.6). Such phenomenon is not present in Figure 6, where there is one peak production, and, after it, the production decreases slowly until it reaches a fixed constant level.

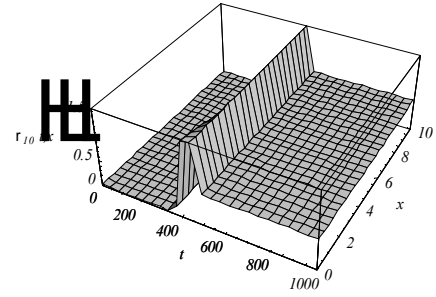


Figure 5: $\rho_{10}(t,x)$ for $\bar{t} = 100$

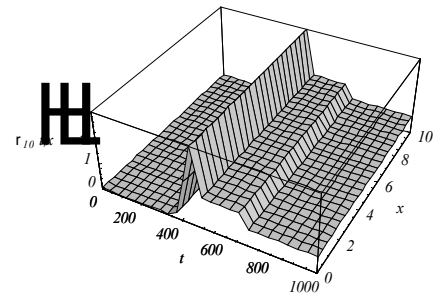


Figure 6: $\rho_{10}(t,x)$ for $\bar{t} = 500$

In Figure 7 and 8, fixing $\bar{t} = 500$ we show how the dynamics of the supply network is influenced by different choices of RSs at nodes 2 and 7.

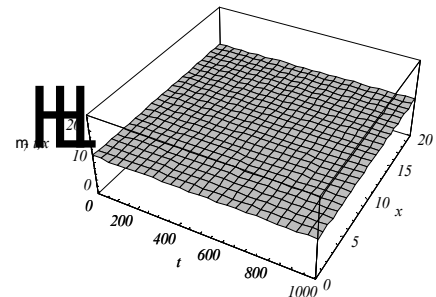


Figure 7: $\mu_7(t,x)$ for $\bar{t} = 500$ using rule SC2 at node 2 and SC3 at node 7

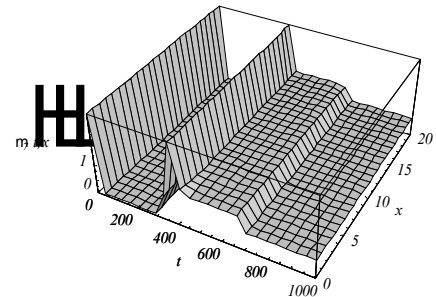


Figure 8: $\mu_7(t,x)$ for $\bar{t} = 500$ using rule SC3 at node 2 and SC2 at node 7

The function $f(\rho_{10}(t, L_{10}), \mu_{10}(t, L_{10}))$, namely the flux on the last point of arc 10, in the case of rules SC2 - SC3, is depicted in Figure 9 for different choices of \bar{t} to understand the final product flows. The obtained results present some interesting features: first, although different values of \bar{t} are used, the flux starts to be different from zero always at the same temporal instant ($t \approx 350$), indicating that the input flow does not influence the production dynamics, that depends only on network characteristics (initial conditions, maximal processing rates, arcs length, and so on); second, shifts of the input flow discontinuity do not foresee translations of $f(\rho_{10}(t, L_{10}), \mu_{10}(t, L_{10}))$. Such phenomenon indicates that, also using a conservation law with a linear function and a transport equation for the production rates, the dynamics on the whole network is strongly not linear.

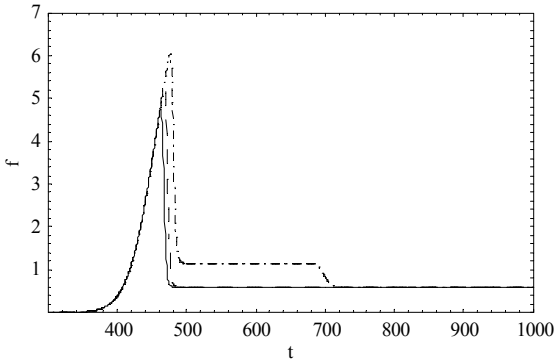


Figure 9: $f(\rho_{10}(t, L_{10}), \mu_{10}(t, L_{10}))$ evaluated for different values of the discontinuity instant: $\bar{t} = 100$ (continuous line), $\bar{t} = 200$ (dashed line) and $\bar{t} = 500$ (dot-dashed line)

The area described by $f(\rho_{10}(t, L_{10}), \mu_{10}(t, L_{10}))$, that can have strong variations for different \bar{t} , represents the number of goods produced at the end of the simulation. In particular, we could ask if there exists a value \bar{t} for which

$$J = \int_0^{\bar{t}} f(\rho_{10}(t, L_{10}), \mu_{10}(t, L_{10})) dt \quad (39)$$

is maximum.

In Figure 10, $J(\bar{t})$ is reported for the following combination of rules at nodes 2 and 7: SC2 - SC2, SC2 - SC3, SC3 - SC2, and SC3 - SC3. We observe that $J(\bar{t})$ almost increases linearly for a wide range of values of \bar{t} (precisely if $\bar{t} \in [200, 700]$), until it reaches a maximum \bar{t}^{\max} , and then it almost decreases in a constant way.

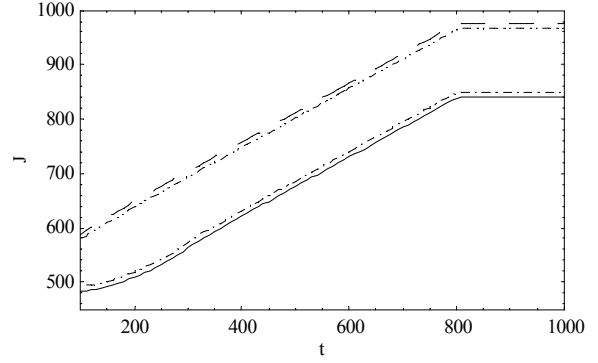


Figure 10: behaviour of $J(\bar{t})$ for different combinations of rules at nodes 2 and 7: SC2 - SC2 (dot dashed line); SC2 - SC3 (continuous line); SC3 - SC2 (dashed line); SC3 - SC3 (dot dot dashed line)

We get that \bar{t}^{\max} is almost insensible to rules at nodes 2 and 7 and its numerical approximation is $\bar{t}^{\max} \approx 830$. The just made analysis strictly depends on the input flow characteristics and network parameters. In general, the behaviour depicted in Figure 8 is a priori unpredictable due to the non linearity of supply networks, as confirmed by other similar simulation.

In Figures 11 and 12, $\rho_{10}(t, x)$ and $\mu_{10}(t, x)$ are represented for $\bar{t} \approx 830$ in the case of SC2 - SC3 rules: $\rho_{10}(t, x)$ is higher with respect to other cases already examined in Figure 5 and 6; $\mu_{10}(t, x)$ is constant and equal to 10. Such result is not surprising since, according to RSs at nodes, the production rates are kept equal to the initial ones on outgoing sub-chains.

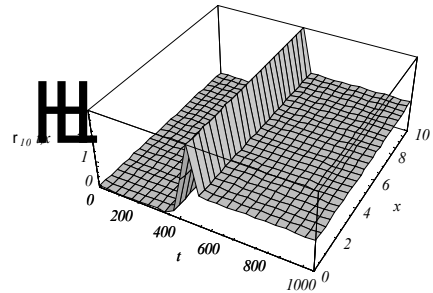


Figure 11: $\rho_{10}(t, x)$ for $\bar{t} \approx 830$

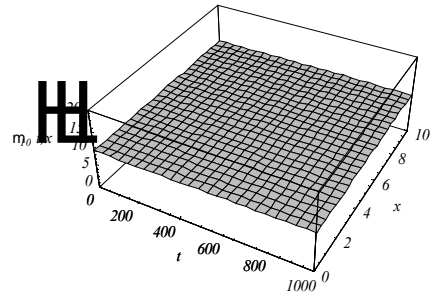


Figure 12: $\mu_{10}(t, x)$ for $\bar{t} \approx 830$

A further remark can be done on the dependence of $J(\bar{t})$ by the distribution coefficient α at node 2. In Figure 13, we represent different pictures of $J(\bar{t})$, evaluated using rules SC2 – SC3, for different values of α . It is evident that, if α grows, $J(\bar{t})$ becomes higher but the value of \bar{t} at which it attains its maximum point has no meaningful variations.

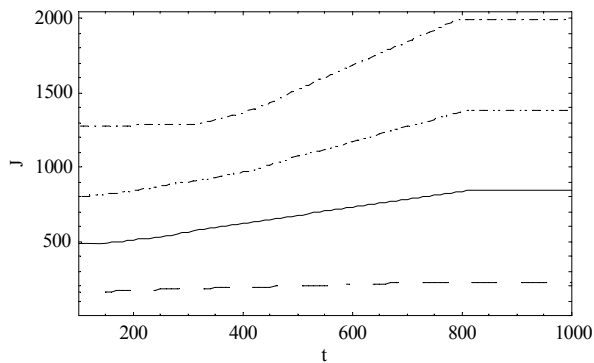


Figure 13: behaviour of $J(\bar{t})$ using SC2 – SC3 for different values of α : $\alpha = 0.1$ (dashed line); $\alpha = 0.3$ (continuous line); $\alpha = 0.5$ (dot dot dashed line) and $\alpha = 0.8$ (dot dashed line)

4. CONCLUSIONS

In this paper, starting from the model proposed in D'Apice et al. 2009, goods flows on a supply network have been studied.

An input flow of piecewise constant type with only one discontinuity has been chosen for simulating the behaviour of a supply network for chips production.

Recent studies on experimental data seems to confirm the correctness of the assumptions underlying the model. In particular, the real flow profiles on each arc are consistent with the shapes of the flux functions.

For such a network it has been proven that an accurate choice of the discontinuity point allows to maximize the total final production. The influence of the supply evolution on RSs at nodes and on the distribution parameter is analyzed.

In future we aim to develop numerical schemes to solve the optimal control problem of choosing an input flow of piecewise constant type in order to obtain an expected pre-assigned network outflow. The idea is to find the minimum of a cost functional measuring the network outflow evaluating its derivative with respect to the switching times (the controls) of the input flows through the evolution of generalized tangent vectors to the control and to the solution of the supply chain model.

REFERENCES

Armbruster, D., Degond, P., Ringhofer, C., 2006. A model for the dynamics of large queueing networks and supply chains, *SIAM Journal on Applied Mathematics*, 66 (3), pp. 896-920.

Armbruster, D., Degond, P., Ringhofer, C., 2006. Kinetic and fluid models for supply chains supporting policy attributes, *Transportation Theory Statist. Phys.*

Armbruster, D., Marthaler, D., Ringhofer, C., 2004. Kinetic and fluid model hierarchies for supply chains, *SIAM J. on Multiscale Modeling*, 2 (1), pp. 43-61.

Bretti, G., D'Apice, C., Manzo, R., Piccoli, B., 2007. A continuum - discrete model for supply chains dynamics, *Networks and Heterogeneous Media (NHM)*, 2 (4), pp. 661-694.

Daganzo, C., 2003. A Theory of Supply Chains, *Springer Verlag, New York, Berlin, Heidelberg*.

D'Apice, C., Manzo, R., 2006. A fluid dynamic model for supply chains, *Networks and Heterogeneous Media (NHM)*, 1 (3), pp. 379-398.

D'Apice, C., Manzo, R., Piccoli, B., 2009. Modelling supply networks with partial differential equations, *Quarterly of Applied Mathematics*, 67 (3), pp. 419-440.

Helbing, D., Lammer, S., Seidel, P., Seba, T., Platkowski, T., 2004. Physics, stability and dynamics of supply networks, *Physical Review E* 70, 066116.

Helbing, D., Lammer, S., 2005. Supply and production networks: from the bullwhip effect to business cycles, in *D. Armbruster, A. S. Mikhailov, and K. Kaneko (eds.) Networks of Interacting Machines: Production Organization in Complex Industrial Systems and Biological Cells*, World Scientific, Singapore, pp. 33-66.

Godunov, S. K., 1959. A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics, *Mat. Sb.* 47, pp. 271-290.

Göttlich, S., Herty, M., Klar, A., 2005. Network models for supply chains, *Communication in Mathematical Sciences*, 3 (4), pp. 545-559.

Göttlich, S., Herty, M., Klar, A., 2006. Modelling and optimization of Supply Chains on Complex Networks, *Communication in Mathematical Sciences*, 4 (2), pp. 315-330.

AUTHORS BIOGRAPHY

CARMINE DE NICOLA was born in Salerno, Italy, in 1972. He graduated in Electronic Engineering in 2002 with a thesis on simulations of processor IAPX 86. He obtained a PhD in Mathematics at the University of Salerno in 2011 discussing a thesis about supply networks modelling and optimization techniques. He is actually a research assistant at the University of Salerno. His scientific interests are about fluid – dynamic models for the analysis of traffic flows on networks, operational research models in airport management, and queueing theory.

His e-mail address is denicola@diima.unisa.it.

ROSANNA MANZO was born in Polla, Salerno, Italy. She graduated cum laude in Mathematics in 1996 and obtained PhD in Information Engineering in 2007 at the University of Salerno. She is a researcher in Mathematical Analysis at the Department of Electronic and Information Engineering University of Salerno. Her research areas include fluid – dynamic models for traffic flows on road, telecommunication and supply networks, optimal control, queueing theory, self – similar processes, computer aided learning. She is author of about 40 papers appeared on international journals and many publications on proceedings. Her e-mail address is rmanzo@unisa.it.

LUIGI RARITÁ was born in Salerno, Italy, in 1981. He graduated cum laude in Electronic Engineering in 2004, with a thesis on mathematical models for telecommunication networks, in particular tandem queueing networks with negative customers and blocking. He obtained PhD in Information Engineering in 2008 at the University of Salerno discussing a thesis about control problems for flows on networks. He is actually a research assistant at the University of Salerno. His scientific interests are about numerical schemes and optimization techniques for fluid – dynamic models, and queueing theory. His e-mail address is lrarita@unisa.it.

SUGAR FACTORY BENCHMARK

Rogelio Mazaeda ^(a), Alexander Rodríguez ^(b), Alejandro Merino ^(c), César de Prada ^(d), Luis F. Acebes ^(e)

^{(a)(b)(c)(d)(e)} Systems and Automation Department, School of Industrial Engineering, University of Valladolid
Center of Sugar Technology (CTA)

^(a)rogelio@cta.uva.es, ^(b)Alexander.rodriguez@autom.uva.es, ^(c)alejandro@cta.uva.es, ^(d)prada@autom.uva.es
^(e)felipe@autom.uva.es

ABSTRACT

This paper describes a simulated benchmark specifically designed for trying out and comparing different decentralized control strategies applicable to large scale complex process plants. The benchmark represents a reduced version of a typical beet sugar factory and basically corresponds to the interrelation of the Evaporation and the Sugar End sections. The underlying dynamic model is full of realistic details and has been derived from first principles, stating all the involved mass, energy and population balances. The ready to be used executable is offered to the interested parties as a standard and documented package accessible from any system implementing the widely used OPC process communication protocol.

Keywords: Sugar production, hybrid models, industrial process, hierarchical control, distributed control

1. INTRODUCTION

The typical beet sugar factory consists of several sections in series, which are respectively concerned with the extraction of the sucrose out of the sliced beets by a diffusion in water process, the elimination of as many impurities as possible in a purification plant, the concentration of the resulting still impure sucrose solution in a cascade of evaporators and finally, the Sugar End or House, where the crystallization of the dissolved sucrose in batch and continuous crystallizers is carried out to deliver the white sugar grains with commercial value.

The preceding cursory description is deceptively simple. Each of the above mentioned sections is a full fledged plant on its own right, hosting many process units, some of them of difficult individual operation (Poel, Schiweck and Schwartz, 1998). Additionally, each specific unit participates in a complex layout with numerous mass and energy recycles implying a tightly integrated environment that makes the overall management of the factory an arduous task.

The Evaporation section (fig. 2.a), for example, is made up of several serially connected units in an arrangement that seeks to improve the efficiency of the factory by reusing the steam served by the boiler. The primary objective here is the elimination of the extra water contained in the incoming fresh juice, but the

obtained steam, with important energy content, is reused in the same section and in other departments of the plant.

On the other hand, the downstream Sugar End, is a very complex installation organized in several stages (in figure 2.b the main "A" stage is depicted), with many individual batch and continuous pans performing the highly uncertain and poorly measured process of sugar crystallization.

The Sugar House and the Evaporation sections (see figure 1) interacts strongly and not only due to the interchange of the stream of concentrated syrup to process. In addition, the crystallizer pans are heavy consumers of the steam which is served by the evaporators cascade.

The overall management and control of these two specific departments constitutes, then, an important challenge. The difficulty of the task is very much compounded by the need to coordinate continuous and semi-batch type units.

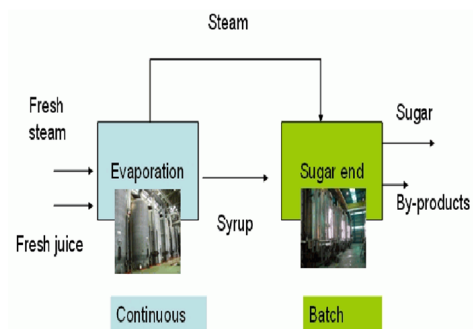


Figure 1: Top level view of the benchmark

The efficient and smooth conduction of the Sugar Factory, and this is true for most process plants, is a complex undertaking. The required solution goes well beyond the regulation of some variables to fixed setpoints at the individual unit level.

Nowadays, in most factories, the needed coordination is heuristically tackled by the operators and technical personnel. Automatic solutions are, of course, possible. In the specific case discussed, where the scheduling of the different batch pans is critical, the control scheme described in Prada, Sarabia, Cristea and Mazaeda (2008) does the job. They use a simplified

model of the section in a coordinating central MPC controller implementing a novel control signal continuous re-parametrization of an otherwise discrete or hybrid problem, to find the optimal solution with a reasonable utilization of computer resources.

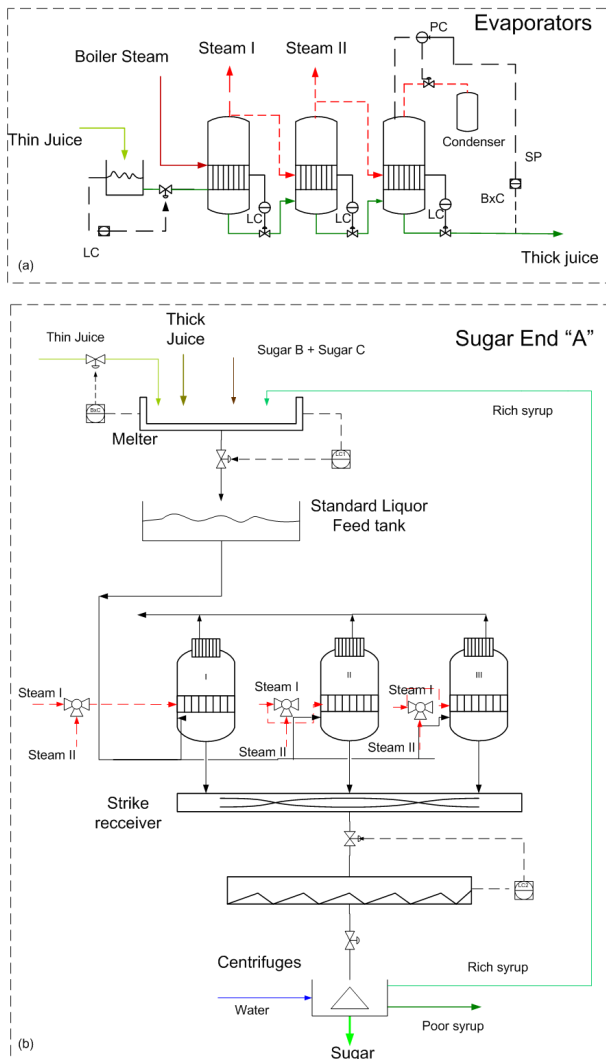


Figure 2: Benchmark detailed view. a) Evaporators cascade. b) Sugar End A stage

In any case, there is an emergent notion in the control community stating that centralized optimizing schemes, in spite of their unquestionable mathematical properties concerning, for example, the rigorous optimality of the solution, should not be the only way to go, especially in the case of large plants. It is argued, for example, that the burden of maintaining a central good enough model is too exacting; and that the associated numerical problem scales badly with the size of the plant.

On the other hand, the structure of the typical process plant, where the main product stream travels serially from installation to installation, with occasional recycle loops, seems to suggest that a decentralized but coordinated or hierarchical control architecture would be a reasonable, maybe suboptimal, way to go. The intuition being that the disturbances appearing, for

example, at the entrance of a plant of several hours processing time, would not imply too much of a difference for the control actions to be currently adopted at the output. It is then possible for the local control to work, and this would mean important advantages like the possibility of a scalable distribution of the control computer job. Additionally, in the long run, it would result in a conceptually simpler problem. A recent review of the subject is to be found in Scattolini (2009).

In this context, a dynamic realistic first principle model of a reduced scaled down version of the Evaporator and Sugar End is going to be offered as a benchmark for testing and comparing different plant-wide control strategies. The benchmark to be described has officially been adopted by the European Network of Excellence entitled Highly-complex and networked control systems, **HYCON2**, (Framework Program 7 Network of Excellence HYCON2, 2011).

In section 2 a description of the process represented in the benchmark is given, section 3 briefly explains the main assumptions and characteristics of the underlying model and gives references to more detailed descriptions, while section 4, for the sake of conclusions, discusses some of the control and plant-wide coordination challenges posed by the proposed simulated plant.

2. BENCHMARK DESCRIPTION

The benchmark proposed consists of the interrelation between the evaporation section and the first stage of a Sugar House. The scaled down version incorporates detailed first principle models of three evaporators in series and the same number of parallel batch crystallizers, a very simplified, static representation of the A stage centrifuges and along with the necessary auxiliary equipment such as buffer tanks.

The first evaporator receives the stream of technical sucrose solution, the thin juice, with a certain flow rate, purity and concentration to deliver the thick juice at a much greater concentration or Brix. In the real situation the purity and Brix would depend on the beet quality and the workings of the previous sections, but here are considered as given. The concentration of the thick juice is enforced by the PID loop controlling the vacuum pressure in the chamber of the last evaporation unit.

The energy for heating the juice in the first evaporator comes from the live steam delivered by the factory boiler but the second and third units re-uses the vapors obtained from the evaporation of part of the water conforming the juice which is processed in the previous one. This reutilization scheme increases the overall factory's efficiency and is possible since each successive unit is operated at a lower pressure, and so temperature, than the upstream unit.

Steam reuse is not limited to the evaporation cascade. An important fraction of the water evaporated from the syrup is diverted to provide heating energy at other sites in the factory. The subsequent Sugar House

department is a particularly heavy steam consumer and the steam demand it represents on the evaporation section is very severe due to the large amounts required and to its intermittent character.

The vapours obtained from the last evaporator are sunk in a barometric condenser. It is common knowledge of the sugar industry that the flowrate of this stream should be minimized since it represents wasted energy.

The difference between the boiler steam pressure at the input and the one that is enforced in the last evaporator chamber by controlling the flowrate to the condenser, drives the downstream vapour flow.

A real factory Sugar House has an architecture consisting in various, usually three, stages: the first or A stage is dedicated to the production of the commercial white sugar crystals and the rest to the exhaustion of the remaining syrup.

The benchmark only represents the first stage. It consists of the parallel array of semi-batch crystallizers followed by a similar disposition of batch filtering centrifuges. The crystallizers receive the so called standard liquor and deliver the massecuite: a viscous slurry consisting of the grown sugar crystal population which is suspended in the resulting syrup or mother liquor.

At the end of their respective strike, each pan discharges the massecuite into a common strike-receiver tank. Next the downstream centrifuges perform the required step of separating the sugar grains from the mother liquor. The purity quality requirements of the white sugar product determine the need of applying water during the centrifuging process for improved filtering. Water helps in expelling the traces of mother liquor from the crystal faces but re-dissolves part of the sugar crystal mass. As a result of this technological setup, each centrifuge offers two type of syrups classified according to their purity: first the so called poor or green syrup, with purity similar to the original mother liquor and then the rich or wash run-off syrup, of higher purity, enriched with the re-dissolved sucrose. The poor syrup gets processed in the following stages and the rich syrup is directly recycled in the A stage.

The standard liquor is conformed in the melter. The main part is the thick juice coming from the evaporators but it also receives the contribution of the above mentioned A rich syrup and of the recovered sugar crystallized in the B and C stages. In the benchmark the melter is simply modelled as a level controlled open tank which assumes the instantaneous and full dissolution of the crystal streams.

The interplay between the workings of the crystallizers and those of the centrifuges is very important for the efficiency of the factory: a bad strike with a not uniform population of crystals would have a poorer filtering capacity in the centrifuges, so it would demand more water with the associated negative impact of the efficiency. To keep the complexity of the benchmark under check, the mentioned compromise is not modelled. In any case, the simplified model of the

centrifuge, allows, as a perturbation, the introduction of more water, always a prerogative of the human operator, to simulate the mentioned effect. It is to be noted that the efficiency impact associated with this action, should be understood also in the sense that the obtained syrups are more diluted, so it would imply and extra evaporation effort in the crystallizers and more processing time to each cycle or alternatively a greater demand of steam from the evaporation section.

The differences in rhythm between continuous and batch operated equipment determines the existence of buffer tanks of the appropriate size in the flowsheet of the plant. The standard liquor tank, in particular, which serves the feed syrup to the pans, should accommodate the peaks in demands from the crystallizers with the continuous supply of standard liquor from the melter. In the typical plant, the operator schedules the workings of the parallel array of pans to keep the syrup inventory in the container between safe limits. Observe that the long and uncertain processing times of the batch units make the task difficult. A similar situation exist in the strike receiver, whose level is controlled by modifying the working rhythm of the batch centrifuges, and so their throughput, but the problem here is easier due to the short cycle time and predictability of their time controlled cycles.

2.1. Evaporator

The evaporators considered are of the Robert type. Each unit has two chambers. The heating chamber or calandria encloses a set of vertical tubes that contains the boiling juice. The heating steam enters the shell of the calandria and the energy needed for boiling is transferred to the juice inside the tubes. The heating steam condenses on the shell walls and finally leaves as condensate water. The juice to concentrate enters at the bottom of the chamber, is heated and rises in the interior of the tubes driven by the resulting vigorous bubbling effect produced while boiling. The more concentrated syrup goes over the rim of the tubes and falls into the central downtake and to the output of the unit. The steam produced from the water evaporation emerges from the juice phase, and reaches the upper space containing the vapour. There is a pipe at the top which leads the vapour resulting from evaporation to the calandria of the next station or to the condenser. There is a complex steam delivery circuit that distributes the vapour stream to other units, especially to the batch crystallizers in the Sugar House.

The juice level in the central downtake must be regulated to assure that the generated vapours are safely sealed in the upper part of the unit and to guarantee the height difference with the input of the downstream evaporator which is needed to drive the juice flow.

2.2. Batch crystallizers

The process of crystallization is possible when the concentration of the solute to crystallize, sucrose in this case, exceeds the concentration defining the solubility

of the substance at the given temperature. The supersaturation, which is conventionally defined in the sugar industry as the ratio of the two mentioned concentrations, should be greater than unity for crystallization to occur. For moderate values of supersaturation, a metastable zone is defined where the growing of existing crystals is possible but the probability of the creation of new grains out of the solution, is negligible. If the supersaturation, however, is increased so that it trespasses the fuzzy defined labile zone frontier, then the nucleation phenomenon turns explosive, a situation to be avoided in industrial crystallization processes. The solubility of the sucrose increases with temperature and with the presence of impurities. In sugar industry supersaturation can be obtained by cooling or by evaporation of the water in excess. In the present benchmark the batch crystallizers use the latter mechanism.

The batch sugar industrial crystallizer serves the purpose of creating the conditions of supersaturation of the technical solution of sucrose, so that a tiny initial population of sugar crystals may steadily grow until it achieves the commercial average size. It is an important technological requirement that the spread of the distribution of sizes in the population is kept as narrow as possible and this implies that supersaturation should be always kept at moderate values in the so called metastable region. The supersaturation conditions are created by striking the right balance between the rates of water evaporation and of the standard liquor which is supplied to replenish the solution of the sucrose that has migrated to the faces of the crystals.

The evaporation is carried out at low, vacuum pressures so as to keep the temperature of the mass at reduced values so avoiding the quality impairing caramelization of sugar. The process is conducted in a semi-batch fashion, and this means that the prevalent conditions are continuously varying along the cycle. The impurities, for example, get accumulated along the cycle, so the mother liquor purity gets progressively lower and this implies a greater difficulty in keeping the right supersaturation. The amount of massecuite in the pan grows from an initial value of roughly half to the full capacity of the pan at the end; and this fact has an important negative impact on the circulation of the mass and on the the heat transfer efficiency of the unit. The difficulties in the conduction of the pan are compounded by the absence of important on line measurements. The supersaturation, for example, that is critical, depends of other variables like the concentration, the temperature and the purity of the solution. Purity is not measured on-line but it is periodically reported by the factory laboratory. The on-line measurement of the concentration of the solution or Brix is problematic and the temperature of the mass is not homogeneous, especially at the end of the strike because of poor circulation, so its determination is also uncertain.

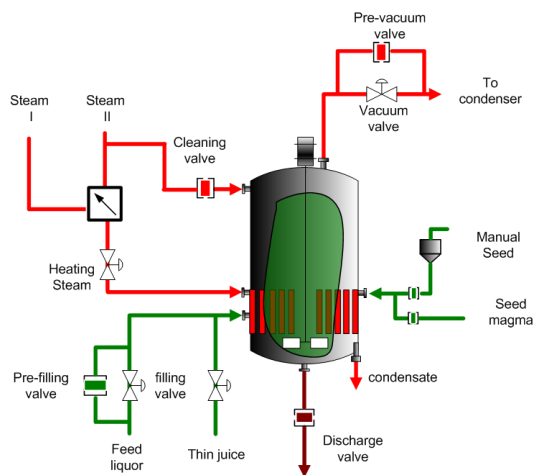


Figure 3: Vacuum pan crystallizer

The vacuum pan is constructed as a cylinder with a diameter and height of comparable dimensions. At the bottom, there is a floating calandria type of heat exchanger. The mass circulates inside the tubes and the central downtake, and the heating steam goes inside the shell.

The calandria is designed so as to bolster the circulation of the mass. The massecuite rises in the tubes as the water component is evaporated. The steam rises through the existing mass and emerges at the surface to enter the vapour occupied phase. The massecuite, which is driven over the tubes rim by the rising bubbles, returns to the bottom via the central downtake. In order to intensify the circulation, a mechanical stirrer is placed at the bottom of the downtake. The stirrer contribution is important mostly at the end of the strike, when the achievable evaporation is lower and the natural agitation provided by the bubbles are probably not enough.

The pan is provided with the necessary valves to regulate the heating steam input to the calandria, the feed syrup to the chamber and the seed magma containing crystal initial population. There are valves that allow the discharge of the mass at the end of the strike, and the control of the evaporated steam out of the chamber to the condenser. There is also a cleaning valve for inputting steam after the product evacuation with the purpose of removing the traces of massecuite.

Note that in the specific unit modelled (fig. 3), it is allowed to individually choose the evaporator effect which is going to be the source of heating steam. Normal practice dictates the use of lower pressure steam (II effect) at the beginning of the cycle when the heating process is more efficient and then, to switch to higher pressure steam (effect I) as the cycle progresses and the mass transfer coefficient rapidly diminishes.

The instrumentation used in the pan allows to measure the following variables: chamber mass temperature and steam space pressure, the level attained by the mass and the steam pressure in the calandria. The unit has a radio frequency (RF) sensor whose on-line readings can be calibrated to somehow represent the

concentration of the slurry: solution plus growing crystals. The electrical current which is drawn by the stirrer motor, could be taken as a indication of the consistency of the mass, and this fact is put to use at the end of the strike, when the RF transmitter measurements are less reliable.

2.2.1. Batch crystallizer program

Each batch crystallizer cycle follows a recipe implemented by sequential program (fig. 4.a) whose main stages are the following:

1. **Loading:** A high capacity valve is fully opened to start the introduction of standard liquor in the chamber. The objective is to load enough syrup so as to completely cover the calandria in such a way as to maximize the circulation process of the mass and improve the heat transfer coefficient. The two PID based loops controlling the pressures in the chamber and in the calandria are both put in automatic mode.
2. **Concentration:** The heating of the mass continues to concentrate the mass with the purpose of reaching the required supersaturation. The heat transfer coefficient and the steam consumption are both, at this moment, very high. The stage ends when the syrup Brix reaches a value, which in view of the most recent standard liquor purity report from the laboratory, would correspond to the right supersaturation.
3. **Seeding:** The seeding stage proceeds by automatically introducing the amount of seed magma which is considered adequate for obtaining the final correct average size.
4. **Growing of the grain:** This is the longest and most important stage of the cycle. The population of crystals in the seed should be made to grow as the sucrose in the solution migrates to their faces along the strike. The supersaturation would tend to decrease as the impurities accumulate in the solution, so standard liquor should be introduced in a controlled way to add the dissolved sucrose needed for compensating this effect. The amount of syrup to introduce would depend on the rate of evaporation, and on the purity currently existing in the pan. In the absence of on-line supersaturation estimation, the stage is conducted by establishing a curve of massecuite Brixes that gives, at each instant in the evolution of the process, the total mass concentration that should be enforced to obtain the right supersaturation. It should be noted, that the readings of the RF sensor takes into account not only the dissolved substances but also the mass of growing crystals. The level attained by the mass in the chamber, which should grow from the value initially loaded to the pan full capacity, is taken as a measure of

the evolution of the strike. So, the feed syrup input flowrate is controlled in the stage by the scheme shown in fig. 4.b, where the setpoint for mass concentration is given by a Brix vs. level curve that should be adjusted by the operator to reflect the changes in standard liquor purity.

The considerable reduction of the heat transfer coefficient in the heat exchanger, taking place as the mass level rises, is compensated, in some degree, by modifying the setpoints of the calandria and of chamber steam pressure regulators.

5. **Tightening Up:** The purpose here is to increase the consistency of the massecuite in preparation for the discharge. There is no further introduction of syrup but the evaporation continues until the electric intensity consumed by the stirrer motor attains a configurable value. The setpoint of the calandria pressure controller is raised to accelerate the process.
6. **Discharge:** The heating steam input is shut down and the vacuum is broken in the chamber by opening the cleaning valve. When the pressure reaches an appropriate high value, the discharge output gates are opened.
7. **Cleaning Up:** Cleaning valve is kept opened with discharge gates closed to get rid of the traces of massecuite which remain contaminating the interior walls.

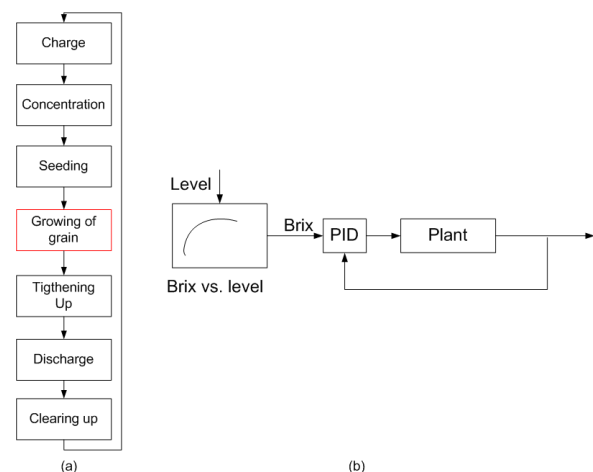


Figure 4: Crystallizer program. a) Main stages. b) Brix controlling loop in growing state

In figures 5 and 6 the evolution along one cycle of the level of the mass and of the mass concentration of crystals are respectively shown, highlighting in each case the instant of activation of some important events.

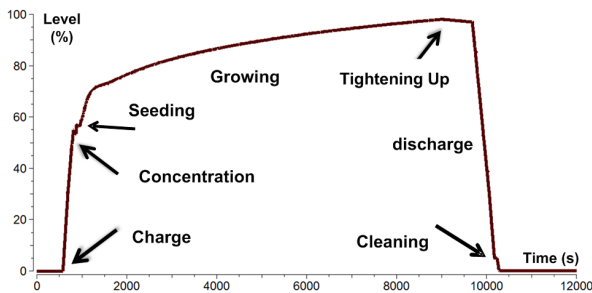


Figure 5: Level evolution

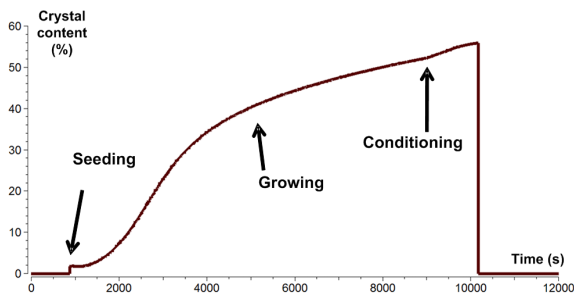


Figure 6: Crystal content evolution

3. BENCHMARK MODEL

The model has been created assembling objects instantiated from a previously existing library of sugar factory components using the Object Oriented (OO) concepts implemented by the **EcosimPro** modelling and simulation tool (ESA International, 2008).

The original purpose of the library was to provide the elements for constructing realistic dynamic simulators specially dedicated to the training of the beet sugar factory control room operators.

The library contains a representation of the main units to be found in the factory. It obviously include classes representing evaporators and batch evaporative crystallizers but it also hosts the auxiliary equipment needed in any process industry such as valves, pumps, tanks and even PID regulators (Merino, Acebes, Mazaeda and Prada, 2009). These ancillary elements are used to create the topology of the specific plant but are also deployed in the definition of the main process units to define its internal structure.

All the models are derived from first principles and are coded in a generic way, exposing numerous parameters, making possible the adaptation of the resulting overall instantiated model to the specific situation at hand.

The models exhibit a hybrid character. They are predominantly made of continuous time dynamical equations but must also respond correctly to discrete events fired, for example, as the program of the crystallizers move from one stage to the next.

The original motivation as a training aid has permeated the modelling effort, guiding the election of the general assumptions of the library. Since all variables and actuators were in principle susceptible of

being accessed by the trainee, the models should meticulously state all the involved mass and energy balances. In the case of the crystallizers the evolution of the mass of grains is tracked by means of the formalism of the population balance equations. The need to be able to simulate whole large factories determined the use of globalized models, wherever possible, to facilitate the numerical integration effort. There is an ample use of non dimensional relations used in the general chemical engineering literature and in sugar studies, and mass and energy transfer rates are put in relation with the characteristics of the processed streams. This provides reasonable starting values for the physically related parameters that can be further tuned to adapt to each specific case. The physico-chemical properties of the main products such as syrup or juice and massecuite had been taken from the existing specialized literature (Bubnik, Kadlec, Urban, and Bruhns, 1995). The characteristics of typical utilities such as liquid water and steam are readily available.

The capacity of the generic model to faithfully reproduce real plant data and to meet the informed qualitative demands of sugar experts had been described elsewhere (Mazaeda, 2010; Merino, 2008; Mazaeda, Prada, Merino and Acebes, 2012). The assembled model here proposed does not exactly emulate any existing plant; but the deployed individual units are the ones that have been calibrated and validated with real data.

The Sugar Benchmark has a very different purpose from the one that originally motivated the design of the underlying OO library. In any case, the model proposed, with its abundance of realistic details, full of special situations, which are needed for training, but that that would be considered as non-essential in almost any other type of application, would stand with respect to the designer of the overall control strategies, in a situation approximately similar to the one he/she would encounter when facing a real world problem.

Detailed descriptions of the evaporator and the vacuum pan crystallizer models can be found in Merino, Alves and Acebes, (2005) and Mazaeda and Prada (2007) respectively. The centrifuges model has been specifically created for the benchmark. It is less involved than the previous models and has a static character. It simply consists of the necessary balances to each component, taking into consideration the dissolution provoked by the amount of water introduced. The exact composition of the poor and rich syrup streams and the humidity of the separated sugar grain product are decided by adjustable parameters.

4. THE BENCHMARK CHALLENGE

The benchmark main purpose is to serve as a testing platform to develop high level strategies with the capacity of guaranteeing an optimal economical behaviour simultaneously dealing with the management of the continuous processes and the scheduling of the batch units. The solutions proposed should be able to cope with the perturbations and uncertainties

represented by the variability of the working conditions and the lack of a complete knowledge of the state of the process.

More specifically, the problem of conducting the whole plant in a smooth way is complicated due to the uncertainty in the batch crystallizers processing times which are very dependent on the characteristics of standard liquor. As an example, in figure 7, the effect that a modification of the feed purity, keeping the Brix vs. level curve fixed, has on the cycle time and on the evolution of other important variables, is shown.

A smoothly managed plant should schedule the workings of the batch pans in such a way as to guarantee the consumption of all the syrup delivered by the upstream continuous section. Figure 8 shows the evolution of the levels in the feed syrup tank and in the strike receiver in this ideal situation.

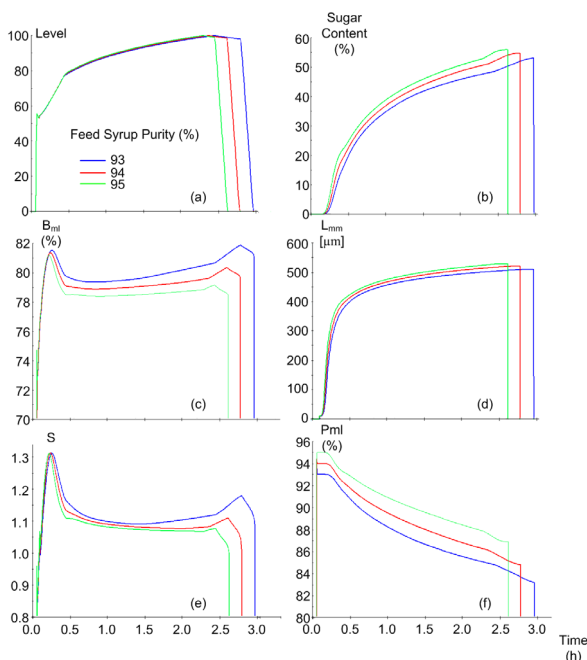


Figure 7: Effect of feed syrup purity on the performance of a crystallizer. a) level. b) Sugar content. c) Brix of mother liquor. d) Average crystal size. e) Supersaturation. f) Purity of mother liquor

In figure 9, a decrease of the flow rate of thick syrup or an increase of purity could lead to the depicted situation, where the level in the feed syrup tank is clearly diminishing with the risk of violating the safety restriction on the required inventories. It should be noted that the availability of steam and the crystallizer cycle time are both very dependent on the Brix enforced at the output of the evaporation section. On the other hand, the amount of water to the centrifuges alters the concentration of feed syrup and also its purity.

More formally, the objectives to be achieved in conducting the plant simulated in the benchmark are the following:

- Minimize the consumption of boil steam while serving the demands of the crystallizers.

- Guarantee the processing of all the incoming syrup. The inventories of buffer units, strike receiver and, fundamentally, the standard liquor tank should be kept between safe limits.
- The quality standards of the produced white sugar should be met.

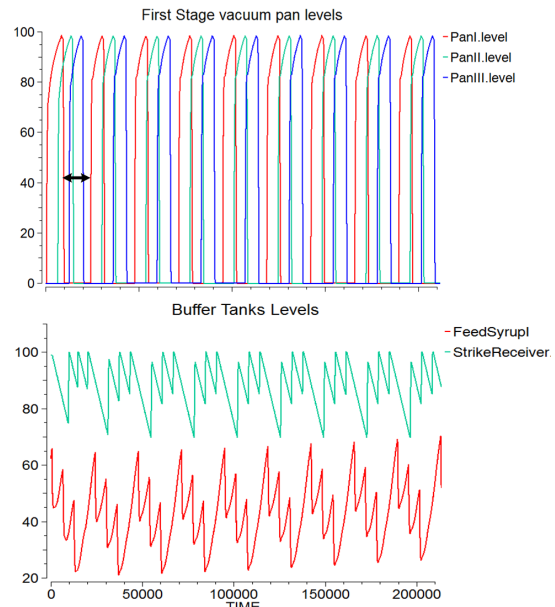


Figure 8: Level in buffer tanks over several cycles

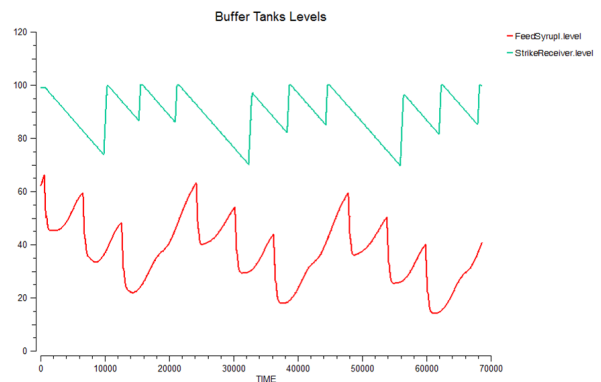


Figure 9: Level in buffer tanks with problems

Table 1: Allowed range for perturbation variables

Parameter	Description	Range
W_{evap}	Mass flow rate into evaporation	10-16 kg/s
B_{mm_evap}	Brix of juice into evaporation	40-45 %
P_{mm_evap}	Purity of juice into evaporation	91-95 %
$W2mc$	Mass flows rate of water to massecuite in centrifuges	0.02-0.025

Table 2: Degrees of freedom for plant-wide control

Variable	Description
$SP_{\text{evap_Brix}}$	Setpoint of Brix control loop at Evaporation output
P_{boiler}	Pressure from boiler to Evaporation effect I
$Load_{\text{vp}}[k]$	$k= 1-3$. Load command for each vacuum pan
$Valv_{\text{cent}}$	Valve controlling centrifuge throughput

The objectives must be attained in the presence of perturbations, whose range of variation is shown in table 1.

The minimum set of variables which are at the disposal of the design solution for the management of the plant are described in table 2.

The variables that can be read from the benchmark are the same ones which are typically sensed in the real factory, namely:

- The levels of all evaporators, vacuum pans and tanks.
- The temperatures of all evaporators, vacuum pans and tanks.
- The Brixes of all syrups involved.
- The evolution of the Brix of the massecuite in each vacuum pan.
- The purity of the syrups involved.
- The values of all the pressures involved: in the chambers of the evaporators and vacuum pans and in the heat exchangers.

It is possible to consider the problem proposed at several levels. In the simpler approach, the crystallizers could be considered to be reasonably well controlled and the plant-wide coordinator should be simply concerned with guaranteeing the availability of steam and syrup as demanded. The third objective of keeping up with the quality requirements of the sugar product would be considered as automatically enforced by the pan's program if the steam demand is served.

But it is also possible a more involved strategy, which deals directly with the control of each crystallizer. This would imply the handling of the values of the Brix vs. level curve, of the setpoints for the pressure of the calandria and of the chamber, among other details. Of course, this other approach would surely be able to achieve a more efficient solution, but would also imply a greater responsibility concerning the quality of the end product.

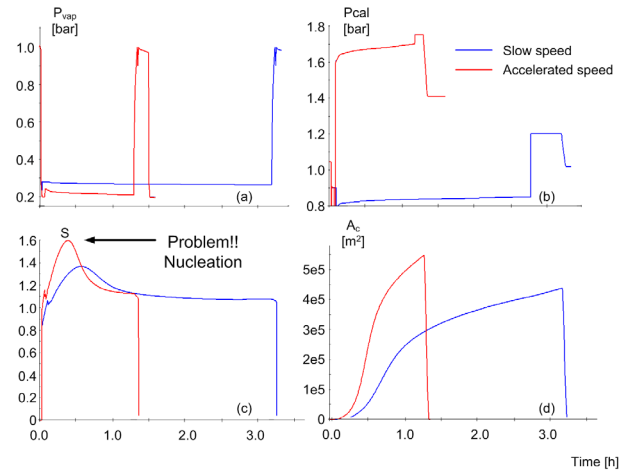


Figure 9: Cycle time acceleration by increasing evaporation rate. a) Vacuum pressure setpoint. b) Calandria pressure setpoint. c) Supersaturation. d) Aggregated area of crystal population

An example of the complex issues involved in applying the second, lower level strategy can be discussed analyzing the figure 9. The setpoints of the controllers of heating pressure to the calandria and of vacuum in the chamber influences the duration of the strike. So it would be legitimate to consider the use of these references as additional degrees of freedom for achieving the plant-wide objectives. It should be bore in mind, however, that an immoderate use of this kind of acceleration is somehow artificial and is limited by the purity of the feed liquor and so should be performed with caution. An unreasonable increase in the calandria pressure would imply an excessive increase of the supersaturation. The reason being, that as the crystallization kinetics is basically unaffected, the existing aggregated crystal area of the sugar population had not reached the value that would be able to sustain a flowrate of crystallization capable of compensating the new rhythm of water evaporation. As a consequence, supersaturation gets inside de labile region with the consequent prejudice to the quality of the product: a wider and reduced average size population. It goes without saying that the new setpoint pressure to fix would also depend of the availability of steam from the evaporation section.

The benchmark gives a suitable platform for testing many other types of interesting applications like, for example, the hybrid identification of the discrete-continuous units, or the explicit handling of the uncertainty by embedding the numerical optimization procedure in a stochastic framework.

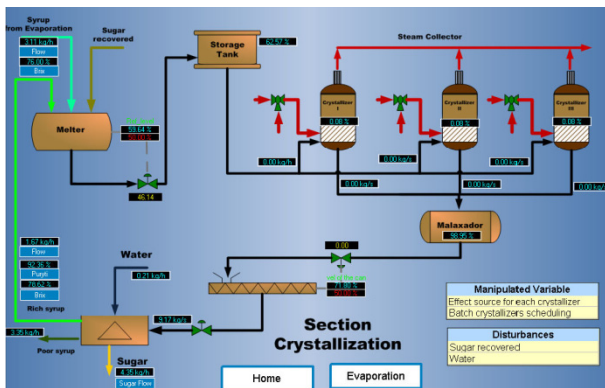


Figure 10: HMI interface to the sugar benchmark

Finally, it should be said that the simulated plant is going to be made available as an executable (Alves, Normey-Rico, Merino, Acebes and Prada, 2005) which implements the OPC protocol (Iwanitz and Lange, 2002; Zamarreño, 2010) and with a graphic interface (Alves, Normey-Rico, Merino, Acebes and Prada, 2006) making possible its standalone operation. In figure 10 a screenshot of the user interface is shown. The use of OPC will additionally facilitate the access to the simulated data from any of the many clients currently supporting that widely adopted standard.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007-2013] under grant agreement n°257462 HYCON2 Network of excellence.

REFERENCES

Alves, R., Normey-Rico, J., Merino, A., Acebes, L.F., Prada, C.d., 2006. Edusca (educational scada): Features and applications. In: *7th IFAC Symposium in Advances in Control Education*.

Alves, R.A., Normey-Rico, J., Merino, A., Acebes, L.F., Prada, C.d., 2005. *OPC based distributed real time simulation of complex continuous processes*. *Simulation Modelling Practice and Theory*, 13, 525-549.

Bubnik, Z., Kadlec, P., Urban, D., Bruhns, M., 1995. *Sugar Technologist Manual. Chemical and Physical Data Manufacturers and Users*. Dr. Albert Bartens.

ESA International, 2008. *EcosimPro User Manual, EL Modelling Guide*. EA International and ESA.

Framework Program 7 Network of Excellence HYCON2, 2011. *Highly-complex and networked control systems*. www.hycon2.eu.

Iwanitz, F., Lange, J., 2002. *OPC: fundamentals, implementation and application*. Hüthig.

Mazaeda, R., 2010. *Librería de modelos del Cuarto de azúcar de la industria azucarera para entrenamiento de operarios*. Ph.D. Thesis, University of Valladolid.

Mazaeda, R., Prada, C. d., 2007. *Dynamic simulation of a sucrose batch evaporative crystallizer for*

operators training. In: *19th European Modeling and Simulation Symposium (Simulation in Industry)*, EMSS 2007.

Mazaeda, R., Prada, C. d., Merino, A., Acebes, L.F., 2011. *Librería de modelos orientada a objetos para la simulación del Cuarto de azúcar: Cristalizador continuo por evaporación al vacío*. *Revista Iberoamericana de Control Automático*, 8, 100-111.

Merino, A., 2008. *Librería de modelos del cuarto de remolacha de la industria azucarera para entrenamiento de operarios*. Ph.D. Thesis, University of Valladolid.

Merino, A., Acebes, L.F., Mazaeda, R., Prada, C.d., 2009. *Modelado y simulación del proceso de producción de azúcar*. *Revista Iberoamericana de Control Automático*, 6, 54-60.

Merino, A., Alves, R., Acebes, L.F., 2005. *A training simulator for the evaporation section of a beet sugar production process*. In: *European Simulation Multiconference. ESM05*.

Poel, P.V.d., Schiweck, H., Schwartz, T., 1998. *Sugar Technology. Beet and Cane Sugar Manufacture*. Dr. Albert Bartens.

Prada, C. d., Sarabia, D., Cristea, S., Mazaeda, R., 2008. *Plant-wide control of a hybrid process*. *International Journal of Adaptive Control and Signal Processing*, 22, 124-141.

Scattolini, R., 2009. *Architectures of distributed and hierarchical Model Predictive Control-a review*. *Journal of process control*, 19, 723-731.

Zamarreño, J., 2010. *Acceso a datos mediante OPC*. Editorial Andavira.

CENTER TITLE HE DESIGNING AND IMPLEMENTING A MODEL TO EXAMINE R&D SECTION'S CAPABILITIES WITH EMPHASIS ON REVERSED ENGINEERING IN CHEMICAL FACTORY

Neda Khadem Geraili^(a), Mona Benhari^(b)

^(a) Department of Technology management, Faculty of Management & Economics, I.A.U
Science & Research Branch, Tehran, Iran

^(b) Computer Engineering Graduated student

^(a) neda_geraili@yahoo.com, ^(b) mona_benh@yahoo.com.

ABSTRACT

Today, research and development and related activities to access new technologies in the industrial world have been challenging activities. R&D units were known as Technology pillars and are unique resources of innovation, creating R&D units and institutes or developing old ones to new effective ones for developing countries are inevitable issues.

There are various methods to access the Technologies and one of the most important ones, especially among developing countries is Reverse engineering which is Consciousness method taken from exist Technology.

This paper also reviews research background of (R & D) and reverse engineering, implements simulation software modeling to achieve the most effective parameters and their relationships and offer strategies for optimal per unit R & D and access to technologies of modern examined. This study contains R & D department of a chemical factory as selected population which is one of the important industrial factories in Iran. In conclusion some key effective factors have been extracted through this simulation. The company can consider them to develop R&D unit and improve the product quality.

Keywords: Research and Development (R&D), Reverse engineering, Technology Transfer, System dynamic.

1. INTRODUCTION

Research, development and management are the pillars of creation, development and utilization of technology and they are the needs of economic and social development in every country, in the developing countries which are seeking industrial and economic self-reliance, policy and planning research programs and development and management of practices impact on these activities are considered as priorities for national, industrial and manufacturing activities. Expansion and enhance of R & D activities, particularly "industrial research", requires understanding the factors affecting research and development process and designing policies and effective activities of such

mechanisms. In competitive conditions of the current industrial world, doing research and development activities is one of the most effective things that managers of economic enterprises can take on and in industrialized countries of the world, the costs allocated to these activities are being increased day by day, on the other hand referring to the research and development without creating the necessary infrastructure units such as: organizational structure, human resources characteristics and management of these organizations causes the failure of research and development process of economic institutions. (Kheradmard 2007)

Process of R & D includes: basic research, applied research and development. By using basic research, the scientific findings are presented in the form of hypothesis, theories and general rules and in the next stage, it would be the applied research which determines the possible applications for basic research findings. The stage of development, scientific knowledge obtained from basic research and applied research is applied in order to provide new and developed products or processes and innovation can occur. The main centers of the R & D process consists of industrial research institutes, academic research institutions and government R & D centers. Reverse engineering is extraction and development of technical information from available products. This method unlike the direct designing process meant to production according to customer requirements and the initial idea is based on the engineering analysis. (Amiri and others 2008)

Many activities in this area are done with different goals daily. Extraction of technical knowledge designed by R & D units in different industries, providing technical documentation of industrial equipment and probably copying the products are some examples in this field. Developing countries to access complicated technologies require a method that fills the technology gap between these countries and the developed countries at the right time and among different methods of having access to technology, reverse engineering is the most appropriate method.

This research was done in the context of systemic thought and to achieve research objectives and answer of research questions, System Dynamics modeling has been used. Modeling dynamic systems is an essential tool in systemic thought that in order to better understand the behavior of the system; imaging and strengthening the conceptive models and displaying the behavior of system are used. Since diversity of effective factors is very significant on issues, in order to obtain acceptable solutions, the least important factors are omitted by using modeling in the study. The general category type of research will also be an applied research.

This study utilizes a model (simulation), which includes relations with variables methods in reverse engineering in R & D units which will be done. This model can be considered as a new process in popular and applied industries. And the places where there is no new scientific and systematic approach in these projects, so this research and simulations can greatly help managers in making decisions and having reasonable controls. Thus, at first there have been gathered interviews and questionnaires with key members collaboration and make a systematic relation between variables and suitable parameters base on model and achieved the desired data will be collected and model validation is done and CLD related to basic parameters can be drawn. Then modeling using Vensim simulation software based on raw data and descriptive statistics are done and the parameter sensitivity analyses also take place.

Finally, based on the model output and parameters sensitivity analysis software, necessary suggestions and performed conclusions are provided.

2. MAJOR RESEARCH AND DEVELOPMENT ACTIVITIES IN THE STAGES OF REVERSE ENGINEERING PROCESS

The experiences of advanced countries suggests that technological progress in these countries owe R & D activities and technological infrastructure more than anything and without implementing such activities there is no way to achieve the desired technology for that, though by buying the product or purchase process, the required parts can be obtained, but even successful technology business by buying technology (research and development activities necessary for building or research strategies to produce copy that is necessary) also requires research activities and industrial engineering services. For example, before attempting to determine how to access to technology, knowledge and action for defining technology needed to meet demands and needs of development projects such countries need some activities which won't be fulfilled without collaboration of involved ones engineering services directly.(Colin Bradley 1998), (Jokar 2008), (Book of technology comercialization)

2.1. Subheadings

Initial Caps, bold, flush left. Use Times New Roman Font and 10 points in size. Start the text on the next line. Please use the "Heading 2" style.

2.1.1. Secondary Subheadings

Initial Caps, bold, indented of 0.7 cm. Use Times New Roman Font and 10 points in size. Start the text on the next time. Please use the "Heading 3" style.

3. SELECTING APPROPRIATE TECHNOLOGY AND REASERCH AND DEVELOPMENT REQUIRED ACTIVITIES

The process of selecting technology and appropriate product includes all actions and activities which are contiguous with the objectives, conditions and specifications and technological needs to determine the most appropriate technology requirements and also the most appropriate strategy to achieve the goals is done by considering the circumstances and technical, economic and legal relations. (Akhbari and others 2008), (Allahyari 2009), (Ebrahimi), (Fort collins and kaufman 1989)

This process typically includes steps such follows:

- 1-Information on market needs consumer preferences and market traction for new products.
- 2-Information about status of competition, situation of required innovation, research and development activities, products and environmental threats.
- 3-Having Information about the required global situation of technology and explanations of their pattern in some period of times by using technology forecasting techniques.
- 4-Identification of technical and scientific facilities of country and possibility of access to materials, energy, and production processes.
- 5- Analysis of investment company status among others in terms of scientific, technical, economic support and technical ability, communication and marketing.
- 6-Strategic planning to determine and select the required technology, according to data collected.
- 7-Designing strategies to achieve product and selecting the most appropriate strategy (according to the results of technical and economic feasibility studies later we will pay attention to this issue.)

Marketing efforts, policy and strategic planning and selecting the required technology to achieve the set of activities that make up except those involved with research and development activities and research and development, so reaching them is not possible except with effective management of information and technical experiences of engineers. The sensitivity of this action is so much in a way if investments policy and managements not to be done accurately and according to practical techniques and if a comprehensive analysis not to be done, the rest innovation activities will be overwhelmed and they eventually might lead to investment plan failure.

4. SELECTING APPROPRIATE TECHNOLOGY AND REASERCH AND DEVELOPMENT REQUIRED ACTIVITIES

(Ghani 2009)], (Houshangnia 2009), (Laghvi 2010),(Mardi 2011)

- 1 - Controlling the possible technical and economic studies, making plans to copy a subset of product or product requirements inside the factory.
- 2 - Researching for understanding mechanisms of functioning components for realization of working mission for products and discovering the relationships between components during operation.
- 3 - Departing the components that we have decided to copy them and registering charts for the products and its details during operation.
- 4 - Providing maps and photos, or piece or pieces that they're going to be copied.
- 5 - Identifying, testing and determining the raw materials used to manufacture any component of product.
- 6 - Identifying, testing and diagnosis process for production of any component of product.
- 7 - Identifying, testing and diagnosis of complementary operation performed to achieve mechanical, physical, metallurgical, chemical properties... which are required.
- 8 - Detecting required machineries and tools for manufacturing and assembly.
- 9 - Compiling technical knowledge for production and assembly and its components.
- 10 - Preparing production drawings, maps, templates, models and tools needed for production and control stages.
- 11 - Monitoring the prototyping operation specifically or under direct supervision.
- 12 - Controlling samples made and matching them with the profile of desired standards.
- 13 - Assessing and analyzing test results and in case of necessary, revision of sizes and tolerances.
- 14 - Designing production and estimating required machinery and equipment.
- 15 - Planning of factory and designing factory production and assembly line.

5. METHODS

This paper aims to achieve a model to evaluate the capabilities and competencies of R & D of a chemical factory based on modeling system. This model is provided to managers and decision makers as a tool and it plays the role of easing in decision process and opting organizational strategies. So according to this view, this article is considered in the category of applied research. On the other hand the type of approach used in the paper is in a way that by dynamic modeling, the effective components on planning present a new model and it solves the problems that current planning methods are facing with. In this study it is tried that modeling (simulation) and all factors and important variables and influencing patterns, trends empowerment of R & D unit for the chemical industry to be

considered with an approach of reverse engineering in order to strengthen R & D unit.(Fort collins and kaufman 1989),(Roussel and saad 1991),(Chiesa and masella 1996)

First stage: gathering effective factors on research and determining the primary assumptions

Second stage: determining the most important factors affecting research and inserting them in the initial model and draw CLD to determine relationships between factors and variables for the basic model.

Third stage: extending the primary model and finding datum and formulas for model and model validation

Fourth stage: testing Software Model (simulated).

(Sterman 2000),(www.system dynamics.org/conferences 1998)

5.1. . *Introducing the basic components in capability of research and development (R&D) of the chemical Factory:*

(Fadaei 2010), (Fort collins and kaufman 1989),(Chiesa and masella 1996)

This section is based on offering presumptions that simply realize modeling process and it is attempted to present the model which can be implemented in Vensim software.

Assumptions made are extracted based on the information of a Chemical factory and interviews with senior managers and library studies and the most effective variables and essential factors are mentioned below:

- 1 - Professional and skilled personnel (manpower)
- 2 - Identification of risk factors
- 3 - Market and customers needs
- 4 - If necessary interaction with industries and marketing and production units
- 5 - System inputs include raw and new materials (new requirements)
- 6 - Technical Knowledge
- 7 - Sufficient capital (cash flow)
- 8 - Cost market
- 9 - Planning
- 10 - Economic blockade
- 11 - Project Manager
- 12 - Skillful and risky management of R & D

5.2. **Model validation test**

In order to demonstrate the model validity, we use real tests K Square test to explain credibility. Since based on calculations done on questionnaires and interviews, the most effective factors and variables in the model were determined, and information relevant to the years 2001 to 2011 are available.

So by putting this information in the model, the information of next year is predicted and they are measured with realities so the credit degree of model could be provided.

It should be mentioned that information and figures of influential factors, including system inputs (raw materials, equipment, technical knowledge), human resources, including (motivation, working

group, spontaneous, education, good environment, planning and project control, management empowerment and the number of projects done in R & D through the questionnaire were determined. Also system inputs, planning and control of projects, skillful management of Research and Development and human resources were considered as independent variables and the numbers of projects R & D were considered as the dependent variables.

The aim is that according to the four-term factors: 1) system inputs, 2) Planning and Project control, 3) strong management, 4) Human Resources with a number of projects to reach an acceptable quality. (Kheradmand 2007), (Rabelo 2004), (Sterman 2000)

Table1: Model validation test by using of real data test & square of KAY :

Hum an Resources	Strong management of R&D unit	Plan ning & Proj ect Control	Syst em Input	Num ber of ende d proj ect in R& D unit	Year
6.45	5.6	7.75	6	2	2001
6.75	5.92	7	6.16	5	2002
6.8	6.08	7.75	6.4	7	2003
6.55	6	8	6.75	2	2004
6.45	5.92	7.5	6.66	1	2005
6.4	5.33	7.25	6.5	1	2006
6.2	5.41	7.25	6.33	3	2007
5.9	5.5	6.75	6.41	0	2008
5.6	5.33	6.75	6.08	0	2009
4.95	5.1	6.5	5.91	7	2010
7.6	7	8	7.6	5	2011
7.6	7	8	7.6	5)fit (Model EI-
2.949	2.872	0.983	2.237	$X^2_{(n-1)}=X^2_{9,0/05}$	X^2

According to figures contained in the questionnaire, the average number of variables and factors extracted from 2001-2011, they have been mentioned in above table and then for credibility and accuracy of model operation, the K Square test was measured by the following formula:

$$X^2_{(n-1)} = \sum \frac{O_i - E_i}{E_i}$$

And the figures of X^2 was extracted according to the above table; X^2 which was provided according to K Square test is $X^2_{0.05,9} = 16.92$ and since X^2 is four variables of model and is less than 16.92, so the modeling assumption is accurate and according to the figures and information listed above, the model is accepted with 0.95 confidence.

5.3. C.L.D draw

After extracting the effective parameters on model, in the second step of modeling, the causal relations between the parameters must be considered, so for this purpose, we extract C.L.D. It is noteworthy that due to the volume of datum and the high number of parameters affecting the issue, single-step extraction of C.L.D is not feasible. therefore for the purpose of understanding it easier, we divided the casual model of the issue into five stages and in each stage a loop of causal model has been drawn; and then finally according to the defined relations, we link each stage to the other stages and each component to the other components to achieve the ultimate causal model of the problem. Finally, the effective parameters of the model simulation, sensitivity analysis and the related results are also presented.

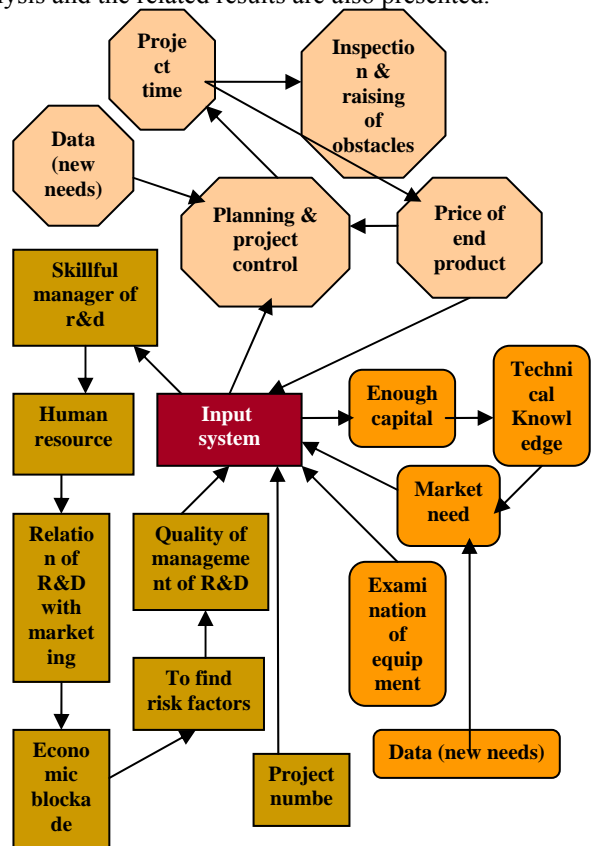


Chart 1: CLD(casual loop diagram),(Model)

(Amiri, farzad and others 2008), (Colin Bradley 1998), (Mardi 2011)

5.4. Draw of chart of stock &flow by vensim software:

(Craig W.kirkwood 1998), (Rabelo and Helal 2004)

Formula tray relation of Regression for Number of ended project in R&D unit:

$$\text{Number} = -0.61 \text{ Input} + 0.07 \text{ Planning control} + 0.72$$

$$\text{Management of R\&D} - 0.34 \text{ Human Resources.}$$

Multiple Graphs relevant to the main parameters of the model

(Kheradmand 2007), (Rabelo 2004)

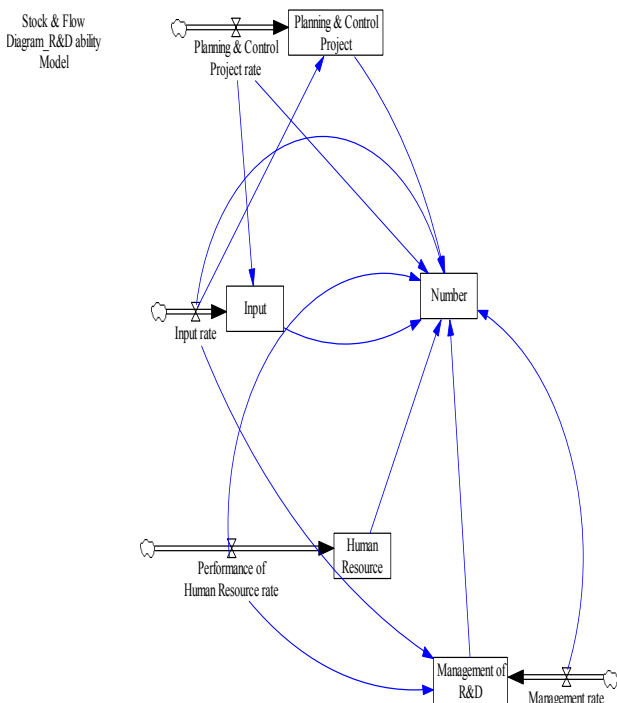
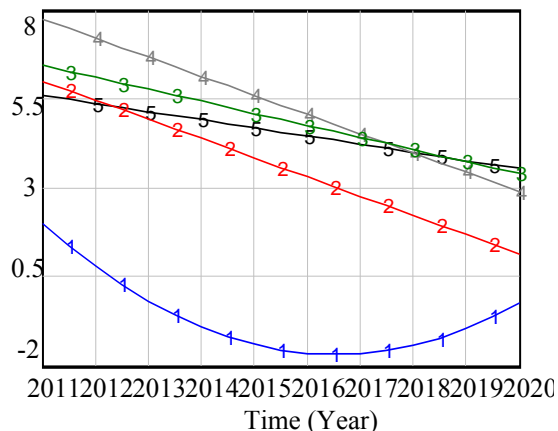


Chart 2: stack and flow

5.5. Multiple Graphs relevant to the main parameters of the model

and Input and Human Resource and Plan

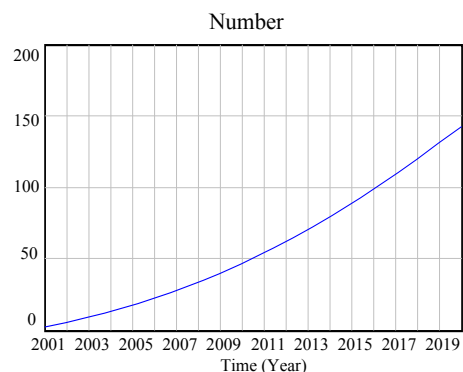


Number : Current
 Input : Current
 Human Resource : Current
 "Planning & Control Project" : Current
 "Management of R&D" : Current

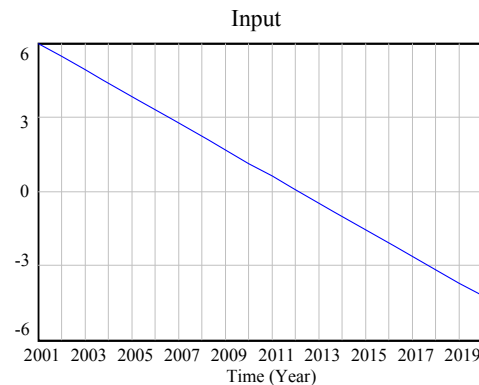
chart3: Multiple Charts of the basic model variable

5.6. Statistical analysis (sensitivity analysis):

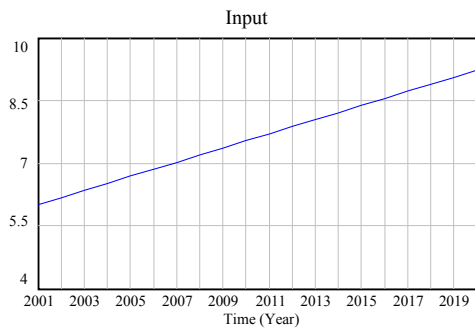
(Houshangnia 2009), (Rabelo 2004), (Sterman 2000)



Number : Current



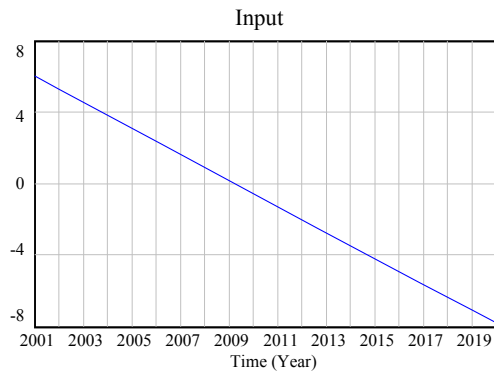
Input : Current



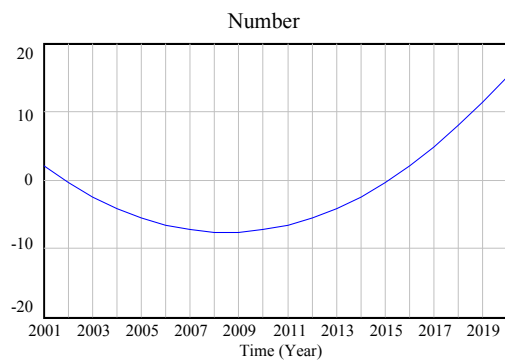
Input : Current

Year	Primary state for system input	Optimistic state for system input	Project numbers in optimistic conditions
2011	0.6	7.7	53.8
2020	-4.2	9.23	142.4

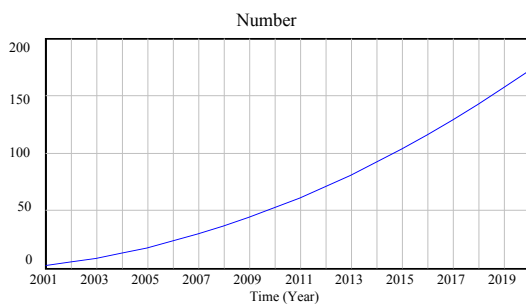
Chart4: Optimistic level conditions for System Input, Coefficient change from -0.61 to +0.1



Input : Current



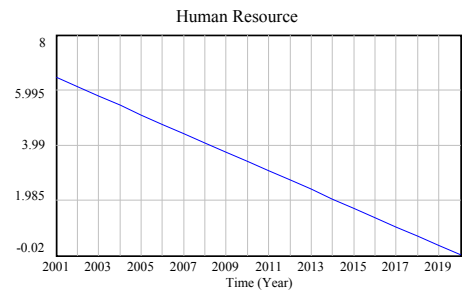
Number : Current



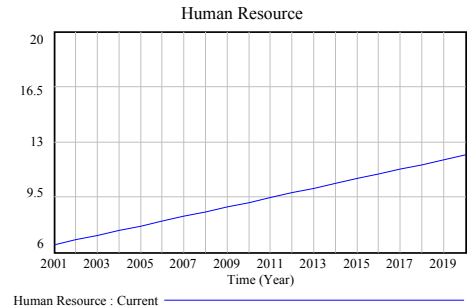
Number : Current

Year	Primary state for system input	Optimistic state for system input	Project numbers in optimistic conditions
2011	0.6	-1.3	-6.6
2020	-4.2	-7.8	15.23

Chart5: Pessimistic level conditions for system input, Coefficient change from -0.61 to -0.80



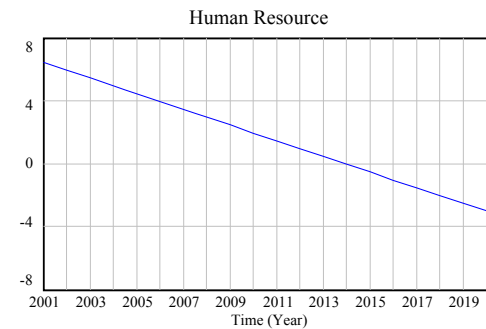
Human Resource : Current



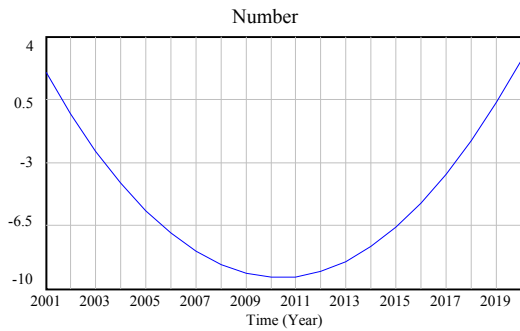
Human Resource : Current

Year	Primary state for human resource	Optimistic condition for human resource	Project numbers in optimistic conditions
2011	3.05	9.4	60.9
2020	-0.01	12.15	171.8

Chart 6: Optimistic level conditions for Human Resource, Coefficient change from -0.34 to +0.3



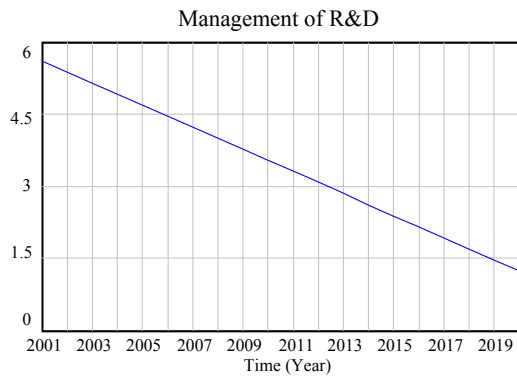
Human Resource : Current



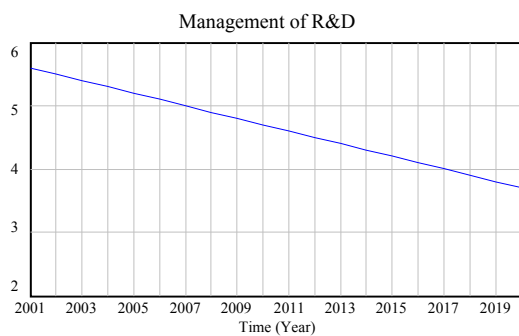
Number : Current

Year	Primary state for human resource	Optimistic condition for human resource	Project numbers in optimistic conditions
2011	3.05	1.45	-9.36
2020	-0.01	-3.05	2.69

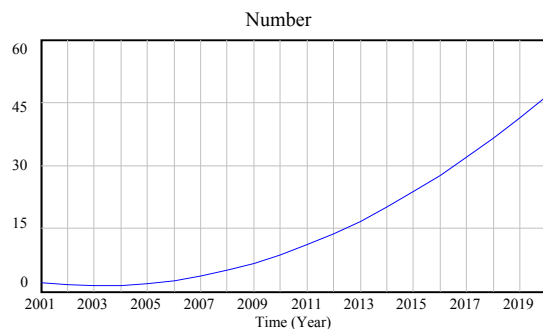
Chart7: Pessimistic level conditions for Human Resource, Coefficient change from -0.34 to -0.5



"Management of R&D" : Current



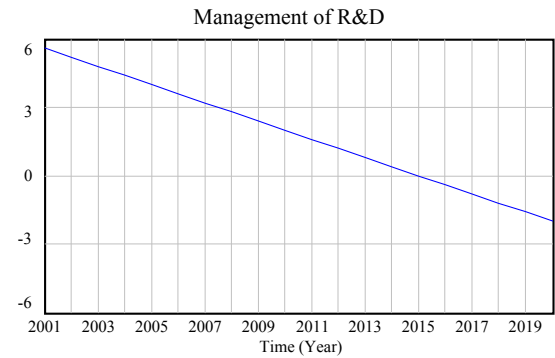
"Management of R&D" : Current



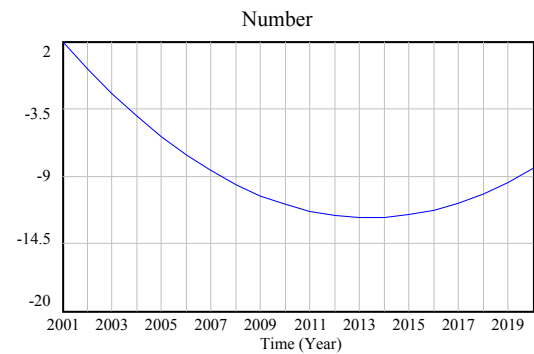
Number : Current

year	Primary state of R&D	Optimistic state of R&D	Project numbers in optimistic conditions
2011	3.3	4.6	10.99
2020	1.23	3.69	46.63

Chart8: Optimistic level conditions for Management of R&D, Coefficient change from 0.72 to 0.85



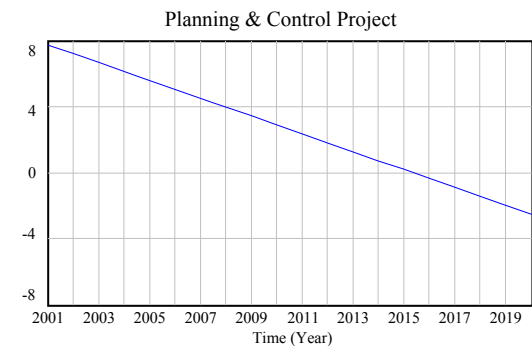
"Management of R&D" : Current



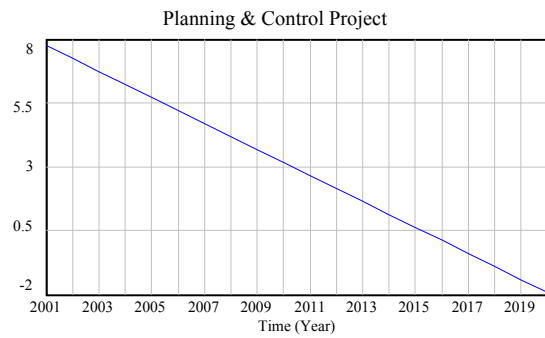
Number : Current

year	Primary state for R&D Management	Pessimistic conditions for R&D management	Project numbers in pessimistic conditions
2011	3.3	1.6	-11.88
2020	1.23	-2	-8.3

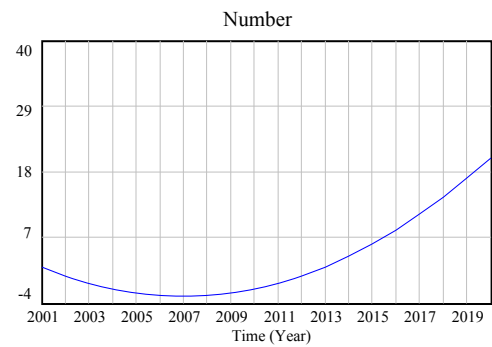
Chart9: Pessimistic level conditions for Management of R&D, Coefficient change from 0.72 to 0.55



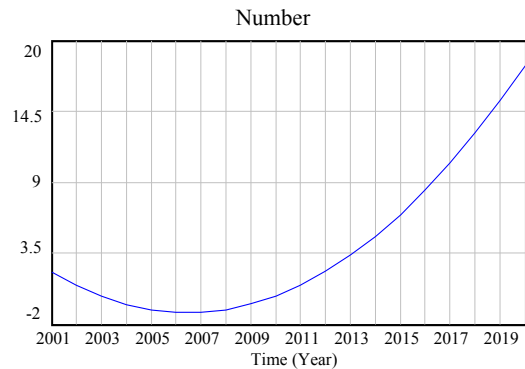
"Planning & Control Project" : Current



"Planning & Control Project" : Current



Number : Current



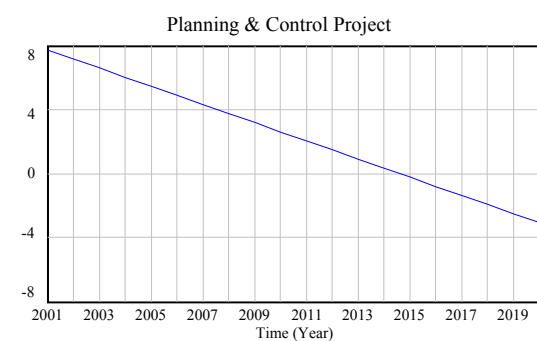
Number : Current

Year	Primary state for planning & project control of R&D	Pessimistic state for planning & project control of R&D	Project numbers in pessimistic conditions
2011	2.39	2.05	-0.7
2020	-2.5	-3.08	20.29

Chart1: Pessimistic level conditions for Planning & Project control of R&D, Coefficient change from 0.07 to 0.04

Year	Primary state for planning & project control	Optimistic state for planning & project control	Project numbers in optimistic conditions
2011	2.39	2.65	0.99
2020	-2.5	-1.9	19.05

Chart10: Optimistic level conditions for Planning & Project control of R&D, Coefficient change from 0.07 to 0.10



"Planning & Control Project" : Current

6. CONCLUSIONS

As noted, the purpose of this paper is presenting a local model to consider capabilities and abilities of Research and Development unit by emphasizing on reverse engineering in a Chemical Factory. Based on the pattern provided, indicators and parameters to measure Features and capabilities of R & D unit are available.

In this model, some of the basic parameters which were measured, analyzed based on sensitivity analysis include:

Number of Projects done in R & D unit, system inputs such as raw materials and new requirements, professional and skilled manpower, skilled management of R & D, Project Planning and Control of research and development project in analyzing the sensitivity of above indices and parameters, some results were obtained. By eliminating restrictions of indices, it means in terms of optimism for the variables, we can mostly observe that growth is very rapid rate for parameters and variables and also features R & D unit and ultimately the growth of parameters lead to significant increase in the rate of the number of projects conducted by R & D unit. Also by considering the conditions of reverse condition it means, the sensitivity test of variables in terms of increase of barriers in the parameters, we can mostly witness decrease of parameters and number of projects done in R & D.

Finally by reviewing the outputs of model software graphs and sensitivity test of parameters and effective factors for model, in this article the capabilities and abilities of Research and development (R & D) of chemical factory have been determined. We can use these indices and variables to achieve specific goals specific and by comparing them relatively with other industries of R & D and or by comparing them with

research and development units of industrial plants located in other countries, the efficiency and effectiveness of R & D unit in the chemical factory to be measured.

Finally, this model can be used to identify desired features and capabilities of R & D and to use required managements to reach them and they be used to successfully control the programs. The model has the ability to generalize and use in most application areas for selection of strategies, policies and research and development projects and access to new technology and we can observe the basic parameters of the model by its effect on the capabilities and abilities of research and development unit.

By observing all tables and charts of the basic parameters we can conclude that in spite of reduced volatility variables and factors in the model during the years 2001-2010, the increase of projects number in R & D unit of chemical factory has had a good condition. This increase is relevant to the urgency and necessity of unfinished projects which are being done by the important factor of R & D Management.

According to the relations of all the basic parameters and doing regression and the corresponding formula model, we can conclude that the most effective factor on the number of projects is the factor of management, planning and control of projects and R & D unit of the chemical factory has a good condition in terms of R & D management, and the reason of this issue in addition to observing relations of variables and formulas coefficients, in addition to the information and graphs is due to necessity and emergency point of unfinished projects and coordination of the main parameters by skillful of Research and development (R & D).

REFERENCES

- Akhbari,mohsen and others,2008,*article of process of developmevt of new products*,tadbir journal number 184
- Allahyari,parinaz,2009, designing a model for organizational R&D units of generator companies, *MS thesis*.
- Amiri,farzad and others,2008, *article of john doe reel of development of new product in environment of generation of globalization level, third international conference R&D*
- Colin Bradley,1998, The application of reverse engineering in rapid product development, *ISI Technical paper*
- Craig W.Kirkwood,1998, System Dynamics Methods: A Quick introduction, college of Business Arizona State University
- Ebrahimi,abdolhamid,*conference of development of new products*, <http://www.irmmc.com/index1.htm>
- Fadaei,marjan, 2010,examination of effects of R&D unit on efficiency of industries, *MS thesis*
- Fort Collins,H.R.Kaufman,1989,*Book of R&D Tactics*
- Ghani,asgar,2009, application of reverse engineering in achieve to technology of wrapped products, *MS thesis*
- Houshangnia,amir pasha,2009, simulation of role of R&D in development process, *MS thesis*
- Jokar,mohamad sadegh,2008, *article of processes of structural for development of new product,(models of innovation in creation technology)*
- Kheradmand,kamran,2007, examination effect of R&D on profitable of industries in Iran, *MS thesis*
- Laghvi,reza,2010, designing of R&D system and execution of it in weaving factories, *MS thesis*
- Mardi,asghar,2011, designing of system structure for R&D units in generator companies, *MS thesis*
- P.A.Roussel K.N.saad ,T.J.Erickson,1991,*Book of Third Generatino on R&D.Harvard Bussiness school press , Boston*
- Rabelo,L. and Helal,M.2004,Analysis of Supply chains using system dynamics, Neural Nets and Eigenvalues
- Sterman,j,2000, Bussiness Dynamics-Modeling & system thinking for a complex world, *McGraw Hill, New York,USA*
- BOOK of Technology comercializationi, the 5-Stage R&D Commercialization Process http://www.1000ventures.com/technology_transfer/tech_commercialization_main.html.
- Vittorio chiesa,Christina Masella,1996,*Paper of Searching for an effective measure of R&D performance, ISI Technical paper*
- www.systemdynamics.org/conferences/1998/PROCEED/abstracts.pdf

ON THE USE OF MINIMUM-BIAS COMPUTER EXPERIMENTAL DESIGNS

Husam Hamad

Electronic Engineering Department
Hijawi College of Engineering Technology
Yarmouk University, Jordan

husam@yu.edu.jo

ABSTRACT

Computer experimental designs are used to generate data in metamodeling of multiresponse engineering systems. Metamodels, which are also called surrogate models, offer more efficient prediction of system responses but add errors when used as surrogates for the simulators. Error sizes depend on computer experimental designs. Only bias errors are incurred in deterministic computer experiments; however, the majority of experiments reported in the literature are not optimized for minimum bias. Box and Draper—the pioneers of the response surface methodology—originated the work on minimum bias designs in the late 1950's. Space-filling designs such as the Latin hypercubes are mainly in current use; sometimes even in response surface models. This work is a practical study via a number of analytical and electronic circuit examples on the use of minimum bias designs for response surface metamodels. Some minimum bias designs in hypercuboidal spaces are also introduced.

Keywords: Experimental design, minimum bias design, Latin hypercube design, response surface models

1. INTRODUCTION

Computer experimental designs are sampling techniques used to determine combinations of design variables to generate metamodels (also known as surrogate models) of complex engineering systems responses. Different sampling techniques are used to generate metamodels using simulation output for system responses for points in the experimental design. For deterministic simulations, errors introduced by the metamodels are systematic, or bias, errors caused by the deficiency of the metamodel in truly representing the response. Contrary to data in practical experiments, no variance-related error components are present in computer experiments.

Experimental designs can be categorized into classical designs and the more recent space-filling designs (Chen et. al 2006). Classical designs such as factorial designs (Myers and Montgomery 1995) and central composite designs (Box and Wilson 1951) are primarily used in response surface modeling methods.

Space-filling designs such as the Latin hypercube designs (Mckay, Beckman, and Conover 1979) aim at uniformly scattering the points over the design variables space.

Different system response complexities require different metamodel types in order to adequately accommodate the underlying behavior and reduce bias errors. Hence, different metamodel types exist depending on the underlying response. Response surface models and kriging metamodel types receive much coverage in the current literature on the design and analysis of computer experiments. Other types, considered to be equally competitive in current usage according to (Simpson et al. 2008) include multivariate adaptive splines, radial basis functions, neural networks, and support vector regressors. The study in (Chen et al. 2006) concludes that no one metamodel type stands out. A similar conclusion is made in (Wang 2003), stating that no one metamodel type is definitely superior to others. According to (Goel et. al 2007), the consensus among researchers is that no single metamodel type can be considered the most effective for all responses. Nonetheless, (Wang 2003) also concludes that kriging and second-order polynomial response surfaces are the most intensively investigated metamodels. Based on a Google Scholar search, the work in (Simpson et al. 2008) concludes that response surface models are the favorite methods in structural optimization disciplines. While (Viana and Haftka 2008) conclude in a Google Scholar search that the distinction between metamodels diminished after an initial popularity for response surface and artificial neural networks techniques; they nonetheless acknowledge that response surface models are the favorite techniques in structural optimization.

The review in (Chen et. al, 2006) on the design and modeling of computer experiments investigated experimental design methods and their relation to the various types of metamodels used in computer experiments. The review presented conclusions from attempts by many researches to determine the most appropriate experimental design for the selected metamodel type. Based on their own computational study tests on the available options, (Chen et. al, 2006) conclude that response surface model designs such as

the central composite designs and the Box-Behnken designs are "good only" for response surface models, while all other experimental designs (space-filling designs such as the Latin hypercube samples) are appropriate for all metamodels other than the response surface models. It is noteworthy to mention here that minimum bias designs were not included in the review by (Chen et. al 2006).

A Google Scholar search similar to those in (Simpson et al. 2008) and (Viana and Hafka 2008) was conducted in this work in April, 2011. The results are shown in Table 1 for response surface models and in Table 2 for kriging metamodels.

Table 1: Search Results Related to Response Surface Models Using Google Scholar

Search Phrase	Number of Publications	
	2000-2011	2005-2011
approximation OR metamodel OR surrogate AND "response surface"	12800	9090
"experimental design" AND "response surface"	15400	12400
"minimum bias design" AND "response surface"	24	14
"Latin hypercube" AND "response surface"	2210	1720

Table 2: Search Results Related to Kriging Metamodels Using Google Scholar

Search Phrase	Number of Publications	
	2000-2011	2005-2011
approximation OR metamodel OR surrogate AND "kriging"	9360	6790
"experimental design" AND "kriging"	1830	1340
"Latin hypercube" AND "kriging"	1370	1130

The results in Tables 1-2 lead to the following general possible interpretations with regard to experimental designs and metamodeling methods:

- Response surface models of the 1950's still compete with the more recent metamodels such as the kriging type. As seen in Table 1, the number of publications with "response surface" in combination with any of the words approximation, metamodel, or surrogate since 2000 is about 12,800. The majority of these publications (9,090) appeared in the last half of the last decade from 2005 to 2011. The corresponding statistics for "kriging" metamodels are 9,360 for the period 2000-2011, with 6,790 of these publications appearing in the period 2005-2011.

- From the other tables entries, of the 15,400 papers since 2000 having the phrase "response surface" AND "experimental design", only 24 papers mention "minimum bias designs" while 2,210 papers talk about "Latin hypercube" designs. What are the reasons for the unpopularity of minimum bias computer designs? During the times minimum bias designs were presented in articles in the late 1950's and early 1960's (Box and Draper 1959; Draper and Lawrence 1965), experiments were conducted in the laboratories, and hence the reasons for avoiding large experimental designs are obvious. However, the recent space-filling designs used in computer experiments of today can have larger sizes than most of minimum bias designs, so reasons attributed to size for ignoring these designs are ruled out.

There are two main objectives for the work presented in this paper:

- To show that minimum bias computer experimental designs *can potentially* give more accurate response surface models than the widely used space-filling designs of comparable size. This is demonstrated via analytical functions and electronic circuits.
- To introduce some minimum bias computer experimental designs for hypercuboidal spaces of dimensions 2 to 6.

The remainder of this paper is organized as follows: section 2 demonstrates through analytical examples the motives for using minimum bias designs. Section 3 deals with error types due to variance and bias, presenting basis which are subsequently applied to cuboidal design spaces to construct minimum bias designs. Some of these designs are then used in the electronic circuit examples of section 4. Conclusions are given in section 5.

2. MOTIVATION

Sample points in a minimum bias experimental design are located in the design region such that the design's moments satisfy certain conditions as outlined in the next section. In this section, analytic examples are used to demonstrate the superiority vis-à-vis prediction accuracy of metamodels based on minimum bias designs (MBD) in comparison to models derived using other experimental designs such as the Latin hypercube (LHC) designs.

Figure 1 shows four experimental designs used to derive a first-order response surface for the response given in Equation (1):

$$y = 5 + 2x_1 - x_2 + 0.5x_1^2 + 3x_1x_2 + x_2^2 ; x \in [-1,+1] \quad (1)$$

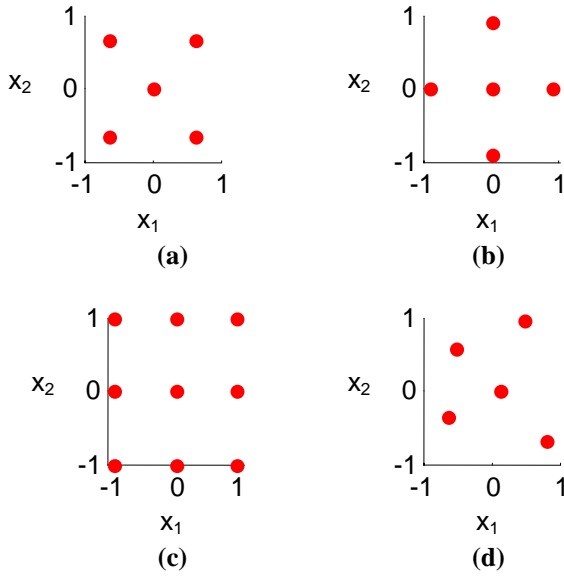


Figure 1: Experimental Designs (a) MBD 1 (b) MBD 2 (c) FAC (d) LHC

Two MBDs are shown in parts (a) and (b); part (c) depicts the standard response surface model design known as factorial design (FAC), and a LHC design is shown in part (d).

Metamodels built using these designs are validated using a 21x21 sample. See Table 3.

Table 3: Validation Results Corresponding to the Four Experimental Designs in Figure 1

Experimental Design	Figure 1 Part	RMSE
MBD 1	a	1.160
MBD 2	b	1.160
FAC	c	1.243
LHC	d	1.381

As shown in the table, the lowest root mean square error (RMSE) is obtained using any of the two MBDs. To demonstrate the relation between RMSEs for MBDs and LHCs, 100 metamodels are fitted using 100 different LHC samples. RMSEs for these metamodels are compared to the RMSE obtained if a MBD is used; see Figure 2. In the figure, the RMSEs shown are normalized to the RMSE for the MBD (the dotted line at 1.0).

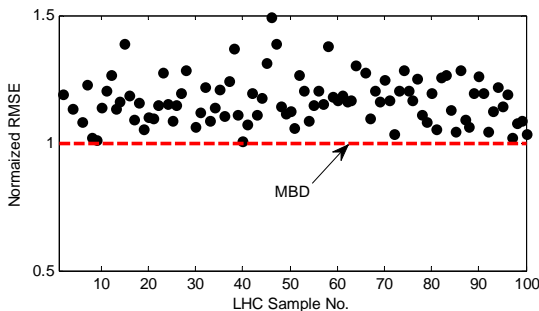


Figure 2: Normalized RMSEs for 100 LHC Designs.

The results depicted in the figure clearly demonstrate the superiority of MBDs. Unfortunately,

this is not always the case. To illustrate, the above metamodeling activities are repeated for the response in Equation (2) (see Figure (3) for function plot):

$$y = \sum_{i=1}^9 a_i (x - 900)^{i-1} ; x \in [905, 995] \quad (2)$$

where $a_1 = -659.23$, $a_2 = 190.22$, $a_3 = -17.802$, $a_4 = 0.82691$, $a_5 = -0.01885$, $a_6 = 0.0003463$, $a_7 = -3.2446 \times 10^{-6}$, $a_8 = 1.6606 \times 10^{-8}$, and $a_9 = -3.5757 \times 10^{-11}$.

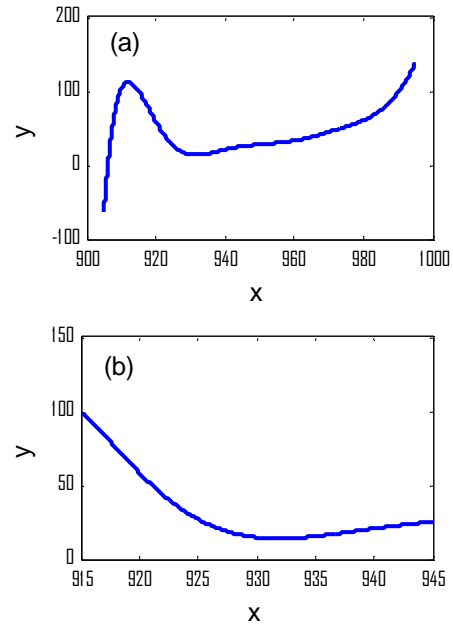


Figure 3: Function Plot for $y(x)$ in Equation (2) for: (a) $x \in [905, 995]$ (b) $x \in [915, 945]$

In Figure (4) RMSEs for 100 different metamodels built using 100 different LHC samples are compared to the RMSE for the metamodel derived using a MBD (each LHC sample has the same size as the MBD).

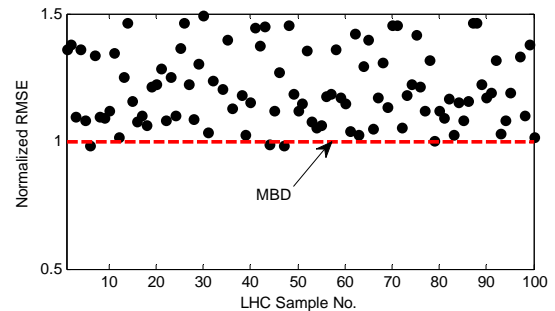


Figure 4: Normalized RMSEs for 100 LHC Designs for the Function in Figure 3(a).

The figure shows that RMSE for the MBD case is not always lower (a few points are below the dotted line

at 1.0 corresponding to the normalized RMSE for the MBD). However, for most of the 100 LHC samples, their RMSEs are worse by comparison to the MBD sample. The reason for the discrepancy between results of the similar metamodeling activities summarized by Figures 2 and 4 are attributed to the underlying response being modeled. As it will be shown, MBDs result in least errors provided the underlying assumptions for deriving MBDs are satisfied. Usually, the derivation assumes that the complexity of the response is such that it is higher than the response surface metamodel that fits it by one order; e.g., the response follows a third-order polynomial if the metamodel fitted is a second-order polynomial. Obviously, as the design variables space narrows down, such assumption about orders becomes more valid (see Figure 3). This is demonstrated in Figure 5, which is similar to Figure 4 except now the design space for the response is narrowed down to $x \in [915, 945]$ from $x \in [905, 995]$ in Equation (2).

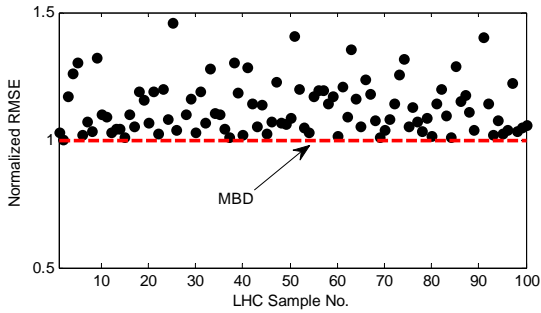


Figure 5: Validation Results for the Function in Figure 3(b).

3. MINIMUM BIAS DESIGNS

There are two sources for errors in metamodels: (i) noise in the experimental design data used to fit the metamodel; and (ii) inadequacy of the metamodel. Accordingly, errors are categorized as: (i) variance, and (ii) bias errors, respectively. In practical experiments, variance errors are the assumed error source while bias errors are the only source of errors in computer experiments.

Standard response surface designs such as the central composite designs are derived ignoring bias errors; i.e., derivations in this case assume that the fitted metamodel adequately represents the response. In minimum bias designs, however, it is customary to assume that the true response is a one-order higher polynomial than the metamodel. Thus, if the metamodel is a second-order polynomial then the MBD is derived assuming a third-order polynomial response.

There is no intention in this paper to provide rigorous mathematical treatment for MBD derivations. Such derivations originated in the pioneering work by Box and Draper in 1950's (Box and Draper 1959), with more recent treatment in (Goel et. al 2008) and (Abdelbasit and Butler 2006). The results are presented in terms of satisfying the necessary and sufficient

conditions for MBD derivation in terms of the so-called design moments.

(Draper and Lawrence 1965) applied the above mathematical conditions in (Box and Draper 1959) to derive MBDs for cuboidal regions. They used parameterized experimental design sets to build first and second-order MBDs. However, many of the tabulated results involve sets with parameters outside the assumed coded design space boundaries. This may be inappropriate in many practical engineering system design problems; for example, negative transistor widths cannot be implemented in practice.

Our work (also for cuboidal design spaces) involves the parameterized experimental design sets mentioned shortly later on in this section. However, solutions for the parameters resulting in practical MBD sets (i.e., with none of the parameters outside the design space) are taken when the mathematical conditions related to design moments are applied. The sets are used to construct second and third-order MBDs.

Consider a k -dimensional space with design (input) variables x_1, x_2, \dots, x_k . It is assumed that the space is coded such that $-1 \leq x_1, x_2, \dots, x_k \leq +1$. Second and third-order MBDs in our work are constructed using combinations of the following sets: $C(0^k)$, $F(\alpha^k)$, and $S(\alpha^a, \beta^{k-a})$. Explanation for this notation is provided in Table 4.

Table 4: Notation Used

Notation	Meaning	#points	Notes
$C(0^k)$	a design point at the center	1	$x_1 = \dots = x_k = 0$
$F(\alpha^k)$	factorial design	2^k	See Table 5 for $k = 3$
$S(\alpha^a, \beta^{k-a})$	All k permutations of factorial designs with a variables at α and $k - a$ variables at β	$k2^k$	See Table 6 for $k = 3$

Table 5: $F(\alpha^k)$ Factorial Design for $k = 3$

$x_1 = \pm\alpha$	$x_2 = \pm\alpha$	$x_3 = \pm\alpha$
$-\alpha$	$-\alpha$	$-\alpha$
$-\alpha$	$-\alpha$	$+\alpha$
$-\alpha$	$+\alpha$	$-\alpha$
$-\alpha$	$+\alpha$	$+\alpha$
$+\alpha$	$-\alpha$	$-\alpha$
$+\alpha$	$-\alpha$	$+\alpha$
$+\alpha$	$+\alpha$	$-\alpha$
$+\alpha$	$+\alpha$	$+\alpha$

Table 6: $S(\alpha^a, \beta^{k-a})$ Design for $k = 3$ with $a = 1$

$x_1 = \pm\alpha$			$x_1 = \pm\beta$			$x_1 = \pm\beta$		
$x_2 = \pm\beta$			$x_2 = \pm\alpha$			$x_2 = \pm\beta$		
$x_3 = \pm\beta$			$x_3 = \pm\beta$			$x_3 = \pm\alpha$		
x_1	x_2	x_3	x_1	x_2	x_3	x_1	x_2	x_3
$-\alpha$	$-\beta$	$-\beta$	$-\beta$	$-\alpha$	$-\beta$	$-\beta$	$-\beta$	$-\alpha$
$-\alpha$	$-\beta$	$+\beta$	$-\beta$	$-\alpha$	$+\beta$	$-\beta$	$-\beta$	$+\alpha$
$-\alpha$	$+\beta$	$-\beta$	$-\beta$	$+\alpha$	$-\beta$	$-\beta$	$+\beta$	$-\alpha$
$-\alpha$	$+\beta$	$+\beta$	$-\beta$	$+\alpha$	$+\beta$	$-\beta$	$+\beta$	$+\alpha$
$+\alpha$	$-\beta$	$-\beta$	$+\beta$	$-\alpha$	$-\beta$	$+\beta$	$-\beta$	$-\alpha$
$+\alpha$	$-\beta$	$+\beta$	$+\beta$	$-\alpha$	$+\beta$	$+\beta$	$-\beta$	$+\alpha$
$+\alpha$	$+\beta$	$-\beta$	$+\beta$	$+\alpha$	$-\beta$	$+\beta$	$+\beta$	$-\alpha$
$+\alpha$	$+\beta$	$+\beta$	$+\beta$	$+\alpha$	$+\beta$	$+\beta$	$+\beta$	$+\alpha$

MBDs for $k = 2 - 6$ (generated by applying the sufficient and necessary conditions in the references mentioned at the beginning of this section to the above design sets) are given in Table 7 for second-order MBDs and in Table 8 for third-order designs. Add one center point $C(0^k)$ for each row in the tables to complete the MBD.

Table 7: Second-Order MBDs

k	$F(\alpha^k)$	$S(\alpha^a, \beta^{k-a})$		
		α	β	a
2	-	0.418	0.759	1
3	-	0.816	0.434	1
4	-	0.868	0.448	1
5	-	0.913	0.460	1
6	0.620	0.973	0.450	1

Table 8: Third-Order MBDs

k	$F(\alpha^k)$	$S_1(\alpha^a, \beta^{k-a})$			$S_2(\alpha^a, \beta^{k-a})$		
		α	β	a	α	β	a
2	0.685	0.255	0.741	1	-	-	-
3	-	0.775	0.252	1	0.378	0.763	1
4	-	0.844	0.305	1	0.202	0.743	1
5	0.724	0.801	0.311	1	-	-	-
6	-	0.951	0.287	1	0.194	0.742	1

Note that the size of second-order MBDs in Table 7 is $1 + k2^k$ points for $k \leq 5$.

4. APPLICATION TO ELECTRONIC CIRCUIT MODELING

In this section two electronic circuits are modeled using MBDs and the results are compared to LHC designs. The two circuits are the amplifier and filter in Figure 6.

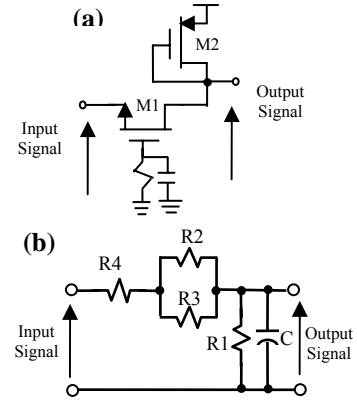


Figure 6: Two Electronic Circuits: (a) Amplifier (b) Filter.

The gain (the ratio of output signal to the input signal) $A_{amplifier}$ of the amplifier, and the maximum gain A_{filter} and bandwidth BW_{filter} of the filter are modeled using the appropriate MBDs in Table 7, with $k = 2$ for the amplifier and $k = 5$ for the filter. Figure 7(a) shows RMSE comparisons for $A_{amplifier}$ for the region $W_1 \in [2, 200]$ and $W_2 \in [2, 200]$, where W_1 and W_2 are the width of the two amplifier transistors M1 and M2 in Figure 6(a). When the space is narrowed down to $W_2 \in [2, 20]$ for W_2 , RMSEs become worse (by comparison to RMSE for the MBD) for more of the 100 LHC samples as demonstrated in part (b) of Figure 7. This is expected as demonstrated earlier for the function in Figure (3).

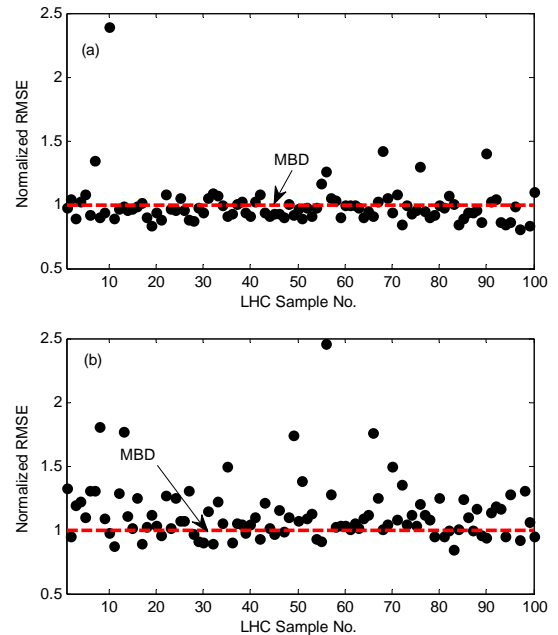


Figure 7: Results for the Amplifier Circuit: (a) for the Region $W_1 \in [2, 200]$ and $W_2 \in [2, 200]$ (b) for the Narrower Region $W_1 \in [2, 200]$ and $W_2 \in [2, 20]$.

Results for RMSEs for the filter circuit are shown in Figure 8. Note that while the results give advantage for the MBD for A_{filter} as shown in part (a); however, part (b) of the figure shows that RMSEs for the LHC samples are lower for BW_{filter} . This is the worst case obtained in our work. Nonetheless, even for this case the RMSE for all 100 LHC samples is nearly 90% on average of the RMSE obtained using MBD as can be inferred from Figure 8(b).

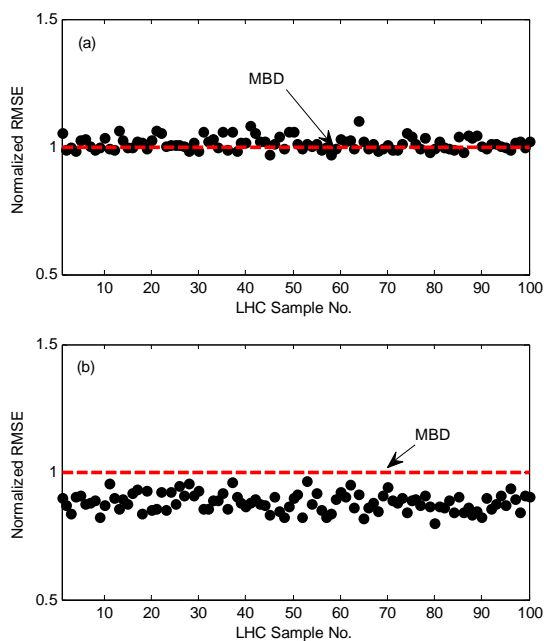


Figure 8: Results for the Filter Circuit: (a) A_{filter}
(b) BW_{filter} .

5. CONCLUSIONS

Metamodels are appropriate surrogates for simulators in the design of complex engineering systems provided that the errors incurred are acceptable. Bias errors due to metamodel inadequacy result in inaccurate metamodels when computer experimental data are used to construct these metamodels. This paper demonstrated that minimum bias computer experimental designs are potentially superior in response surfaces by comparison to space-filling designs such as the popular Latin hypercube samples. Also, the paper introduced minimum bias designs for normalized hypercuboidal spaces. The list of these designs is by no means exhaustive, and more work is needed to expand the list for higher-dimension spaces and higher-order minimum bias designs.

REFERENCES

Abdelbasit, K.M., Butler, N.A., 2006. Minimum bias design for generalized linear models. *The Indian Journal of Statistics*, 68, 587-599.

- Box, G.E.P., Draper, N.R., 1959. A basis for the selection of a response surface design. *Journal of the American Statistical Association*, 54, 622-654.
- Box, G.E.P., Wilson, K.B., 1951. On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society, Series B*, 13, 1-45.
- Chen, V., Tsut, K-L., Barton, R.R., Meckesheimer, M., 2006. A review on design, modeling and applications of computer experiments. *IIE Transactions*, 38, 273-291.
- Draper, N.R., Lawrence, W.E., 1965. Designs which minimize model inadequacy: cuboidal regions of interest. *Biometrika*, 52, 111-118.
- Goel, T., Haftka, R.T., Shyy, W., Queipo, N.V., 2007. Ensemble of surrogates. *Structural and Multidisciplinary Optimization*, 33, 199-216.
- Goel, T., Haftka, R.T., Shyy, W., Watson, L.T., 2008. Pitfalls of using a single criterion for selecting experimental designs. *International Journal for Numerical Methods in Engineering*, 75, 127-155.
- Mckay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21, 239-245.
- Myers, R.H. and Montgomery, D.C., 1995. *Response surface methodology: process and product optimization using designed experiments*. New York: Wiley.
- Simpson, T.W., Toropov, V., Balabanov, V., Viana, F.A.C., 2008. Design and analysis of computer experiments in multidisciplinary optimization: a review of how far we have come – or not. *American Institute of Aeronautics and Astronautics*, 1-22.
- Viana, F.A.C., Haftka, R.T., 2008. Using multiple surrogates for metamodeling. *Proceedings of the 7th ASMO-UK/ISSMO International Conference on Engineering Design Optimization*, Bath (UK). 1-18.
- Wang, G.W., 2003. Adaptive response surface method using inherited Latin hypercube design points. *Journal of Mechanical Design*, 125, 210-220.

AUTHOR BIOGRAPHY

HUSAM HAMAD is an associate professor in the Electronic Engineering Department at Yarmouk University in Jordan. He is the Vice Dean of Hijjawi College of Engineering Technology at Yarmouk. He received his B.S. in Electrical Engineering from Oklahoma State University in 1984, M.S. in Device Electronics from Louisiana State University in 1985, and PhD in Electronic Systems Engineering from the University of Essex, England, in 1995. He was a member of PHI KAPPA PHI Honor Society during his study in the U.S. His research interests include modeling, analysis, simulation and design of electronic systems and integrated circuits, metamodel validation, electronic design automation, and signal processing.

A PRACTICAL GUIDE FOR THE INITIALISATION OF MULTI-AGENT SYSTEMS WITH RANDOM NUMBER SEQUENCES FROM AGGREGATED CORRELATION DATA

Volker Nissen^(a), Danilo Saft^(b)

^(a) ^(b) Ilmenau Technical University, Faculty of Economics, Institute for Commercial Information Technology,
Chair of Information Systems for Services (WI 2),
Postfach 100565, 98684 Ilmenau, Germany

^(a) volker.nissen@tu-ilmenau.de, ^(b) daniilo.saft@tu-ilmenau.de

ABSTRACT

This article describes a scalable way to initialise a simulation model with correlated random numbers. The focus is on the nontrivial issue of creating predefined multidimensional correlations amongst those numbers. A multi-agent model serves as a basis for practical demonstrations in this paper, while the method itself may be interesting for an even wider audience within the modelling and simulation community beyond the field of agent-based modelling. In particular, we show how researchers can create streams of correlated random numbers for different empirically-based model parameters when just given aggregated statistics in the form of a correlation matrix. An example initialisation procedure is demonstrated using the open source statistical computing software “R” as well as the open source multi-agent simulation software “Repast Symphony”.

Keywords: MAS Parameterisation, Correlated Random Numbers, R-Project, Repast

1. INTRODUCTION

The simulation of a model may sometimes require a large amount of parameters, which influence its outcome (significantly). In a subset of these cases, the parameters may be interdependent in such a way that the initialisation of a model needs two or more parameters to correlate in a predefined manner. A procedure to generate and utilise such numbers will be explained in the following. We use the example of an agent-based model, since one of our main research areas is the field of agent-based economics. In this research, we regularly find illustrative scenarios to which the concept presented in this paper is applicable. Note that while the statements here will be kept limited to agent-based models for scientific validity, these explanations can easily be transferred to the initialisation of other types of simulation models.

In a variety of multi-agents systems, the model to be simulated may consist of a large number of heterogeneous agents. Heterogeneity can come in the form of different spatial positions of individual agents, different network connections, opinions, etc. In general, each of these agents possesses a set of parameters with

different initialisation values. Researchers may want to relay data acquired from the real world (e.g. through measurement series, questionnaires or statistical archives) to initialise their agents with according parameter values for reasons of testing, prognosis, or simply for validity.

In the case of social or economic simulations, an agent may possess variables such as income, reputation, job satisfaction, household size, etc. Scientists may however not in all cases be lucky enough to find real-world-data at a level that is as detailed as their desired simulation setup may require. They, therefore, may need to evade to more aggregated forms of simulation, matching the aggregation level of the empirical data available. This option can be unsatisfactory as important details of the micro-level to be simulated and/or the emerging micro-macro-links within such a simulation might need to stay unaddressed.

One alternative is testing different random number distributions where detailed data is missing. This latter approach holds chances, but also challenges, such as the possible availability of empirical data that cannot give numbers on a detailed level, but gives aggregated distributions and correlations of different variables found in an empirical study. Table 1 (Oreg, 2006, p. 88) shows an instance of the results such empirical surveys yield. The values shown in table 1 were put forth by Oreg in 2006 and will serve as a data example throughout this paper. Table 1 gives descriptive statistics and correlations between parameters important to individual behaviour in the context of organisational change. The statistics were extracted from a series of questionnaires given to individuals within a company undergoing several organisational adjustments. The researches recorded variables thought to be important to an individual's opinion formation about an organisational change. They derived a static statistical model of interdependencies of individual properties (tb. 1, variables 1 to 9), the resistance an individual develops towards an organisational change (tb. 1, variables 10 to 12) and the behavioural outcome its opinion has on its specific job (tb. 1, variables 13 to 15).

From a multi-agent modelling perspective, it becomes possible to analyse the *dynamic* behaviour of a simulated company as a whole by modelling individuals

Table 1: Exemplary descriptive statistics and correlations for the variables of an empirical study by Oreg (2006, p. 88)

Variable	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Age	45	12	1													
2. Manager	0.54	0.50	.17*	1												
3. Dispositional resistance to change	3.19	0.75	.19*	-.04	1											
4. Improvement in power-prestige	2.91	0.74	.11	.09	.05	1										
5. Improvement in job security	2.81	0.68	.20**	.11	.02	.48**	1									
6. Improvement in intrinsic rewards	3.12	0.67	.10	.14	.18*	.68**	.30**	1								
7. Trust in management	3.82	1.39	-.01	.09	-.03	.32**	.25**	.31*	1							
8. Information	3.93	1.34	.13	.19*	.04	.13	.05	.08	.45**	1						
9. Social influence (against the change)	3.71	1.44	-.01	-.13	-.05	-.27**	-.06	-.09	-.15	-.03	1					
10. Affective resistance	3.02	1.15	-.09	-.01	.31**	-.43**	-.31**	-.32**	-.33**	-.02	.33**	1				
11. Behavioural resistance	2.30	1.19	.01	.17*	.12	-.30**	-.19*	-.24**	-.30**	.03	.26**	.60**	1			
12. Cognitive resistance	4.21	1.11	.03	-.04	-.03	-.55**	-.24**	-.52**	-.52**	-.09	.23**	.46**	.50**	1		
13. Job satisfaction	5.74	0.92	.15	-.04	.00	.17*	.06	.21**	.17*	.19*	-.17*	-.15*	-.18*	-.02	1	
14. Intention to quit	2.42	1.26	-.15	.03	-.06	-.17*	-.16*	-.16*	-.24**	-.19*	.13	.23**	.15	.15*	-.49**	1
15. Continuance commitment	3.71	1.09	-.01	-.15	.33**	.10	-.03	.10	.02	-.04	.02	-.01	.11	-.12	-.11	.10

* $p < .05$, ** $p < .01$.

with heterogeneous specific opinions and the (direct or indirect) influence individuals have on each other, e.g. through social interactions and information exchange (cmp. tb. 1, variables 8 and 9) or other behaviour affecting organisational “neighbours”. In this example, the goal of researchers in the field of multi-agent-based simulation (MABS) would be to better understand the process of organisational change or analyse the effect of certain formal and informal hierarchies and network structures on company performance. In fact, such investigations are taking place within our own MABS research of which a first part has already been published (Nissen and Saft, 2010). We utilise the real-world study of “resistance to change” behaviour in organisations to initialise our own multi-agent simulation of a virtual organisation in order to better understand how resistance to change spreads and can be influenced by management. This requires (here: agent-based) modelling at a more detailed level than the aggregated statistical data in table 1 provides. To this end we can however use the values in table 1 to yield specific initialisation data for a (large) number of simulated members of a virtual organisation. To initialise each agent in our simulation model with its own personality values, it is necessary to extract sequences with specific random numbers while accounting for the multidimensional correlations given in table 1 in order to yield correct number sequences. This challenge is trivial only when one needs to generate two correlated series of random numbers, i.e. when the agents in the simulation only possess pairs of two correlated properties.

2. A LITTLE BIT OF MATH

For only two numbers (parameters) to be correlated, there is a simple approach to retrieve two correlated random numbers from a set of uncorrelated random numbers:

$$y_1 = \sigma_1 * x_1 \quad (1a)$$

$$y_2 = c * \sigma_2 + \sqrt{(1 - c^2)} * \sigma_2 * x_2 \quad (1b)$$

where x_1 and x_2 are two uncorrelated random numbers from a given distribution, σ_1 and σ_2 are their standard deviations, and c is the desired correlation coefficient between y_1 and y_2 ; i.e. the resulting correlated random numbers.

With more than two sequences of correlated random numbers to generate, one can use a variety of mathematical approaches that are more or less difficult to go through manually. In our case, an Eigenvector-decomposition was employed as we found it to be a process that is robust and offers good performance. There also is the option of using the so-called Cholesky-decomposition (Lloyd and Trefethen, 1997, pp. 172 - 178) which will however not be explained further here. Given a correlation matrix C (see columns “1” through “14” in table 1) one can define a matrix

$$V = E_i \text{Diag}(\sqrt{\lambda_i}) \quad (2)$$

where E_i are the eigenvectors of C and λ_i are the eigenvalues of C .

With a matrix I_u consisting of formerly uncorrelated random numbers, we can derive a matrix

$$I_c = I_u V^T \quad (3)$$

where V^T is the transpose of V .

I_c then contains random numbers with correct correlations.

I_u can consist of any number of random values where each line can represent the initialisation values for a single agent and each column stands for one of the correlated parameters of an agent. This offers great flexibility since one only needs to choose the number of rows in the original matrix I_u as big as the number of agents one wishes to instantiate/simulate. I_u should already include the general distribution properties such as – in our case – the mean and standard deviations given in table 1.

3. A PRACTICAL GUIDE FOR THE UTILISATION OF CORRELATED RANDOM NUMBER SEQUENCES USING “R” AND “REPAST”

While the pattern to generate correlated random numbers shown in section 2 can be time-consuming to do by hand, it is a very easy process once one employs supporting software. The popular and well-documented open source program “R” (Hrshikesh, 2010; Jones, 2009) is able to make the necessary calculations (for all practical purposes implied here) in just a fraction of a second. The tool, along with many additional packages, can be downloaded freely for a variety of platforms (The R Project, 2010).

Below, we will present exemplary step-by-step R code to generate random number sequences for the three correlated parameters “improvement in power-prestige”, “improvement in job security”, and “trust in management” listed in table 1. The code is to generate correctly correlated random numbers for 2500 agents as virtual “employees” of a simulated organisation. Note that the code pattern is scalable to a large number of parameters and agents.

The first step is to generate a matrix of uncorrelated random numbers for each agent, already taking into account the correct mean and standard deviation properties (in this case assuming the Gaussian distribution from table 1):

```
//create a matrix with 3 columns for 3
//parameters and 2500 rows for 2500
//agents:
Iu <- matrix(, ncol=3, nrow=2500)

//fill in random values for “power and
//prestige”,...:
Iu[,1] <- rnorm(2500, 2.91, 0.74)

//...then “job security”,...:
Iu[,2] <- rnorm(2500, 2.81, 0.68)

//... and finally “trust in management”:
Iu[,3] <- rnorm(2500, 3.82, 1.39)
```

Next, the correlations for these parameters need to be filled into another matrix C :

```
//create a squared matrix and fill in the
//correlations for each of the three
//correlated parameters, using the order
//“power-prestige”, “job satisfaction”,
//and “trust in management”:
C <- matrix(, ncol=3, nrow=3)
C[1,] <- c(1, 0.48, 0.32)
C[2,] <- c(0.48, 1, 0.25)
C[3,] <- c(0.32, 0.25, 1)
```

R offers a simple command to calculate both the eigenvalues and eigenvectors of a matrix. We will save the result of this operation in an object E . The parameter “symmetric” refers to C being a symmetrical matrix so that only the lower triangle of the matrix needs to be used in the calculations:

```
E <- eigen(C, symmetric=TRUE)
```

The call to $E\$vectors$ will then give us the eigenvectors of C and $E\$values$ will return the eigenvalues accordingly. We use the command “diag” to construct a fictive matrix diagonal from the square roots of the three eigenvalues of C . This call is necessary for a valid multiplication. Note that for the calculation of the square roots to be valid, all eigenvalues of C must be positive. This will however implicitly be the case for valid correlation matrices. We can then create the matrix V according to equation 2 given in section 2:

```
V <- E$vectors %*% diag(sqrt(E$values))
```

Finally, we can multiply our formerly uncorrelated random number matrix I_u with the transpose of V in order to receive a matrix I_c with 2500 rows each containing three correctly correlated values in three columns, where (in our example) the first column stands for the parameter “power and prestige”, the second for “job security”, and the last one for “trust in management”:

```
Ic <- Iu %*% t(V)
```

The result of this code can be saved in a CSV-File. In order to do this, we can use the “write.table” command included in R. “write.table” takes several parameters of which the first is the matrix to write into a file and the second is the file’s name on disk. The parameter “sep” defines a character by which the values of the matrix are to be separated in the output file. We use “col.names=FALSE” and “row.names=FALSE” to specify that we do not wish to export any column or row names. In order to not put any values of the matrix in quotes, we set “quote=FALSE”. In case a value is not set in the matrix (which, following the aforementioned steps, ought to be irrelevant in our case), we can specify a string value written to the file in its place. Here, for instance, we could use the Java-compatible “NaN” string for “not a number” by setting the parameter “na” accordingly:

```
write.table(Ic, "C:/filename.csv", sep=",",
col.names=FALSE, row.names=FALSE, quote=FALSE,
na="NaN")
```

We can now use the random number sequences in this file for use in an external program, in our case the “Recursive Porous Agent Simulation Toolkit” Repast Symphony (North et al., 2006). It is a Java-based open source software with seemingly growing popularity in the MABS-research community (Barnes, 2010) and is available as a free download (REPAST, 2010).

Repast employs a so-called Context Creator to initialise simulations. Within this class, agents can be

created and parameters may be set before the simulation begins. Skipping over most of the code of our initialisation routine, we will again present exemplary code to assign the generated correlated values to each agent using the Context Creator. In our example, we wish to create a virtual organisation with 2500 employees, each having different, but correlated parameters as explained above. We utilise a CSV-reader class that simply returns the matrix saved by R as a two-dimensional Java Double array¹.

```
Double[][] correlatedRandomNumbers =
CSV_Reader.readFile("C:/filename.csv");
```

All we then need to do is to create our agents and read out the correlated random values in the correct order:

```
//stylised iteration:
for (int i=0; i<2500; i++) {
    EmployeeAgent ea = new EmployeeAgent();
    ea.powerPrestige =
correlatedRandomNumbers[i,0];
    ea.jobSecurity =
correlatedRandomNumbers[i,1];
    ea.trustInMgmt =
correlatedRandomNumbers[i,2];
    ...
}
```

The approach itself is very flexible and scalable to a large number of agents and correlated parameters. Performance tests were conducted on an Intel Core2-Duo pc with 4GB memory and a hard disk spinning at 5400rpm. Figure 1 shows a 3D-mesh-plot portraying the execution times for the generation of 10.000 to 50.000 correlated vectors (i.e. agents) for respectively 10 to 100 parameters for each individual. The upper part of figure 1 displays execution times for the calculations themselves without disk output to a CSV file. The bottom plot shows execution times including the disk output. The data shows that even for the calculation of 100 correlated parameters for 50.000 agents it only takes slightly over two seconds to retrieve all necessary values using a code analogous to the example code listed above. Also, the computation complexity seems to rise only linearly with a rising number of agents and parameters which makes this approach interesting for large scale simulations with either a large number of agents or parameters in one simulation, or a large number of simulations running in parallel (e.g. for parameter optimisation purposes). Note that there is, however, a performance-bottleneck where the files need to be written to disk. Therefore, using many parameters or agents, one should refer to Lang (2005) or JRI (2011) for a way to directly call R-functions from within Java-

¹ Several similar Java-based CSV-readers are also available online.

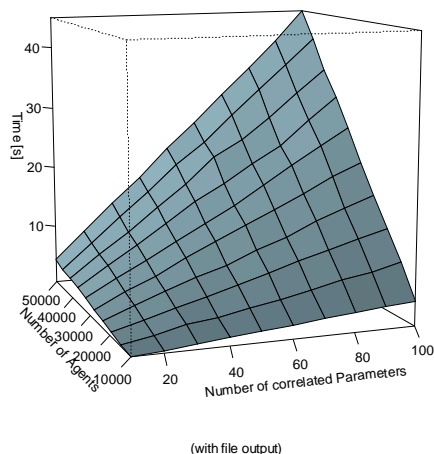
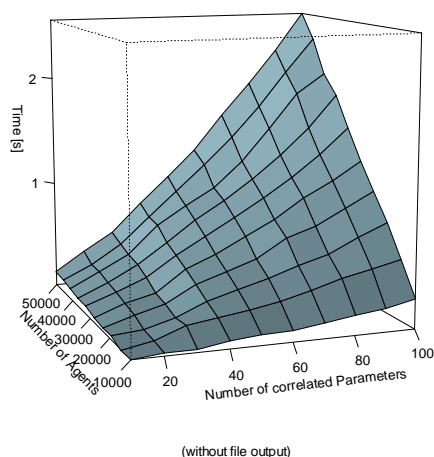


Figure 1: Computation Time of Generating Random Correlated Number Sequences for Varying Numbers of Agents and Parameters with (bottom) and without (top) Disk Output to a CSV File.

based applications such as Repast. We will explain this process briefly in the following.

The so-called Java-R-Interface (JRI, 2011), amongst other interfaces available, is able to send commands to an instance of R running in the background of a Java-based application. Since simulations in the MAS-tool Repast Symphony can be programmed in the Java language, JRI can be easily implemented for use in such multi-agent-simulations.

JRI is available as part of a package-extension of R called “rJava”, which was originally designed to send data and commands in the opposite direction, i.e. from R to Java. The quickest method to utilise only the functionality of the Java-R-Interface, nevertheless, is to install the complete “rJava” package within R. Once installed, the downloaded folders within R’s own extension library will contain the JRI Java-archive for implementation as a library in any Java project, too. Note that the subsequent setup of the JRI library files can be difficult. For a step by step guide based on

Eclipse/Repast, please refer to Shah (2009). Once JRI is available for use in Repast, one can extend the Context Creator class of a simulation analogously to the case of CSV files described above. However, here one would directly initialise agents using the calculations made in R. For our example case, the necessary code is listed below:

As a first step, it is important to make the JRI library available to the Repast simulation project and include it in the import statements of the simulation part needing to access R, i.e. the ContextCreator.java class file in our case:

```
import org.rosuda.JRI.REXP;
import org.rosuda.JRI.Rengine;
import org.rosuda.REngine.*;
```

We are then using the initialisation routine of the Context Creator class (e.g. the “*build()*” method) to access R routines. Here, we first create an object of type *Rengine* which is the main instance passing commands and data between Java and R. The constructor of this class can pass various arguments to the instance of R to be used. Please refer to the JRI documentation (JRI, 2011) at this point in order to adjust this step for your requirements. In all cases, the “*waitForR()*” function of the newly created *Rengine* object should be called to make sure that the R thread finished its program start sequence before calculations can begin. Not including a call to this method may lead to Java exceptions being raised at this step and the failure of Repast’s Context Creator initialisation routine:

```
//creating a new instance of R for
//calculations:
Rengine re = new Rengine(new
String[]{""}, false, null);
//important: waiting for the R instance
//to finish loading:
if (!re.waitForR()) {
    System.out.println("Cannot load R");
}
```

The *Rengine* type now offers the method “*eval(String command)*” (amongst numerous others) to execute and evaluate a command passed over to R in the form of a simple string parameter. This method returns an object of type *REXP* (R expression), which can subsequently be used to output or further interpret results of the command sent via “*eval*”. The following code listing demonstrates several calls to send commands to R analogous to the example R-code already given above. We retrieve the final calculation of matrix I_c within the object *ex* of type *REXP*:

```
//creating instance for return values:
REXP ex;
//executing R example code as stated in
//beginning of section 3 of this paper:
re.eval("Iu <- matrix(, ncol=3,
nrow=2500)");
re.eval("Iu[,1] <- rnorm(2500, 2.91,
0.74)");
re.eval("Iu[,2] <- rnorm(2500, 2.81,
0.68)");
re.eval("Iu[,3] <- rnorm(2500, 3.82,
1.39)");
re.eval("C <- matrix(, ncol=3,
nrow=3)");
re.eval("C[1,] <- c(1, 0.48, 0.32)");
re.eval("C[2,] <- c(0.48, 1, 0.25)");
re.eval("C[3,] <- c(0.32, 0.25, 1)");
re.eval("E <- eigen(C,
symmetric=TRUE)");
re.eval("V <- E$vectors %*%
diag(sqrt(E$values))");
//catching the final result of Ic in
//"ex" for further handling:
ex =re.eval("Ic <- Iu %*% t(V)");
```

Now we only need to create an array just as one would when using CSV files. The *REXP* type has several routines for formatting the results, e.g. an “*asString()*” method for outputting its contents to the Java console. Here, we use the “*asMatrix()*” method that interprets the results as a two-dimensional array of *Double* values. We can then again use this array to iterate through it, assigning the parameter values to each of our agents:

```
Double[][] correlatedRandomNumbers =
ex.asMatrix()
//stylised iteration:
for (int i=0; i<2500; i++) {
    EmployeeAgent ea = new
    EmployeeAgent();
    ea.powerPrestige =
    correlatedRandomNumbers[i,0];
    ea.jobSecurity =
    correlatedRandomNumbers[i,1];
    ea.trustInMgmt =
    correlatedRandomNumbers[i,2];
    ...
}
```

Accessing R through the JRI library is a very efficient method to compute the necessary calculations for the initialisation of a Repast simulation. Further performance tests using this method have shown that there is no loss of performance present when using JRI rather than R itself to create large numbers of correlated random values. Performance results are comparable to those shown in the top part of fig. 1.

4. CONCLUSIONS AND EXTENSIONS

This article dealt with the question of how to generate scalable sets of correlated random numbers for the initialisation of agent-based simulations. The authors found this to be a question both important and difficult as many empirical studies provide aggregated descriptive statistics, including correlations, while there is also a necessity for detailed simulations at the level of single and interacting individuals to explore certain issues especially when dealing with complex and/or emergent systems (Nissen and Saft, 2010, p. 113). The process of extracting correctly correlated sequences of random numbers for each agent is nontrivial and literature on this topic, especially in the form of practical guides for researchers in the (multi-agent) simulation community without a deep mathematical background, is scarce. The reader therefore was provided with a step-by-step guide for how to create a matrix containing MAS initialisation data in the form of correlated random number sets for each agent as well as with a stylised example code for the wide-spread Repast simulation software in order to access those values indirectly via file output or directly using the so-called JRI-package. This paper therefore serves as a demonstrative guide for a wide audience of researchers in the simulation community. It provides a time-saving way as well as a quick access for newcomers to the creation of correlated random number sequences for MAS parameterisation.

The steps shown here can further be enhanced by using additional packages, e.g. the R-Commander package for R, providing quick and easy access to basic R operations via a graphical user interface (Fox, 2010). There are also other options to call R directly from Java and Java-based software (Lang, 2005), eliminating the need to use CSV Files for storage. Such approaches are beneficial since one is not only able to outsource initialisation calculations, but any complex set of calculations that should rather be executed in a professional environment such as R.

REFERENCES

Barnes, D. J.; Chu, D., 2010. ABMs Using Repast and Java, *Introductio to Modeling for Biosciences*, Springer London, 79-130.

Fox, J., 2010, *The R Commander: A Basic-Statistics GUI for R*, Available from: <http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/> [accessed 01 May 2011]

Hrishikesh D., 2010. Advances in Social Science Research Using R. *Lecture Notes in Statistics*. Springer.

Jones, O.; Maillardet, R., Robinson, A., 2009. *Introduction to Scientific Programming and Simulation Using R*. Chapman & Hall/CRC, Boca Raton (Florida, USA).

JRI (without date), *JRI* Available from: <http://www.rforge.net/JRI/> [accessed 01 May 2011]

Lang, D.T., 2005, *Calling R from Java*, Available from: <http://www.omegahat.org/RSJava/RFromJava.pdf> [accessed 01 May 2011]

Lloyd N., Trefethen, D.B., 1997. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics

Nissen, V.; Saft, D., 2010. Social Emergence in Organisational Contexts: Benefits from Multi-Agent Simulations. In: Madey, G.R.; Sierhuis, M.; Zhang, Y. (Eds.): *Proceedings of the Agent-Directed Simulation Symposium*, San Diego: SCS, 2010, 106 – 113 (CD).

North, M.J.; Collier, N.T.; Vos, J.R., 2006. Experiences creating three implementations of the repast agent modeling toolkit. *ACM Trans. Model. Comput. Simul.*, 16(1), 1-25

Oreg, S., 2006. Personality, context, and resistance to organisational change. *The European journal of work and organisational psychology*, (15), 73-101

REPAST Developers Group, (2010), *Repast Home Page*, Available from: <http://repast.sourceforge.net> [accessed 01 May 2011]

Shah, M., 2009, *R and Java – JRI using eclipse*, Available from: <http://mithil-tech.blogspot.com/2009/11/r-and-java-jri-via-eclipse.html> [accessed 01 May 2011].

The R Project for Statistical Computing, (2010), *The R Project for Statistical Computing*, Available from: <http://www.r-project.org/> [accessed 01 May 2011]

MULTI-AGENT MULTI-LEVEL MODELING – A METHODOLOGY TO SIMULATE COMPLEX SYSTEMS

Jean-Baptiste Soyez^(a), Gildas Morvan^(b), Rochdi Merzouki^(c), Daniel Dupont^(d)

^(a,b,c,d)Univ. Lille Nord de France, 1bis rue Georges Lefèvre 59044 Lille cedex, France

^(a,c)LAGIS UMR CNRS 8146 École Polytechnique de Lille Avenue Langevin 59655 Villeneuve d'Ascq, France

^(b)LG12A, Univ. Artois Technoparc Futura 62400 Béthune, France

^(a,d)HEI Hautes Études d'Ingénieurs 13 rue de Toul 59046 Lille cedex, France

^(a)Jean-Baptiste.Soyez@polytech-lille.fr, ^(b)gildas.morvan@univ-artois.fr, ^(c)rochdi.merzouki@polytech-lille.fr,
^(d)daniel.DUPONT@hei.fr

ABSTRACT

This article deals with the conception, modeling and simulation of complex systems, represented at different *levels* of analysis with respect to the agent-based modeling (ABM) paradigm and more precisely on the generic meta-model IRM4MLS. A methodology, using IRM4MLS, is proposed to save computational resources in multi-level agent-based simulations, representing only the relevant elements. It means that the structure of agents can be modified during simulation, by temporarily aggregating, removing or approximating their characteristics to maximize their life-cycles.

Keywords: multi-agent modeling, multi-level modeling, simulation, influence/reaction model

1. INTRODUCTION

This article deals with the conception, modeling and simulation of complex systems, represented at different levels of analysis. A level can be defined as a point of view on the studied system and its relations to other points of view. Therefore, in most real world applications, a level encapsulates the processes executed at a given spatio-temporal scale; the term multi-scale is also used. This work relies on the agent-based modeling (ABM) paradigm and more precisely on the generic meta-model IRM4MLS introduced by Morvan et al. (2010), based on the influence/reaction principle (Ferber & Muller 1996).

According to Quijano and al. 2009, three important issues in *classic* (or flat) ABM can be solved by multi-level agent-based models (MAM). (1) Some complex systems cannot be understood without integrating knowledge that is ontologically distributed over multiple levels of organization. In other words, system behavior cannot, using the available knowledge, be described with a purely emergent (or bottom-up) approach. Examples of such models can be found in Morvan and al. 2008 and Morvan and al. 2009. (2) Many distributed systems are characterized by a macroscopic behavior that becomes remarkable when the number of entities is important. Indeed, “*agentifying*” this behavior can be useful (Servat and

al. 1998). (3) Simulated entities can be reflexive, i.e., able to reason on social facts (similarities with other agents, group involvement, etc.) such as in Gil Quijano and al. 2009 and Pumain and al. 2009.

In this article, a methodology using IRM4MLS, is proposed to tackle complexity issues of MAM which have to simulate many entities with complex interactions. An important aim of this methodology, among others, is to save computational resources, representing only the relevant elements of a simulation when they are needed, i.e., lighten the representation of agents and their life-cycles. The main idea is to identify agent characteristics (state variables, available actions and decision processes, etc.) that can be modified at run-time. It means that the structure of agents can be modified during simulation, by temporarily aggregating, removing or approximating their characteristics to maximize their life-cycles.

2. RELATED WORKS

The works on multi-level agent-based modeling are related to other approaches that view simulations of complex systems as *societies of simulations*.

The High Level Architecture (HLA) is a general purpose architecture for distributed simulations computer simulation systems. Using HLA, computer simulations can interact (that is, to communicate data, and to synchronize actions) to other computer simulations regardless of the computing platforms. The interaction between simulations is managed by a Run-Time Infrastructure (RTI).

Holonic multi-agent systems (HMAS) can be viewed as a specific case of multi-levels multi-agent-systems (MAS), the most obvious aspect being the hierarchical organization of levels. However, from a methodological perspective, differences remain. Most of holonic meta-models focus on organizational and methodological aspects while MAM is process-oriented. HMAS meta-models have been proposed in various domains, e.g., ASPECTS (Gaud and al. 2008) or PROSA (Van-Brussel et al. 1998.). Even if MAM and HMAS structures are close, the latter is too constrained for the target application of this work.

Multi-Resolution Modeling (MRM) (Davis and al. 1993) which is the joint execution of different models of the same phenomenon within the same simulation or across several heterogeneous systems, can inspire our approach if the different models are at different levels. The consistency symbolizes the amount of essential information lost during the passing between models and it is a good tool to test the quality of this approach.

Navarro and al. (2011) present a framework to dynamically change the level of detail in an agent-based simulation. That is to say, represent in detail only which is needed during simulation, to save CPU resources and keep the consistency of the simulation. But this framework is limited because all levels form a meshed hierarchy, without the possibility of having two different levels at the same scale and communication between levels is not explicitly defined.

In the section 3, the main concepts used in this article (ABM and IRM4MLS) are introduced. This presentation emphasizes our vision of agent architecture in the context of IRM4MLS. In the section 4, 3 methods to save computational resources (RAM and CPU) in models based on IRM4MLS are introduced. In the conclusion, we sum-up the main interests of our method and its perspectives.

3. MAIN CONCEPTS

3.1. Agent-based modelling and simulation

The multi-agent formalism is straightforward. Each entity of the studied system possesses an equivalent entity (or *computational agent*) in the computer representation. It is then easier to apprehend than mathematical models.

Jacques Ferber (1995), p. 12 gives a definition of what is an agent: “We call agent a *physic or virtual entity which*

- a) can act in an environment,
- b) can directly communicate with others agents,
- c) is pushed by a set of tendencies (represented by some goals or a satisfaction function or a survival one),
- d) possesses its own resources,
- e) is able to perceive (but in a limited way) its environment,
- f) has a partial representation of its environment (eventually none),
- g) has some capacities and proposes some service,
- h) can eventually reproduces itself,
- i) whose behavior tend to satisfy its objectives, considering his resources and available capacities, also considering its perception, its representations and the received

communications.” (translation from french by authors)

An agent can be seen as an entity composed of a non-observable internal state (e.g., its beliefs, desires and intentions in the BDI architecture) and an observable external state (e.g., its position, velocity, acceleration and direction in situated multi-agent systems). In multi-agent based simulations, the execution of agents is scheduled (cf. Fig. 1).

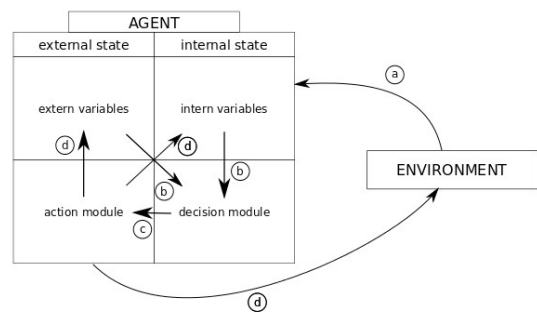


Figure 1: Basic Structure of an Agent and his life-cycle

The life-cycle of an agent can be described as follows:

- a) the agent senses its environment to construct percepts,
- b) the decision module selects the action to perform
- c) from its internal and external states
- d) the agent acts (in the influence reaction model, produces influences in its environment.

Drogoul and al. (2002) highlight that in ABM, 3 different kinds of agent are used: “*real agents*”, that can be observed in the studied system, “*conceptual agents*”, i.e., formalizations real agents with respect to MAS concepts (communications, interactions, etc.) and finally “*computational agents*”, i.e., executable implementations of conceptual agents on the target simulation platform. This precision is important: in the following a *same real agent* will be successively represented by *different computational agents*.

3.2. IRM4MLS

IRM4MLS is a MAM meta-model proposed by Morvan and al. 2011. It relies on the influence / reaction model (Ferber & Muller 1996) its extension to temporal systems, IRM4S (Michel 2007). Beside its generality, an interesting aspect of IRM4MLS is that any valid instance can be simulated by proposed algorithms (Soyez and al. 2011). Only the main aspects of IRM4MLS are presented in this section. Readers

interested in a more exhaustive presentation may refer to referenced publications.

A MAM is characterized by a set of levels, L , and relations between levels. Two types of relations are considered in IRM4MLS: influence (agents in a level l are able to produce influences in a level $l' \neq l$) and perception (agents in a level l are able to perceive the state of a level $l' \neq l$). These relations are respectively formalized by two digraphs, $\langle L, E_I \rangle$ and $\langle L, E_P \rangle$ where E_I and E_P are sets of edges, i.e., ordered pairs of elements of L . The dynamic set of agents at time t is denoted $A(t)$. $\forall l \in L$, the set of agents in l at t is $A_l(t) \subseteq A(t)$. An agent acts in a level iff a subset of its external state belongs the state of this level. An agent can act in multiple levels at the same time.

Environment is also a top-class abstraction. It can be viewed as an agent with no internal state that produces “natural” influences in the level (Fig. 2).

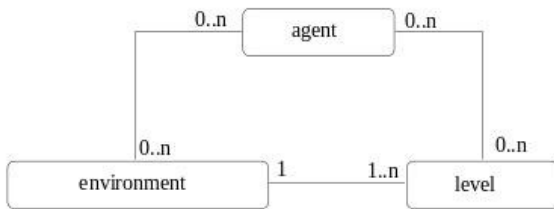


Figure 2 : Central Concepts of IRM4MLS (cardinalities are specified with the UML way)

The scheduling of each level is independent: models with different temporalities can be simulated without any bias. On an other hand, it permits to execute only the relevant processes during a time-step.

A major application of IRM4MLS is to allow microscopic agents (members) to aggregate and form-up lower granularity agents (organizations). It has to be noticed that level does not necessarily means scale: it can be useful to create multiple levels at a same scale.

4. THE METHODOLOGY

4.1. General principle

Agent-based modeling possesses many advantages, but it can be difficult to be run in a simulation because the agent execution is often greedy in computer resources, particularly in used CPU time and memory.

A simple idea is to keep active in the simulator the minimum number of variables and processes and update these elements with the lower frequency (only when it's necessary). All unused elements, at a given moment, by the simulator can be write on the hard disk and delete of the simulation. When the simulator need these elements it can read and integrate them again. This is useful only if the gain spawn by the smaller quantity of data to be managed and it smaller updating frequency is more important than the loss of time

generated by the reading and writing of data on the hard disk.

Also, it's not necessary to always know with the same precision some data, that's why these data can be aggregated (Lucia 2010) or approximate. There are many methods or algorithms to apply these three mechanisms : approximation, aggregation or reading/writing data on hard disk and free random access memory. We do not propose new algorithms to dynamically determine which active data can be processed by these methods.

This work takes place in the InTrade european project, which deals with the Autonomous Intelligent Vehicles (AIV) traffic flow in the major european ports. ScannerStudio is a real time simulator created for InTrade. This makes us work on the real time problematic where the proposed methods to lighten a simulation have a strong sens. The example cited below are inspired and can be implemented in ScannerStudio.

4.2. Agents resources

Agent structure is rather an heavy thing. Modelers of a simulation, generally, conceive agents with the four parts described above, which integrate all elements which can be needed at one moment of the simulation. We propose to dynamically adapt the structure of an agent to only represent what is necessary at a given time. This includes the used agent resources or the processes whom the modelers want to observe. Description of the life cycle and the structure of an agent and it modeling composed of four parts permit us to apprehend the dependencies between these parts. A modeler can see, during the modeling or the simulation phases, that these parts can be decomposed into relatively independent sub-parts. And to take a decision the sub-parts of the decision module don't called the same variables (internal or external), the sub-parts of the decision module don't activate the same actions and each action modify a different set of variables.

In his book Calvez (1990) mentions some decomposition methods which can be applied to specify a system to model it. Applying these decomposition methods, which can be functional, structural or modal, on a agent permit to determine the sub-parts of the four parts of this agent. Functional decomposition decompose the agent mission in a number of independent functions. For example an intelligent vehicle has a moving, diagnostic, pick up and delivery functions. Structural decomposition is based on the material components of agent, isolating groups of independent components. For example, a vehicle traction system can be isolated from it odometer. Modal decomposition can be seen in system whose elements posses several functioning modes. This can be illustrated by a vehicle with several modes to describe it state: ready, degraded, failure.

Once these we used these methods and isolated variables and process, we create, for this independent group, as many level as many expressed needs, in necessary resources to model these groups. These

groups can contain heterogeneous agent if their representation possess similar needs. By this way, during the simulation an entity of the studied system, will be represented by agents situated on different levels.

This part of the article illustrate that some agents at different levels are not obligatory situated at different scales but at different level of representation, more or less detailed or supporting more or less functionality. The following figure illustrate that fact, showing three agents at different levels, obtained with functional decomposition, these agents can potentially represent the same vehicle. During the simulation a vehicle can be represented by an agent of one of the three levels according to the simulation needs.

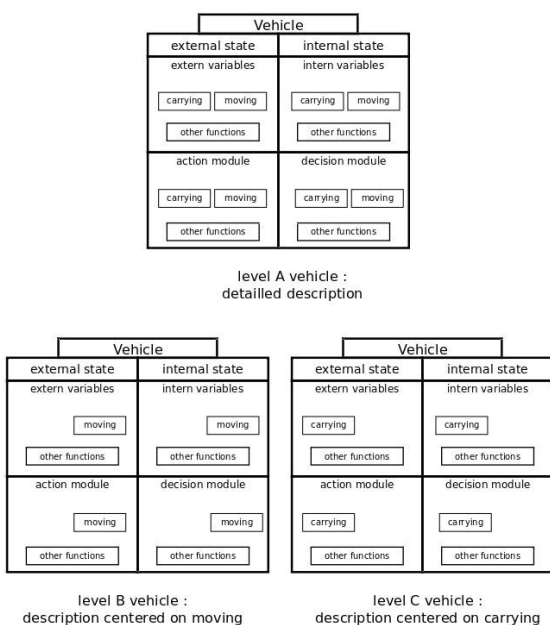


Figure 3: 3 Levels of Agents at the Same Scale

4.3. Relevant elements in a multi-level simulation

4.3.1. The organizations viewpoint

When agents of a single scale group themselves, it's possible to create in the simulation the formed organization and the corresponding agent at an smaller level of granularity. This is interesting to represent characteristics proper to the organization no deductible only from the agents member characteristics. Also, it permit to approximate or aggregate the life of the members and their interactions in the organization. Representation of some agents with bigger granularity is contained in the organization agent. That's make these member agents useless for a time and they can be delete of the simulation.

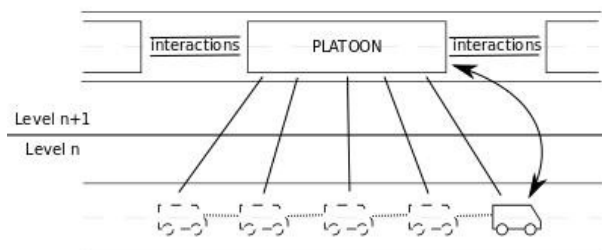


Figure 4: Organization Viewpoint driven implementation

The above figure illustrates this case. Five vehicles grouped themselves to form a platoon. That gives birth to a platoon agent representing this group. Because the platoon agent can approximate the variables and the results of the communications between members of the platoon the member agents representing following vehicles are no more needed in the simulation and can be delete. The member agent representing the leader vehicle is needed because there is no function, in the platoon agent, which can approximate it routing. So it have to share it routing with the platoon agent (this interaction is represented by the double arrow).

4.3.2. The composing viewpoint

When modelers wish to represent some entities of the system at a given scale, it's no more necessary to keep the personification of groups formed by these entities because this induce a redundancy of data and process. This is illustrated by the next figure with the last shown example. The modelers wish to observe the platoon vehicles individually, that requires the existence of the agent vehicles in the simulation. Once these agents have been created the agents representing the groups they form (here a platoon) can be delete. At that moment the life of the organization and it interactions with others ones are supported by the vehicles agents.

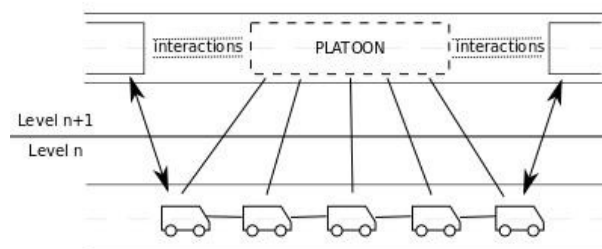


Figure 5: Members Viewpoint driven Implementation

4.4. Level temporality

We use the framework of temporalities simulation system (Zeigler 2000). No constraints is imposed on the scheduling mode (time or discrete events), but the fact that the scheduling is distributed between level. This

distribution makes more sense than a synchronization on others “levels” like the agent one (Weyns 2003) or the system one (Michel and al. 2003), which are not adapted for our problematic.

Levels can have different temporal dynamics. Independently of the other levels, it's interesting to give a level a dynamic temporal whose time step is higher as possible. This is made in order to update the dynamic state of this level as less often as possible, respecting the wish of the modeler. That is to say, giving to the agents the longest possible life cycle which stay coherent with the rest of the simulation. IRM4MLS is a structured interactions model. Morvan and al. (Morvan and al. 2010) propose an algorithm adapted to IRM4MLS which manages the coupling between levels with different temporal dynamics. That permits to apply easily the proposed methods above.

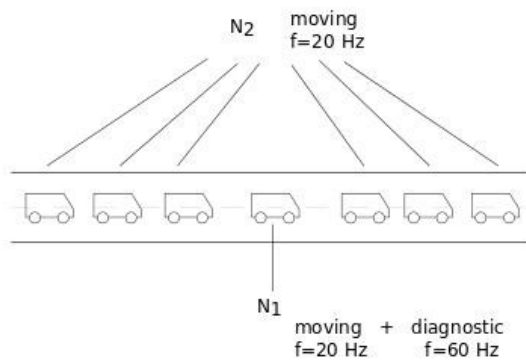


Figure 6: Example of Multi-Agent Multi-Level modelling with Different Temporalities

The preceding figure permits us to mention different constraints which fix the life cycle term of agents in a same level. Here we speak about the frequency of a level expressed in Hertz. This indicates how many times a second, it's necessary to execute the updating processes of the dynamic state of a level. Let's imagine that all functions of an agent possess a minimal frequency beyond which the simulation of this function is no more realistic. If a level permits to his agent to dispose of functions with different frequencies, it adopts the higher one, to keep a correct simulation of the functions with this frequency.

That's why in the example of the figure 6 the frequency of the level $N1$ is equal to 60 Hz because the diagnostic function of the modeled vehicles needs this minimal frequency.

The other constraint comes from the interactions between levels. If we continue with the preceding example, let's say that the level $N1$ needs a minimal frequency equal to 20 Hz, the modeler could allocate this frequency to $N2$. However if the $N1$ level is influenced by a $N2$ and has to calculate the reaction induced by these influences at a frequency higher to 20 Hz (logically less or equal to 60 Hz) it is possibly necessary to allocate a higher frequency to $N2$. So it is necessary to dynamically modify the frequency of a

level N and adapt it to the changing needs (for example, when a level with a higher frequency receives influences from N is created) of the simulation and give back to his level his minimal frequency, defined during the implementation phase, when no needs are expressed.

5. CONCLUSION

In this article we presented a methodology permitting to give a bigger representation power to a multi-levels MAM representing a big number of agents which maintain complex multi-levels interactions.

It's necessary to start by decomposing the agents at a same given scale and for each obtained representation we allocate to it at a different level. Thus in function of needs a same entity passes from a level to another one. After that, we define the elements which can be simplified and supported by organizations and conversely which elements of the organization can be supported by its members. At last, it's convenient to fix the minimal frequency for each level, defined before, specifying which frequency have to adopt a level when it interacts with another one according to the interactions between levels.

However in this article we do not describe in detail the mechanisms to approximate, aggregate or stock outside of the simulation data, at some times of the execution. Also we do not evoke the problem which can be the loss of significant information related to these mechanisms, neither get round them. To test the global coherence of our simulation we should measure the consistency of it, as expressed in (Davis et al. 1993).

Our future work will consist in creating a framework adapted to IRM4MLS, with a strong formalism, which permits dynamic changes of level of detail during the simulation.

REFERENCES

- Angelo, L., 2010. Multi-scale methods and complex processes: A survey and look ahead. *Computer and Chemical Engineering* 34 : 1467-1476.
- Calvez, J.-P., 1990. Spécification et conception des systèmes : une méthodologie. Editions Masson, ISBN 2-225-82107-0.
- Davis, P. and Hilestad, R., 1993. Families of Models that Cross Levels of Resolution : Issues for Design, Calibration and Management. *In proceeding of the 25th Winter Simulation Conference (WSC'93)*.
- Drogoul, A, Vanbergue, D, Meurisse, T, 2002. Multi-Agent Based Simulation: Where are the Agents. *Proceedings of MABS*.
- Ferber, J, 1995. *Les Systèmes Multi-Agents: Vers une Intelligence Collective*. InterEditions.
- Ferber, J, Müller, J-P., 1996. Influences and reaction: a model of situated multiagent systems. *In 2nd International Conference on Multi-agent systems (ICMAS-96)*, pages 72–79.

- Gil Quijano, J, Hutzler, G, Louail, T, 2009. De la cellule biologique à la cellule urbaine : retour sur trois expériences de modélisation multi-échelles à base d'agents. In *Actes des 17èmes Journées Francophones sur les Systèmes Multi-Agents (JFSMA '09)*.
- Michel, F, Gouaïch, A, Ferber, J, 2003. Weak interaction and strong interaction in agent based simulations. *Lecture Notes in Computer Science*, 2927 :43-56,.
- Michel, F, 2007. Le modèle irm4s. de l'utilisation des notions d'influence et de réaction pour la simulation de systèmes multiagents. *Revue d'Intelligence Artificielle*, 21 : 757-779.
- Morvan, G, Veremme, A, Dupont, D, Jolly, D, Charabidze, D, 2008. Vers une modélisation multi-niveaux. *Acte de la 7ème conférence de Modélisation et Simulation MOSIM*, Paris, France, 1, 167-174.
- Morvan, G, Veremme, A, Dupont, D, Jolly, D, 2009. Modélisation et conception multiniveau de systèmes complexes : stratégie d'agentification des organisations. *Journal Européen des Systèmes Automatisés*, 43 :381-406, 2009.
- Morvan, G, Veremme, A, Dupont, D, 2010. Irm4mls: the influence reaction model for multi-level simulation. In *11th International Workshop on Multi-Agent-Based Simulation*.
- Navarro, L., Flacher, F. and Corruble, V., 2011. Dynamic Level of Detail for Large Scale Agent-Based Urban Simulations. *Proc. of the 10th Int. Conf. On Autonomous Agents and Multiagent Systems (AAMAS 2011)*, 701-708, May, 2-6, 2011, Taipei, Taiwan.
- Servat, D, Pierrer, E, Treuil, J-P, Drogoul, A, 1998. Towards Virtual Experiment Laboratories: How Multi-Agent Simulations Can Cope with Multiple Scales of Analysis and Viewpoints, pages 205-217. *Springer-Verlag Berlin Heidelberg*.
- Pumain, D, Sanders, L, Bretagnolle, A, Glisse, B, Mathian, H, 2009. West The Future of Urban Systems: Exploratory Models. In: Lane, D, Pumain, D, Van der Leeuw, S, West, G, eds. *Complexity Perspectives in Innovation and Social Change*. Netherlands : Springer, 331-360.
- Soyez, J-B, Morvan, G, Merzouki, R, Dupont, D, Kubiak, P, 2011. Modélisation et simulation multi-agents multi-niveaux. *Studia Informatica Universalis*.
- Van Brussel, H , Wyns, J, Valckenaers, P, Bongaerts, L, Peeters, P, 1998. Reference architecture for holonic manufacturing systems: PROSA *Computers in Industry* 37, 255-274.
- Weyns, D, Holvoet, T, 2003. Multi-Agent System Technologies, chapter Model for Simultaneous Actions in Situated Multi-agent Systems. Number 2831 in: *Lecture Notes in Artificial Intelligence*. Berlin Heidelberg: Springer-Verlag, 105-118.
- Zeigler, BP, Kim, TG, Praehofer, H, 2000. Theory of Modeling and Simulation. *Academic Press*, 2nd edition.

INDOOR PEDESTRIAN NAVIGATION SIMULATION VIA A NETWORK FRAMEWORK

John M. Usher^(a) and Eric Kolstad^(b)

^(a,b)Dept. of Industrial & Systems Engineering, Mississippi State University

^(a)usher@ise.msstate.edu, ^(b)ewk6@msstate.edu

ABSTRACT

The Intermodal Simulator for the Analysis of Pedestrian Traffic (ISAPT) is being developed for the purpose of modeling pedestrian traffic within intermodal facilities, such that designers may evaluate the impact of building design on the Level of Service provided. The navigational and decision-making behaviors that influence a pedestrian's travel rely heavily on the functional attributes defined within the facility model. An associated network delineates the principal path structure between nodes that represent waypoints and locations of pedestrian resources such as ticket counters, food service vendors, and restrooms. This paper describes the network framework employed by ISAPT and illustrates the necessity of key features it affords to the simulation. Several examples are given that demonstrate various applications of the capabilities provided by the framework in the context of intermodal facilities.

Keywords: pedestrian traffic simulation, micro-simulation, behavioral modeling.

1. INTRODUCTION

In the process of facility engineering and construction, designers must evaluate overall building design considerations in the context of their effective impact on Level of Service and other factors. The Intermodal Simulator for the Analysis of Pedestrian Traffic (ISAPT) system models each pedestrian's behavior individually within the context of a facility defined by its architecture and available resources that offer services to the travelers within it. Using agents to represent the individual pedestrians, the collective behavior of the crowd emerges from the interactions between them. Pedestrian navigation choices are simulated using a probabilistic approach that takes into account those factors considered by humans as they move about their environment.

The overall trip of each pedestrian is governed by an agenda that is generated when they enter the system. This agenda defines the tasks they would like to accomplish and requires real-time decision-making based on the environmental conditions they encounter within the facility. The navigational and decision-making behaviors that influence the pedestrian's travel rely heavily on the functional attributes defined within

the facility model. A network of nodes delineates the principal path structure with the nodes representing waypoints and locations of resources (services) such as ticket counters, food service vendors, and restrooms.

This paper describes the network framework employed by ISAPT and illustrates the necessity of key features it affords to the simulation. Several examples are provided that demonstrate various applications of the capabilities provided by the framework in the context of intermodal facilities.

2. BACKGROUND

Capturing realistic pedestrian behavior in simulation is useful for evaluation and planning in building design (Daamen, Bovy, and Hoogendoorn 2001), urban design (Jiang 1999), design of the area around an outside memorial (Monteleone et al. 2008), land use (Parker et al. 2003), marketing (Borgers and Timmermans 1986), facility operational assessment (Daamen, Hoogendoorn, and Campanella 2009), city wide regional planning (Raney et al. 2002), and evacuation evaluation (Sagun, Bouchlaghem, and Anumba 2011).

Popular schemes for simulating pedestrian crowds include cellular automata, social forces and rule-based systems – each of which has certain tradeoffs (Pelechano, Allbeck, and Badler 2007). The first of these relies on a grid cell-based division of space for pedestrian travel, occupancy and consideration of movement alternatives (Kirchner et al. 2003). Using a force strategy attempts to distribute motion via physically motivating forces such as forward movement, steering and avoidance of obstacles.

Rule-based systems enact logical choices when a certain set of conditions or overall criteria have been met. System models are somewhat divided as to whether choices are examined in terms of a continuous space of motion alternatives, as with social forces systems modeled as repulsive forces (Helbing, Farkas, and Vicsek 2000; Williams and Huang 2006), or these are focused on several discrete combinations of base navigation factors during a given time step – for example, discrete spatial regions in cellular automata (Kirchner et al. 2003; Blue and Adler 2001) or representative combinations of direction and speed (Bierlaire, Antonini, and Weber 2003; Antonini, Bierlaire, and Weber 2006). When simulated movement choices are not viable or have already resulted in an

undesirable state, however, it is up to the individual system to allow for braking force, stop-and-wait conditions, and other actions to enable individual pedestrians (and the system overall) to re-adjust, for example, after one or more collisions have been encountered.

3. SYSTEM DESCRIPTION

The ISAPT simulation system is designed to simulate pedestrian traffic within intermodal facilities (e.g., airports, train stations, etc.). Based on the overall travel schedule for the facility, the system generates a dynamic pedestrian population representative of those arriving, departing, and connecting transit points. Since pedestrian traffic in a facility is not limited to travelers only, ISAPT also provides the capability to simulate what is termed “non-travelers” representing persons that enter the facility for the express purpose of picking up or dropping off travelers. At this time, the system does not consider the contribution of employee traffic within the facility.

ISAPT is implemented as a 3D OpenGL-based application written in the C++ programming language with an objective of supporting cross-platform use. The system simulates the behavior of each individual pedestrian, employing probabilistic navigation at the local level and route based planning at the strategic level. Each pedestrian moves in continuous 3D space, planning their trip inside the facility determined by an agenda defining a list of activities they intend to complete during their visit. An example activity set would include check-in at the ticket counter, checking of bags, passing through security, obtaining food, a visit to the restroom, and waiting in the gate area prior to the boarding call. Agendas are defined for each individual within the facility, with the system simulating and collecting data on the traffic flow and resource utilizations that arise as the result of the pedestrians executing their agendas. For additional details, the reader is referred to associated publications (Usher and Strawderman 2010; Usher, Kolstad, and Liu, 2010).

4. SYSTEM FRAMEWORK

As opposed to grid-based approaches, pedestrians in ISAPT are able to move freely in any direction within the navigable space of the facility model. A structured network of nodes within the facility provides a navigational framework defining generalized paths interconnecting the available resources of the facility. The resulting network provides a basis for performing route based planning to decide the overall path a pedestrian will follow in order to perform the tasks on their agenda. Pedestrians are not bound to the paths initially selected, but instead are permitted to re-assess their activity schedule based on the conditions they encounter as they proceed with their tasks in the facility. Each pedestrian must thus possess a certain awareness of the potential paths defined by the network, along with current conditions on that path and resources

within visual range (e.g., congestion, queue lengths, etc.). It is the job of the system framework to provide these capabilities.

Conceptually, network nodes within the system can be divided into three categories: (1) those that actively represent physical entities within the system, (2) those primarily used to define path connectivity, and (3) those that provide needed support for the simulation itself (e.g., data collection). A general node object acts as a base prototype to represent all these types of nodes and its respective attribute values define its specific purpose, capabilities, and influence on pedestrian behavior. A node is not limited to a particular function, but can provide a flexible range of components to serve multiple purposes. For example, a user can define a single node to represent a physical queue that is part of a pedestrian’s path and collects data on its operation.

The sections that follow discuss this framework along with the nodes and attributes that define the network that provides these capabilities.

4.1. Physical Entities

As indicated earlier, a pedestrian’s trip within a facility is determined by their current agenda, which results in a list of activities they intend to complete during their visit to the facility. Each activity has one or more corresponding resource locations where it may be performed (e.g., ticket counters, food service areas, boarding gates, retail stores, restrooms, etc.). These resources are represented by *resource nodes* that act as primary objects on the underlying network used as waypoints along potential routes planned by pedestrians. Figure 1 shows a queue formation in front of a ticket counter, involving two node-based structures common to many simulations. The queue itself is comprised of a series of directionally linked nodes along a navigation route. Each ticketing location is represented by a resource node requiring a variable service time to complete.

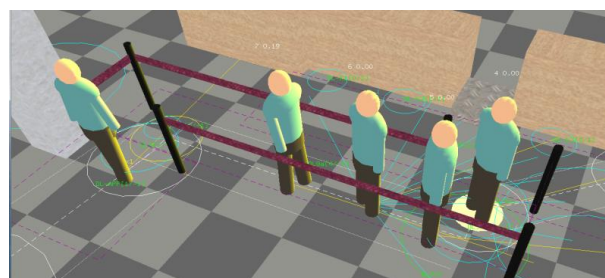


Figure 1: Simulated Pedestrians Arriving at Queue, in Front of Ticket Counters

4.2. Path Connectivity

When a pedestrian first enters the system, they are assigned a planned initial route from their origin to final destination (e.g., the boarding gate for their departing flight). This route is defined by a set of nodes they intend to visit from start to finish. An extensive set of intermediary nodes connects the space between resource nodes, acting as a navigation framework that provides a

structural description of the building's thoroughfares to enable pedestrian route-based navigation. Consider the simple layout shown in Figure 2 defining the location of several resources (ticket counter locations K1-K2 and C1-C2). If a pedestrian were approaching at the lower left with intent to visit the ticket counter area, they would typically be observed to follow an indirect path (such as the curving dashed line) that is somewhat impeded due to obstacles along the way. Without more detailed connectivity information, a navigational decision might suggest a planned travel path (shown as a straight dashed line) that does not realistically represent typical behavior for an actual pedestrian. *Intermediate navigation nodes* are thus arranged within the network as needed to increase the spatial resolution of successive waypoints located between resource locations. Figure 2 shows several dotted-outline nodes added to the graph for secondary navigation detail.

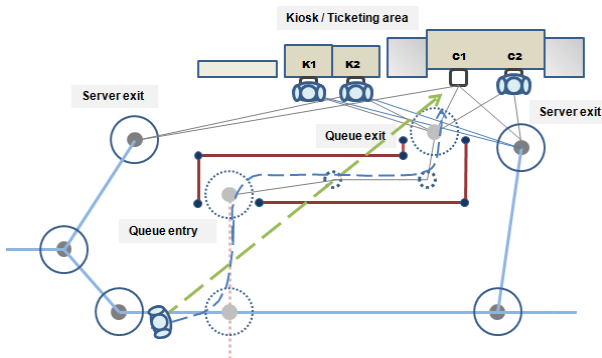


Figure 2: Nodes of an Airport Ticket Counter Area

As can be seen in Figure 2, it is possible that multiple connecting links (pathways) will be associated with a given node. We refer to these nodes as *decision nodes*, as they will trigger pedestrian decision logic in association with path-based routing through that node. A decision node in effect is an intermediate navigation node with two or more alternative links that may be considered for outbound traffic. The routing logic will assess which path is best utilized at the point when such a node is approached. This becomes necessary in situations where:

- Pedestrians have the choice of several branching routes through (sometimes crowded) areas,
- A pedestrian approaches an area with multiple queue lines for a resource (i.e., alternative vendors in a food court), and
- Exit paths lead to variously distanced resources with different priorities.

In the first situation, the pedestrian will evaluate the congestion and approximate travel time for each region ahead and make an informed path decision. Where multiple queues are observed, they will estimate the waiting times for the available options and select one they believe will provide the best service. Similarly, the choice amongst multiple paths and available resources will be based on priority and time required to obtain

them. Estimates made by the pedestrians in each of these cases may take into account region-based data maintained by associated *data collection nodes* (discussed below).

4.3. Simulation Support

At this point, we have additional nodes that provide simulation support which are used for data collection and compilation of statistics. These nodes may appear within the system either as stand-alone nodes (i.e. not connected as part of the navigation structure) that define regions such as a corridor area of interest for traffic analysis, or as part of the overall node network.

The data collection region can be defined as either a radial extent or a rectangular zone of a certain width and length. The region encompassed by any one node is permitted to overlap with other nodes, and can automatically collect data on all behavior within its region without interference.

A representative list of the types of data that can be requested to be logged for each pedestrian is given in Table 1. Some of this data is available to pedestrians for decision-making purposes, e.g. in the form of an observed queue line length and estimated wait time. Any pedestrian estimates, however, are subject to variability in the form of added noise representing a degree of uncertainty.

Table 1: Information Logged in Association with Data Collection Nodes

- | |
|--|
| <ul style="list-style-type: none"> • Tacked location over time • Time spent in regions • Pedestrian speed • Distance traveled • List of node visited (travel history) • Travel time between nodes • Wait time in queue(s) • Service time • Levels of needs (e.g., hunger, thirst, etc.) • Decisions made |
|--|

4.4. Node Attributes

The attributes of a node define its features, parameter values, behavioral effect, and limitations for a given region. To enable navigational or other decision-making use, node attributes can also be marked to indicate an association with a larger structure.

The common attributes of a node are its *name*, *location*, and *connectivity*. The name of a node provides a unique identifier used for general reference, while its location provides a relative locale within the 3D geometry representing the facility. A node's connectivity describes its relation to other nodes in the navigation network via incoming and outgoing links.

While a node's general function is to support path-based routing and navigation behavior, more specialized attributes are available to help represent varied service-related resources found within intermodal centers. These act together to simulate larger system processes

involved with services such as ticket counters, security gates, food courts, restrooms and so on.

The principal functional node attributes consist of:

- Active region extent (rectangular or circular)
- Navigable zones for nodes and pathways
- Conditional entry requirements
- Resource type notation (used primarily in decision logic)
- Behavioral influence factors
- Data collection within an associated region

An overview of the use of these attributes and their influence on pedestrian behavior and general system operation is discussed below.

4.4.1. Active Region and Navigable Zones

Each node in the system occupies a define space, which pedestrians must cross into for purposes of navigation and/or resource use. With an active route plan in mind, pedestrians will follow their path traveling from one node to the next. As they do, their local movements are guided primarily by collision avoidance as they progress towards their next objective (target node). The actual path traveled by the pedestrian is also a function of the presence of other pedestrians in proximity. If each pedestrian were asked to follow such a path exactly, this would result in unnatural behavior and traffic congestion along the route, with further convergence towards waypoints (intermediate nodes) that represent the navigational targets, as shown in Figure 3a.

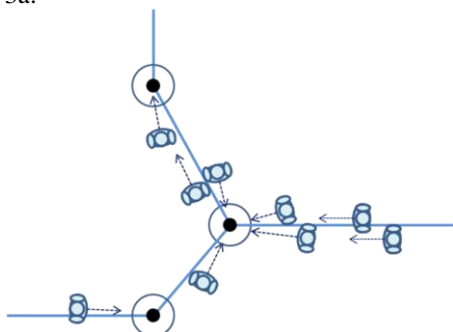


Figure 3a: Pedestrians Following in Proximity to Network Structure (tendency to converge at nodes)

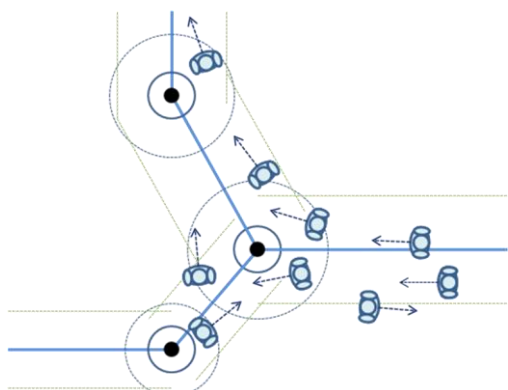


Figure 3b: Network Structure with Navigational Extent Region Hints Added for Nodes and Connecting Paths

To overcome this problem, the system permits the user to define greater navigable extents along paths and within the vicinity of node regions. A connected route may be considered to have a collective region of space encompassing its nodes and pathways that represents an area of more likely travel under certain constraints. Pedestrians will generally attempt to stay within a reasonable *proximity* of the path segments between and regions surrounding waypoints. Figure 3b's dashed *navigational zones* (above) can be interpreted as regions of relatively preferred movement.

When a pedestrian moves within the *active range* of an upcoming target node (i.e. shown in the figure as a larger radius around the node), they are considered to have effectively reached that node for navigation purposes en-route. At this point, the next node on the pedestrian's path becomes its waypoint and travel continues towards the next location on the current route, or into the node's *physical limits* if it is an intended resource they plan to utilize. Similarly, there exists a buffer zone of space around inter-connecting paths (specific to certain node connections) where travel en-route is relatively more preferred, although not strictly confining. The result is a more natural representation of traffic through the facility (Figure 3b, above) as pedestrians are encouraged to go with the overall flow and direction of the network structure, and may take shortcuts past recognized waypoints. The factors that encourage this behavior are adjustable for individual nodes and connections (e.g. for a single-occupancy resource vs. a wide-open area, or a corridor junction with a diverse range of traffic relative to a queue line) and act to direct pedestrian focus when approaching resources.

4.4.2. Conditional Entry

These properties permit the user to limit entry into a region defined by a node based on the region's capacity, conditions, and/or specific requirements of the pedestrian expressed in terms of their attributes. Being able to limit access based on the current occupancy is important in situations such as forcing a pedestrian to wait in line before approaching a ticket counter until the area in front of the agent is open.

At times a pedestrian's attributes will influence their navigation decisions and access to resources. A passenger may, for example, need to have presented an ID and acquired a ticket (ahead of time or at a kiosk/counter) prior to receiving their boarding pass, which in turn is necessary to proceed through security and enter the terminal areas. It is also possible to require that node attributes match a certain value or range to be valid, and for overlapping regions to impose combined limits.

4.4.3. Resource Type

In today's modern intermodal facilities, there are many resources available to the traveler that provide a wide variety of services. In some instances a pedestrian must make decisions based on information related to the type

of resource and the current conditions of their environment. To enable the decision logic to ascertain the function of a larger structure, or mark parts of it that are relevant for a particular use, attributes defining the resource type and its service parameters can be assigned. For example, marking a node as the entrance to a queue would indicate its connection with upcoming servers to the route decision logic for use in resource acquisition, and provide key details when a pedestrian wants to choose which queue to enter based on observed statistics.

4.4.4. Behavioral Influence (Region of Effect)

These attributes act to modify pedestrian behavior by adjusting the relative influence of certain navigational factors, decision-making factors, and/or by altering the logic utilized. For example, when a pedestrian enters a queue, given the need to form a line and stand next to others, they will effectively reduce the size of their personal space. Their criteria for obstacle avoidance also changes in that they are more willing to stand right next to (or even touch) an obstacle such as a wall, column, or turnstile. The distance they look ahead to avoid obstacles is similarly reduced. To accommodate the capability to alter pedestrian behavior, each node can specify a set of factor adjustments that are active while a pedestrian is present within that node's active region, thereby altering one or more of the area-based influences and attenuating the pedestrian's characteristic response. Once the pedestrian leaves the node, the pedestrian's factors return to their original settings. Another example would be a region with a distracting element (i.e., window display, food vendor, etc.) that might influence passing pedestrians by causing them to slow down and/or adjust navigation behavior. In addition, a turn-taking behavior could be enforced to prevent a traffic jam in an area with limited space.

4.4.5. Data collection

A designated node may likewise be tasked to actively record data, based on pedestrians that pass within its node boundaries. The node may either be linked within the navigation network or exist as a stand-alone node.

Given that the overall attributes of a node define what it can do, it is viable to have a single node that performs multiple functions. For example, facility resources are often modeled using nodes that represent physical entities that provide services, but these nodes likewise place conditions on who can use them while influencing the behavior of those present, and collecting data on their usage patterns. This effectively allows ISAPT to:

- Enforce occupancy requirements to facilitate arrival/departure and parking locations (for seating, etc.),
- Influence navigational decisions within node regions and connected routes,

- Enable pedestrians to obtain resource benefits according to a user-specified service time distribution,
- Modify decision factor influence and/or behavior within proximity to a node,
- Provide the planning module with current data updates for dynamic route planning and resource use,
- Gather region-based statistics,
- Represent the dynamic input sources for arriving populations, and
- Modify the navigation network to include additional structures e.g. for temporary use (such as in dynamic queue formation)

5. NAVIGATIONAL DECISION-MAKING

Given the availability of a navigation network of nodes with such attributes, the following traversal criteria (relative to current node status) may be actively considered by the pedestrian when determining their an ongoing course of action:

- When the proximate node is a point with multiple decision choices, what are the currently navigable nodes connected to resources of interest?
- If the next planned node acts as a server, is its region clear and available for use?
- When an upcoming node is a queue entry node, which potential resources are available at the end?
- When the next node is a queue entrance, how many people are now in the queue (i.e. to estimate wait time)?
- Which related resource nodes should be obtained [in sequence] to properly satisfy a resource need (e.g., selection of ticketing Kiosks, Counters, Kiosk-to-Counter-to-TSA, Counter-to-TSA, etc.)?
- Is it viable to traverse a node region when it is conditional use zone (permitting occupancy only under certain conditions, such as capacity limits, pedestrian attribute, etc.) or while it may have a region of effect (whereby certain travel factors are affected)?

Given these capabilities, the ISAPT system is able to support the simulation of a wide variety of services offered within intermodal facilities along with the pedestrian traffic that results.

6. IMPLEMENTATION EXAMPLES

This section provides several practical examples of how the nodes may be collectively assembled to define areas of functionality common to intermodal facilities.

6.1. Ticket Counter

A simple ticket counter, as seen in Figures 1 and 2, can be constructed as follows. An initial node at the

entrance of a queue (marked as a “queue entry” node, with a capacity of 1) is associated with a region-based statistics node for collecting data on the queue for later analysis as well as real-time use for pedestrian decision making (e.g., line length and estimated wait time). The queue entry node connects a sequence of nodes within the line barriers to an exit node that is linked to one or more ticket resource server nodes. A free-form extendable line might likewise be configured for the entrance node to handle additional visitors beyond a strictly enclosed path. A resource node is used to represent each service point at the counter area (kiosk terminal, ticket agent, baggage scanning). The attributes of each resource node define their respective characteristics such as capacity, service time distribution, task handling ability and resource type. Real-world scenarios such as shown in Figure 1 may require a node region space in front of the counter to enforce conditional entry zone, enacting a *waiting* behavior that prevents people in line from traversing the limited corridor space until a server (resource node) becomes available. These nodes may also be linked to other resource nodes nearby (e.g. if additional processing is required for check-in following kiosk use, or if luggage must be scanned at another station) along with navigation nodes defining potential points for departure.

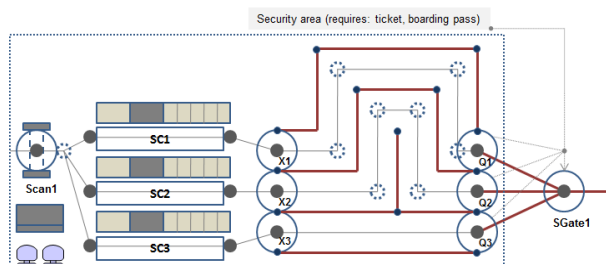


Figure 4: Security Processing Area

6.2. Security Gate

The security gate area in Figure 4 depicts a structure with several queues. A node defining the security area’s surrounding region (dotted line) will not permit pedestrian entry without a valid ticket and boarding pass. The main entry node, SGate1, is connected to several queue lines. The three nodes at the entrances to the queues (Q1-Q3) provide detail on their current occupancy and maximum capacity for decision-making purposes. Q3 might have an added conditional entry requirement, e.g. for flight crew status or passengers needing assistance in processing. At the end of each line, a conditional entry node (SC1-SC3) is used to define a capacity limited service area where persons empty their belongings into bins and then wait for their turn to pass through a common node (Scan1) of capacity one that represents a personal screening device.

6.3. Food Court

A food court might be comprised of multiple queue lines attached to food and drink resources, which have exit points linked back to the larger court area. The

court area in turn may have individual seating resources (e.g. tables) each with several available positions that a pedestrian may occupy for a time (as required to eat), and an overall region-based occupancy may be enforced by a capacity-limited node surrounding the court. While individual seating occupancy depends on a given pedestrian’s requirements, the seats at the same table (when occupied by a group) may further enact an overall waiting behavior or preferred seat selection.

6.4. Restroom

Larger enclosed-spaced resources such as restrooms – where concern is focused primarily on capacity rather than individual dynamics – would typically require a single resource node associated with the entrance (enforcing some capacity limit for total occupancy). This node would often be tagged such that pedestrians entering do not remain within the node’s physical extent, but rather “park” outside of the observable simulation space while service is underway. To a simulation observer, the pedestrian avatar disappears as they pass through the door and then later reappears as they exit after completing service. A linked exit node – slightly offset from the entrance location to encourage flow of traffic due to simulation limits with narrow openings – would thereafter return the individual pedestrians back to the system.

Such a configuration enables the system to simulate services that do not require physical modeling or graphical representation of the service being performed, i.e., that are able to be represented as a self-contained “black box”. A similar setup might be used with stores along a terminal where the customer will enter or exit through a doorway.

6.5. Dynamic Queues

The waiting behavior at the end of demarked queue-like zones (such as cordoned lines leading up to a ticket counter) or near certain resources that have particular demand (e.g., a water fountain), may not be sufficiently represented with a single node region (of limited or unlimited capacity) or a simple group of nodes. While there are many cases when a small crowd will gather around a resource waiting for it to become available, it is common that a temporary line formation of sorts will occur. In our observation and more commonly, these lines often take two basic forms: the *free-form* style line that arcs away from the resource in one direction or another (as in Figure 5), and a more linearly *stacked* variety where the line begins to form side-to-side and outward as space permits. ISAPT allows either type of line to form starting at an arbitrary node configured as its root, with control over the directional parameters, spatial extent and permitted line length. A series of temporary nodes dynamically added to the graph structure act to enforce regular directed line formation along with their own specific evolving directional flow. Node connectivity changes are communicated to pedestrians en-route for ongoing route adjustment and re-targeting, as needed.

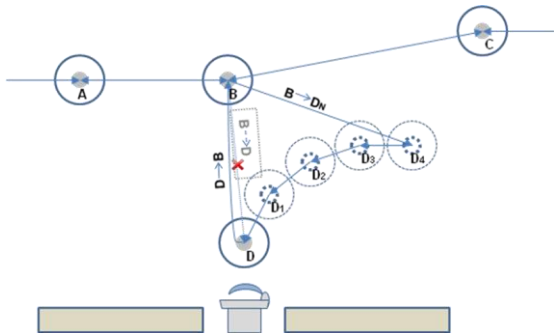


Figure 5: Dynamic Queue Formation

7. SUMMARY

ISAPT provides the means to enact a crowd-based simulated environment by modeling agent-based behavior via individuals' reactions to their environment, based on observed real-world behavior. Data obtained in both experimental trials and on-site studies of real-world facilities have been incorporated in the modeling process – including route-based decision logic. The location and availability of resources in a structured node network are central to how pedestrians interact with the larger system, in obtaining their planned objectives and choosing a preferred course of action from moment to moment. The flexible array of node-based attributes in the ISAPT system is designed to enable a broad range of possible resource configurations and behaviors, without imposing substantial limits on design. Apt resources matching real-world structures and facility-specific data may be constructed with a model-appropriate set of attributes for varied scenarios and activity levels. This gives rise to a modeling platform that is highly adaptable and able to easily represent a wide range of intermodal facilities.

We are presently involved in site studies and modeling of several transportation facilities and experimental scenarios, focused on inter-modal facilities and pedestrian traffic through common areas. The network-based attributes that have been discussed here are subject to further expansion, as it becomes apparent that certain modeling tasks and varied facility types benefit from added functionality. Current and future applications will utilize these to define their own unique system of interconnected resource nodes to model additional facilities.

ACKNOWLEDGMENTS

This project is sponsored by the U.S. Department of Transportation (Grant No. DTOS59-06-G-00041).

REFERENCES

- Antonini, G., Bierlaire, M., and Weber, M., 2006. Discrete choice models of pedestrian walking behavior. *Transportation Research Part B: Methodological*, 40(8), 667-687.
- Bierlaire, M., Antonini, G. and Weber, M. 2003. Behavioral dynamics for pedestrians. In *Moving through nets: the physical and social dimensions of travel*, Elsevier.

- Blue, V.J., and Adler, J.L., 2001. Cellular automata microsimulation for modeling bidirectional pedestrian walkways. *Transportation Research Part B: Methodological*, 35(3), 293-312.
- Borgers, A., and Timmermans, H., 1986. A model of pedestrian route choice and demand for retail facilities within inner-city shopping areas, *Geographical Analysis*, 18 (2) 115-128.
- Daamen, W., Bovy, P.H.L and Hoogendoorn, S.P., 2001. Modelling Pedestrians in Transfer Stations. In Schreckenberg, M., and Sharma, S.D. eds., *Pedestrian and Evacuation Dynamics*, Duisburg: Springer Verlag, 59-74.
- Daamen, W., Hoogendoorn, S.P. and Campanella, M.C., 2009. Ticket reservation posts on train platforms: an assessment using the microscopic pedestrian simulation tool Nomad. *Proc. of Transportation Research Board*, Jan. 11-14, Washington, D.C., paper no. 09-1853 CD-ROM.
- Helbing, D., Farkas, I., and Vicsek, T., 2000. Simulating Dynamical Features of Escape Panic. *Nature*, 407, 487-490.
- Kirchner, A., Namazi, A., Nishinari, K. and Schadschneider, A., 2003. Role of Conflicts in the Floor Field Cellular Automaton Model for Pedestrian Dynamics. *2nd International Conference on Pedestrians and Evacuation Dynamics*, 51-62.
- Monteleone, M. F., Veraart, N., Chough, S., and Aviles, W., 2008. Pedestrian Simulation Modeling Study for World Trade Center Memorial. *Transportation Research Record: Journal of the Transportation Research Board*, 2073, 49-57.
- Parker, D.C., Manson, S.M., Janssen, M.A., Hoffmann M.J., and Deadman, P., 2003. Multi-agent systems for the simulation of land-use and land-cover change: a review. *Annals of the Association of American Geographers*, 93(2), 314-337.
- Pelechano, N., Allbeck, J.M. and Badler, N.I., 2007. Controlling Individual Agents in High-Density Crowded Environments. *Eurographics/ACM SIGGRAPH Symposium on Computer Animation*.
- Raney, B., Cetin, N., Vollmy, A., and Nagel, K., 2002. Large Scale Multi-agent Transportation Simulations, *Computer Physics Communications*, 147, 559-564.
- Sagun, A., Bouchlaghem, D., and Anumba, C., 2011. Computer simulations vs. Building guidance to enhance evacuation performance of buildings during emergency events, *Simulation Modelling Practice and Theory*, 19, 1007-1019.
- Usher, J.M., and Strawderman, L., 2010. Simulating Operational Behaviors of Pedestrian Navigation. *Computers and Industrial Engineering*, 59, 736-747.
- Usher, J.M., Kolstad, E., and Liu, X., 2010. Simulation of Pedestrian Behavior in Intermodal Facilities. *International Journal of Agent Technologies and Systems*, 2(3), 66-82.

RECONFIGURABLE HUMAN-SYSTEM COSIMULATION

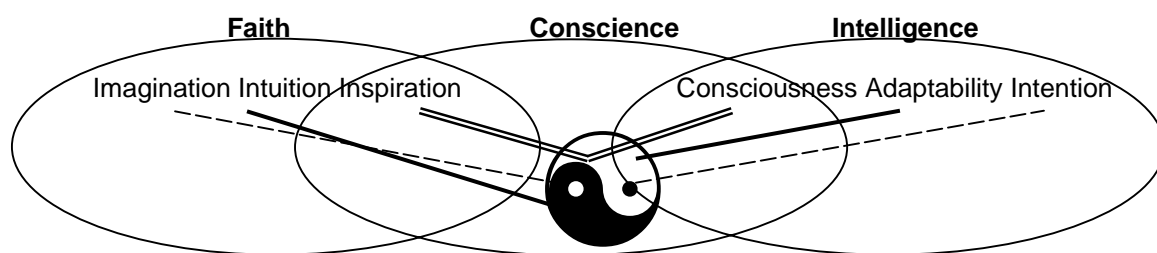
Tudor Niculiu^(a), Maria Niculiu^(b)

^(a)University Politehnica București, Faculty of Electronics, Telecommunications, and IT

^(b)University București, Faculty of Foreign Languages and Literatures

tudor-razvan@ieee.org

mariaofficialfac@yahoo.com



ABSTRACT

Abstraction power is the crucial difference between human and any other natural being. *Divide et Impera et Intelligere* applies the hierarchically expressed abstraction. Abstraction can simplify or reflect. Religion had to learn us about God's existence in our being. Philosophy has to learn us about essence, existence, and being. Conscience represents the essence of our existence as being; therefore it tells us that God is in ourselves, for ourselves, and among ourselves. Further, we have to be, in order to search for our essence researching our existence. We try to model the Conscience to simulate the Intelligence, reaching for the reconfigurable human-system cosimulation. The alliance between arts, sciences is vital and demonstrates the insolvability of the nowadays Spirit-Matter dichotomy, and of all secondary dichotomies, actually functionally generated by the Space-Time dichotomy that is necessary to the human *evolution*.

Keywords: Faith, Intelligence, Conscience, Abstraction

1. INTRODUCTION

The Faith experiment took place in the Middle Age by spiritual and chivalrous search, mediated by Masonic buildings. The Cathedrals were the symbol of the coming *revolutions* that intended to institute the *intelligent Faith* as basis of the human society. The concentration of the mind on the reasonable control of the Adaptability followed the spiritual revolution that tried to bring into individual and social conscience the human choice for evolution without disregarding the Eternity or the knowledge of the Way.

Reconfiguration continues the ideas of hard-soft cosimulation, intending to extend the soft flexibility to hard, as parallel soft gets closer to hard performance.

Experimented ways to reconfigurable design are Field-Programmable Gate Arrays for circuits (Miller 1993) and reconfigurable networks for systems (Rabaey 1997). Our project extends the reconfigurability to the simulation itself.

First, by a self-aware simulation, for that we build a knowledge hierarchy corresponding to the simulation hierarchy, we intend to get a self-control of the simulation process. Then, by expressing both simulation and knowledge hierarchies in the reference system of the basic hierarchy types that correspond to essential views in language/ system theory (Keutzer 2000) derived from the main partition of our real life, we aim to create the context for a self-organization of the simulation.

Reconfigurable computing architectures (Ștefan 2010) complement the existing alternatives of spatial custom hardware and temporal processors, combining increased performance and density over processors, with flexibility in application. If one of the imposed properties is considered as not being fulfilled after applying a technique, using a model and suitable methods for measure and reconfiguration, different strategies permit altering one of the techniques/ models/ methods.

The process repeats for the initial description or the one resulted from prior insufficient improvement. This calls for an intelligent choice of the intelligent system that assists/ automates the reconfiguration. The methods are recursive to handle the different components in the system's description.

Measurement functions (Lupu 2004) control the continuation process of the reconfiguration, suggesting bringing reconfiguration in the context of software and hardware, as the strategies can be expressed object-oriented/ categorical and understood mathematically.


```

class Reconfiguration ...
Description reconf (Description descr,
                    Bool increment, Bool integrated)
{ techs :=  $\emptyset$ ; models :=  $\emptyset$ ; meths :=  $\emptyset$ ; good := false;
while (not good) {
    tech := selTech (descr, techs, models, meths);
    if (not tech in techs) {techs.add (tech); models:= $\emptyset$ };
    if (not model in models) {models.add (model);
                                methods :=  $\emptyset$  };
    reconf := model.detSpec(Description);
    meth := model.selMeth (reconf, meths)
    if (not meth in meths) meths.add (meth);
    if (integrated) {
        (good, enough) :=meth.measure (reconf);
        while (not enough) {reconf :=
                                improveLoc (reconf);
        (good,enough):=meth.measure(reconf)}}
    else (good, reconf) :=
        improveGlob (reconf, meth.measure(reconf));
    if (increment) descr :=
        model.returnToDescription (reconf) };
return model.returnToDescription (reconf) }.

```

Representation is a 1-to-1 mapping from the universe of systems (objects of simulation) to a hierarchical universe of models - a representation can be inverted. A model must permit knowledge and manipulation, so it has two complementary parts/ views: description and operation.

If models correspond to classes, in a formal approach, specifications are instances; for language-like models, specifications are expressions.

Hierarchy types open the way to simulate intelligence as adaptable consciousness by integrating the system and the metasystem. Hierarchy is the syntax of abstraction.

There are different kinds of abstraction that need different types of hierarchy. Most abstractions are simplifying the approach, what is compulsory for complex object-systems.

Knowledge and construction hierarchies cooperate to integrate design and verification into simulation. Object-oriented concepts are symbolized to handle data and operations formally. Structural representation of behavior manages its realization.

Classes abstract the form, symbols the contents, and partitions simplify the approach. All these enable the simulation hierarchy to assist construction, verification, optimization, and testing, being managed completely by pure reason, by discrete formalisms/ simulations.

The natural limit of complexity is caused by the essentially sequential approach, whereby the real limit of computability results from the discreteness of our reason. Understanding and construction should use correspondent hierarchy types, i.e., a reflexive kind of abstraction has to be expressed by the knowledge hierarchy type.

2. POWER OF ABSTRACTION

Metaphor is a popular instance of abstraction. God is the absolute abstraction. And if we remember that *liberty is understood necessity* (Georg Wilhelm Friedrich Hegel), we could detail the metaphorical thesis:

God is the evolution goal of our faithful intelligence.

We can reduce abstraction to its simplifying types (classes, symbols, modules, construction) hoping to get to the absolute liberty, i.e., considering God, the simplest item of the Reality, totally unconstrained. But we can simulate/ construct/ live/ work associating a knowledge hierarchy to everything we do, aiming to understand constructively the most complex absolute necessity, defining God.

The power of abstraction is human's gift to surpass the natural limits, extending pure reason to real intelligence. As any other dichotomy pair, faith and intelligence can evolve convergent to integration, or can destroy one another if they are not linked together constructively. *Divide et Impera et Intellige* has three parts as *Alle guten Dinge sind drei*. Mathematics develops from three basic structure types, usually integrating them: algebra, order, and topology. We divided our existence in three collaborating parts: arts, sciences, and Engineering, correspondent to our world of beauty-loving ideas, our world of truth-searching efforts, and our presently exaggerated world of good-aiming constructions.

*Einstweilen bis den Bau der Welt
Philosophie zusammenhält,
erhält sich das Getriebe durch
Hunger, Furcht und Liebe.*

Friedrich Schiller

Mathematics (the most accessible art) discovers and studies structure types: (algebra, topology, order), correspondent to (construction, orientation, understanding), and rarely separately used, example of correct and complete integration to be followed by Science and Engineering. *Art is for art*, so it's defining itself, looking for the Beauty. (Hofstadter 1979)

Physics (the paradigmatic science) should integrate its fundamental forces theories, and as chapters, all natural and social sciences, leading them to really apply mathematics. Social sciences study a universe, as complex and nondeterministic as the natural one, so mathematics is at least as important to them as for natural ones, and science would also better inspire mathematics. Science raises the fear and the research inspired by it to more abstract domains, so it is defined hierarchically, as *Fear of God*, looking for the Truth.

Engineering has to be closely related to mathematical approach and integration of parts, not only to mathematical techniques, as to scientific courage and multiple views, not only to scientific results.

As reality contains the abstract ideas, even if physics could explain everything discretely, the power of continuum cannot be forgotten, i.e., analog engineering cannot be neglected in modeling and simulation. (Zeigler 2000)

Paying attention only to the Good in our life is most dangerous, as this part of the Reality is defined by its complement, so it is not better than this, if not closely constrained by Art & Science.

Das schöne wahre Gute

Johann Wolfgang von Goethe

is compulsory while we evolve to *God-alike humans*.

Hierarchy is a network that can represent any mathematical structure type (algebraic, topological, order). Hierarchies are leveled structures, which represent different domains. A level is an autonomous mathematical structure, containing abstract/ concrete entities, linked by level scoped relations.

Abstraction relates the levels: this induces an order relation between levels, partial, concerning entities, and total, regarding the levels. Beyond the hierarchical point of view, the system can be formalized as an autonomous domain, structured by metahierarchical relations, building a level in a higher order hierarchical system.

Hierarchic structures exhibit two complementary processing strategies: top-down and bottom-up. Coexistent interdependent hierarchies structure the universe of models for complex systems, e.g., hardware/software ones. They belong to different hierarchy types, defined by abstraction levels, autonomous modules, classes, symbolization and knowledge abstractions.

Abstraction and hierarchy are semantic and syntactical aspects of a unique fundamental concept, the most powerful tool in systematic knowledge; this concept is a particular form of *Divide et Impera et Intellige*; hierarchy results of formalizing abstraction.

Hierarchies of different types correspond to the kind of abstraction they reflect (\uparrow the abstraction goal):

- Class hierarchy (\uparrow concepts) \leftrightarrow virtual framework to represent any kind of hierarchy, based on form-contents, modularity, inheritance, polymorphism.
- Symbol hierarchy (\uparrow metaphors) \leftrightarrow stepwise formalism for all kind of types, in particular also for hierarchy types.
- Structure hierarchy (\uparrow strategies) \leftrightarrow stepwise managing of all (other hierarchy) types on different levels by recursive autonomous block decomposition,
- Construction hierarchy (\uparrow simulation) \leftrightarrow simulation (design/ verification/ optimization/ testing) framework of autonomous levels for different abstraction grades of description.
- Knowledge hierarchy (\uparrow theories) \leftrightarrow reflexive abstraction, aiming that each level has knowledge of its inferior levels, including itself. This hierarchy type offers a way to model conscience.

The first idea is to (re)consider that reality is more than nature, as the continuum of IR is more powerful than the discrete universe of IN. The second analogy is that integer beauty is not enough to comprehend the Reality. The third argument is that reason is less than our real thoughts, as the cardinal of $|Q$ is \aleph_0 , while cardinal of IR is infinitely superior.

Although $|Q$ is dense in $|R$, so pure reason could converge to reality, the complexity problem limits the computability. (Zhong 2003)

The essential limit of the discrete computability, as of the computable intelligence, results from the self-reference, demanded by the integration of level and metalevel needed for consciousness.

A hierarchical type is necessary to represent conscious knowledge. The classical activities in complex systems simulation, that regard different levels of the construction or knowledge hierarchy, can be expressed symbolically then represented object-oriented and simulated structurally:

- Complex simulation needs consistent combination of mathematical domains and an intelligent compromise between consistence and completeness.
- Intelligence simulation implies a hierarchical approach of different types. Any application of it can be imagined as an educational system to discover models for conscience and understanding.
- Constructive type theory permits formal specification and simulation, generating an object satisfying the specification.

The formalism for hierarchy types is the theory of categories. (Ageron 2001) Even if for the moment other aspects can neither be constructive or intuitive, they should not be neglected.

For example, there are much more real things than those reasonably imagined, although between any two real numbers there is a rational one - not intuitive.

We know that if there is no cardinal between that of the countable sets and that of the continuous ones, then there exists no other logical value than true and false, what simply hurts the human in his love for nuances.

This could be avoided only if we believe - not constructive – that an intermediary level between natural reason and Reality exists, as the wise think there is between humans and God: *angels*-Andrei Pleșu.

*Faith, Intelligence and Conscience are ☯ in our life
Way-Truth-Life*

3. CONSCIENT EVOLUTION

Intelligence = (Consciousness, Adaptability, Intention) and Faith = (Inspiration, Intuition, Imagination) are complementary parts of the human mind, separated by the Conscience = (Consciousness, Inspiration), a non-deterministic interface between the non-conscious faith and the conscious intelligence.

Both intelligent simulation and simulation of intelligence demand transcending the present limits of computability to simulability, by an intensive effort on extensive research to integrate essential mathematical and physical knowledge guided by philosophical goals.

The historical experiment of the pure reason should have ended long time ago. Human thoughts cannot be explained or handled by our adaptability-based reason, even if non-deterministic or parallel. Reason has to extend to intelligence in the context of faith. An obvious way is to integrate consciousness, then intention and imagination to intelligence, then to extend this to inspiration and intuition.

Hierarchy types reveal their comprehensive constructive importance based on structural approach, symbolic meaning, object-oriented representation. The power to abstract is the crucial difference between human and other natural living beings.

1. *Intelligence* and *Faith*, like any dichotomy, can converge to integration or can destroy one another if not associated by *Conscience*
2. *Function* is a transformation that can be mathematically formalized, or physically instantiated as temporal behavior. *Structure* is a set of properties that characterize a mathematical or physical space. The properties can be constant or variable in time, reflecting static or dynamic structures. *Architecture* controls both of them. *Simulation* is the relation between function and structure. Structured set = (Set, structure)
3. *Language/ system* is a generic form of a mathematical/ physical *model*, resulting of an inversion-able simulation object representation
4. *Hierarchy* is a functional/ structural concept that fulfils mathematically/ physically the concept of abstraction. Hierarchy is syntax of abstraction
5. *Abstraction* is a human defining capacity that enables him to think.
6. The simplifying abstraction concentrates on a superior level the information that is considered essential for the current simulation approach. Reducing the informational complexity has in view to clear the operation and to ease its formalism; it can be only quantitative, but also qualitative.
7. The reflexive abstraction, expressed as knowledge hierarchy type, tries to understand itself better at higher levels, by understanding more of the inferior levels
8. *God* is in us - as faith is part of our definition, with us - by the others, and for us – the spiritual evolution that is first conditioned, then assisted, to be followed by the social one
9. Against the danger of dichotomy, we concentrate in 3 different ways on the unique Reality (*Plato*): Art for the art - to look for the essential Way, Science with God's fear - to search for the existential Truth, and Engineering - to understand the Being and to concentrate more on the Spirit in our Life

Formal hierarchical descriptions contribute to a theoretical kernel for self-organizing systems. A way to begin is hierarchical simulation. A way to confirm is the object-oriented reconfigurable simulation.

Essential relations are sketched before searching conscience models enabling intelligent simulation: Conscience is self-awareness of individual faith and intelligence, as well as of the relation to the local context (society) and to the global one (Universe/ Reality).

To appear it needed self-knowledge, what could have resulted from community conscience featured by an eternal human structure, e.g., from the past, shepherds, farmers, sailors, Africans, Amerindians, ... Each individual recognized himself in his cohabitants, being most adaptable and having a lot of intuition.

The common measure evolution implies the construction of correspondingly intelligent agents to manage the lower stages and to concentrate on the higher ones. Industry built the agricultural mechanization, and also the concentration on economics.

Human = human (Humanity);
 $human \in Faith \times Intelligence \rightarrow Faith \times Intelligence$;
Humanity = (humans Set, evolution-oriented Structure).
 $evolution \in Hunger, Fear, Love) \times (Engineering, Science, Art) \rightarrow (Engineering, Science, Art)$
 $Mathematics \subset Art = Human :: beauty-oriented activity (Science, Engineering)$
 $Physics = (natural \cup social) Science = Human :: truth-oriented activity (Art, Engineering)$
 $Engineering = Human :: good-oriented activity (Art, Science)$

The history of the common measure could be synthesized along the following line:

... ← Philosophy ← ... ← human Culture ← specific Knowledge ← material Economics ← brute Force.

Evolution is a multiple *Divide et Impera et Intellige* for conscience, associated to generating the *components* lacking of the mind at start, then assisted by them:

individual-social-universal conscience (subjective-contextual-objective) → *inspiration* ↓
space-time (structure-behavior) → *imagination* ↓
discrete-continuous (natural-real) → *intention* ↓
beauty-truth-good (art-science-Engineering).

The convergence process of evolution demands struggle against time, with structure as ally. Structure is sometimes too conservative, so it has to be reconfigured, at abstract levels, e.g., a plan, as at concrete ones.

The adaptability-based Reason cannot explain or control thoughts, even if sequential is extended to unlimited parallel/ nondeterministic. Anyway, these desired operational properties can be found mainly in the right side of the human mind.

Further, the difference between continuous and nondeterministic sequential is positive. Therefore, the Reason has to be Faith-dependent completed to Intelligence. A being needs more than Intuition and Adaptability to surpass the Matter by Spirit; only the integration of Intuition and Adaptability by Conscience can explain the Human being. We propose the thesis:

$$\text{Conscience} = \text{closure to } (\text{knowledge} \circ \text{simulation})^{-1} \text{ of} \\ \text{Conscience} \\ \text{initially Conscience} = \text{Consciousness}$$

The idea can be formally sustained in the category theory. Informal arguments follow. The essential limit of discrete computability, inherited by computational intelligence, is generated by the necessity for self-reference to integrate the level knowledge with metalevel knowledge in Conscience modeling.

4. DISCRETE TO CONTINUOUS

Mathematics develops the countable natural to the uncountable real numbers closing to the inverse, on its three integrated ways: algebra, order, and analysis.

Physics uses particles or fields in various chapters. All other sciences are chapters of physics, inheriting and developing the inheritance. At the limits of reasonable understanding, quantum physics tries to balance the knowledge and the unknown, without success.

Engineers have always considered digital a mere ingenious abstraction of analog. Presently, we talk about electronic computers, but the nowadays trend is to copy from the living Nature, i.e., emulation of the advantages of living beings, to achieve complex duties unconsciously.

Reality does not reduce to Nature, as $\text{card}(\text{IN}) < \text{card}(\text{IR})$ (Cantor). Reason is the closure of the Nature relative to the primary operations, as \mathbf{Q} results from the closure of IN to the inverse operations of addition and multiplication. However, the Reason is dense in Reality – as IR is the analytical closure of \mathbf{Q} , $\text{IR} = \{\lim_{n \rightarrow \infty} (q_n) \mid (q_n) \in \text{IN} \rightarrow \mathbf{Q}\}$.

Reality extends beyond Nature and Reason, for the quality of the quantity, and also regarding the power of transforming operations. IR closes \mathbf{Q} to the inverse of power rising – the last arithmetic operation resulted by recurrence of the prior one, which can be pursued by Reason, e.g., algorithmically.

Further, closing to inclusion order, the set of all subsets of countable sets is the uncountable IR, the power of continuum. To get to complex numbers is a matter of imagination.

Reason closes Nature to the inverse of natural operations. Reality closes Reason to the inverse of reasonable operations.

Conscience needs continuous feedback, not only discrete recurrence,. Social and individual conscience are mostly divergent nowadays, i.e., we only performed *Divide et Impera*, neglecting *et Intellige*. It's high time to correct this!

Formalizing the reflexive abstraction by the knowledge hierarchy type and the simplifying abstraction mainly by the simulation hierarchy type, it follows that:

$$\text{Consciousness} = \text{knowledge} \circ \text{simulation} \\ (\text{Consciousness})$$

This fixed-point relation suggests to model conscience by association of a knowledge level to any hierarchical level of the simulation process.

To solve the fixed-point problem we build a metric space where knowledge \circ construction is a contraction - the elements implied in the construction get closer to one another in the formal understanding of the formal construct.

If, even in the sketch, we consider general functional relations between the essential parts of the faith-assisted intelligence, it results:

$$\text{Consciousness} = \text{knowledge} (\text{intention} (\text{Inspiration}, \\ \text{simulation} (\text{imagination} (\text{Intuition}, \text{Consciousness}))))$$

A generic modeling scheme defines the model universe as a mathematical theory or a design paradigm. Any entity has behavior (relations to other entities) and structure (internal relations). Behavior can be functional (context-free) or procedural (context-dependent).

Evidently, the anterior relations are oversimplified in order to move towards intelligent simulation. Although we claim they are intuitive and hope they are inspired, to begin, we neglect the essential but too far from reason to understand intuition and inspiration.

An algorithm is an entity that can be computer simulated, so it represents computability, behavior-oriented (understanding, verifying, learning) / structure-oriented (construction, design, plan).

The algorithmic approach is equivalent to the formal one: If a sentence of a formal system is true, then an algorithm can confirm it. Reciprocally, for a verification algorithm of the mathematical sentences, a formal system can be defined, that holds for true the sentences in the set closure of the algorithm's results towards the operations of the considered logic.

David Hilbert's formal systems, Kurt Gödel's construction algorithm, Alonzo Church's λ -calculus, Stephan Kleene's recursive functions, Emil Post's combinational machines, Alan Turing's machines, Noam Chomsky's grammars, Alexander A. Markov's normal algorithms, are the best-known (equivalent) formalisms for sequential reason-based computability.

The alternative ways followed to extend the computability concept are suggested by approaches known from German literature, which is philosophy-oriented, trying to express essential ideas that link to the unconscious part of our mind.

They respectively concentrate on the mental world of the good managed by Engineering, the physical world of the truth researched by science, and Plato's ideal world of abstractions discovered by arts.

1. Faust (Johann Wolfgang von Goethe): heuristics - risking competence for performance, basing on imagination, confined to the mental world.
2. Das Glasperlenspiel (Hermann Hesse): unlimited natural parallelism - remaining at countable physical suggestions, so in the Nature.
3. Der Zauberberg (Thomas Mann): hierarchical self-referential knowledge - needing to conciliate the discrete structure of hierarchy with the continuous reaction, hoping to open the way to Reality.

Recurrence is confined to discrete worlds, while abstraction is not. This difference suggests searching for understanding based on mathematical structures that order algebra into topology. Intelligence in evolution is the faculty to transform abstract, natural/ artificial objects, and representations, in the correspondent worlds of arts, science and engineering.

Transform = analyze/ synthesize/ modify, especially hierarchical reflexive: ideas about ideas, how to get to ideas, objects to transform objects, representations on representations, how to build/ understand representations.

Evolution is linked to the initial design of mental faculties for surviving of the whole system, but also to the space-time context for communication between intelligent agents.

Recurrence of structures and operations enables approximate self-knowledge (with improved precision on the higher levels of knowledge hierarchies). A continuous model (Traub 1999) for hierarchy levels, without losing the hierarchy attributes, would offer a better model for conscience and intelligence.

A possible interpretation of knowledge hierarchies is: real time of the bottom levels - corresponding to primary knowledge/ behavior/ methods, is managed at upper levels - corresponding to concrete types/ strategies/ models, and abstracted on highest levels - corresponding to abstract types/ theories/ techniques.

Knowledge is based on morphisms that map the state-space of the object-system onto the internal representation of the simulator. An intelligent simulator learns generating and validating models of the object-system. Therefore: representations for design and verification should be common; the algebraic structures on which the different hierarchy types are based on should be extended to topological structures; the different simulation entities should be symbolic, having attributes as: type, domain, function.

Knowledge-based architecture separates representation from reasoning. A topology on the space of symbolic objects permits grouping items with common properties in classes. A dynamically object-oriented internal representation results, that can be adapted to the different hierarchy types.

Topological concepts, as neighborhood, or concepts integrating mathematical structures, as closure, can be applied in verification and optimization, for objects as classes.

The simulation environment prepares a framework for representing entities and relations of the system to be simulated, as general knowledge about the simulated universe.

Knowledge-based architecture, both at environment and simulation component level, ensures flexibility of the framework realization, by defining it precisely only in the neighborhood of solved cases.

For representation, this principle offers the advantage of open modeling. The user describes models, following a general accepted paradigm that ensures syntactic correctness, leaving the meaning to be specified by user-defined semantic functions that control the simulation.

For example, a module in an unfinished design can be characterized by constraints regarding its interaction to other modules; the constraints system is a model, open to be interpreted, thus implemented, differently, adapting to criteria in a non-monotonic logic.

Mathematics contains structures that suggest to be used for self-referent models. The richest domain in this sense is functional analysis (Rudin 1973) that integrates algebra, topology and order:

- contractions and fixed points in metric spaces
- reflexive normed vector spaces
- inductive limits of locally convex spaces
- self-adjoint operators of Hilbert spaces
- invertible operators in Banach algebra.

Let $(U, \{H_i \in S_h\})$ be a universe, structured by different hierarchies H_i and S_h the set of hierarchies defined on universe U .

Then $H = (Rel_{eq}, \{(Level_j, Structure_j) \mid j \in S_l\}, Rel_{ord}, \{A_j \mid j \in S_l\})$ is a generic hierarchy, with: S_l the set of hierarchy levels, Rel_{eq} the equivalence relation generating the levels, $Structure_j$ the structure of level j , Rel_{ord} the (total) order relation defined on the set of hierarchy levels, and $A_j \subset Level_{j-1} \times Level_j, j \in S_l$ the abstraction relation. U is a category, e.g., containing Hilbert spaces with almost everywhere-continuous functions as morphisms, enabling different ways to simulate self-awareness. A hierarchical formal system can be defined:

Considering self-adjoint operators as higher-level objects of the knowledge hierarchy, these levels can approach self-knowledge in the context of knowledge about the inferior levels as of the current one, and having some qualitative knowing about the superior levels. The correspondence problem, i.e., associating the knowledge hierarchy to the simulation hierarchy, is managed by natural transformations over the various functors of the different hierarchies regarding the simulated system. To complete the simulation of the intelligence's components, intention is first determined by human-system dialog.

$(U, \{H_i \in S_h\})$, $\text{card}(U) > \aleph_0$ // hierarchical universe
 $\Sigma = F \cup L \cup A \cup K$ // functional objects
 $F = \{f \mid f: U^* \rightarrow U\}$ // global functions
 $L = \{f \mid f: \text{Level}_j^* \rightarrow \text{Level}_j\}$ // level structures
 $A = \{f \mid f: \text{Level}_j^* \rightarrow \text{Level}_{j+1}\}$ // abstractions
 $K = \{f \mid f: \text{Level}_j^* \times \text{Level}_{j+1} \rightarrow \text{Level}_{j+1}\}$
// knowledge abstractions
 $I = \Sigma^* \cap R$ // initial functions
 $R = \{r \mid r \in \Sigma^* \times R^* \rightarrow \Sigma \times R\}$ // transformation rules.

5. SEMIOTICS \subset SYNTAX \times SEMANTICS

Transferring an ontological approach, communication through language requires the distinction of three levels:

- 1) the level of reality;
- 2) the level of cognitive representation of this reality;
- 3) the level of material representation - text, signs, images etc

When we acknowledge an object in association with a certain sign, than marks are created in our brain in virtue of which the simple appearance of the same sign will *evoke a thought or reference* directed to this object as the impressions stored in the memory were reactivated –see Figure1 (Ogden and Richards 1930).

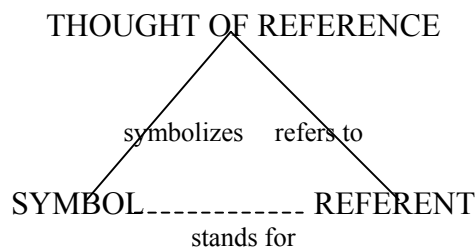


Figure 1: Semiotic triangle

The solid lines in this triangle are meant to represent the causal relations of *symbolization* (remembrance, evocation) and *reference* (memory, perception).

Opposite to these, the dashed line signifies that the relation between symbol (word) and referent (reality), linguistically the most important, is barely *imputed*. The immediate conclusion is that the multiple perspectives in multilinguistic endeavors are (at best) locally, temporarily and partially resolvable.

Assuming the *referent* (reality) is an existing entity for all the interlocutors, they may still have different *thoughts* (concepts) associated to the same referent, depending on their social-geographical personal universe or past experiences.

It is therefore a difficult task for the translator to find the most suitable *word* for the most similar *reference* in the target language.

As difficult as the above situation may seem, it can get even worse: the situation in which the *referent* exists in one language - and has both a reference and a symbol accordingly *attached* - and is inexistent in other(s). Language can here not overpass its limits, e.g., *snow* for the languages of the hot climate countries.

Without giving up anything essentially human, culture, social or natural togetherness, different approaches, humans have a lot in common: philosophic desire, comprehension of the own hierarchy in the context of the other two, free life based on understanding the necessities, constructive fear of the unknown, and especially the love for creation. Except the three cultural ways, that permanently *Divide et Impera et Intellige*, there is no other. (Niculiu 2008)

We need Consciousness to return intelligently to Faith

People of one choice exist, in all senses of the word. They either comprehend all the alternative ways and their convergence, or, in the context of natural love for philosophy and interest for the other selectable directions, put more passion in one direction. Of the first category are temporary elected, in different convergent hierarchical modes, the social leaders, of the second, the institutional directors.

Both kinds of leaders are more philosophical than their cohabitants, even if the ones master the strategic perspective given by an attained peak, while the others have the joy of the courage to climb into profoundness.

The elected artists permanently reconfigure a system of laws, to be beautiful by intelligibility, true by consistence, and good by human understanding. The elected physicists, pure or from different correlated scientific domains all collaborating with mathematics and engineering, govern by research strategies with Gods Fear. The elected engineers critically construct and criticize constructively.

For any social role, the elected concentrate, respectively, on Faith (mathematicians), Intelligence (physicists), and Conscience (engineers).

There always exists a human, called No.1 or the Philosopher, depending on the stability of the times, cloudy or clear Sky. He will always lead directly the elected or the philosophers, who will know to educate and learn optimally the humans of all ages, including themselves.

We have to start. Otherwise, it is no hurry.

Intellige is to link, to understand, and to be aware. In Latin: *intellego* = to understand, to feel, to master, to gather in mind. Artificial has a derogatory sense; however, the root of the word is art. Arts remind of liberty, as *Arts for arts*.

Artificial is at first sight the complement of natural. Our ideas transfer us to places that are neither natural nor artificial. Maybe artificial means something natural created by the human being and Nature is an extension of our body. However, we feel to be superior to Nature, as to our body: we can think. (Penrose 1994)

Why are only humans creating arts, why do they need to know more, and why do they construct other and other natural things they have not found in the Nature? We learned the arts have to discover the Beauty, that science has to look for the Truth, and that engineering invents things to help us, caring for the Good.

Arts are free, and even when they return to Reason, as mathematics, they bring results, that could before just be seen by Intuition, to send by Inspiration and Imagination to Intelligence.

Physics reaches and gets conscious of Reasons limits, both by the quantum theory and by the too complex phenomena, e.g., society and human. It looks like there is no difference for the intelligence that is useful to one of the ways.

An example, that confirms that they simply represent different approaches to understand and develop the (presently natural) Reality, is *architecture*, which we cite in each of them.

To conclude: Intelligence is more than Reason, to make us feel as beings superior to Nature, what also means that we have to respect Nature more:

Spiritus Sanus in Mens sana in Corpore sano

Therefore, there is something else in the Intelligence, which allows us to consider ourselves humans, human groups, peoples, beings on the Earth, or conscious beings in the physical Universe. We also feel that there is something essential beyond the physical – the metaphysical (*Plato*). More, there is something exterior to the human intelligence, without that we could not fight the Time to evolve. We have to feel complete, even if we need education and permanent work in communication with the other humans, of the past, the present, and the future.

We need Conscience to link Faith to Intelligence

We have to remember the abstractions that assisted us to go further. We said complete human to someone complete in a context, what implicitly supposes the power to go beyond the context. This is the story of the integers (*integer* = perfect, complete): they have a beautiful complete theory, however, do not forget to build the rational numbers to feel as close as needed to any real number. Nevertheless, they realize this is not enough, rewarded by the conscience of the continuous Reality – infinitely more powerful than the discrete/countable one. To IR, we get by the perfect circle that is beyond the power of Reason. For example, we plan to realize artificial intelligence, to have a friend that is conscious of the problems to solve together.

For the moment, there is no artificial intelligence. However, we learn to be conscious of the computer limit to process only rational numbers. This means it uses a sequence $(x_n)_{n \in \mathbb{Q}}$ that converges to $\sqrt[n]{a}$ (*Newton*), what reminds us of the density of IQ in IR.

6. INTELLIGENT SYSTEMS

The reasoning of systems capable of reflexive abstraction, i.e., intelligent, starts by describing the problem, and is controlled by problem solving strategies; these derive from the approach principles contained by a knowledge level superior to that of the current simulation.

The principles are structured/ typified corresponding to the higher level => hierarchy types. For the classical representation problem space = (states, actions), problem solving means the process starting from an initial state to look for an operation set that leads to a result state. Solving strategies structure the process to look for the solution (goal-project-concept).

Intelligent systems demand a *cosimulation* of the parts belonging to different domains, e.g., hardware & software, in the context of unified representation for design and verification.

Unified simulation of hard-soft systems is imposed by the incompatibility or non-optimality that results by the initial partition of the system, as by the inefficiency of traversing the design-verification cycle for a fixed partition.

Unified simulation methodologies eliminate the rigid partition constraint: It implies planning and learning, i.e., the possibility of communication between different levels of the knowledge hierarchy. Intelligent simulators can learn by iterative generation and validation of models, possibly interactive.

The objective of the *human-machine dialog* is to advance toward the simulated intelligence by transmitting the knowledge between human and his mental/ physical extensions in a common language. The input dialog is oriented toward learning.

Knowledge bases on a morphism that applies the behavior of the object-system on the internal model of the simulator. The output-dialog on the result specifications is oriented toward planning. The dialog can be extended to the internal unfinished zones, to maintain the integrity of the hard-soft simulation.

Further, communication concordant to the human-machine dialog principles can be also extended from assuring the interface problems between the knowledge hierarchy (planning/ learning) and similar activities corresponding to the hierarchy types that are based on simplifying abstraction forms.

The three different essential ways to approach this goal have common central themes: learning and planning, knowledge representation, and functional constraints.

- Concept-symbol analogy: concept representation and symbol operation try to simulate the mental processes.
- Structural analogy: the activity of brain is emulated by neural networks, cellular automata, genetic algorithms, membranes or quantum computing.
- Hierarchic-parallel analogy: thinking is considered a collective phenomenon that is produced by constitutive phenomena parallel and recurrently.

The limits of the knowledge domain for intelligence simulation are *reconfigurable*: learning can guide the representation - semantics and architecture of the system, and functional constraints can formalize the cognitive constraints in the spatial-temporal reasoning context.

The informatics extensions of the contemporary human impose the knowledge of a third language to use his artificial mental or physical extensions, next to the mother language for context integration and to the one surpass the context - nowadays, American English.

The evolution of the programmable systems from punctual activities as answer to explicit orders, to autonomous activities, supposes a knowledge-based high-level symbolic object-oriented dialog, to awake the consciousness by explicit selections, and the adaptability by assisted decisions. As result of formal version of (part of) the natural language, a high-level language for intelligent dialog has to inherit:

- syntactical regularities (studied by computational linguistics) and semantic correspondences (studied both in language philosophy as in AI),
- regularities of the cognitive processes (studied by cognitive psychology and intellectics),
- relations with the I/O system (perception/ action) of the individual intelligent agent, and with the interactions of the individual intelligent agents (social relations) in the intelligent system.

Formalized conform to information theory, syntax and semantics offer a representation of a world preexistent to the dialog. The resulted formal system has to be correct - any formula corresponds to a fact and any formal computation to a real reasoning, and complete – any real fact corresponds to a formula and any real reasoning to a formal computation.

Consequently, understanding is simulated by the evolution of the representation language in the symbol hierarchy. This approach is that of the classical artificial intelligence. Its limits proceed from restriction to logic sequential mathematical discrete reasoning, what results in the incapacity to represent conscience, intention, intuition, i.e., intelligence. The regularities of the cognitive processes are represented as inferential strategies common to the dialog partners: inference is not just deductive, but mostly inductive.

The evolution of the programmable systems from punctual activities as answer to explicit orders, to autonomous activities, supposes a knowledge-based high-level symbolic object-oriented dialog, to awake the consciousness by explicit selections, and the adaptability by assisted decisions. As result of formal version of (part of) the natural language, a high-level language for intelligent dialog has to inherit:

- syntactical regularities (studied by computational linguistics) and semantic correspondences (studied both in language philosophy as in artificial intelligence),
- regularities of the cognitive processes (studied by cognitive psychology and intellectics),
- relations with the I/O system (perception/ action) of the individual intelligent agent, and with the interactions of the individual intelligent agents (social relations) in the intelligent system.

Consequently, understanding is simulated by the evolution of the representation language in the symbol hierarchy. This approach is that of the classical artificial intelligence. Its limits proceed from restriction to logic sequential mathematical discrete reasoning, what results in the incapacity to represent conscience, intention, intuition, i.e., intelligence. The regularities of the cognitive processes are represented as inferential strategies common to the dialog partners: inference is not just deductive, but mostly inductive.

CONCLUSIONS

Conscience simulation demands transcending the present limits of computability, by an intensive effort on extensive research to integrate essential physical and mathematical knowledge guided by philosophical goals. Even mathematics will have to develop more philosophy-oriented to approach intuition. Simulability is computability using the power of continuum. There are positive signs for this from analog electronics, control systems, and mechatronics. Real progress towards this way of computation needs unrestricted mathematics, integrated physics and thinking by analogies. Evolution implies the separation of faith and intelligence, so we have to better understand both, integrating them to human wisdom, to be divided further to get more human. Metaphorically phrased, our searches and researches should have as axioms:

- *God is Unique.*
- *His ways are Uncountable*
- *His plans are Hierarchical.*

Philosophy is not a specialty but a human right. There have to be schools to prepare the teachers of philosophy for the other humans. These schools have to develop also respect for those that look for the Way on one of the three alternative paths that correspond to the fundamental partition (arts, science, engineering). Because recently the essential *Divide et Impera* do not *Intellige*, the only philosophers are the masters in:

- Arts – especially mathematicians, and others that, aware or not, compose mathematically
- Science – physicists, and those that do not forget their science is a chapter of physics
- Engineering – mostly those working in domains that attain the limits of the pure Reason.

Mathematics is one of the arts. The music is at least as beautiful and expressive, but mathematics does not demand an extraordinary talent, allows a reasonable dialog about it, and has well-defined reconfigurable limits of that it is aware. Mathematics has to be educated as soon as possible and has not to be confounded with its handcraft. The music gets more often out of its character. The two arts evolved together: *Johann Sebastian Bach, Antonio Vivaldi, Joseph Haydn* were musically gifted mathematicians, who preferred the liberty of the music to the bands of the Reason.

The Reason, as initial zone, makes mathematics more sure but less charming than the other arts that can refer directly to the Reality: literature, music and sculpture. The visual arts are too dependent of the Nature because seeing is the most used sense for the human natural being. The mathematics school is continuous, whereby sculpture, literature, and music can generate sooner higher singular peaks: *Michelangelo, Shakespeare, Beethoven*, by an exponential/ other highly nonlinear continuity. Arts are free. But mathematics first expressed reasonably that Reality could only be approached by Reason

Physics is the Science. The other natural and social sciences are its chapters, even if they are not yet aware of it, or just try to return to their riverbed by intermediary specialties instead of integrative bridges. As any artificial system, the society is structured on natural bases, and it develops by natural laws. The modern age forced these laws towards Reason, and recently they got out of control. The social laws got also unreasonable. Physics is essential for the constructive reconfiguration of the Faith.

Engineering is most frequently both art and science, and is as important as arts and sciences in the fundamental partition of the Reality needed for evolution. However, it is more dangerous than its alternative approaches, of which it has to be strictly bridled. Reasons are twofold: Its result, called *technology*, is defined by its complement – so it is not superior to this. It does not impose spiritual proximity between the creator and the user – so it can be applied in a complete different scope than it was generated. However, any engineering is the homonymous complement of a special science that collaborates with mathematics, therefore, integrated sciences into physics and mathematics remaining among arts solve the case.

REFERENCES

- Ageron, P., 2001. Limites inductives point par point dans les categories accessibles, *Theory and Applications of Categories*, 7 (1), 313-323.
- Hofstadter, D., 1979. *Gödel, Escher, Bach - The Eternal Golden Braid*, Washington DC: Vintage.
- Keutzer, K., et al., 2000. Orthogonalization of Concerns & Platform-based System-Level Design, *IEEE Transactions on CAD of Integrated Circuits and Systems*, 19 (12), 1523-1543.
- Lupu, C., 2004. Locality Measured by Contour Patterns - A Topographic Model. *Proceedings of 15th IASTED International Conference on Modelling and Simulation*, pp. 50-54. March 1-3, Marina del Rey (California, USA).
- Miller, R., et al., 1993. Parallel Computations on Reconfigurable Meshes, *IEEE Transactions on Computers*, 42(6), 678-692.
- Niculiu, M., Niculiu, T., 2009. European Spirit Evolution by Multicultural Harmony. *Proceedings of 6th International Symposium on Personal and Spiritual Development in the World of Cultural Diversity*
- Niculiu, T., 2008. *Object-oriented Symbolic Structural Intelligent Simulation*, București: Matrix Publishers.
- Ogden, C., Richards I., *The Meaning of Meaning*, 3rd ed. New York, 1930
- Penrose, R., 1994. *Shadows of the Mind, Consciousness and Computability*, Oxford: University Press
- Rabaey, J., 1997. Reconfigurable Computing, *Proceedings of IEEE International Conference on Acoustics, Speech & Signal Processing*, pp. 127-132. March 28-31, München.
- Rudin, W., 1973. *Functional Analysis*, New York: McGraw Hill.
- Ștefan, G., 2010. *Ethos, Pathos, Logos*, București: ALL.
- Traub, J.F., 1999. A Continuous Model of Computation, *Physics Today*, 5(5), 39-43.
- Zeigler, B., Praehofer, H., Kim, T., 2000. *Theory of Modeling and Simulation*, Oxford: Academic Press.
- Zhong, N., Weihrauch, K., 2003. Computability Theory of Generalized Functions, *Journal of Automated Computer Machinery*, 50(6), 574-583.

MARIA NICULIU is MA (European Intercultural Communication Strategies) and MS (Mechanics Engineering). Her MA thesis (2011) is about *Von intensiv zu nachhaltig* – from intensive to sustainable. She graduated the Faculty of Literatures and Foreign Languages, German/ French/ Dutch - Bucharest University (2009) with a paper on *Sprachenvielfalt und Mehrsprachigkeit, was taugt die Europäische Union als Standort der Mehrsprachigkeit* - Multilingualism and what makes the European Union as model for a multilingual society. After the MS from the Faculty of Mechanics Engineering of University *Politehnica* of Bucharest (1987), she enriched her technical and economical background - member of the Romanian Authorized Accountants and Financial Experts (2008), and directed her interests towards a multicultural approach of the European social and economical evolutions, based on linguistic and behavior simulation.

TUDOR NICULIU is Professor at the Electronics, Telecommunications, and Information Technology Faculty of the *Politehnica* University in Bucharest, and Senior Researcher at the Center for New Electronic Architectures of the Romanian Academy. He is looking for hierarchical integration of different domains, to understand intelligence by simulating it, and to apply it to intelligent simulation. Since 1991, he teaches and researches at the same institution, PhD 1995, MS 1985. Before, he was Senior Researcher at the R&D Institute for Electronic Components in Bucharest, researching and designing hierarchical simulation of analog integrated circuits. He studied Mathematics at University of Bucharest (MA 1994). He published 12 books, 25 journal articles, and 69 international conference papers. He is IEEE Senior Member of CAS, Computer, and SMC Societies, as Fellow of International Institute for Advanced Studies in Systems Research & Cybernetics.

AN ASYNCHRONOUS PARALLEL HYBRID OPTIMIZATION APPROACH TO SIMULATION-BASED MIXED-INTEGER NONLINEAR PROBLEMS

K.R. Fowler^(a), T. Kopp^(b), J. Orsini^(c), J.D. Griffin^(d), G.A. Gray^(e),

^(a,b,c)Department of Mathematics and Computer Science, Clarkson University
^(d)SAS

^(e)Sandia National Labs

^(a)kfowler@clarkson.edu, ^(b) kopptr@clarkson.edu, ^(c) orsinijw@clarkson.edu, ^(d) joshua.griffin@sas.com,
^(e) gagray@sandia.gov

ABSTRACT

To address simulation-based mixed-integer problems, a hybrid algorithm was recently proposed that combines the global search strengths and the natural capability of a genetic algorithm to handle integer variables with a local search on the real variables using an implementation of the generating set search method. Since optimization is guided only by function values, the hybrid is designed to run asynchronously on a parallel platform. The algorithm has already been shown to perform well on a variety of test problems, and this work is a first step in understanding how the parallelism and local search components influence the search phase of the algorithm. We show that the hybridization can improve the capabilities of the genetic algorithm by using less function evaluations to locate the solution and provide a speed-up analysis on a standard mixed-integer test problem with equality and inequality constraints.

Keywords: genetic algorithm, pattern search, asynchronous, mixed-integer nonlinear programming

1. INTRODUCTION

The need for reliable and efficient optimization algorithms that do not require derivatives is common across engineering disciplines. In general, the optimal design process requires such algorithms to work in conjunction with simulation tools, resulting in what is known as black-box optimization. For example, the simulation may require the solution to a system of partial differential equations that describes a physical phenomenon. These problems are challenging in that optimization must be guided by objective function (and possibly constraint) values that rely on a computer simulation, without any additional knowledge other than the output from the simulation itself. The simulation may be computationally expensive and add undesirable features to the underlying problem such as low amplitude numerical noise, discontinuities, or hidden constraints (i.e. when the program simply fails to return a value due to its own internal solver failure). Derivative-free optimization (DFO) methods have been developed, analyzed, and demonstrated successfully

over the last several decades on a wide range of applications (Conn, Scheinber, and Vincente 2009). Because DFO methods only rely on function values, parallelism is often straightforward and, in the case of expensive simulation calls, can make otherwise intractable problems solvable.

Hybrid DFO algorithms have emerged to overcome inherent weaknesses and exploit strengths of the methods being paired (Talbi 2004; Raidl 2006; Alba 2005). Often, the hybrid algorithms are designed to address problems that could not otherwise be solved. In this work, we focus on the parallelism of a hybrid evolutionary algorithm with a local search that was designed for simulation-based mixed-integer problems with nonlinear constraints (Griffin, Fowler, Gray, Hemker, and Parno). The performance of the hybrid was demonstrated on a suite of standard test problems and on two applications from hydrology (Gray, Fowler, and Griffin 2009; Gray, Fowler, and Griffin 2010; Griffin, Fowler, Gray, Hemker, and Parno) that were known to be challenging for a wide range of DFO methods (Fowler, Kelley et al 2004; Fowler 2008). Some of those challenges, which are not unique to environmental engineering, included discontinuous optimization landscapes, low amplitude noise, and multiple local minima. Specifically, in (Fowler, Kelley et al 2004), a comparison of derivative-free methods on the hydrology applications showed that a genetic algorithm (GA) performed well in terms of identifying the correct integer variables but then failed to achieve sufficient accuracy for the real variables. On the other hand, given a reasonable initial iterate with respect to the integer variables, the local search methods showed fast convergence. These observations motivated the pairing of the GA with a generating set search approach, referred to as Evolutionary Algorithms Guiding Local Search (EAGLS). The resulting algorithm pairs the binary mapping of the genetic algorithm to handle integer variables with asynchronous, parallel local searches on only the real variables. The new method has strong global search aspects and can still maintain high accuracy from the local search phase.

Previous studies focused on the ability of EAGLS to solve a variety of MINLPs with varying difficulties

in constraint formulations and problem size. In (Griffin, Fowler, Gray, Hemker, and Parno), EAGLS was able to solve a water supply hydrology application that previously could not be solved without significant parameter tuning of either of the two software tools that were merged to create the hybrid. Little work has been done to understand how the asynchronous parallelism that is inherent in the implementation impacts the search phase of the algorithm. This work is a first attempt at using parallel performance measures to understand the algorithms strengths and weaknesses.

For this work, we consider objective functions of the form $f: \mathbb{R}^{n_r+n_z} \rightarrow \mathbb{R}$ and mixed-integer nonlinear optimization problems of the form

$$\min_{p \in \Omega} f(p). \quad (1)$$

Here n_r and n_z denote the number of real and integer variables and, $x \in \mathbb{R}^{n_r}$, $z \in \mathbb{Z}^{n_z}$. In practice, Ω may be comprised of component-wise bound constraints on the decision variable in combination with linear and nonlinear equality or inequality constraints. Often, Ω may be further defined in terms of state variables determined by simulation output. We proceed by first reviewing the genetic algorithm, the generating set search method, and software that are hybridized to form the new algorithm. We then present numerical results and outline future directions.

2. EAGLS

2.1. Genetic Algorithms

The EAGLS approach combines a genetic algorithm and a generating set search approach. GAs (Goldberg 1989; Holland 1975; Holland SIAM) are one of the most widely-used DFO methods and are part of a larger class of evolutionary algorithms called population-based, global search, heuristic methods (Goldberg 1989). GAs are based on biological processes such as survival of the fittest, natural selection, inheritance, mutation, or reproduction. Design points are coded as “individuals” or “chromosomes”, typically as binary strings, in a population and then undergo the above operations to evolve towards a better fitness (objective function value).

A simple GA can be outlined with:

1. Generate a random/seeded initial population of size n_p
2. Evaluate the fitness of individuals in initial population
3. Iterate through the specified number of generations:
 - a. Rank fitness of individuals
 - b. Perform selection
 - c. Perform crossover and mutation
 - d. Evaluate fitness of newly-generated individuals
 - e. Replace non-elite members of population with new individuals

During the selection phase, better fit individuals are arranged randomly to form a mating pool on which further operations are performed. Crossover attempts to exchange information between two design points to produce a new point that preserves the best features of both ‘parent points’. Mutation is used to promote a global search and prevent stagnation at a local minimum. Termination of the algorithm is typically based on a function evaluation budget that is exhausted as the population evolves through generations.

Often, GAs are criticized for their computational complexity and dependence on optimization parameter settings, which are not known a priori (Dejong and Spears 1990; Grefenstette 1986; Lobo, Lima, and Michalewicz 2007). Parameters like the population size, number of generations, as well as the probabilities and distribution indices chosen for the crossover and mutation operators affect the performance of a GA (Reed, Minsker et al. 2000; Mayer, Kelley, et al. 2002). Also, since the GA incorporates a randomness to the search phase, multiple optimizations are often useful to exhaustively search the design space. However, if the user is willing to spend a large number of function evaluations, a GA can help provide insight into the design space and locate initial points for fast, local, single search methods. The GA has many alternate forms and has been applied to a wide range of engineering design as shown in references such as (Karr and Freeman 1998). Moreover, hybrid GAs have been developed at all levels of the algorithm and with a variety of other global and local search DFO methods. See for example (Blum, Aquilera, et al. 2008; Talbi 2004; Raidl 2006) and the references therein.

The EAGLS software package was created using the Non-dominated Sorting Genetic Algorithm (NSGA-II) software, which is described in (Deb, Pratap et al. 2002; Zitzler, Deb and Thiele 2000; Deb 2000; Deb and Goel 2001). Although a variety of genetic algorithms exist, the NSGA-II has been applied to both single and multi-objective problems for a wide range of applications and is well supported. In particular, it is designed to be used “off-the-shelf” which made it a good candidate for hybridization.

2.2. Generating Set Search and APPS

Asynchronous Parallel Pattern Search (APPS) (Hough and Kolda 2001; Kolda 2004) is a direct search methods which uses a predetermined pattern of points to sample a given function domain. APPS is an example of a generating set search (GSS), a class of algorithms for bound and linearly constrained optimization that obtain conforming search directions from generators of local tangent cones (Lewis, Shepherd et al. 2005; Kolda, Lewis et al. 2006). In its simplest form, the method evaluates the objective function on a stencil of points and if a better point is found, the stencil is moved to that point, otherwise the size of the stencil is reduced. Optimization is terminated either based on a function evaluation budget or when the stencil becomes sufficiently small. The basic GSS algorithm is:

Let x_0 be the starting point, Δ_0 be the initial step size and $\mathcal{D}\{d_i\}_{i=1}^{2n_r}$ be the set of positive spanning directions.

While not converged Do:

1. Generate trial points $Q_k = \{x_i + \tilde{\Delta}_k d_i | 1 \leq i \leq \mathcal{D}\}$ where $\tilde{\Delta}_k \in [0, \Delta_k]$ denotes the maximum feasible step along d_i .
2. Evaluate trial points (possibly in parallel).
3. If $\exists x_q \in Q_k$ such that $f(x_q) - f(x_k) < \alpha \Delta_k^2$
Then $x_{k+1} = x_q$ (successful iteration)

Else $x_{k+1} = x_k$ (unsuccessful iteration) and $\Delta_{k+1} = \frac{\Delta_k}{2}$ (step size reduction)

The majority of the computational cost of pattern search methods is the $2n_r$ function evaluations, so parallel pattern search (PPS) techniques have been developed to perform function evaluations simultaneously on different processors (Dennis and Torczon 1991; Torczon 1992). For example, for a simple two-dimensional function, consider the illustrations in Figure 1 taken from (Gray and Fowler 2011). First, the points $f, g, h,$ and i in the stencil around point c are evaluated. Then, since f results in the smallest function value, the second picture shows a new stencil around point f . Finally, in the third picture, since none of the iterates in this new stencil result in a new local minima, the step size of the stencil is reduced.

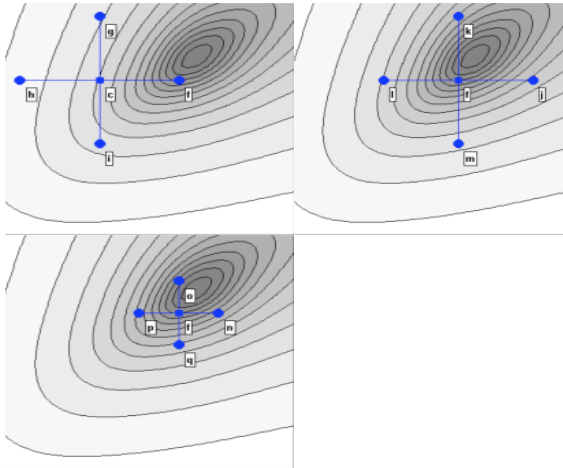


Figure 1: Illustration of the steps of Parallel Pattern Search (PPS) for a simple two-dimensional function. On the upper left, an initial PPS stencil around starting point c is shown. In the upper right, a new stencil is created after successfully finding a new local min f . On the bottom left, PPS shrinks the stencil after failing to find a new minimum

Note that in a basic GSS, after a *successful* iteration (one in which a new best point has been found), the step

size is either left unchanged or increased. In contrast, when the iteration was unsuccessful, the step size is necessarily reduced. A defining difference between the basic GSS and APPS is that the APPS algorithm processes the directions independently, and each direction may have its own corresponding step size. Global convergence to locally optimal points is ensured using a sufficient decrease criterion for accepting new best points. A trial point $x_k + \Delta d_i$ is considered better than the current best point if

$$f(x_k + \Delta d_i) - f(x_k) < \alpha \Delta^2, \quad (2)$$

for $\alpha > 0$.

Because APPS processes search directions independently, it is possible that the current best point is improved before all the function evaluations associated with a set of trial points Q_k have been completed. These results are referred to as *orphaned points* as they are no longer tied to the current search pattern and attention must be paid to ensure that the sufficient decrease criteria is applied appropriately. The support of these orphan points is a feature of the APPS algorithm which makes it naturally amenable to a hybrid optimization structure. Iterates generated by alternative algorithms can be simply be treated as orphans without the loss of favorable theoretical properties or local convergence theory of APPS.

2.3. Why EAGLS works

EAGLS combines the NSGA-II with the APPSPACK software (Gray and Kolda 2006). APPSPACK is written in C++ and uses MPI (Gropp, Lusk et al. 1996; Gropp and Lusk 1996) for parallelism. Function evaluations are performed through system calls to an external executable which can write in any computer language. This simplifies its execution and also makes it a good candidate for inclusion in a hybrid scheme. Moreover, it should be noted that the most recent version of APPSPACK can handle linear constraints (Kolda, Lewis, and Torczon 2006; Griffin, Kolda and Lewis 2008), while a software package called HOPSPACK builds on the APPSPACK software and includes a GSS solver that can handle nonlinear constraints (Griffin and Kolda 2010a; Plantega 2009). To implement EAGLS, as in (Griffin and Kolda 2010b), a preliminary version of HOPSPACK was used.

The EAGLS algorithm is designed to exploit parallelism. A goal of a parallel program is to ensure that all available processors are continuously being used. However, in practice this is often not the case. To understand this more fully in our context, consider a hypothetical black-box bound constrained optimization problem that has two real variables and an objective function with an evaluation time of at least one hour; further, we assume the user has 128 nodes with 2 processors each. There are a number of advantages that come from the use of parallelism in this context.

- Most local search algorithms (even asynchronous parallel ones) have a cap on the

maximum number of processors they can effectively use. For our example problem, APPS will generate at most 4 trial-point per iteration. For the first hour APPS is called 252 processors will be idle. The user would need to start by hand 64 different instances of APPS centered at unique starting points, to fully exploit the computational power at hand with APPS alone.

- Most algorithms are synchronous by design, and parallel versions typically run in a “batch” mode. For example, a genetic algorithm requires all points in the current generation be evaluated before creating the next. Suppose a parallel GA uses a population of size 256 and submits all 256 points to be evaluated in parallel. Before the second iteration can begin, all 256 points must be evaluated; if all evaluations are complete but one, then the entire optimization processes is halted until this final evaluation is completed, even if this remaining evaluation takes hours longer to complete. Thus synchronous parallel algorithms necessarily move at the rate of the slowest evaluation.

The downsides described in the preceding bullets are actually advantageous for hybrid algorithms. Rather than attempt to redesign APPS so that it will submit more points in each iteration or invent a new asynchronous genetic algorithm that seeks to update multiple generations asynchronously, we simply tie multiple algorithms together loosely, pooling the resources in such a way that any unused resources can be shared. In the case of EAGLS, a single GA is run, and remaining idle processors are used to perform local searches. However, because local searches are often much faster than a GA at finding a local minimum, priority is given each generation to the local searches in the evaluation queue until that iterations local search evaluation budget has been expended. An immediate consequence and benefit of the EAGLS structure is that there is virtually no cap on the maximum number of processors that can be utilized for a given problem. At the same time, even with a few extra processors, significant wall-clock gains can be achieved, as the local search can be used to quickly find the global minimum once the GA is sufficiently near.

2.4. EAGLS Algorithm

EAGLS uses the GA's handling of integer and real variables for the global search, and APPS's handling of real variables in parallel for local search. Note that a MINLP could be immediately reduced to an integer programming problem if there was an analytic formula that provided x^* where

$$x^* = \arg \min_x f(x, z)$$

given an integer variable z . Though for a general MINLP, such a formula may not exist, local searches can be used (in parallel) to repeatedly

replace (x, z) pairs in the GA population pool with (\hat{x}, z) , where \hat{x} is an improved estimate of x^* provided by a local search. The GA still governs point survival, mutation, and merging as an outer iteration, but, during an inner iteration, individual points are improved via APPS applied to the real variables, with the integer variables held fixed. For simplicity, consider the parallel synchronous EAGLS algorithm with k local searches:

1. Evaluate initial population in parallel
2. While *not converged* Do
 - a. Choose a subset \mathcal{L} of k points from current population for local search
 - b. Simultaneously run k instances of APPS centered at points in \mathcal{L}
 - c. Replace respective points with their optimized values
 - d. Perform selection, mutation, crossover
 - e. Evaluate new GA points in parallel

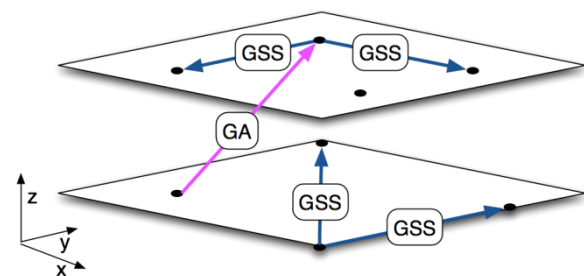


Figure 2: In EAGLS the genetic algorithm optimizes over both integers and real variables, while local search instances work solely within a given integer plane (Griffin, Fowler, Gray, Hemker, and Parno).

To select points for the local search, EAGLS uses a ranking approach that takes into account individual proximity to other; better points (see Figure 2). The goal of this step is to choose promising individuals representing distinct integer subdomains. The EAGLS algorithm allows the local search and the GA to run simultaneously using the same pool of evaluation processors. For the most part, the GA and each local search run asynchronously. However, after each GA generation, a new batch of local searches are created and given priority in the evaluation queue. This implies that given an adequate number of local search instances, the GA generations and local search generation will necessarily be nested, as the number of local search trial-points will always be greater than the number of available processors in the evaluation queue. This forces the GA to wait until the local search generation depletes its current evaluation budget prior to proceeding. Once the GA population has been evaluated, the local

searches begin and operate asynchronously. To avoid re-evaluating points, all function values are stored in cache. The external parallel paradigm is nearly identical to that used in (Griffin and Kolda 2010b; Gray, Griffin et al. 2008). Whenever an improved point is found with respect to the real variables, the corresponding population member is immediately updated. See Figure 3 for a short point-flow sketch of this process.

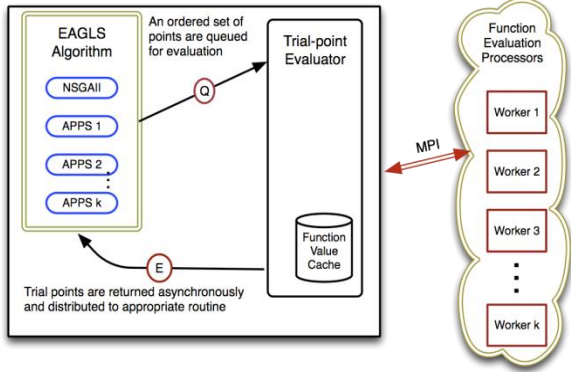


Figure 3: The EAGLS user can decide the population size and the number of local searches in an input file. The algorithms are run asynchronously in parallel with the local searches periodically inserting new improved points into the current GA population.

3. NUMERICAL RESULTS

3.1. Test Problem

To evaluate the parallelism of EAGLS we consider two studies. In the first, we fix all the optimization algorithm parameters and increase the number of processors used. In the second, we fix the number of processors to 16 and vary only the number of local searches while all other optimization parameters are held fixed. We use a classical mixed-integer test problem taken from (Kocis and Grossman 1988) that was proposed to study process synthesis applications with the outer approximation method. While this may seem to be simple, it is representative of the MINLPs encountered in process design and engineering. Thus, understanding how the parallelism and local search components of EAGLS affect its solution will aid in our ability to more efficiently solve similar MINLPs. The decision variables are $p = (z_1, z_2, z_3, x_1, x_2)^T$ with bound constraints given by

$$p \in \Omega = \{p \mid z_1, z_2, z_3 \in \{0,1\}, x_1, x_2 \in [0,10]\}.$$

We seek to minimize the objective function $f(p)$ where

$$f(p) = 2x_1 + 3x_2 + 1.5z_1 + 2z_2 - 0.5z_3 \quad (3)$$

subject to the following constraints,

$$\begin{aligned} c_1(p) &= x_1^2 + z_1 - 1.25 = 0 \\ c_2(p) &= x_2^{1.5} + 1.5z_2 - 3.00 = 0 \\ c_3(p) &= x_1 + z_1 - 1.60 \leq 0 \\ c_4(p) &= 1.333x_2 + z_2 - 3.00 \leq 0 \\ c_5(p) &= -z_1 - z_2 + z_3 \leq 0. \end{aligned} \quad (4)$$

The constraints on both the integer and real variables make this problem challenging. For constraint handling, we use the ℓ_1 and the ℓ_1 -smoothed penalty function where the constraint violation is incorporated with the objective function to form a corresponding merit function (Griffin and Kolda 2010b). Although the problem is small dimensionally, it is non-convex and some of the sub problems obtained by fixing the integer variables contain a unique local minimum which is challenging for standard MINLP solvers to avoid, as shown in (Kocis and Grossman 1988). Thus, this problem was ideal for testing the integer capabilities of EAGLS (Griffin, Fowler, Gray, Hemker, and Parno) and thereby was chosen here to study the asynchronous parallel local search capabilities. The known solution has a function value of 7.667 and the local minimum has a value 7.931. To add computational expense to each function evaluation and test the asynchronous nature of the algorithm, we add a random pause between one and three seconds to each function evaluation. This approach was used to test parallel optimization approaches in (Hough, Kolda, and Torzan 2001; Griffin and Kolda 2010b).

3.2. Algorithmic Parameters and Platform

Since the solution to the test problem is known, we stop when the best point found is within 1% of the known solution. We provide the other relevant optimization parameters in Table 1. The numerical experiments were performed on a 102 processor Beowulf blade cluster (IBM e1350) with 3.0 Ghz Intel Xeon processors and Myrinet Networking.

Table 1: Optimization Parameters

Parameter	Value
Population size	40
Number of Generations	250
Real Crossover Probability	0.9
Real Mutation Probability	0.5
Binary Crossover Probability	0.9
Binary Mutation Probability	0.0125
GSS Contraction Factor	0.5
GSS Sufficient Decrease Factor	1e-9
GSS Step Tolerance	1e-5
Maximum Generation Evaluations	840
Maximum Function Evaluations	3000

3.3. Varying Number of Processors

Since the GA has stochastic optimization parameters and APPS is asynchronous, EAGLS is not a deterministic method, thus each optimization experiment was run five times and average values are reported. This approach has been used in numerous studies for APPS (Griffin and Kolda 2010b). Average run times and number of function evaluations required for convergence are shown in Figure 4 as the number of processors doubles from 2 to 64. For these experiments EAGLS used 8 local searches. Since there are only $n_r = 2$ real variables, for each local search APPSPACK would not see increased speed up beyond $2n_r = 4$ processors for a total of 32 while the additional processors can be used to evaluate the GA population. The figure on the left shows the speed-up one would expect. The figure on the right is interesting in that the number of function evaluations increases with the number of processors. This is because as APPSPACK is run on more processors, the algorithm may move the stencil to a new location if a point is found with a lower function value but older points are not deleted from the queue if sufficient processors are allocated. So if a point from an older stencil does return a lower function value, the algorithm would move back to that location and continue. Note that because significantly more processors are being used, the computational time still shows linear speed-up despite the increased number of function evaluations.

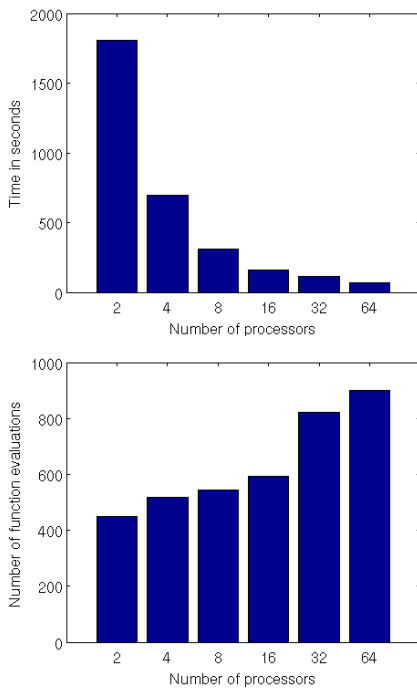


Figure 4: Computational time and number of evaluations required as the number of processors varies. Run times are shown in the upper picture and number of function evaluations are shown in the lower picture.

3.4. Varying Number of Local Searches

To further understand how the asynchronous nature of APPSPACK impacts the search phase of EAGLS, we vary the number of local searches. For these experiments, 16 processors were used and all optimization algorithmic parameters were fixed except the number of local searches, which was varied from 4 to 8. We also consider the case of no local searches, which means EAGLS is simply a genetic algorithm with function evaluations performed in parallel. Figure 5 shows the average run times and number of function evaluations needed for convergence.

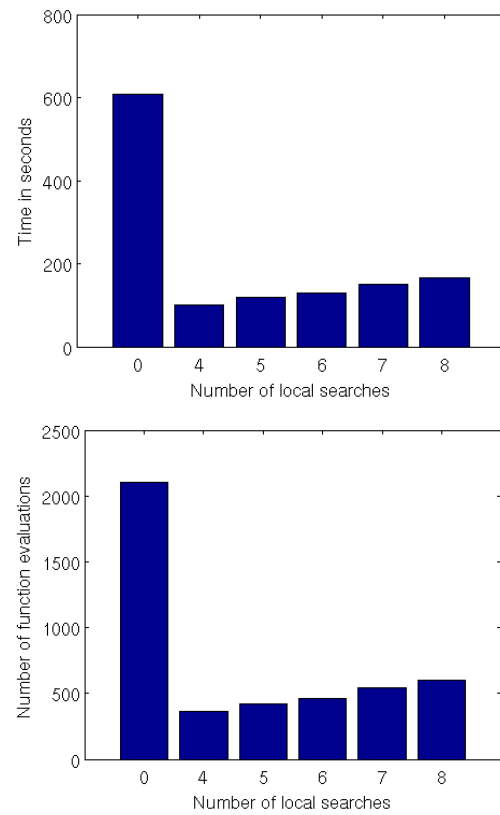


Figure 5: Computational time and number of function evaluations required as the number local searches varies. Run times are shown in the upper picture and number of function evaluations are shown in the lower picture.

The local searches have a significant impact on the optimization history using roughly one fifth of the computational effort of the GA alone. As the number of local searches increases, the number of function evaluations increases as one would expect but it is not significant. This is due in part to the fact that the algorithm is terminating based on proximity to a known solution. Future work will include exploring the behavior on larger dimensional problems which may show more dynamic results in terms of the optimal

number of local searches, but for this work we are staying in the context of simulation-based MINLPs which typically are not too large. We should further note that this test problem does have a feasible local minimum with a function value of roughly 7.931, and EAGLS avoided convergence to this suboptimal point in all trials.

4. CONCLUSIONS

These experiments are the first step in understanding an asynchronous hybridization of a genetic algorithm with a local search based on a generating set search method for mixed-integer problems. This approach has extended the APPSPACK software to handle integer variables, improved its global search capabilities, and added parallelism and a local search to the NSGA-II software package. The tests done here are promising in showing that using local searches can help accelerate the convergence of the GA but also indicate that there is a complex interaction among algorithm parameters. The GA is well-known to be sensitive to parameter settings and the addition of an asynchronous local search with additional parameters warrants a more extensive study to better guide users. Future work will include a sensitivity study similar to that in (Matott, Bartlett et al. 2006) to understand the interaction and main effects of the optimization settings.

ACKNOWLEDGMENTS

This work was made possible by support from the American Institute of Mathematics.

REFERENCES

- Alba, E. (2005). *Parallel Metaheuristics*. John Wiley & Sons, Inc.
- Blum, C., Blesa Aquilera, M. J., Roli, A., & M., S. (2008). *Hybrid Metaheuristics*. Springer.
- Conn, A., Scheinberg, K., & Vicente, L. N. (2009). Introduction to Derivative Free Optimization. *SIAM*.
- Deb, K. (2000). An efficient constraint handling method for genetic algorithms. *Computer Methods in Applied Mechanics and Engineering*.
- Deb, K., & Goel, T. (2001). Controlled Elitist Non-dominated sorting genetic algorithms for better convergence. *Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization {EMO} 2001*.
- Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A Fast and Elitist Multi-Objective Genetic Algorithm: {NSGA-II}. *{IEEE} Transactions on Evolutionary Computation*.
- Dejong, K., & Spears, W. (1990). An Analysis of the Interacting Roles of Population Size and Crossover in Genetic Algorithms. *First Workshop Parallel Problem Solving from Nature*. Springer-Verlag, Berlin.
- Dennis, J. E., & Torczon, V. (1991). Direct search methods on parallel machines. *SIAM J. Optim.*
- Fowler, K. e. (2008). A Comparison of Derivative-free Optimization Methods for Water Supply and Hydraulic Capture Community Problems. *Adv. Water Resour.*, 743-757.
- Fowler, K., & C.T., K. (2004). Solution of a Well-Field Design Problem with Implicit Filtering. *Opt. Eng.*
- Goldberg, D. (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison Wesley.
- Gray, G., & Fowler, K. (2011). Traditional and Hybrid Derivative-free Optimization Approaches for Black-box Optimization. In *Computational Optimization and Applications in Engineering and Industry*. Springer.
- Gray, G., & Griffin, J. (2008). *HOPSPACK: Hybrid optimization parallel search package*. Livermore, CA: Sandia National Labs.
- Gray, G., & Kolda, T. (2006). Algorithm 856: APPSPACK 4.0: Asynchronous Parallel Pattern Search for Derivative-Free Optimization. *ACM TOMS*.
- Gray, G., Fowler, K., & Griffin, J. (2010). Hybrid Optimization Schemes for Simulation Based Problems. *Procedia Comp. Sci.*, 1343-1351.
- Grefenstette, J. (1986). Optimization of Control Parameters for Genetic Algorithms. *IEEE Trans. Sys. Man Cybernetics*.
- Griffin, J., & Kolda, T. (2010). Asynchronous parallel hybrid optimization combining DIRECT and GSS. *Optim. Meth. Software*.
- Griffin, J., & Kolda, T. (2010). Nonlinearly-constrained optimization using heuristic penalty methods and asynchronous parallel generating set search. *Appl. Math. Res. eXpress*.
- Griffin, J., Fowler, K., Gray, G., Hemker, T., & Parno, M. (n.d.). Derivative-free Optimization via Evolutionary Algorithms Guiding Local Search (EAGLS) for MINLP. *Pacific Journal of Optimization*.
- Griffin, J., Kolda, T., & R., L. (2008). Asynchronous Parallel Generating Set Search For Linearly-Constrained Optimization. *SIAM J. Sci. Comp.*
- Gropp, W., & Lusk, E. (1996). *User's Guide for mpich, a Portable Implementation of MPI*. Mathematics and Computer Science Division, Argonne National Lab.
- Gropp, W., Lusk, E., Doss, N., & Skjellum, A. (1996). A high-performance, portable implementation of the MPI message passing interface standard. *Parallel Comput.*
- Holland, J. (1975). *Adaption in Natural and Artificial Systems*. University of Michigan Press.
- Holland, J. (1975). Genetic algorithms and the optimal allocation of trials. *SIAM J. Comput.*
- Hough, P., T.G, K., & Torczon, V. (n.d.). Asynchronous Parallel Pattern Search for Nonlinear Optimization. *SIAM J. Sci. Comput.*, 2001.
- Karr, C., & Freeman, L. (1998). *Industrial Applications of Genetic Algorithms*. CRC Press.
- Kocis, G., & Grossman, I. (1988). Global Optimization of Nonconvex Mixed-Integer Nonlinear

- Programming (MINLP) Problems in Process Synthesis. *Ind. Eng. Chem. Res.*
- Kolda, T. (2004). *Revisiting Asynchronous Parallel Pattern Search*. Livermore, CA: Sandia National Labs.
- Kolda, T., Lewis, R. M., & Torczon, V. (2006). Stationarity results for generating set search for linearly constrained optimization. *SIAM J. Optim.*
- Lewis, R., Shepherd, A., & Torczon, V. (2005). *Implementing generating set search methods for linearly constrained minimization*. Williamsburg, VA: Department of Computer Science, College of William & Mary.
- Lobo, F., Lima, C., & Michalewicz, Z. (Eds.). (2007). *Parameter settings in evolutionary algorithms*. Springer.
- Matott, L., Bartlett-Hunt, S., & Rabideau, A. F. (2006). Application of Heuristic Techniques and Algorithm Tuning to a multilayered sorptive barrier system. *Environmental Science & Technology*.
- Mayer, A., Kelley, C., & Miller, C. (2002). Optimal design for problems involving flow and transport phenomena in saturated subsurface systems. *Advances in Water Resources*.
- Plantega, T. (2009). *HOPSPACK 2.0 User Manual (v 2.0.1)*. Livermore, CA: Sandia National Labs.
- Raidl, G. R. (2006). A unified view on hybrid metaheuristics. *{HM06:} Third International Workshop on Hybrid Metaheuristics*.
- Reed, P., Minsker, B., & Goldberg, D. (2000). Designing a competent simple genetic algorithm for search and optimization. *Water Resources Research*.
- Talbi, E. (2004). A taxonomy of hybrid metaheuristics. *J. Heuristics* 8, 541-564.
- Torczon, V. (1992). *PDS: Direct Search Methods for Unconstrained Optimization on Either Sequential or Parallel Machines*. Houston, TX: Rice Univ.
- Zitzler, E., Deb, K., & Thiele, L. (2000). Comparison of Multiobjective Evolutionary Algorithms: Empirical Results. *Evolutionary Computation Journal*.

RECONSTRUCTION OF CLINICAL WORKFLOWS BASED ON THE IHE INTEGRATION PROFILE “CROSS-ENTERPRISE DOCUMENT WORKFLOW”

Melanie Strasser ^(a), Franz Pfeifer ^(a), Emmanuel Helm ^(a), Andreas Schuler ^(a), Josef Altmann ^(b)

^(a) Upper Austria University of Applied Sciences, Research & Development,
Softwarepark 11, 4232 Hagenberg, Austria

^(b) Upper Austria University of Applied Sciences,
School of Informatics, Communications, and Media, Softwarepark 11, 4232 Hagenberg, Austria

^(a) [{melanie.strasser, franz.pfeifer, emmanuel.helm, andreas.schuler} @fh-hagenberg.at](mailto:{melanie.strasser, franz.pfeifer, emmanuel.helm, andreas.schuler}@fh-hagenberg.at),

^(b) josef.altmann@fh-hagenberg.at

ABSTRACT

Process-oriented workflow management in healthcare is a prerequisite to deliver high quality of care and to decrease treatment costs. Clinical workflow management systems (WfMS) cover the definition, execution and reconstruction of healthcare processes, which aims to identify main causes of high medical costs. Nevertheless, almost all existing clinical WfMS share one common drawback: they build upon proprietary hospital information systems (HIS). This paper presents a HIS independent workflow reconstruction approach based on the IHE integration profile "Cross-Enterprise Document Workflow" (XDW). XDW addresses the management of clinical, cross-enterprise workflows and makes use of a specific document, the Workflow Document. This document tracks every step of a workflow (e.g. ePrescription) including control information as well as input and output documents. Therefore it is ideally suited to reconstruct a fine-grained clinical workflow including associated documents and metadata. Furthermore this approach enables a nominal-actual comparison between a clinical workflow and a process definition.

Keywords: Integrating the Healthcare Enterprise, Cross-Enterprise Document Workflow, Clinical Workflow Management System, Reconstruction of Clinical Workflows

1. INTRODUCTION AND MOTIVATION

In our previous work (Strasser et al. 2011; Altmann and Mayr 2011) we focused on the definition and execution of administrative and clinical processes and the reconstruction of patient pathways based on Integrating the Healthcare Enterprise (IHE) (IHE International Inc. 2011) and Business Process Model and Notation 2.0 (BPMN) (Object Management Group Inc. 2011). First results showed that this approach has some limitations due to insufficient prospects on the part of IHE. First of all, the definition and execution of processes was based on selected IHE integration profiles, such as Cross-

Enterprise Document Sharing (XDS), Patient Demographics Query (PDQ), Patient Administration Management (PAM) and Patient Identifier Cross-referencing (PIX). These integration profiles are well suited to define administrative processes but are not usable in the context of clinical processes. Moreover, the reconstruction of patient pathways was based on PAM audit messages stored in the Audit Record Repository (ARR). Although the entries stored in the ARR are suited to reconstruct a patient's way through a healthcare facility, a complete reconstruction of a clinical workflow failed.

In this work we focus on the reconstruction of clinical workflows based on the Cross-Enterprise Document Workflow (XDW), an IHE integration profile which handles the workflow of documents in a clinical context. The presented approach is part of the research project IHEplorer (IHEplorer 2011), whose main objective is to support hospital operators and clinical process managers with a set of tools to monitor, analyze and visualize clinical transactions and workflows.

2. METHODS

This section describes the essential methods and standards used by the XDW-based process reconstruction approach.

2.1. Integrating the Healthcare Enterprise

Integrating the Healthcare Enterprise (IHE) is an international initiative by healthcare professionals and industry to improve the integration and interoperability of medical information systems with standardized descriptions of medical use cases and the systematic use of well established communication standards like Health Level 7 (HL7) and Digital Imaging and Communications in Medicine (DICOM).

IHE issues technical guidelines called integration profiles that describe clinical use cases with actors, which represent software systems or software components, and standard-based transactions,

representing the communication between IHE actors. Integration profiles provide instructions for software manufacturers to develop interoperable software systems (IHE International Inc. 2011).

2.2. Cross-Enterprise Document Workflow

The IHE integration profile Cross-Enterprise Document Workflow (XDW) focuses on the management of cross-enterprise clinical workflows and makes use of a specific document, the Workflow Document. This document administrates all documents related to one clinical workflow and handles the changing of document states (Zalunardo and Cocchiglia 2011).

The Workflow Document is a structured document, characterized by tasks, representing a single step in the workflow. Every task results in the creation of a new document or represents a document state change. Each task has the same structure, based on three elements (see Figure 1): control information, input and output data. The control information element contains metadata needed to describe the specific step (e.g., author, date and time, organization). The input could be data or references to documents needed to perform the current step. The output of a task is a reference to one or more documents created during this step.

The structure of the Workflow Document is kept general and extensible to take account of further use cases. The current revision of the XDW integration profile describes a large number of different use cases to cover as much scenarios as possible. The use cases are simplifications of real life scenarios, e.g. ePrescription and eReferral.

The XDW integration profile is currently submitted for public comment and is a supplement to the IHE IT Infrastructure Technical Framework 7.0 (IHE International Inc. 2010).

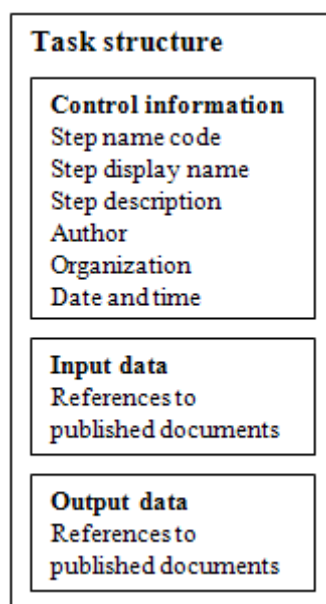


Figure 1: The structure of a task in the XDW Workflow Document

3. RECONSTRUCTION OF CLINICAL WORKFLOWS BASED ON XDW

This chapter presents a novel clinical workflow reconstruction approach based on the IHE integration profile XDW. The functionality can be used in addition to the PAM-based reconstruction approach (Strasser et al. 2011) or standalone. Finally, the reconstructed clinical workflow can be compared to an existing BPMN 2.0 process definition.

3.1. Process definition and execution

The initial point of the presented reconstruction approach is a process definition based on BPMN 2.0, describing a clinical pathway which in turn is executed on a WfMS. Each action performed during process execution creates a new entry in the XDW Workflow Document according to the structure presented in section 2.2. The execution of a process leads to a comprehensive record of actions performed.

As the main field of application of the XDW integration profile are document-based workflows such as eReferral or ePrescription, the structure of the XDW Workflow Document contains references to other documents created during process execution as well as information about the authors and their organizations.

A typical example of an action which creates a new entry in the Workflow Document is the creation of an electronic prescription (see Figure 2). A prescription depends on clinical information which is often provided by means of one or more documents. Therefore, the input section of the appropriate task in the Workflow Document is used to store references to the according input documents. Moreover, the result of the prescription placement task is a new document. This information is stored in the output section of the according task.

For better comprehension and traceability of the approach this paper exclusively focuses on the IHE pharmacy process ePrescription which is described and illustrated with a sequence diagram in the XDW integration profile (Zalunardo and Cocchiglia 2011). Figure 2 shows the ePrescription use case as BPMN 2.0 process definition.

3.2. Workflow reconstruction

Workflow Documents are updated every time a new task is executed, so all documents created during a patient's treatment are referenced in the input- and output section of a task. The tasks in the Workflow Document can be sorted chronologically by using date and time of the control information.

Due to the fact that XDW Workflow Documents are well-formed and valid XML documents, it is possible to use a standard mechanism to display and transform the documents. Extensible Stylesheet Language Transformation (XSLT) is a declarative, XML-based language used for the transformation of XML documents (W3C 2007).

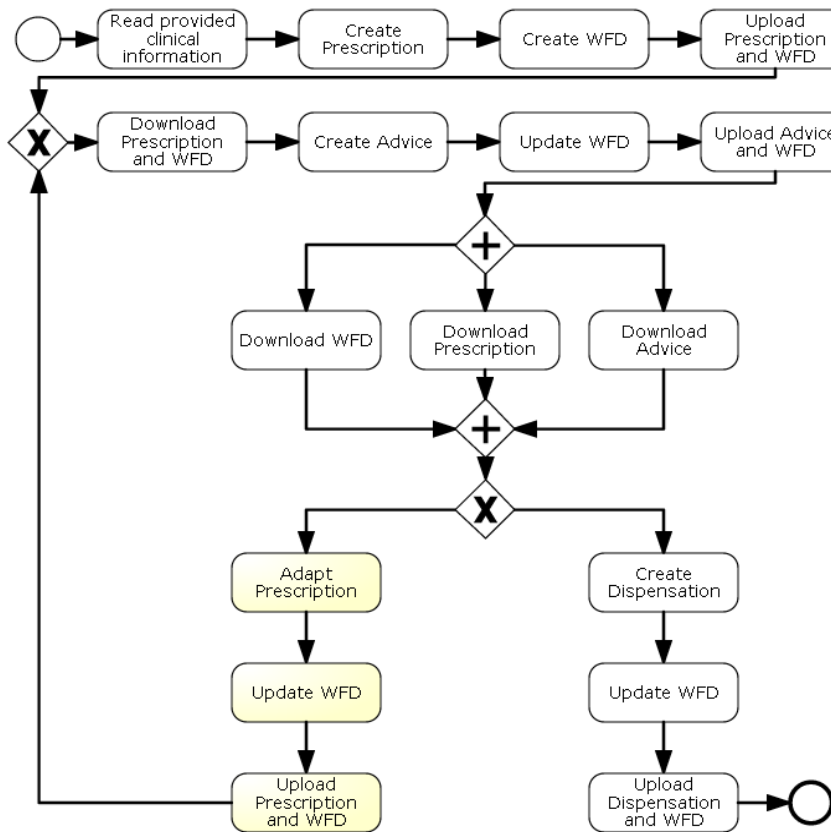


Figure 2: The ePrescription use case as BPMN 2.0 process definition

In this context XSLT is used to transform the XDW Workflow Document into a BPMN 2.0 process (see Figure 3).

The transformation process relies on the correct implementation of the Workflow Document. Each task in the Workflow Document (e.g. Create Advice) is mapped to multiple BPMN 2.0 activities; there are one or more activities to download clinical documents (e.g. Download Prescription and Workflow Document), one or more activities to create or update clinical documents (e.g. Create Advice) and finally the upload of the modified documents (e.g. Upload Advice and Workflow Document). The activities are consecutively numbered to guarantee unique identifiers. For better readability it is important to define meaningful display names in the tasks code system, because these names are used to title the activities in the reconstructed process. In terms of codes and display names, there is still a huge amount of work to be done by the XDW technical committee, as there is currently no common code system available.

Depending on the author of a task we distinguish between user tasks and service tasks. On the one hand an author can be human like a nurse or a doctor, so the task is obviously an user task, but on the other hand tasks may also be executed by a program e.g. a HIS, which identifies a task as a service task. To access the underlying IHE infrastructure with service tasks,

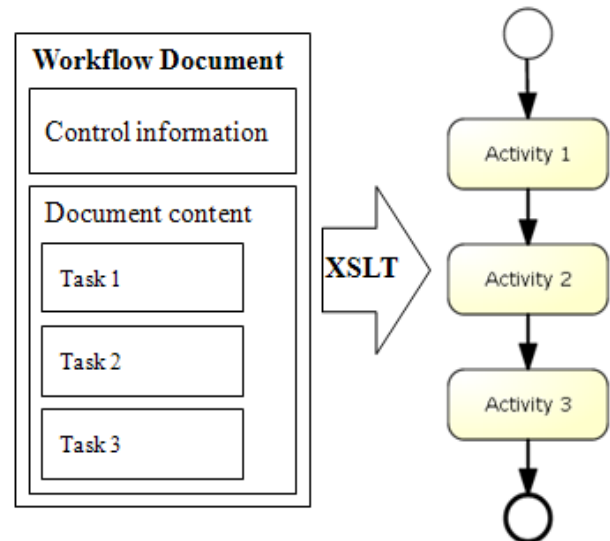


Figure 3: XSLT transformation to a BPMN 2.0 process

existing actors of the Open Source framework Open Health Tools (Open Health Tools Inc. 2011) are used.

To meet the requirements of the BPMN 2.0 standard definition at least two events must be implemented. The start and end events enclose the entire sequence flow. All activities between them are ordered chronologically in a straight line (see Figure 4).

Since a reconstructed BPMN 2.0 document represents only one patient pathway, it is linear and without any gateways.

The reconstructed patient pathway can be visualized in any BPMN 2.0 editor, because it is available in a standard format.

Furthermore functionality for actual-theoretical comparison of BPMN 2.0 documents is provided.

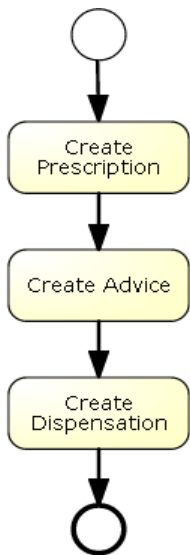


Figure 4: Reconstructed ePrescription process

3.3. Process comparison

In order to compare an document-based, clinical workflow with an existing, predefined BPMN 2.0 process definition one needs to have a tool to match two BPMN 2.0 documents. Currently there is no such tool available, so we decided to implement one on our own.

The tool enables a nominal-actual comparison, by examining all the tasks listed in an existing process definition and comparing them with the display names given in the Workflow Document. Subsequently the resulting information can be used to emphasize the actual workflow in the process definition.

Figure 5 shows the ePrescription process definition with the highlighted reconstructed workflow.

4. RESULTS

The active cooperation with L. Zalunardo, the main author of the XDW integration profile, resulted in the creation of a conceptual design of the HL7 v3 CDA Workflow Document. The Workflow Document draft was introduced to the latest version of the XDW integration profile.

The Workflow Document is ideally suited for document-based workflows, tracking document states, like document creation and document updates. All workflow steps are summarized in one standard-based, structured document. Moreover, the current status of a clinical workflow can be determined with the Workflow Document.

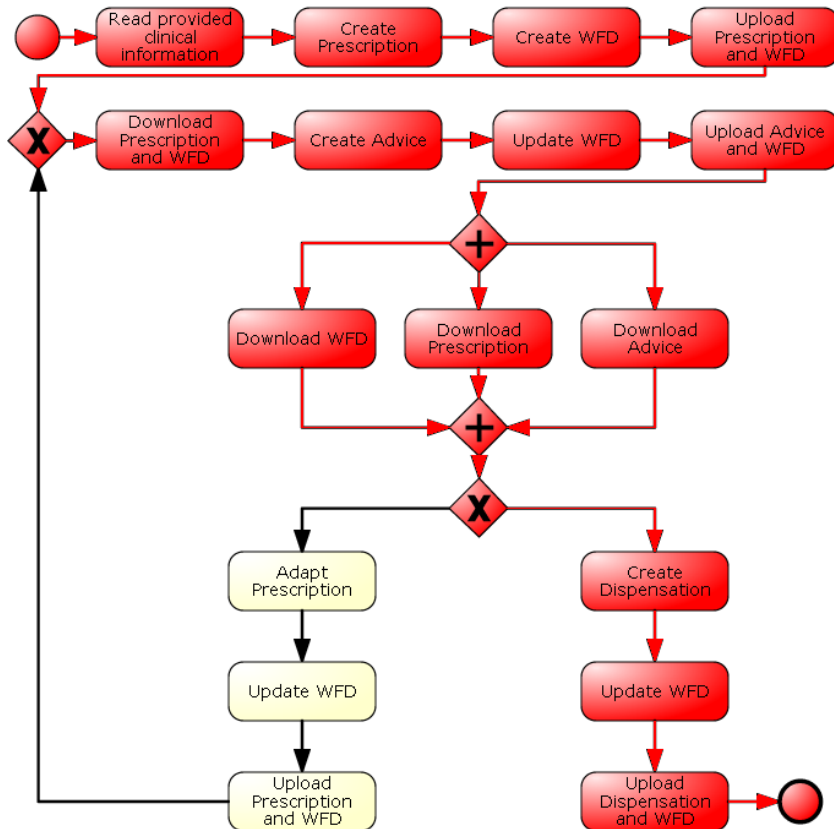


Figure 5: A reconstructed BPMN 2.0 process definition based on a XDW Workflow Document

A further result is the detailed workflow reconstruction based on the Workflow Document (see section 3.2) which can be used as an extension to the previous process reconstruction work. As every task in the Workflow Document contains control information as well as input and output documents, a fine-grained, document-centered process reconstruction is available.

As reconstructed workflows are available in the standard format BPMN 2.0, new possibilities for further processing open up. This approach, for example enables a nominal-actual comparison based on two BPMN 2.0 documents, e.g. a clinical workflow and a clinical process definition (see section 3.3). The result of the calculation, another BPMN 2.0 document, highlights the executed workflow in the process definition. This functionality enables delta analysis, as there are often discrepancies between the defined and the executed process.

5. CONCLUSION

The presented approach shows the reconstruction of clinical workflows based on the IHE integration profile XDW. In the following, we want to point out the advantages of this new approach in comparison to the previous PAM-based reconstruction approach.

Both approaches are hospital information system independent, because the reconstructions build upon IHE integration profiles. Certainly both approaches can be combined to receive further details about an executed workflow.

The PAM-based reconstruction approach shows some limitations, mainly because IHE doesn't provide enough integration profiles to reconstruct comprehensive, clinical workflows. The existing integration profiles are restricted to the definition of administrative healthcare processes. Furthermore the PAM-based approach can't reconstruct holistic workflows without using the event logs of the WfMS. The assignment of clinical documents to a certain workflow is a time-consuming calculation and sometimes infeasible because of multiple parallel patient workflows.

In contrast to the PAM-based reconstruction approach, XDW enables a complete reconstruction of a clinical workflow with an unique assignment of clinical documents to a specific workflow step. Furthermore the presented approach allows a fine-grained reconstruction, because the Workflow Document describes every workflow step in a high level of detail with control information and associated input and output documents. Moreover the reconstructed workflow is available in the standardized format BPMN 2.0.

Next to the reconstruction, functionality was developed to make a comparison between an executed workflow and a process definition. To achieve further information about a certain workflow, the PAM events in the ARR and the tasks in XDW Workflow Document might be combined.

ACKNOWLEDGMENTS

The project IHE Explorer is funded by the Austrian Research Promotion Agency, Tiani-Spirit GmbH, OÖGKK, X-Tention Informationstechnologie GmbH and the OÖ Gesundheits und Spitals AG.

REFERENCES

- Altmann, J. and Mayr, H., 2011. *e-Health: Die IT-Basis für eine Integrierte Versorgung*. Linz: Wagner Verlag; Auflage 1.
- IHE Explorer, 2011. Research Project of the Upper Austria University of Applied Sciences. Available from: <http://ihexplorer.fh-hagenberg.at> [June 2011]
- IHE International Inc., 2011. *Integrating the Healthcare Enterprise*. Available from: <http://www.ihe.net> [June 2011]
- IHE International Inc., 2010. IHE IT Infrastructure (ITI) Technical Framework, Volume 1 (ITI TF-1) Integration Profiles. Available from: http://www.ihe.net/Technical_Framework/upload/IHE_ITI_TF_Rev7-0_Vol1_FT_2010-08-10.pdf [June 2011]
- W3C, 2007. *XSL Transformation (XSLT) Version 2.0*. Available from: <http://www.w3.org/TR/2007/REC-xslt20-20070123/> [June 2011]
- Object Management Group Inc., *Documents Associated With Business Process Model and Notation (BPMN) Version 2.0*. Available from: <http://www.omg.org/spec/BPMN/2.0> [June 2011]
- Open Health Tools Inc., 2011, *Improving the world's health and well-being by unleashing health IT innovation*. Available from: <http://openhealthtools.org/> [June 2011]
- Strasser, M., Pfeifer, F., Helm, E., Schuler, A., Altmann, J., 2011. Defining and reconstructing clinical processes based on IHE and BPMN 2.0. Accepted as full paper at the XXIII International Conference of the European Federation for Medical Informatics. August 28-31, Oslo.
- Zalunardo, L. and Cocchiglia, A., 2011. *IHE IT Infrastructure (ITI) Technical Framework Supplement, Cross-Enterprise Document Workflow (XDW)*. Available from: ftp://ftp.ihe.net/IT_Infrastructure/iheitiyr9-2011-2012/Technical_Cmte/Profile_Work/XDW/IHE_XDW_v16_23_03_11_draft.doc [June 2011]

AUTHORS BIOGRAPHY

Melanie Strasser finished her studies in „Information Engineering and Management“ at the Upper Austria University of Applied Sciences in 2009. Since her master thesis in 2008 she is a scientific researcher at the Research Center Hagenberg. Her research focuses on eHealth and IHE.

Franz Pfeifer is a research associate at the Upper Austria University of Applied Sciences, Campus

Hagenberg. His research interests are medical informatics and digital image processing.

Emmanuel Helm is a researcher at the Research Center Hagenberg and a student at the Upper Austria University of Applied Sciences. His research work concentrates on IHE focusing on the IHE compliant transmission of discrete and continuous data.

Andreas Schuler finished his studies in “Software Engineering” at the Upper Austria University of Applied Sciences in 2011. Since September 2010 he is part of the research project IHEplorer and his work focuses on healthcare process management.

Josef Altmann is head of the department „Communication and Knowledge Media“ of the Upper Austria University of Applied Sciences. Moreover, he is head of the research project IHEplorer. His research interests are in the area of component and service oriented software development as well as in the area of data and information integration.

A METHODOLOGY FOR DEVELOPING DES MODELS: EVENT GRAPHS AND SHARPSIM

Arda Ceylan ^(a), Murat M.Gunal ^(b)

^(a) Institute of Naval Science and Engineering
Turkish Naval Academy, Tuzla, Istanbul, Turkey

^(b) Department of Industrial Engineering
Turkish Naval Academy, Tuzla, Istanbul, Turkey

^(a) aceylan@dho.edu.tr ^(b) mgunal@dho.edu.tr

ABSTRACT

In this paper, a methodology for fast development of Discrete Event Simulation (DES) models is presented. The methodology simply works in two stages. In the first stage the modeler builds a Conceptual Model (CM) of the system to be modeled. A CM is represented as an Event Graph (EG). EGs are used to document the events and their causes in a system. In the second stage the CM is translated to an Event Based DES model. To fulfill this task we developed a DES library, SharpSim, using C# (CSharp) programming language. This paper gives an introduction to our methodology. We provide an insight into SharpSim and EGs, and illustrate a modeling example.

Keywords: Discrete Event Simulation Library, Event Scheduling, Event Graph, Simulation Software.

1. INTRODUCTION

Due to its popularity in simulation world, plenty of Discrete Event Simulation (DES) software has been developed and this trend is likely to continue in the future. There are two main types of simulation software; those which aim at non-programmers (e.g. a graphical user interface which provides drag and drop facilities to build a model by simple mouse clicks) and others which require programming skills (e.g. extending a given source code library to write a full program). First type of software is Commercial-Off-The-Shelf (COTS) such as Arena, Simul8, and Flexsim and they reach a wider user community than the other type does. It is in fact for this reason why the first type dominates the market. The obvious difference of the two types is the user friendliness; one requires the knowledge of how the software is used, and the other requires special expertise, e.g. programming. It is noteworthy that it is dangerous to strictly separate the two types since most COTS software today provides limited programming features.

In either type of simulation software, a DES is approached by a variety of worldviews. A worldview is described as a “modeling framework that a modeler uses to represent a system and its behavior” (Carson,

1993). Simulation software adopts one of these worldviews. DES worldviews in the literature can be categorized into:

- Process Interaction
- Activity Scanning
- Three Phase
- Event Scheduling

Process Interaction focuses on processes which can be described as “set of events” (Rooder, 2004). In this approach, entity flows play the main role where flows include all states of objects. The process is described as “a time-ordered sequence of events, activities and delays that describe the flow of a dynamic entity through a system” (Carson, 1993). Process Interaction is popular and widely used since it is easier to conceive and implement, but “deadlock problem” (Pidd, 1998) stands as the weak point. This approach is used commonly by COTS simulation software, such as Automod. Another common approach, Flow Transaction, is a derivative of Process Interaction. Arena, ProModel and Witness are some of popular software using this approach (Abu-Taieh and Sheikh, 2008).

In Activity Scanning, all activities are scanned in each time step and initiated up to their conditions. It is also called as Two Phase. Three Phase approach is a variant of Activity Scanning. It is more tedious to model, but faster since only conditional activities are scanned at each step.

Event Scheduling requires the identification of events and their impacts on system’s state variables. This approach is most efficient but can be complicated to conceptually represent when the model size is big.

In this paper, we particularly focus on Event Scheduling world-view. As a first step of our interest we review the methods for conceptual modeling, such as Event Graphs (EG). Secondly, we present a new DES library developed in C#: SharpSim. Additionally we give a general idea about some basics of DES and pertinent general purpose DES software in use, and then position the SharpSim in this picture. Finally, we

provide an EG of M/M/n queuing system and a short tutorial on how a SharpSim model can be built.

2. A BRIEF REVIEW OF CONCEPTUAL MODELING METHODS AND EVENT GRAPHS

Conceptual modeling in DES is an active research area and there is still no consensus among simulation modelers on its representation, although Onggo (2009) is an attempt in which unified conceptual modeling is discussed. There are a variety of methods to conceptualize the problems in hand in terms of logical flow of objects and events in the system. We review three methods here.

The first method is the most commonly used; Process Flow Diagrams (PFD). PFDs focus on the flows of entities in a system and are used by most COTS simulation software. A PFD is created by simply placing drag-and-drop objects to represent processes and links between these processes to represent interactions between processes. The modeler, in a way, treats him or herself as an entity and follows the processes which transform an entity.

The second method is Activity Cycle Diagrams (ACD) for conceptualizing the logical flows of objects in the system. In an ACD, life cycle of entities in the system is shown. In their life time, entities changes state and interact with each other. Entity states alternate from active to dead states. Simulation time moves forward and entities of the system spend time in these states. Active states represent activities which different types of entities can cooperate. Once an entity enters an active state, its duration can be determined, generally by taking a sample from a probability distribution. However some conditions must be satisfied for an entity to be in an active state, for example, if there is a server available and there is a client waiting in a queue, a customer entity enters to a service active state. Dead state is the opposite of an active state that is when an entity is idle or waiting for something to happen. This generally means a waiting area. Unlike an active state, duration of a dead state cannot be determined in advance since the time an entity spends in dead state is bound to preceding and succeeding activities.

Finally, Event Graphs (EG) are used to conceptualize a system by focusing on its events. EGs work well with Event Scheduling approach since “Event Graphs are a way of representing the Future Event List logic for a discrete-event model” (Buss, 2001). There are two main components of EG; nodes to represent events and edges to represent transitions between events. Figure 1 shows the basic structure of EG (Roader, 2004).

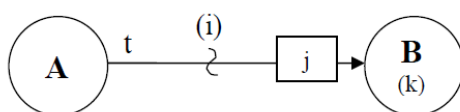


Figure 1: A Basic Event Graph

The translation of the EG in Figure 1 is as follows; “If condition (i) is true at the instant event A occurs, then event B will immediately be scheduled to occur t time units in the future with variables k assigned the values j”.

3. SHARPSIM OVERVIEW

In the second stage of our methodology an Event Graph is translated into computer code to build a simulation model. SharpSim is developed for this purpose. SharpSim is an open-source Discrete Event Simulation (DES) code library developed in C# (The code can be accessed at <http://sharpsim.codeplex.com>). It implements Event Scheduling world-view which involves three main classes; Simulation, Event, Edge and 3 secondary abstract classes; Entity, Resource and Stats. The objects instantiated from these classes are used to implement the EG drawn, as described in the first stage of our methodology. SharpSim is appropriate for multi threading. This is particularly helpful for animation, for example a simulation model running as a thread can communicate with animation classes, e.g. updating screen objects periodically. In this section we briefly explain how SharpSim works.

3.1. Simulation Class

Simulation class is the core of SharpSim. It includes the main Event Scheduling algorithm and the thread that executes the model. Number of replications and seed number for random number generation are the parameters of this class.

There are four properties of simulation class and their descriptions are as follows;

- Future Event List (FEL): This collection involves the set of events that will be executed in the future. An instance of the event that is to be scheduled is inserted into FEL. FEL is sorted by the due time of the events, e.g. earliest event is on top of the list. Note that the event scheduling simulation algorithm scans FEL repeatedly until no events exist in FEL. After the execution of an event, it is removed from the list.
- Clock: The clock variable keeps the simulation time. It is handled in Run method and proceeds to the execution time of the next event.
- Events: This collection involves the set of events instantiated at the beginning of simulation and provides easy manipulation of events. Note that the events of a model are instantiated in the model class that is coded outside of SharpSim library.
- Edges: This collection involves the set of edges instantiated at the beginning of the simulation and provides easy manipulation of edges as in the Events list.

Simulation class includes two main and two supplementary methods. Main methods are described below. The two supplementary methods, Create Events and Create Edges, are useful for reading event and edge details direct from an Excel input file. Events and Edges are instantiated and added to Events and Edges collections.

- Run: The Event Scheduling algorithm is handled in this method. It involves a loop for each replication and another embedded loop for each event in the future event list. The first loop iterates for a number of replication times while the second embedded loop iterates till termination event is executed. In the second loop, first the clock is set to next event's execution time, then event is executed and at the end of the execution it is removed from future event list.

- Start Simulation Thread: This method is used to start simulation thread which is created when a simulation object is instantiated.

3.2. Event Class

Event is an activity which causes a state change. The set of events together with edges forming a system is created at the start of the simulation and according with interrelations among events and edges new events are cloned and added to future event list during simulation.

The constructor of this class has four arguments; event id, event name, priority, and event due time. If an instance of an event class is created with an event due time, the event is inserted to FEL directly. When more than one event has the same execution time, a second parameter is needed to decide which event will be executed first. Priority provides this secondary regulation. It is crucial to assign priorities on events particularly in complex systems. Properties of this class are explained below;

- Execution Time: Each event has an execution time. The execution time of an event is mostly set during the simulation.

- Parameter: This property is used to implement parameter passing on edges in EGs. When an event is executed, a parameter, either a single value such as an integer or an object such as a customer object, can be set into the next event. With this mechanism, for example, individual entities can be transferred from event to event.

- Queue: This property is used to keep the entities that are waiting to be scheduled into the FEL.

There is one method in the Event class, Event Executed, which is a delegate method associated with the next event. This is the point where C# event handling mechanism meets with the simulation's events. When the due time of an event comes this method is called to schedule the next event linked to the current event being executed. The event schedule occurs if the condition on the edge is true. It clones a new independent event from the following event, provides parameter passing between edge and cloned event, and sets its execution time and finally insert cloned event into the FEL.

3.3. Edge Class

Edge is a link between two events. It defines relations between events and accordingly flow of the system. Scheduling of events is decided up to edge conditions. Furthermore, execution times of newly cloned events are set according with edge's next event time value. The constructor of this class has three parameters; name of the edge, source event, and target event. Target events subscribe to source events. There are three properties of the edge class; next event time, attribute, and condition. Next event times can be deterministic or stochastic. Attribute is a variable which is set when parameters are passed between events. Condition is the condition of scheduling an event.

The modeler can create entities and resources by inheriting from the entity and resource classes. Additionally Stats class provides an easy output manipulation for the simulation.

4. M/M/N SERVICE SYSTEM SIMULATION

In this section, we aim to describe how a model of an M/M/n service system can be built using our methodology. As stated earlier, the first stage requires drawing an event graph of the system to be modeled. M/M/n queuing system's event graph is drawn in Figure 2. There are four events and six edges in this graph.

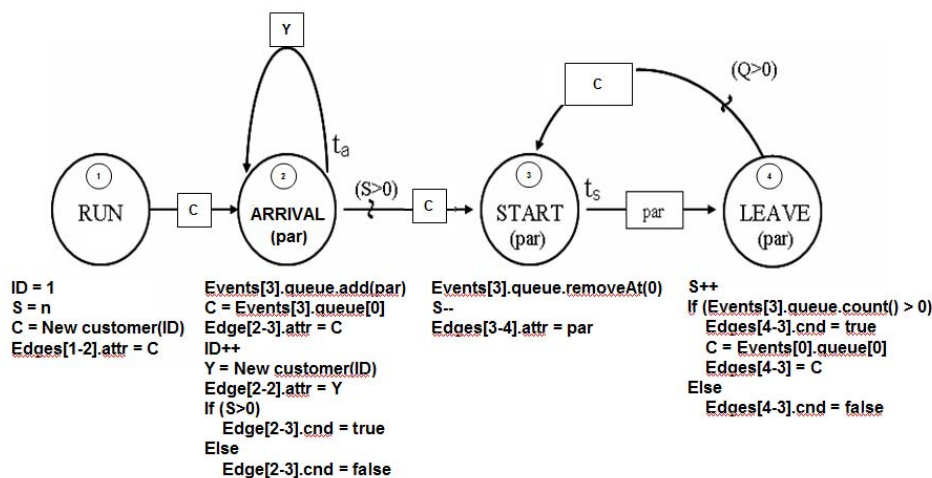


Figure 2: Event Graph of M/M/n

The Event Graph (EG) in Figure 2 has four events which represent start of simulation, arrival of customers, start of service, and end of service events in a queuing system. The variables are ID (arriving customers' ID number), S (number of available servers), and C (Customer Entity).

The explanation of this EG is as follows; When the Run event occurs, set the ID to 1 (first customer's ID), the S to n (there are n servers) and create C (an instance of Customer object). Setting the attribute of Edge [1-2] to C is required to pass the C object to the next event.

When the Arrival event occurs you first need to add the receiving Customer object to the next event's queue. Later you pull a Customer entity from the Start Event's queue and set this to the edge's attribute. Likewise, you need to create a new Customer instance and set it to the self loop attribute. Finally, you need to set the condition on edge between Arrival and Start events based on S (number of available servers).

When the Start event occurs, first customer in the Start event's queue is removed, since it is time for that customer to be served. Number of available servers decremented by one and the receiving customer entity is set as the parameter on the edge. This is to transfer the customer to the leave (end of service) event.

Final event is Leave event. Executing a leave event means that a customer finished the service and therefore number of available servers (S) must be incremented by one and a new Start event must be scheduled. Scheduling a Start event is possible if number of customers waiting in the queue (Q) is non-zero. Q is the queue count of Start event (Event[3]).

4.1. Building a SharpSim model

To build a simulation model in SharpSim, a C# project must be created. You need to create a Windows Forms Application project in a C# compiler such as Microsoft C# Express Edition 2008. The project must include the SharpSim library. SharpSim Library is a DLL file although full source code is provided and can be added to the project.

On the default form in your project, generally named as Form1, a variable of type Simulation, five variables of type Event, and five variable of type Edge are defined. You need to add a button and a richtextbox components on to the form to create a basic user interface. The model will run when the button is clicked and an output will appear in the text area.

On the button's click event, firstly you need to instantiate your simulation model by calling Simulation class's constructor. The constructor has three parameters; track list to show a default output screen, replication number, and seed number. Secondly, events are instantiated. An event has an ID, name, and priority parameters. Note that these simulation events have no due time given since all four events instantiated, and shown in Figure 2, are dynamic event. This means that their due time will be known during the simulation. On the other hand, the fifth event is the Termination event,

which triggers when to stop the simulation, and has a due time. Termination event's due time is the replication length of the simulation.

After creating the events, you need to add State Change Listeners. State change listeners are related to C# event handling mechanism and help connect SharpSim events with C# form events, for example when the Run event occurs in SharpSim, Run method of the form is executed. An instance of Edge is instantiated for every edge in Figure 2. The Edge class has three parameters; name, source event, and target event. The modeller can set the time distribution and the distribution parameters on an edge by setting its ".dist" and ".mean" properties.

After the definitions, a stats collection line can be written, such as the delay time between arrival and leave events. The delay between these two events means the total time of customers in the system. And finally, the run method of the simulation instance is called which causes the simulation to start.

4.2. State Change Handlers

When a SharpSim event occurs, its corresponding method in the form is also executed. These methods are coded in the model file and inside these methods there are state change related codes, such as incrementing state variables and creating new entities.

Inside a state change handler, it is essential to write code inside "evt.EventExecuted += delegate(object obj1, EventArgs e){ ...}" block. For example for the Run event in Figure 2, write a "public void Run" method and inside the method write the delegate line and then set the ID variable to 1 which means that the very first arriving customer's ID will be 1, set the S variable to 2 which means that we have initially 2 servers, create a new customer instance, and set the edge between Event 1 and 2 parameter value to this new customer.

For every event in the model, there must be a State Change handler method in the code.

4.3. Running the model

After buiding a SharpSim model as described above, the project is built and run. Clicking the button on the default form will start the simulation. This means that an instance of Simulation class, instances of events and edges will be created. Since the starting event is Run and scheduled to time 0, the model will start with this event. Execution of Run event will then cause scheduling an arrival event and in turn new arrivals will be created and so on.

In the text area, text outputs will appear as simulation runs. Note that any message, simulation related or not, can be written to the text area on the form inside the state change handlers.

5. CONCLUSION

In this paper we presented a methodology for building DES models. The methodology incorporates Event Graphs, as a conceptual modeling tool, and SharpSim, a new Discrete Event Simulation (DES) library. The DES library, SharpSim, is created using C# language and allows modelers to build DES models by programming in C#.

The SharpSim library is an implementation of event scheduling simulation approach. It aims at translating event graphs into simulation models easily. SharpSim is open source and can be downloaded at <http://sharpsim.codeplex.com>. The website also includes tutorials on modeling examples.

REFERENCES

- Abu-Taieh, E.M.O., El Sheikh, A.A.R. (2008). Methodologies and Approaches in Discrete Event Simulation
- Abu-Taieh, E.M.O., El Sheikh, A.A.R. Commercial Simulation Packages: A Comparative Study
- Buss, A. (2001). Technical Notes, Basic Event Graph Modeling.
- Carson, J.S. 1993. Modeling and Simulation Worldviews.
- Pidd, M. (1998). Computer Simulation in Management Science. Chichester, UK: John Wiley & Sons.
- Roeder, T.M.K. (2004). An Information Taxonomy for Discrete Event Simulations
- Rossetti, M. D. (2008) "JSL: An Open-Source Object-Oriented Framework for Discrete-Event Simulation in Java", International Journal of Simulation and Process Modeling, vol. 4., no. 1, pp69-87, DOI: 10.1504/IJSPM. 2008. 020614
- Evans, W.A., 1994. Approaches to intelligent information retrieval. *Information Processing and Management*, 7 (2), 147–168.
- Schruben, L. (1983). Simulation Modeling with Event Graphs. Communications of the ACM, Volume 23, Number 11.
- Tocher, K. D. 1963. The art of simulation. London. English Universities Press.
- Onggo, BSS. 2009. Towards a unified conceptual model representation: a case study in healthcare. Journal of Simulation, 2009, 3, 40-49.

AUTHORS BIOGRAPHY

Arda Ceylan has received his MSc Naval Operations Research degree at the Institute of Naval Science and Engineering in Turkish Naval Academy.

Murat Gunal is an assistant professor at the department of Industrial Engineering in Turkish Naval Academy.

REVISITATION OF THE SIMULATION METHODOLOGIES AND APPLICATIONS IN MANUFACTURING

. Radha Ramanan^a and Ihsan Sabuncuoglu^b

a. Assistant Professor, Mechanical Engineering Department, National Institute of Technology Calicut, Calicut – 673 601. Kerala, India

b. Professor, Department of Industrial Engineering, Bilkent University, Ankara, Turkey

^(a)radha_ramanan@nitc.ac.in

^(b)sabun@bilkent.edu.tr

ABSTRACT

Manufacturing is one of the largest application areas of simulation. For the purpose of understanding where, how and why the simulation is used in the manufacturing, this survey classifies the manufacturing system into two broad areas viz. manufacturing system design and manufacturing system operations. The two broad areas are further subdivided for this study. The survey discusses the evolution of the subdivisions before detailing the need of simulation in each of the sub divisions of the manufacturing systems. Finally, a discussion is made in order to understand where the research is heading for and identifying the future directions.

Keywords: simulation, manufacturing system design, manufacturing system operations

1. INTRODUCTION

Simulation involves the development of descriptive computer models of a system and exercising those models to predict the operational performance of the underlying system being modeled. Simulation has been one of the most widely used tools for manufacturing (Banks *et al.* 2005).

The basic components of manufacturing include product design, manufacturing/production, planning and control. The product design functions include, conceptualization, function identification, modeling and CAD, material selection, design for manufacturing and dimension and tolerance setting. The manufacturing operation includes processing, assembly, material handling, inspection and test. The planning function includes material requirement planning, capacity planning, process planning. The control function includes production scheduling, inventory control and tool management. For the sake of convenience, in this paper we consider the production, planning and control components all together as manufacturing operations.

Exact solutions are available for most of the manufacturing systems. In spite of it, simulation remains as a widely used tool in finding a solution to a problem. This paper focuses on the application of simulation technology to *manufacturing system design and manufacturing system operations*. System design generally involves making long term decisions such as facility layout and system capacity/configuration. As such, models are typically created and used for single design exercise, and model run time is not a significant factor during the simulation process (Smith, 2003). On the other hand *manufacturing system operations* focus on day-to-day activities within the company and are typically made by lower-level managers. Decisions made at this level help to ensure that daily activities proceed smoothly and therefore help to move the

2. MANUFACTURING SYSTEM DESIGN

In general, manufacturing system design problem (MSDP) encompasses the problem of facility location, plant layout, materials handling system design, assembly line balancing, and other ancillary functions necessary for the production of products. We discuss below the above sub divisions in detail.

2.1 Location Problems

The generic term of facility is used to denote a large variety of entities such as warehouses, plants, antennas, hospitals and other industrial or public structures. The problem is to choose a set of points where these facilities are located so that the sum of location costs and transportation costs are minimized and satisfy the needs of all or part of the customers. The complexity stems from a multitude of qualitative and quantitative factors influencing location decisions as well as the intrinsic difficulty of making trade-offs among those factors. In general, the location problems are formulated as un-capacitated facilities location problem or simple facility location problem and capacitated facility location problem.

The optimization problems defined above are mainly handled by deterministic and static approaches and these studies resulted in a number of valuable contributions to the area. There are a few studies such as, Hidaka and Okano (1997), Kurt and Scott (2007) that utilize simulation to investigate these trade offs. In general, the researchers employ the simulation tool to understand what if scenario. Simulation is used here either because data are not available or because of interactions that exist among many variables, such as customers, warehouse location, delivery time, transportation cost, fixed costs etc. involved in the decision making process.

2.2 Facility / plant Layout

Facility/plant design applications may involve modeling many different aspects of the production facility, including equipment selection/layout, control strategies (Push pull logic), material handling design, buffer sizing, etc. In general, the overall objective of facilities design is to get the inputs (material, supplies, etc.) into, through and out of each facility in the shortest time practicable, at acceptable cost. The material flow pattern becomes the basis for an effective arrangement of physical facilities. The facility layout problem is either formulated as a static layout problem or dynamic layout problems with optimizing the transportation or material handling cost as the primary objective. With this objective in mind different mathematical models have been proposed in the literature (Amine et al. 2007 and Balakrishnan and Cheng, 1998). Simulation has found a large number of applications in the facility layout problems. Specifically, it is used for better understanding and visualizing the complexity of the problems as well as evaluating the system performance for alternative layouts. The complexity increases with increasing number of planning periods, stochastic flow patterns, stochastic demand patterns, unequal size of facility, different product mixes, etc. Some of the simulation studies that are found in the literature are Greasley (2008) and Harrell and Gladwin (2007). In these studies simulation is predominantly used as an interactive modeling and analysis tool to measure the performance of the system in terms of the work-in-process, bottlenecks, routing complexity, the machine setup, machine down time, capacity etc.

2.3 Material handling System Design

The material handling system includes two highly inter-related sub-problems: (i) design of the material flow network that provides the resource interconnections; (ii) sizing of the transporter fleet, and allocation of the inter-group moves to these transporters. Sub-problem (ii) vehicle routing problem.

The stochastic nature of some of the input factors such as demand, processing time, material flow, production schedules, travelling time, etc does not only increase the complexity of the problem but also necessitates a need for simulation or other stochastic optimization tools. Thus the material handling system design offers wide scope for simulation to test the different variables playing crucial role in the design of material handling systems and their interaction effect. The input factors (or variables) required for simulation of the material handling systems may be, the type of material flow, the type of material handling equipment, level of automation, machine schedules, shift patterns, the travelling distance, demand rate, desired throughput rate. Thus the material handling design offers wide scope for simulation to test the different variables playing an important role in the design of material handling system and also its interaction effects.

2.4 Assembly Line Balancing

Assembly lines consist of successive workstations at which products are processed. Workstations are defined as places where some tasks (operations) on products are performed. Products stay at each workstation for the cycle time, which corresponds to the time interval between successively completed units. There are a large number of methods proposed to solve these problems in practice. Bhattacharjee and Sahu (1990) discuss the complexity of the assembly line balancing problem. Some of the factors, such as, work content, cycle time, standard deviation of elemental times, TF-ratio, etc., which are responsible for the complexity of the line balancing problem, are identified and their effect on the complexity of the problem is discussed. Since the problem is NP hard, a number of heuristics are proposed to solve this problem (Sabuncuoglu, Erel and Tanyer (2000)). Boysen, Fliedner and Scholl (2008) provide a classification of ALB problems.

While the ALB problems are generally formulated by static and deterministic models, the stochastic nature of demand, the transport times, processing times, set up times etc. necessitates the tools such as simulation. Su and Lu (2007), Mendes et al. (2005), Bukchin et al. (2002), Hsieh (2002) propose simulation to obtain optimum results for ALB problems. The researchers predominantly use simulation packages to evaluate the performance of the system and identify bottlenecks.

3. MANUFACTURING SYSTEM OPERATION

Operational decisions focus on day-to-day activities within the company and are typically made by lower-level managers. Decisions made at this level help to ensure that daily activities proceed smoothly and therefore help to move the company toward reaching

the strategic goals. Examples of operational decisions include scheduling, handling employee conflicts, and purchasing raw materials needed for production. System operation involves making decisions on a much shorter time schedule. As such, the model is generally used much more frequently, and simulation run time is a more significant factor in the software / package selection and model design process. The classifications made here for the purpose of study are operations scheduling, lot-sizing and operating policies such as push/pull systems.

3.1 Scheduling

Scheduling is the allocation of resources to tasks in order to ensure the completion these tasks in a reasonable amount of time.

The objective of scheduling is to determine the job schedules that minimize (or maximize) a measure (or multiple measures) of performance. Literature has shown that only a few instances of the scheduling problems as polynomially solvable. The majority of the problems are NP hard. Some of the recent literatures (Mejtsky (2007), Metan and Sabuncuoglu (2010)) are reviewed to understand the need of simulation in scheduling. The roles of simulation in these applications are: to test the proposed heuristics in different scenarios or operating conditions, estimate the performance of schedules, identify bottlenecks or critical resources in the schedules, and generate input data for other heuristic or meta-heuristic algorithms to arrive at an optimized objective function values. As stated by Sabuncuoglu and Goren (2009), the future applications of simulation in scheduling still lie in the area of estimation and testing alternative solutions or schedules generated by scheduling algorithms in stochastic and dynamic environments. Simulation will also be used to determine appropriate scheduling or dispatching policies for manufacturing systems. In the recent years, robust optimization and scheduling have become very popular. Simulation has a potential to be used as a surrogate measure in these applications.

3.2 Lot-sizing

The lot size is the amount produced for each machine set up or the aggregate order size. Two very important dimensions of performance relate to inventory levels and customer delivery performance. The objective is to minimize total costs for the planning horizon while satisfying all demands, without backlogging. The literature is replete with a lot of mathematical models right from linear programming, integer programming, branch and bound procedures, dynamic programming, exact formulations like Wagner and Whitin algorithm.

Karimi, Fatemi, and Wilson (2003) describe the eight characteristics that affect the complexity of the lot

sizing problems. There are a number of review papers that study the lot-sizing under different classifications.

The roles of simulation in lot-sizing are: to develop the inputs for the heuristics, understand the bottlenecks, understand the different operating conditions, impact of scheduling, understanding the capacity constraints etc. The complexity is sought to be modeled and a robust approach is made to minimize the impact of uncertainties using simulation. Researchers in the future will concentrate more on doing robust design of the demand and integrating the lot sizing with scheduling as the work in this area is also limited, but the need is highlighted by many researchers.

3.3 Control Logic

In a typical manufacturing system a job moves from workstation to workstation. The control logic for managing this movement through the system can be based on push logic, pull logic or some combination. Special modeling features are required to accommodate each class, additional flexible constructs are required to represent the specific details and exceptions of the lower level control logic.

There are no known available mathematical formulations for the control logic. Simulation seems to be the best way forward to evaluate the performance of the system. The input variables for control logic required may be the set up time, the number of transporters, demand rate etc and depend on the model construct. Enns and Suwanruji (2006) have summarized one group of recent simulation studies comparing replenishment strategies. Time-phased planning, implemented using DRP and MRP logic, continuous-review reorder point and single-card Kanban systems. There are at least two types of performance measure of interest, one related to the inventory level and the other to delivery performance. A tradeoff between these two types of measure exists. Therefore the problem is one of obtaining the desired performance across multiple performance measures (such as inventory level and delivery performance) through the selection of multiple interacting decision variables (such as lot size, reorder point)

Some of the simulation studies that are found are Enns (2007), Jula and Zschocke (2005), Krishnamurthy and Claudio (2005), Treadwell and Herrmann (2005). Simulation is used here to understand the system performances with respect to capacity, storage space, number of transporters and simultaneously collect data for decision making such as prioritizing and routing depending on the replenishment rate based on delivery performance.

Loading of a facility requires complete tracking of all the resources and facilities, tracking of

the schedule of the events to occur, the operator allocation and subsequent delivery of the material. Owing to its complexity of many interacting factor no results can be claimed as optimal. Optimizing a control logic phase of the manufacturing operations has immense scope of future research

4. Discussions and Future Directions

In this paper, the simulation studies in the manufacturing area are analyzed. A fairly comprehensive review is presented for the design and operational problems. The recent developments and applications of simulation are also discussed by identifying the future research directions.

This survey indicates that manufacturing is one of the prime application areas of simulation. At the same time, simulation is one of the indispensable tools for manufacturing. Design problems are usually viewed as tactical or strategic decision problems that contain lots of randomness. Hence, stochastic simulation with appropriate output data analysis is generally required to estimate the long term or steady state performance of the systems. The general purpose simulation languages available in the market place today are quite sufficient to answer the design questions. In these applications, simulation is mostly used in the off-line mode as a stand-alone decision tool to enforce the decisions made by analytical or other models. Since the time is not the main constraint in this decision making environment, computationally demanding simulation optimization procedures can be used to make better decisions. Because the implication of false or incorrect conclusions from a simulation study can be disaster for a firm which has to make long-term design decisions.

In contrast, operational issues span relatively short time horizons. Hence, deterministic simulation (or stochastic simulation with a few random variables) is normally sufficient. Output data analysis and other statistical issues are not the main concern in these applications.

Since for the operational problems, simulation is used as an on-line tool, its integration to the existing decision support system is an important issue. Depending on the type of the application, web-based and/or distributed simulations may also be employed to improve the effectiveness of simulation studies. Virtual reality is also a challenge for the real-time applications of simulation in future studies.

REFERENCES

- Balakrishnan J. and C.H. Cheng, "Dynamic layout algorithms: A state-of-the-art survey", *Omega* 26 (4) (1998), pp. 507–521.
- Banks, J., J. S. Carson, B. L. Nelson, and D. M. Nicol, 2005. *Discrete-event Simulation*, Prentice-Hall, Upper Saddle River, NJ, New Jersey: Prentice-Hall, Inc.
- Bhattacharjee T. K. and Sahu S., 1990. Complexity of single model assembly line balancing problems *Engineering Costs and Production Economics* 18(3), pp 203-214.
- Boysen N., Flidner M. and Scholl A., 2008. Assembly line balancing: Which model to use when? *International Journal of Production Economics*, Volume 111(2), pp 509 – 528.
- Bukchin J., Ezey M. Dar-El and Jacob Rubinvitz, (2002) "Mixed model assembly line design in a make-to-order environment, *Computers & Industrial Engineering*, volume 41(4), pp 405-421
- Enns S.T and Suwanruji Pattita, 2006. Observations on material flow in supply chains. *Proceedings of Winter simulation conference*, Monterey, California December 03 - 06, 2006, Pages: 1446 – 1451.
- Enns Silvanus T. 2007. "PULL" replenishment performance as a function of demand rates and setup times under optimal settings" *Winter simulation conference*, Washington D.C. December 09 - 12, pp. 1624-1632
- Harrell, Charles and Gladwin, Bruce, 2007. Productivity improvement in appliance manufacturing, *Proceedings of Winter simulation conference*, Washington D.C. December 09 - 12, 2007 pp 1610-1614.
- Hidaka, Kazuyoshi and Okano Hiroyuki, 1997. Simulation-based approach to the warehouse location problem for a large-scale real instance, *Proceedings of the 1997 Winter Simulation Conference*, Atlanta, USA.
- Greasley A., 2008. Using simulation for facility design: A case study. *Simulation Modelling Practice and Theory*, 16 (6), pp 670-677.
- Karimi, B. Fatemi Ghomi S.M.T. and Wilson J.M. 2003, The capacitated lot sizing problem: a review of models and algorithms, *Omega* 31, pp. 365–378.
- Krishnamurthy Ananth and Claudio David, 2005. Pull systems with advance demand information, *Proceedings of the 37th conference on Winter*

simulation 2005, Orlando, Florida December 04 - 07, 2005, pp. 1733 – 1742.

Kurt De Maagd and Scott Moore, 2007. Using Classifiers to Solve Warehouse Location Problems”, *Proceedings of the 40th Hawaii International Conference on System Sciences*.

Mendes A.R., Ramos A.L., Simaria A.S. and Vilarinho P.M., 2005. Combining heuristic procedures and simulation models for balancing a PC camera assembly line, *Computers & Industrial Engineering* **49**, pp. 413–431.

Metan, G., Sabuncuoglu, I. and Pierreval, H., 2010, Real time selection of scheduling rules and knowledge extraction via dynamically controlled data mining, *International Journal of Production Research (forthcoming)*.

Mejtsky George Jiri, 2007, A metaheuristic algorithm for simultaneous simulation optimization and applications to traveling salesman and job shop scheduling with due dates. *Proceedings of Winter simulation conference*, Washington D.C. December 09 – 12.

Sabuncuoglu, I., Goren S., 2009. Hedging production schedules against uncertainty in manufacturing environment with a review of robustness and stability research. *International Journal of Computer Integrated Manufacturing* Vol. 22 (2), pages 138 – 157.

Sabuncuoglu I., Erel E. and Tanyer M., 2000. Assembly line balancing using genetic algorithms. *Journal of Intelligent manufacturing*, 11 (3), pp 295-310

Smith, Jeffrey S, 2003. Survey on the use of simulation for manufacturing system design and operation, *Journal of Manufacturing systems*, 22(2), pp. 157-171.

Su, Ping and Lu Ye 2007, Combining Genetic Algorithm and Simulation for the Mixed-model Assembly Line Balancing Problem, *Third International Conference on Natural Computation (ICNC 2007)*.

Treadwel A. I Mark and Herrmann Jeffrey W. 2005. A kanban module for simulating pull production in arena. *Proceedings of the 37th conference on Winter simulation 2005*, Orlando, Florida December 04 - 07, pp. 1413 – 1417.

Bruzzone, A.G., Longo, F., 2005. Modeling & Simulation applied to Security Systems. *Proceedings of Summer Computer Simulation*

Conference, pp. 183-188. July 24-28, Philadelphia (Pennsylvania, USA).

Mercer, P.A. and Smith, G., 1993. *Private view data in the UK*. 2nd ed. London: Longman.

Bantz, C.R., 1995. Social dimensions of software development. In: J. A. Anderson, ed. *Annual review of software management and development*. Newbury Park, CA: Sage, 502–510.

Holland, M., 2004. *Guide to citing Internet sources*. Poole, Bournemouth University. Available from: <http://www.bournemouth.ac.uk> [accessed 15 July 2005]

Das, A., 1992. Picking up the bills, *Independent*, 4 June, p. 28a.

Agutter, A.J., 1995. *The linguistic significance of current British slang*. Thesis (PhD). Edinburgh University.

AUTHORS BIOGRAPHY

T. Radha Ramanan after his graduation in Mechanical Engineering pursued his post graduation in Industrial Engineering and PhD at NIT Trichy. His doctoral dissertation was in the area of Flow shop scheduling applying Artificial Neural Networks. He worked for a few years in industry before joining the academics. His areas of interest are operations management, supply chain management, production planning and control, and technology management. His research interests include sequencing and scheduling, lotsizing, simulation modelling etc. Four of his articles are published in referred international journals such as International Journal of Production Research, Journal of Intelligent Manufacturing, International Journal of Advanced manufacturing Technology. He is a life member of ISTE (Indian Society of Technical Education)

Ihsan Sabuncuoglu is Professor and Chair of Industrial Engineering at Bilkent University. He received the B.S. and M.S. degrees in Industrial Engineering from Middle East Technical University and the Ph.D. degree in Industrial Engineering from the Wichita State University. Dr. Sabuncuoglu teaches and conducts research in the areas of simulation, scheduling, and manufacturing systems. He has published papers in IIE Transactions, Decision Sciences, Simulation, International Journal of Production Research, International Journal of Flexible Manufacturing Systems, International Journal of Computer Integrated Manufacturing, Military Operations Research, Computers and Operations Research, European, Journal of Operational Research, International Journal of

Production Economics, Production Planning, Control, Journal of Operational Research Society, Journal of Intelligent Manufacturing, Computers and Industrial Engineering. He is on the Editorial Board of International Journal of Operations and Quantitative Management, International Journal of Systems Sciences, Journal of Operations Management, and International Journal of Computer Integrated Manufacturing. He is an associate member of Institute of Industrial Engineering and Institute for Operations Research and the Management Science.

Insights into the Practice of Expert Simulation Modellers

Rizwan Ahmed ^(a), Mahmood H. Shah ^(b)

^(a) Lahore Business School University of Lahore

^(b) Lancashire Business School, University of Central Lancashire

^(a) rizwanahmed@uol.edu.pk, ^(b) mhshah@uclan.ac.uk

ABSTRACT

In this study we report the result of an empirical study investigating simulation modelling practices and processes of expert modellers in business and industry. The results suggest that most of the participants do not have a clearly defined or a formal process for developing their models, rather a set of key steps or stages depending on certain contextual factors and personal style. A number of contextual factors such as the problem domain, the scope of the problem, the size and complexity of the model, may affect the way a modeller goes about developing his/her simulation models. Generally a three phased approach is identifiable which can be named as problem definition, model development, and model usage. Model documentation largely depends on model life, client requirement, and type of model being developed. Maintenance and reuse of model is generally not practiced, given most of the models developed are of short to medium term use; however, experience and knowledge is something that is reused.

Keywords: business process modelling, simulation modelling practice, simulation context, simulation modelling process

1. INTRODUCTION

We present the results from an interview study that investigates the practices of business process simulation modellers in order to discover their underlying process of model development. Twenty expert simulation modellers selected from industry and academia described their simulation contexts and practices.

Business process modelling & simulation (BPMS) generally lacks a rich body of literature reflecting on the modelling and simulation practices of modellers in real world. Successful application of modelling and simulation may depend very much on the personal practices of a simulation modeller (Willemain 1994). A huge number of case studies and personal anecdotes of successful application of simulation in different areas of business and industry can be found in simulation and modelling literature, however, little can be found in these studies as to how these modellers go about developing their models and simulation.

Modellers in business and industry develop their models under a variety of constraints and contexts. The contextual factors may have an effect on the way modellers go about developing their models (Robinson 2002, Salt 2006). The problem domain, the scope of the problem, simulation language/technique/package used, the size and complexity of the problem simulated are some of the contextual factors which may affect a modeller's approach to model development. Therefore it is important to reflect on how simulation context and practice relate with each other.

Quite a few surveys have been reported in BPMS literature aiming to explore characteristics of modellers (Murphy and Perera 2001, Hollocks 2002), and practice (Melao and Pidd 2003, Cochran 1995), nevertheless, there is rare accounts of in-depth studies of modelling & simulation practice. These quantitative studies have provided useful indicators to understand characteristics of modellers and their backgrounds, nevertheless, these studies may not provide an in depth view of practice. One of the prominent in depth study of simulation modelling practice has been conducted by Willemain (1994, 1995), that explores the way expert modellers develop their models. Willemain (1994) studies the practices of expert modellers and suggests that practical guidelines for model formulation should be developed for novices in order to become experts. Foss et al (1998) reports a field study of industrial modelling process. Foss et al. (1998); interviewed 10 expert modellers and explored their process of simulation model development and proposed guidelines for improving simulation practice. This study empirically investigates as to how expert modellers develop their simulation models and how their context may affect their simulation practice.

We believe that investigating the practices of expert modeller will enable further understanding of simulation practice and underpin the simulation methodology research.

The paper has been organised in 6 sections. Section 2 gives an overview of the research methodology, Section 3 summarises study participants and their contexts. Section 4 discusses participants' simulation practice and processes, Section 5 provides a discussion on the results and Section 6 concludes the paper.

Table 1: Participants and modelling contexts						
<i>Summary of Education and Professional Roles</i>						
Education summary	PhD	14	Professional role summary	Consultant (C)	9	
	Masters	3		Researcher (R)	5	
	Bachelor	3		C/R	6	
Experience	Avg. Experience	8.5 years				
<i>Summary of Model life, size, complexity, and Modelling Techniques</i>						
Model Life	Short-term	8 (40%)	Long-term	2 (10%)	Long/Short-term	10 (50%)
	Modelling Technique	DE: 8 (40%)	SD: 3 (15%)	Both DE and SD: 9 (45%)		
Size	Small: 3 (15%)	Medium: 14 (70%)	Large: 3 (15%)			
Complexity	Low: 3 (15%)	Medium: 12 (60%)	High: 5 (25%)			
<i>Summary of Types of Models</i>						
Aims of models : Insights, cost and schedule, forecasting, Resource planning, allocation and evaluation Process improvement, Quality assurance, Understanding, Process performance monitoring and measurement, Process design						
Application area: Process change, improvement, and optimisation, Planning, Technology adoption, Project management, Education and training, Project control and operational management						
Problem domain: Safety control systems, Oil and gas pipelines, Mining, Supply chain and logistics, Airport processes, Call centres, Manufacturing, Financial services, Defence (weapons, vehicles), Telecom, Retail, Road and traffic, Health care, Software development processes, Scientific (physical, bioinformatics)						
Key: C=Consultant, R=Researcher, DE = Discrete event, SD = System dynamics, HB = Hybrid models, SB = State based						

2. METHODOLOGY

This study follows a preliminary survey of 17 expert modellers (Ahmed et. al. 2008) which was an adaptation of Willemain's survey. Insights from this survey instigated our interest in exploring the context and practices of expert modellers in depth. The results from survey allowed construction of a framework of ideas, relevant to the context and practices of simulation modellers, explored in this study.

We wanted to study the context and practices of expert modellers in-depth and generally in a structured manner, therefore, we used semi-structured interviewing technique. Answers to the following research were explored with the participants:

RQ1: *What are the modelling contexts of business process simulation modellers?*

RQ2: *What are the modelling practices of business process simulation modellers?*

A pool of interview questions was prepared, consisting of some main open ended questions and several auxiliary questions which were to be asked depending on the flow of interview. A questionnaire consisting of open ended questions was sent to the participants a week prior to conducting the interviews. We also prepared an interview script document, which was used during the interview to ensure a generally uniform way of conducting interviews with all the participants.

We also conducted an intensive pilot study to evaluate the interviewing instrument. This pilot study was conducted in two phases; first, pre-testing the interview questions validity and second, piloting the interview sessions. In the pre-testing, four participants evaluated each question for its understandability and

relevance on an initial draft and questions were improved on the basis of feedback by participants. Piloting the interview sessions with four other participants to evaluate the research instrument helped assessing the appropriateness of the structure and flow of the interview questions. It also helped testing and improving interviewing approach and provided valuable practice for the main set of interviews. The use of audio recording equipment was also evaluated. Moreover, it helped determine the time necessary for interviews.

3. THE PARTICIPANTS & THEIR CONTEXTS

The participants in this study consist of both simulation practitioners and researchers. There are 20 participants in total coming from USA, UK, Germany, Spain and South Africa. Table 1 provides a summary of participants' contexts. A thorough discussion on participants' contexts has been provided in an earlier paper (Ahmed & Robinson 2007), however, here we will provide a summary of their contexts.

The participants consisted of three groups; researchers (R), consultants (C), and researchers cum consultants (C/R); inclusion of both groups gives an insight both into the industry and academia. Table 1 shows that there are 14 participants with a PhD, 3 participants with Master degrees, and 3 participants hold Bachelor degree. This suggests that the participants in this study are highly educated and most of them had some modelling education as part of their professional or research degrees. The average experience of the participants in simulation is 8.5 years. This suggests a high level of simulation experience amongst the participants.

The types of model developed by the participants have been classified with regard to their aims, application area, problem domain, size, complexity, and term of use.

Most the participants develop process simulation models to study, plan, control, and manage the issues of cost, quality, and resources as shown in Table 1. Table 1 shows that they mainly develop simulation models that fall in the application areas of process improvement, process understanding, project planning and management, technology adoption, and project/process control and operational management. Moreover, the participants have developed simulation models in the problem domains of airport processes, passenger flow, cargo, logistics, supply chain management, mining, oil and gas pipelines, call centres, manufacturing, telecom, financial sector, banks, healthcare policy planning, defence, and software development processes.

Table also shows that most models developed by the participants are for short-term use, however, on rare occasions they have also developed models for longer term use. The model's life of use may have an effect on the practices of simulation modellers (Ahmed & Robinson 2007), which will be described in the upcoming sections.

Most of the participants have experience of working both with discrete event and continuous techniques. Only 3 participants have experience of using continuous simulation exclusively while 8 participants have worked exclusively with discrete event simulation. The participants use different tools for developing simulation models; Witness and Extend for discrete event and Vensim for system dynamics are the most popular tools amongst these participants. Participants claim that choice of simulation tool may have a positive or negative effect on the simulation practice of a modeller (Ahmed & Robinson 2007).

They mostly develop simulation models of small and medium size. Also most of the participants develop simulation models of low or medium complexity. Most of the participants also believe that simulation model size and complexity are related, i.e. the bigger the simulation model, the higher the complexity will be, however, some participants also noted that a small model may also be very complex depending on the nature of a problem (Ahmed & Robinson 2007).

4. SIMULATION MODEL DEVELOPMENT PROCESS

In this section we present an analysis of the simulation model development process of the participants. There are 35 themes identified from the interview transcripts which are relevant to simulation modelling processes. Each participant described his/her simulation modelling process at varying levels of detail. Each participant's simulation modelling process has been summarised in a process matrix in Table 2.

Most of the participants described their process in a linear fashion, emphasising that there is always a fair amount of iteration in their process. The main process activities described by the participants are problem communication with the client, defining simulation objectives and questions, problem understanding and analysis, definition of inputs and outputs from the

simulation model, model design, construction, verification and validation, and experimentation.

Table 2 shows that some of the participants tend to use software engineering terms such as requirements, requirements analysis, basic and detailed design, and testing. S2 describes a spiral approach to simulation model development and S8 describes an evolutionary and iterative approach. S4, S5 and S10 describe a process similar to the waterfall model of software development, with steps such as requirements gathering, analysis, design, implementation, and testing (validation and verification). S7 said that he/she has a completely ad-hoc approach to simulation model development with no specific process steps. S3, S4, S5 and S9 described their process in much more detail than the others. S3 and S4 develop highly complex models and S5 develop large models; perhaps this could explain the detailed nature of their process. Also S3 and S4 have experience of working both with discrete event and continuous simulation. S11, S13 and S14 described their process in a highly detailed manner. S15 and S19 described their process at a very low detail.

In Table 2 we summarise findings about the simulation modelling process practice of the participants.

Apparently the simulation modelling process of the participants can be categorised into three phases as Problem Definition, Model Development, and Model Usage and Experimentation. Following we describe findings related to each phase and subsequently some other related themes.

4.1. Modelling Process Phase I: Problem Definition

1. Only three participants mentioned simulation user identification as a step in their process. The user can be the client or some other person in the organisation who needs results from the simulation study. They claim that establishing who the user of the simulation is very important to increasing confidence in the study results. This is because without close interaction with the user, a simulation study may not be of any value to its users. Moreover it is also important to identify the domain or subject matter experts with whom the simulation modeller may need to liaise during the model development.
2. Most of the participants indicate that the identification of simulation goals/objectives and simulation questions is one of their earliest steps in a simulation study.
3. Some of the participants used the term "requirements gathering" while talking about simulation goals and questions. This is perhaps because of their software engineering background.
4. Some participants (S7, S8, S12, S15, S20) do not spend much time on analysis and design, rather they identify simulation goals, gain a basic understanding of the problem and develop a simple and small simulation model straightaway, adding details as they go; a rapid approach.

Table 2: Simulation Modelling Process matrix of the participants

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	
Phase - I: Problem Definition																					
1									X												
2									X					X							
3									X					X							
4				X				X										X			
5	X	X		X				X						X						X	
6	X	X		X		X		X						X						X	
7				X				X						X							
8				X				X						X						X	
9	X	X		X				X						X						X	
10	X	X		X				X						X						X	
11				X				X						X						X	
12	X			X		X		X		X				X						X	
13				X		X		X						X							
14				X		X		X						X							
15				X				X						X							
16	X			X				X						X						X	
17				X				X		X				X							
18				X				X		X				X						X	
19				X				X		X				X						X	
20				X				X		X				X						X	
21				X				X		X				X						X	
22				X			X	X						X						X	
Phase-II: Model Development																					
23				X		X		X		X				X							
24				X		X		X		X				X						X	
25				X		X		X		X				X						X	
26				X		X		X		X				X						X	
27	X	X		X		X		X		X				X						X	
28	X	X		X		X		X		X				X						X	
29	X	X		X		X		X		X				X						X	
30				X		X		X		X				X						X	
Phase III: Model Usage and Experimentation																					
31				X		X		X		X				X						X	
32				X		X		X		X				X						X	
33	X	X		X		X		X		X				X						X	
34				X		X		X		X				X						X	
35				X		X		X		X				X						X	

5. Most of the participants emphasised developing a firm understanding of the problem and capturing the scope of the problem. They talked about identifying the factors contributing to a system/process, understanding relationships between different factors/variables, and confirming those relationships with the client/user.
6. Some of participants emphasised that diagramming methods should be used to illustrate relationships between various factors. This would not only enhance problem understanding but also helps validating the problem understanding with the client.
7. Most of the participants say that identification and definition of inputs and outputs of a simulation model is very important and should be started in the earliest stages of a simulation study.
8. Two participants mentioned conceptual modelling as part of their simulation process. Conceptual modelling in the general simulation literature is said to consist of detailed analysis of the problem and designing the simulation. Analysis would be a detailed account of all the activities performed for problem understanding, identification of variables and the relationship between them. Robinson (2004) defines a conceptual model as, “*a non-software specific description of the simulation model that is to be developed, describing the objectives, inputs, outputs, content, assumptions and simplifications of the model*”.
9. Four participants mentioned checking technical feasibility; i.e. whether simulation is an optimum tool for answering the problem. Moreover, simulation may not be needed to solve certain simple problems; in such cases simulation would prove to be rather an expensive solution.
10. S8, S9, S11, S14, S16, S18, and S20 emphasised on prototyping or building an initial simple abstraction of the whole problem explicitly talked about prototyping. These participants think that building a prototype and then getting feedback from the client helps validate problem understanding and also in checking the feasibility of simulation tool.
11. Only one participant, R6, mentioned planning as a step in the simulation modelling process. S16 generally developed very big and highly complex defence simulation models with a team of people; perhaps this is the reason that he/she mentioned planning as an important step.
12. Simulation tools can positively or negatively impact the efficiency and performance of simulation modellers, according to S6, S7, and S17. None of the other participants mentioned tool selection as a part of their process.

4.2. Modelling Process Phase II: Model Development

13. Only a few of the participants mention simulation model design as part of their process. Only six participants talk about design as a process step;

three of these participants claim to be developing big and highly complex models. The results from our preliminary survey (Ahmed et. al. 2008) indicate that simulation model design is considered to be an issue, however, only a few participants in this study indicate that they do model design any formally. One possible explanation, as mentioned by S2, that the nature of simulation modelling does not require to devise a design prior to constructing the model; because most of the time in the early stages of modelling, neither client nor modeller understand the problem for which the model is to be designed; therefore it is difficult to design a model for which requirements are not clear. Another possible explanation could be that most of the simulation projects developed by these participants are small or medium which take a few days, weeks or months to develop; for such small projects as S8 says, it is not feasible to spend too much time on formally designing the simulation model.

14. All participants talked about building or constructing the simulation model using some simulation tool or programming language. Verification of the model is performed as the model is constructed. Most of the participants say that the whole simulation should not be constructed in one go, rather the validation of the model with the customer should be performed as parts of model are completed. During verification or validation, the modeller may discover some bug or problem with the model and may have to go back to develop further understanding of the problem. Almost all the participants emphasise that a modeller must provide sufficient comments in code or comment boxes while developing the model. This is crucial to understanding the model in case the modeller or some other person has to change the model at some later time.
15. Most participants consider validation and verification as equivalent to evaluation. Evaluation is driven more by customer satisfaction than any other factor. Moreover, some participants refer to model validation and verification in numerous ways such as testing, calibration and validation and verification.

4.3. Modelling Process: Phase III: Model Use and Experimentation

16. Most of the participants explicitly mention experimentation as part of modelling process. They describe that designing the experiments, analysing the results and presenting the results to the client are important tasks for conducting experiments with the simulation models.

4.4. Client contact and rapid development

17. Most of the participants emphasise heavy client contact. It is important to note that those who have emphasised heavy client contact are consultants or

researchers cum consultants. This is perhaps because in a research environment there is usually no client; therefore, the researchers do not mention heavy client contact as an important part of their process.

18. Most of the consultants indicate that in the commercial world it is very important to deliver a solution to the client very *rapidly*; because processes have to be adapted according to changing business need. If a simulation study takes months or years to deliver the results, it may not be of use to the client because during that time the business would have changed even further. Moreover, when the client is spending money on a simulation study, he/she wants to see the results instantly. Therefore, a simulation modeller must involve the client heavily and adapt his/her modelling process according to the client needs in order to deliver the results and recommendation quickly.

4.5. Individual Nature of Simulation Practice

19. All the participants say that they typically develop simulation models alone. However, they have to interact with the client, model users or the domain experts to understand the problem and collect data. Most of the participants say that sometimes they have worked and collaborated with other modellers; however, it seldom happens that they work on the same model concurrently. Only S16 says that he has worked and managed simulation model development where multiple people worked on the same model. However, in that case the project was an enormous defence simulation on which around 200 people worked. In other cases, as for instance S2 and S10 say, they worked with other modellers in a managerial role. S5, S9, S10, S11, S12, S13, say that they have worked on simulation projects in teams; however, in such situations roles such as simulation modeller, data collector, and process-mapper/system-engineer were well defined.
20. The participants give different reasons as to why simulation modellers tend work alone on a simulation study. One reason is that the nature of the simulation problems and the nature of modelling itself that do not require many people to work on the same project. Having more than one person introduces a time overhead because all the people involved have to have a similar level of understanding; S12 says this makes a project inefficient. S13 and S14 believe that having more than one person developing the same model introduces the problem of version/modification control and integration. In the view of S12, S15, and S16, the biggest problem in teamwork is the communication between different team members. S16 states that communication becomes even more problematic if the team members come from

different educational and professional backgrounds.

4.6. Documentation practices

21. Most of the participants think that the best documentation for a simulation model is to put comments in the code or the comment boxes provided by the simulation tool rather than producing formal documents.
22. As shown in Chapter 6 (Table 6.10), most of the participants say that simulation goals and objectives should be clearly stated in the documentation (in comments or in formal documents) and be agreed upon with the client. However, a few of the participants also think that the scope of the model should also be defined in the documentation.
23. Some of the participants recommend that model inputs and outputs should also be defined so that the model can be well understood in future if needed.
24. Some of the participants think that the relationships between data items (inputs and outputs) should also be documented along with an influence/process diagram or using some other diagram methods. An overview of model structure or model working is also necessary to understand the model.
25. Most of the participants say that they produce reports or presentations of the simulation results which are presented to the client. These reports or presentation include the report of experiments, the scenarios and assumptions under which experiments have been run, analysis of results and recommendations from the analysis.

4.7. Others

26. Model reuse for a similar problem is not important for most participants. This is because they think that a model developed at one point in the past may be not depict the real world as it is now; as R3 says "*the business changes so much that the objects become out of date; I wonder if they are updatable*". However, some of the participants mention that the experience and learning gained from simulation projects is reused in subsequent projects. This finding is similar to what is found in literature that reuse in simulation is difficult therefore not much practiced (Robinson et al. 2004). Two of the participants, S4 and S6, mention that they reuse parts of their existing models. However, some participants said that it is the experience that is reused in subsequent simulation projects.
27. Majority of the participants do not emphasize simulation model maintenance. Only S9 explicitly mentions maintenance as part of the process; no one else discuss maintenance as part of their process. This is perhaps because majority of the models developed by the participants are of short-term use.

5. DISCUSSION

These results provide a general picture of a model development practice of the participants and the type models they develop under a variety of contexts.

The results indicate that most of the participants develop their models alone supporting the literature finding of Robinson (2002); however, for relatively larger projects a number of people may be working in different aspects such as problem understanding, data collection, model construction and validation and verification.

Most participants do not produce design prior to constructing their models. A possible explanation as mentioned by one of the participant is that the nature of simulation modelling does not require to devise a design prior to constructing the model; because most of the time in the early stages of modelling, neither client nor modeller understand the problem for which the model is to be designed; therefore it is difficult to design a model for which requirements are not clear. Another possible explanation could be that most of the simulation projects developed by these participants are small or medium which take a few days or weeks; for such small projects as S8 says, it is not feasible to spend too much time on formally designing the simulation model. However, for large models designing prior to model development and adapting the design during development is a must

Maintainability of models is not an issue for majority of our respondents; however maintainability will inevitably become an issue if these models are to be capable of being evolved so that they remain useful in the long term. Our literature review suggests that the maintainability of models has not been given much attention in the general simulation literature; similarly in this study only a few participants indicate that they are concerned about maintainability. Maintenance and documentation are low priority issues. Another potential reason could be that perhaps the simulation models developed are too small (though they say they mostly build medium sized models as we have no agreed measure of size); or large but conceptually too simple to be documented and maintained. Another reason could be that most simulation models may not be used in the long term, therefore documentation and maintenance is not a problem. The participants believe that extent of documentation and maintainability varies in each individual case depending on the contextual factors such as client requirements, budget, time, expertise, and simulation model size and complexity. Issues of simulation model documentation and maintenance are also seldom discussed in the general simulation literature. Foss et al. (1998) say that most simulation models are poorly documented and are therefore rarely reused. The models evolve and are redefined over the period of time, and the managers who use the models may change their minds about priorities. Foss et al. (1998) further state that poor documentation makes it very hard to maintain the models. However, it is generally believed that reusing simulation models is difficult and less cost effective than building a new one

from scratch (Taylor et al. 2004, Robinson et al 2004). On the other hand, the importance of maintenance, reuse, and documentation has been highlighted by Gass(1987) for large scale models.

Most of our participants also suggest that their models are rarely reused, however, the participants from military simulation background say that they emphasise model reuse. This finding supports the view of Salt (2006) where he suggests that defence modellers are obsessed with reuse while civilians do not bother reusing their models. One of the main reasons suggested by the participants is that reusing a simulation model is difficult; because most often simulation model represent a reality in business process at a given time but as the time passes the reality changes therefore an old model of that reality is of little use after the reality changes. However, the knowledge and experience gained from an old model can be reused in a new project. A similar view is held by the authors in (Taylor et al. 2004), Robinson et al 2004).

Gass (1987) suggests that the evaluation of models encompasses both validation and verification activities along with an assessments of the models' quality, usability, and utility. However, the results of our study suggest that this form of evaluation does not have a formal position in the simulation modelling practice of our participants. In general, simulation modelling literature seems to emphasize validation and verification activities; therefore, most possibly modellers consider this to be equivalent to evaluation. However, the extent of evaluation largely depends on contextual factors such as requirements. Another reason for not conducting a holistic evaluation by our participants could be that most models are used by themselves and results are provided to the client. Therefore, evaluating models in the aspects other than validation and verification is not of importance. However, if a model is to be handed over to the client; perhaps, evaluating usability and documentation is given some conscious consideration.

The results from this study also suggest that majority of the participants in this study don't seem to be using a highly defined formal process framework for their simulation modelling practice. However, most of them seem to have some specific steps, perhaps, unconsciously infused in their simulation modelling practice.

Simulation modelling in commercial context involves people, technology and tools. A well-defined process is believed to provide a framework where tools, technology and people collaborate, to enhance productivity and quality (Humphrey and Kellner 1989). Humphrey (1997) states that a good process brings discipline in human activities and improve the quality of software. It is the process that can effectively help engineers to produce high quality products, with reduced time, and control over cost (Cugola & Ghezzi 1998). This suggests that a good simulation modeling process may also improve quality and increase the productivity. However, it is rare to find such studies in simulation modelling literature where relation has been

drawn between simulation model quality, modeller's productivity and use of a disciplined process.

On the other hand, Shannon (1975) says that simulation modelling is both art and science; producing art needs creativity (Kneller 1965), therefore simulation modelling needs creativity. Many simulation modellers believe that simulation is a creative accomplishment and if it is constrained by a process, creativity may suffer (Powell 1995). Paul et al. (2003) says, "*One can instantly see that fixed structure to develop simulation models will not be able to cope with all the situations at all times*". This suggests that consider the context of a simulation modellers is important for applicability of a simulation modelling process.

Simonton (2002) suggests that creativity can be considered a constrained stochastic process; that is creativity is not completely random or stochastic, rather loosely bound in the rules of the domain for which creativity is needed. Johnson-Laird (1988) says that there can be many criteria of creative processes on which a creator may rely; some of those criteria will be common to many practitioners while others may depend on individual aptitude and style. This suggests that creative process does not consist of only stochastic random activities but there is some structure in the creative process. Ferguson et. al. (1997) suggest why discipline is needed alongside creativity:

"In most professions, competent work requires the disciplined use of established practices. It is not a matter of creativity versus discipline, but one of bringing discipline to the work so that creativity can happen."

However, it seems that generally the simulation modellers are more interested in the end product and less in the process of creating that product. In simulation, where the world is driven by time constraints, commercial pressures, and competition, weakness in the modelling process may bring up many issues. Therefore, Gass (1987) suggested:

"We need to get away from the crutch that modelling is an art. Guidelines need to be proposed, methodologies for validation and evaluation need to be formalized and applied; and the concept that modelling is a profession with standards must be brought into education and on-the-job training activities of the coming generation of analysts."

Eriksson (2003) suggest that a model's quality is questionable if it is constructed without a disciplined approach. It can be argued, therefore, that the creative principles of simulation modelling can be incorporated in a disciplined framework for simulation model development. A disciplined simulation modelling process that provides room for creative aspects of simulation is likely to produce good simulation models

efficiently. Therefore if a process consolidated from real world simulation practice of expert modellers may provide discipline for productivity and quality and liberty and flexibility for creativity.

A number of simulation modelling processes have been reported in the literature for example Robinson (2004), Law and Kelton (2000), Shannon (1998), Nordgren (1995), however, they are based on author's personal experience of simulation model development. No such process has been reported in the literature that entails a simulation modelling process based on an empirical study of expert modellers' contexts and practices. It would be interesting to consolidate a process from real world practices of expert modellers and compare it with the processes reported in literature.

6. CONCLUSION

Studying the simulation contexts and practices of experts helped understanding the way they develop their models modellers. Most of the participants do not seem to have a very well defined and a formal simulation modelling process. However, most of them seem to have some key steps or stages in their process of simulation model development. Generally a three phased process has been identified from the participants which can be named as problem definition, model development, and model usage. This study identifies some general trends in the simulation model development practice of the participants. It would be hard to generalise the results across the business process modelling and simulation community, however, it gives us some indication as to how people develop their models when their models are small/medium and their model's complexity is low/medium and when models developed for short-term use.

This study does not provide a uniform view of simulation practice in business and industry but some trends and indications on which future studies can be built to further underpin our understanding of the simulation practice real world. Conducting studies in each niche (e.g. defence, manufacturing, healthcare, retail, logistics etc.) of simulation modelling will help further understanding the state-of-the-art and state-of-practice in discipline specific area. Moreover, in-depth studies of various aspect of simulation modelling process (e.g. problem understanding, model design, documentation) will help understand and improve simulation practice. Furthermore, the findings from this study also encourage us to consolidate a simulation modelling process based on the empirical data collected from expert simulation modellers, which will be reported in future publications.

REFERENCES

- Ahmed, R., Hall, T., Wernick, P., Robinson, S. and Shah, M., (2008). *Software process simulation modelling: A survey of practice*, Journal of Simulation Modelling 2 (2) : 91- 102
- Ahmed, R., and Robinson, S. (2007). *Simulation in Business and Industry: How simulation context can*

- affect simulation practice? Spring simulation (SpringSim) multi-conference, Business and Industry Symposium, 24-29 March 2007, Norfolk, USA
- Arthur J D, Sargent R G, Dabney J D, Law A M, Morrison J D (1999). *Verification and validation (panel session): what impact should project size and complexity have on attendant V&V activities and supporting infrastructure?* Proceedings of the 31st Winter Simulation Conference, pp: 148 - 155
- Banks J. (2001). *Panel Session: Education For Simulation Practice --- Five Perspectives.* Proceedings of the 33rd Winter simulation conference Arlington, Virginia pp. 1571 - 1579
- Chwif, L., Barretto, M.R.P. and Paul, R.J. (2000), *On simulation model complexity.* Proceedings of the 32nd Winter Simulation Conference archive. Dec. 10-13, 449 - 455
- Cochran, J.K., Mckulak, G.T. and Savory, P.A. (1995), *Simulation project characteristics in Industrial settings.* INTERFACES. 25:104-113
- Cugola G, Ghezzi C. (1998). *Software Process: A Retrospective and a Path to the Future.* Software Process Improvement and Practice. Vol. 4:101-123
- David A. (2001). "Model Implementation: A State of the Art." European Journal of Operational Research. Vol. 134:459-480
- Eriksson DM, (2003). *A Framework for the Constitution of Modelling Processes: A Proposition.* European J. of OR, vol. 145, pp202-215
- Ferguson P., Humphrey WS., Khajenoori S. Macke S. and Matvya, A. (1997). *Results of Applying the Personal Software Process.* Computer. Vol. 30(5):24 -31
- Foss, B.A., Lohmann, B. and Marquardt, W. (1998). *A Field Study of the Industrial Modeling Process.* Journal of Process Control, Vol. 8(5/6):325-338
- Gass, S.I. (1987), *Managing the Modelling Process: A Personal Reflection.* European Journal of Operations Research. 31:1-8
- Hollocks, B.W. (2001), *Discrete-event simulation: an inquiry into user practice.* Simulation Practice and Theory, 8:451-471
- Humphrey WS. (1997). *Introduction to the Personal Software Process.* Addison-Wesley Publications, Harlow, UK.
- Humphrey WS. and Kellner MI (1989). *Software Process Modelling: Principles of Entity Process Models.* 11th Int. Conference on Software Engineering 15-18 May.
- Johnson-Laird PN. (1988). *Freedom and Constraint in Creativity, in R.J. Sternberg (ed) The Nature of Creativity: Contemporary Psychological Perspectives.* Cambridge: Cambridge University Press. 1988 pp 202-219
- Kellner IM., Madachy R. and Raffo D. (1999). *Software process simulation modelling: Why? What? How?* J. of Systems and Software, Vol. 46(2/3):91-105
- Kneller G. (1965). *The Art and Science of Creativity.* Holt, Rinehart and Winston Inc. London
- Law, D.R. (1998), *Scalable means more than more: a unifying definition of simulation scalability.* Proceedings of the 30th Winter Simulation Conference, 781-788
- Law AM. and Kelton WD. (2000). *Simulation Modeling and Analysis.* 3rd ed. McGraw-Hill, New York.
- Nance RE. and Sargent R.G. (2002). "Perspective on the evolution of simulation", Operations Research, INFORMS, 50(1):161-172
- Melão N. and Pidd, M. (2003), *Use of business process simulation: A survey of practitioners.* Journal of Operational Research Society 54:2-10
- Nordgren WB. (1995). *Steps for Proper Simulation Project Management.* Proceedings of the 27th Winter Simulation Conference Arlington, Virginia, United States. pp: 68-73
- Murphy, S.P., and Perera, T.D. (2001), *General applications: Simulation practice: key enablers in the development of simulation.* Proceedings of the 33rd Winter Simulation Conference
- Page E H, Nicol D M, Balci O, Fujimoto R M, Fishwick P A, L'Ecuyer P and Smith R (1999). *Panel: Strategic directions in simulation research.* Proceedings of 31st Winter Simulation Conference, pp: 1509-520
- Paul R.J., Eldabi T. and Kuljis J. (2003). *Perspectives on Simulation in Education and Training: Simulation Education is no Substitute for Intelligent Thinking.* Proceedings of the 35th Winter Simulation Conference: New Orleans, Louisiana. pp. 1989 - 1993
- Pidd M. (1996). *Tools for thinking: modelling in management science.* Chichester, John Wiley & Sons Ltd.,
- Powell SG. (1995). *The Teachers Forum: Six Key Modelling Heuristics.* INTERFACES, Vol. 25(4):114-125
- Robinson S. (2002). *Modes of Simulation Practice: Approaches to Business and Military Simulation.* Simulation Modelling Practice and Theory, Vol. 10(8):513-523
- Robinson S. (2004). *Simulation: The Practice of Model Development and Use.* Wiley, Chichester, UK
- Robinson, S., Nance, R.E., Paul, R.J., Pidd, M. and Taylor, S.J.E. (2004), *Simulation Model Reuse: Definitions, Benefits and Obstacles.* Simulation Modelling Practice and Theory. 12:479-494
- Salt, J.D. (2006), *Modes of practice in military simulation.* Proceedings of OR 48, 11-13 Sept. 2006. UK

- Shannon RE. (1975). *Systems Simulation: The Art and Science*. Prentice-Hall.
- Shannon RE. (1998). *Introduction to the Art and Science of Simulation*. Proceedings of the 30th Winter Simulation Conference Washington, D.C., United States pp:7-14
- Simonton DK. (2002). *Creativity as a Constrained Stochastic Process*. in R.J. Sternberg, E.L. Grigorenco, & J.L. Singer (eds) *Creativity: from Potential to Realization* Washington, D.C. American Psychological Association. pp 83-101
- Taylor, S.J.E., Lendermann, P., Paul, R.J., Reichenthal, S.W., Strayburger, S. and Turner S.J. (2004), *Panel on Future Challenges in Modeling Methodology*. Proceedings of the Winter Simulation Conference, 319-327
- Willemain TR. (1994). *Insights on Modelling from a Dozen Experts*. Operations Research, Vol. 42 (2):213-222
- Willemain TR. (1995), *Model Formulation: What Experts Think About and When*. Operations Research, Vol. 43(6): 91

MODELLING RESILIENCE IN CLOUD-SCALE DATA CENTRES

John Cartlidge^(a) & Ilango Sriram^(b)

Department of Computer Science
University of Bristol
Bristol, UK, BS8 1UB

^(a) john.cartlidge@bristol.ac.uk, ^(b) ilango@cs.bris.ac.uk

ABSTRACT

The trend for cloud computing has initiated a race towards data centres (DC) of an ever-increasing size. The largest DCs now contain many hundreds of thousands of virtual machine (VM) services. Given the finite lifespan of hardware, such large DCs are subject to frequent hardware failure events that can lead to disruption of service. To counter this, multiple redundant copies of task threads may be distributed around a DC to ensure that individual hardware failures do not cause entire jobs to fail. Here, we present results demonstrating the resilience of different job scheduling algorithms in a simulated DC with hardware failure. We use a simple model of jobs distributed across a hardware network to demonstrate the relationship between resilience and additional communication costs of different scheduling methods.

Keywords: cloud computing, simulation modelling, data centres, resilience

1. INTRODUCTION

Cloud computing—the online utility provision of hardware and software computing infrastructure and applications—necessitates the demand for data centres (DC) on an ever-increasing scale. The largest now fill purpose-built facilities approaching one million square feet.¹ Already, DCs are so large that manufacturers (including IBM, HP, Sun) do not have the capability to build and destructively test models on the scale of the final production systems. Hence, every day, massively parallel, tightly-coupled, complex and sometimes heterogeneous data centres are put to service having undergone insufficient pre-testing; while it is still possible to test individual node servers and other standalone hardware, the complex interactions between the components of the DC under normal and abnormal operating conditions are largely unknown. Whereas in other engineering domains this problem has been addressed with robust industry-standard simulation tools—SPICE for integrated circuit design (Nagel 1975), or computational fluid dynamics for the aeronautics industry—a well established realistic (rigorous) simulation framework of cloud computing facilities is lacking.

There are two important reasons why this is the case. Firstly, there is no uniform definition of what a cloud computing infrastructure or platform should look like: where Amazon uses virtualization (DeCandia *et al.* 2007), Google uses MapReduce (Dean and Ghemawat 2008). Secondly, it is a hard problem: a realistic simulation tool should include real network models (fibre channel, Gbit ethernet), disk models (disk arrays, solid-state, caching, distributed protocols and file systems), queueing models for web servers, etc. As such, while it is our long-term goal to develop a set of simulation tools that can be used to aid the development of cloud DCs, as an initial step we present a tractable problem using a simplified model.

DCs for cloud computing have now reached such a vast scale that frequent hardware failures (both temporary and permanent) have become a normal expectation. For example, if a DC contains 100,000 servers and the average commodity server life expectancy is 3 years, we expect a server to reach the end of its natural life every 15 minutes; considering temporary failures and failure of other components makes failures occur even more frequently. Thus, when a job is submitted to the cloud, the physical hardware available at the start of the job cannot be guaranteed to be there at the end:

With such high component failure rates, an application running across thousands of machines may need to react to failure conditions on an hourly basis (Barroso and Hölzle 2009)

To avoid frequent job failures, redundancy is necessary. The cloud computing design paradigm builds on achieving scalability by performing scale-out rather than scale-up operations, i.e., increasing resources by using additional components as opposed to using more powerful components. For this reason, jobs are generally split into parallel tasks that can be executed by (potentially) several services. For resilience purposes, the tasks can be multiply copied and run in parallel threads on different hardware (Hollnagel, Woods, and Levson 2006). Thus, as long as a “backup” copy exists, individual task failures will not degrade a job's overall resilience.

However, redundancy inherently generates extra work, requiring more space, greater computational

¹ <http://www.datacenterknowledge.com/special-report-the-worlds-largest-data-centers>

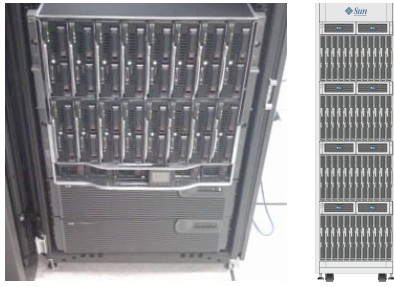


Figure 1: Example hardware: (a) HP C7000 chassis holding 16 blade servers; (b) Sun Pegasus C48 server rack, containing 4 chassis \times 12 blade servers.

effort and increased communication costs. There is clearly a trade off here: how much redundancy and how to schedule redundancy—where to physically locate copies of the same code in the DC to minimise the chances of failure—versus increased communication cost and computational effort.

In this paper, we conduct an initial foray into the analysis of this trade off, using a simple simulation model to analyse the relationships between scheduling, redundancy, network structure and resilience. In Section 2 we introduce cloud-scale data centres and the problem of failure resilience. Section 3 outlines the simulation model we use, before detailing our experimental set-up in Section 4. Section 5 presents the results of our experiments, which are discussed in Section 6. In Section 7 we outline our future plans to extend this work, before summarising our conclusions in Section 8.

2. BACKGROUND

2.1. Cloud Data Centres

Cloud Computing transitions DCs from co-located computing facilities to large resources where components are highly connected and used in an interlinked way. Computations are broken down into services, allowing for easier scale-out operations. From the physical perspective, DCs are structured regularly in a hierarchical design: a warehouse scale DC is made up of aisles of racks, each rack being a vertical frame to which a number of chassis can be mounted; each chassis containing an arrangement of thin computer mother-board units: the blade-servers that make up the DC's computing infrastructure. Each blade server in turn hosts Virtual Machines (VMs) running cloud services. Figure 1 shows example chassis and rack components.

With Cloud Computing, the level of interconnectivity and dependency between services across the DC is so high that Barroso and Hölzle (2009) coined the term “warehouse-scale computers”. This introduces various aspects of complexity to DCs. Firstly, many of the protocols in place scale worse than linearly, making conventional management techniques impractical beyond a certain scale as complex interactions between services lead to unpredictable behaviour. Secondly, DC design has reached a stage

where test environments are no longer larger, or even of the same order of magnitude, as the final products. Cutting edge DCs are believed to have more than half a million cores,² but even one order of magnitude less would make a physical test environment too expensive to be practical. Hence, it is difficult to impossible to test the chosen configurations before going into production, which can lead to costly errors.

This highlights the need for predictive computer simulations to evaluate possible designs before they go into production: with simulation studies it is possible to rapidly evaluate design alternatives. However, for simulating cloud-scale computing DCs there are currently no well-established tools.

The literature includes some early-stage cloud simulation models. For a consumer centric view of the cloud, there is CloudSim (Buyya, Ranjan, and Calheiros 2009). CloudSim's design goal is to compare the performance of services on a cloud with limited resources against their performance on dedicated hardware. To aid the vendor perspective, we have previously developed SPECI (Simulation Program for Elastic Cloud Infrastructures) for modelling scaling properties of middleware policy distribution in virtualised cloud data centres (Sriram and Cliff 2011). This paper explores aspects of resilience modelling that we aim to develop as a component in a set of simulation tools for data centre designers.

2.2. Failure, Resilience and Communication Cost

As economies of scale drive the growth of DCs, there are such a large number of individual independent hardware components that the average life expectancy will imply that component failure will occur continually and not just in exceptional or unusual cases. This expected near permanent failing of components is called *normal failure*. For practicable maintenance, failed components are left *in situ* and only replaced from time to time; it may also be imagined that entire racks are replaced once several servers on it have failed. However, despite normal failure, resiliency must be maintained. Furthermore, the cloud design paradigm of solving jobs using smaller tasks or services that are typically spread across several physical components further increases the risk of normal failure affecting any given job. As cloud vendors seek to provide reliable services, requiring the maintenance of guaranteed levels of performance and dependability, resilience has become a new non-functional requirement (Liu, Deters, and Zhang 2010). To this end, cloud applications such as BigTable, Google's massively parallel data storage application, have in-built management systems for dealing with failures (Chang *et al.* 2008).

Hardware failure can occur anywhere in the physical hierarchy of the data centre: power outages can disable an entire DC; faulty cooling system behaviour can force an aisle to be shutdown to avoid overheating; racks, chassis and blades have individual power

² <http://www.zdnet.com/blog/storage/googles-650000-core-warehouse-size-computer/213>

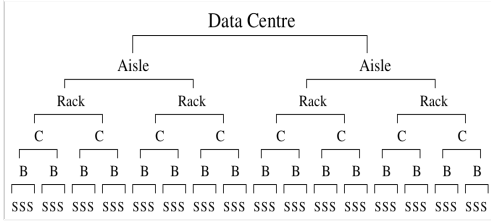


Figure 2: Data centre tree schematic. We describe this as an $h-2-2-2-3$ hierarchy (2 racks per aisle, 2 chassis per rack, 2 blades per chassis and 3 services per blade). The full DC contains as many aisles as necessary.

supplies which can fail; and individual VMs can suffer from instability in software and require an unplanned reboot. Thus, with growing DC scales, resources can no longer be treated as stable; and interactions no longer static but rather exhibiting dynamic interactions on multiple descriptive levels.

To counter normal failure, redundancy must be introduced. This happens by spinning off parallel copies of all tasks. Thus, when any task from the original copy fails, a redundant copy is available to replace the service that has gone “missing”. Hadoop, for example, is an open-source software for reliable, scalable distributed computing and is used by Yahoo!, Facebook and others, on clusters with several thousand nodes.³ It includes HDFS file system, which as default creates 3 copies (redundancy 3).⁴

When considering parallel execution of tasks rather than file storage, service redundancy causes extra load through the additional execution of tasks. The execution load grows linearly with the numbers of redundant copies, but in addition, there will be some form of load associated with parallel threads periodically passing runtime data that we describe as communication cost. This paper uses a simulation model of parallel job execution to explore the trade-off between resilience and communication cost as failure, redundancy and scheduling types vary. For model simplicity we focus on computational redundancy and ignore disk and I/O redundancy.

3. SIMULATION MODEL

3.1. Network Tree Hierarchy

We model the interactions between networks of VM cloud services that exist in a hierarchical tree-structure (refer to Figure 2). Network structure is configurable and we use several tree hierarchies. Throughout this paper, however, unless otherwise stated assume a fixed hierarchy $h-8-4-16-16$. That is, each aisle has 8 racks, each with 4 chassis containing 16 blades, with each blade running 16 cloud services. This structure was chosen to model realistic hardware, such as the 16-blade

HP C7000 chassis and 4-chassis IBM rack shown in Figure 1.

3.2. Jobs, Tasks and Redundancy

We assume that all jobs to be run in the cloud can be parallelized into T independent task threads. We make this simplifying assumption on the basis that one of the major draws of cloud infrastructures is the elasticity of rapid scaling and de-scaling through parallelization. In our model, J jobs are run on the DC, with each job, \mathbf{J} , consisting of T independent parallel tasks. While tasks can be parallelised, they are not entirely independent otherwise they would constitute a new job. Thus, tasks must periodically communicate with each other, passing runtime data when necessary. To pass runtime data, tasks within a job communicate at fixed time intervals. Normally, if any one task within a job fails, the entire job will fail. To mitigate this, redundancy can be introduced by running R duplicate copies of tasks in parallel. Then, job \mathbf{J} will fail *if and only if* all redundant copies of an independent parallel task fail. Such redundancy introduces failure resilience.

Let \mathbf{J} denote a job consisting of T tasks, each having R redundant copies. Then, \mathbf{J} can be written in matrix notation, with T rows and R columns:

$$\mathbf{J} = \begin{pmatrix} j_{1,1} & j_{1,2} & \cdots & j_{1,r} & \cdots & j_{1,R} \\ j_{2,1} & j_{2,2} & \cdots & j_{2,r} & \cdots & j_{2,R} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ j_{t,1} & j_{t,2} & \cdots & j_{t,r} & \cdots & j_{t,R} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ j_{T,1} & j_{T,2} & \cdots & j_{T,r} & \cdots & j_{T,R} \end{pmatrix} \quad (1)$$

Job failure occurs when all tasks in a given row fail. More formally:

$$fails(\mathbf{J}) \Leftrightarrow \exists t \in T, \{ \forall r \in R, fails(j_{t,r}) \} \quad (2)$$

Throughout this paper, we denote experiments running J jobs, each with T tasks and R redundancy as a $\{J, T, R\}$ configuration, with sum total tasks $\#T = J \times T \times R$.

3.3. Scheduling Algorithms

Jobs and tasks can be placed onto a DC network in an infinite variety of ways; using schedules that range from the simple to the complex. In this work, we are interested in deriving general relationships between job scheduling methods and the effects they have on communication cost and resilience. Since we cannot hope to assess the relative behaviours of *every* scheduling algorithm, to aid analytical tractability, we selected a small subset purposely designed to be simple. The intention is not to test intelligent, complicated, real-world algorithms, but rather to tease out general behaviours of these simple algorithms so that we can equip ourselves with better knowledge to design intelligent industrial algorithms in the future. To this end, we define the following three scheduling algorithms:

³ <http://wiki.apache.org/hadoop/PoweredBy>

⁴ <http://www.hadoop-blog.com/2010/11/how-to-change-replication-factor-of.html>

- Random: Uniformly distribute tasks across the DC, independent of job or redundancy group.
- Pack: Use the minimum amount of DC hardware to run all jobs. Place tasks from consecutive redundancy groups for all jobs on consecutive DC services.
- Cluster: Place all tasks belonging to the same redundancy group on the smallest piece of hardware that they fit (e.g., on one blade)⁵. Uniformly distribute redundancy groups across the DC.

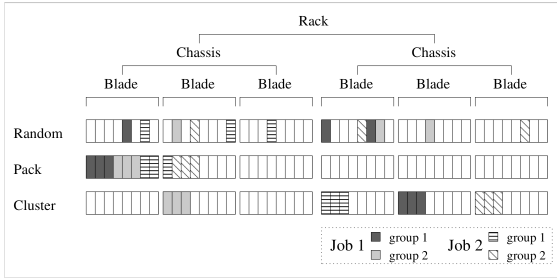


Figure 3: Job scheduling for a $\{J=2, T=3, R=2\}$ configuration on an $h-2-3-8$ hierarchy subset. Top: Random uniformly distributes the $\#T=12$ tasks across the DC. Middle: Pack schedules tasks onto the minimum physical hardware set, in this case 2 blades on 1 chassis. Bottom: Cluster schedules full job copies onto the same physical hardware, while uniformly distributing copies across the DC.

Figure 3 shows a schematic example of each scheduling algorithm. Random, *top line of figure*, assigns tasks to the DC using a random uniform distribution over all DC services. Random schedules tasks independently, taking no account of which job or redundancy group a task belongs. Conversely, Pack preserves geographical co-location of individual tasks according to job and redundancy groupings, *middle*. Tasks are sequentially scheduled using neighbouring services until each hardware is filled. Finally, Cluster uses a combined approach, *bottom*. Similar to Pack, Cluster places all tasks belonging to a job redundancy group on the same physical hardware. However, redundancy groups themselves are uniformly distributed across the DC. In aggregate, these trivial scheduling algorithms form a simple strategy spanning-set from which we aim to tease out general rules for improving failure resilience.

3.4. Network Communication Costs

As explained in Section 3.2, the model assumes that tasks within a job need to communicate at fixed time intervals, passing runtime data between parallel threads. Table 1 shows inter-task communication costs within

⁵ In the case that no hardware has enough free space to fit the entire task-group, deploy as many tasks as possible on the hardware with the largest free space, then deploy the remaining tasks as “close” (lowest communication cost) as possible.

Communication	Relative Cost
Inter-Service	$C_S = 10^0$
Inter-Blade	$C_B = 10^1$
Inter-Chassis	$C_C = 10^2$
Inter-Rack	$C_R = 10^3$
Inter-Aisle	$C_A = 10^4$

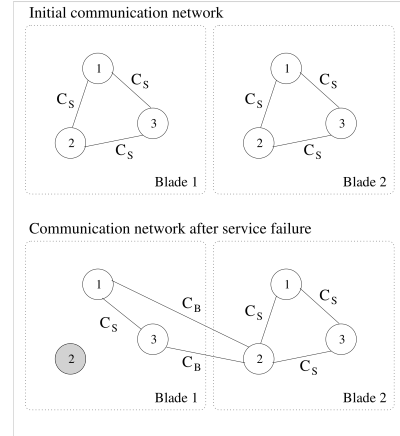


Figure 4: Communication network costs. Top: tasks communicate with the nearest copy of every other task. Bottom: when a task fails, communicating tasks find the nearest alternative. When task 2 fails, communication costs increase from $6C_S$ to $4C_S + 2C_B$. Refer to Table 1 for cost values.

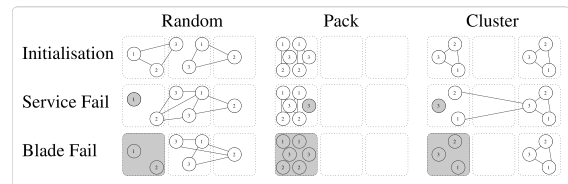


Figure 5: Hardware failure. Top: initial communication networks resulting from alternative scheduling methods. Middle: individual service failure produces minor restructuring of communication networks, including the addition of more costly inter-blade edges. Bottom: blade failure produces major network reconfiguration, while Random and Cluster recover Pack results in job failure.

the DC. Intuitively, cost increases with physical distance between tasks, increasing in magnitude each time it is necessary to traverse a higher layer in the DC tree (Figure 2). From Table 1, tasks communicate with themselves and other tasks on the same service with zero cost. For tasks on different services on the same blade the communication cost is $C_S=10^0$; between different blades $C_B=10^1$; etc.

These costs were chosen to give a qualitatively intuitive model of costs: clearly, communication costs between tasks running on the same physical chip are

many orders of magnitude lower than between tasks located in different aisles of a data centre. While more accurate estimates of relative costs are possible, we believe the simple relationship defined in Table 1 adequately serves the purposes of this paper, since we are only interested in qualitative differences between scheduling algorithms, rather than accurate quantitative relationships.

3.5. Hardware Failure

Hardware failures directly affect task communication. Figure 4 highlights a schematic example of the effects of a single service failure. Initially, tasks communicate with the “nearest” copy of every other task; where “nearest” is defined as the least costly to communicate. Top: communication takes place between tasks running on the same blade. Middle: after task 2 fails on blade 1, tasks 1 and 3 on blade 1 begin inter-blade communication with the “nearest” alternative copy of task 2. The resulting network communication load increases from $6C_S$ to $4C_S+2C_B$.

Within the model, hardware failure can occur at any level in the physical hierarchy tree. Figure 5 demonstrates example effects of failure on the underlying communication network. Tasks form an initial communication network, *top*. Individual service failure, *middle row*, results in some restructuring of the communication network; with the addition of more costly inter-blade links for Random and Cluster. Hardware failure of an entire blade server, *bottom row*, has a more profound effect on the network. While Random and Cluster find a new rewiring, there no longer exists a full task-set for Pack, thus resulting in job failure.

4. EXPERIMENTAL DESIGN

Using the model described in Section 3, we perform a series of empirical experiments to observe the effect that different job scheduling algorithms have on resilience and communication cost in a data centre with hardware failures.

4.1. Assumptions

To keep the model tractable we make some simplifying assumptions.

Time: Simulations have a fixed time length. Jobs are scheduled before the simulation clock begins, then run for the entire length of the simulation.

Jobs: Jobs consist of a set of tasks that can be run in parallel with no inter-dependencies or I/O requests, other than periodic passing of runtime data. Consider, for example, computationally intensive batch jobs such as overnight computation of market data for financial institutions, or CFD simulations for the aeronautics industry or Met Office. For all tasks comprising a job, if at least one copy of the task succeeds, then the job completes successfully; refer to equation (2).

Communication Cost: Tasks within a job need to communicate with a copy of all other tasks at a constant

rate. Communication costs increase with physical distance; see Table 1.

Network Utilisation: The DC is effectively infinite in size, enabling us to ignore the dynamics of full utilisation.

Failure: Failure can occur at any level in the hierarchical tree. Failure events are drawn from an exponential distribution.

4.2. Configuration

Hierarchy Tree: Unless otherwise stated, all experimental runs use an *h-8-4-16-16* tree hierarchy. These are realistic values based on current consumer hardware (refer to Section 3.1). Where alternative tree architectures are used, we use the notation *h-5* and *h-10* as shorthand for *h-5-5-5-5* and *h-10-10-10-10*, respectively.

DC size: To approximate unlimited resources, we scale the size of the data centre, $|DC|$, to equal twice the size needed to run all jobs, that is:

$$|DC| = 2 \times \#T = 2 \times J \times T \times R \quad (3)$$

In an alternative configuration, data centre size is fixed. Under these conditions, set:

$$|DC| = 20 \times J \times T \quad (4)$$

Communication Costs: Communication costs are set equal to Table 1 Refer to Section 3.4 for a discussion.

Scheduling: Jobs are scheduled using the algorithms Random, Pack and Cluster (as detailed in Section 3.3).

Hardware Failure: We set the proportion of hardware that will fail, f_{hw} , during the length of a simulation run to 1%, 5% or 10%. Note, however, that a failure event will cascade down the hierarchy tree, such that failure of a chassis will cause all blades and services running on the chassis to fail. Thus, the overall proportion of a DC that will fail during a simulation run will be larger than the value of f_{hw} . These failure rates may appear to be high. However, it is our intention to model resilience under extreme conditions that cannot be observed readily in operational environments. When a hardware failure event occurs, a discrete distribution is used to select the type of hardware failure. The relative probability of a given type of hardware, h_{type} , failing is calculated as the relative proportion of that hardware in the data centre, h_{type}/h_{all} . Although this distribution is simplistic, it provides the intuitive result that the more common a type of hardware, the more likely it is to fail.

5. RESULTS

Here, we present simulation results for all scheduling experiments. Figures plot mean values of repeated trials, plus or minus 95% confidence interval. Thus, where error bars do not overlap, differences are

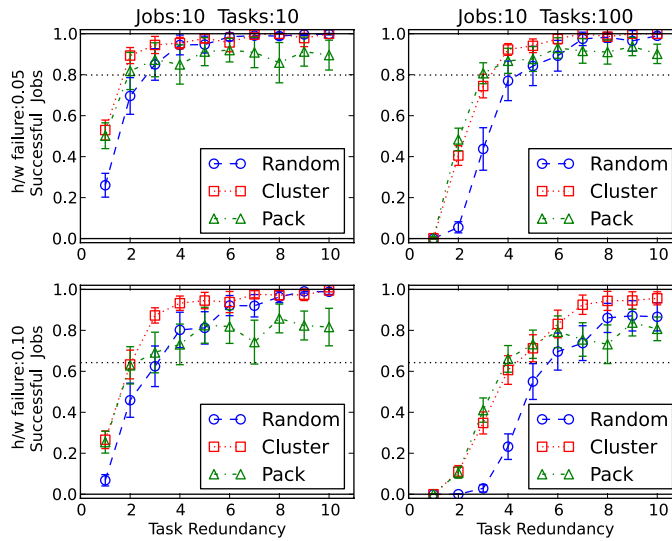


Figure 6: Resilience of scheduling algorithms in a fixed-size data centre with tree hierarchy $h-8-4-16-16$. As hardware failure increases, *top to bottom*, resilience falls; as tasks per job increases, *left to right*, resilience falls. Overall, Cluster is more resilient than Random and Pack across all conditions. Error bars show 95% confidence intervals. The dotted horizontal line plots the mean percentage of DC services surviving at the end of a run: the resilience that jobs with $T=1$ and $R=1$ will tend toward.

statistically significant. The simulation experiments were run in parallel and distributed across a cluster of 70 linux machines and the number of repetitions varies between 30 repetitions to over 100 repetitions. Confidence intervals remain relatively large due to the stochastic nature of the failure process: particular failure events can have widely ranging effects. It should be noted that occasionally the entire DC fails during simulation. When this occurs, the run is rejected so these catastrophic outliers do not skew results. This is reasonable since DC failure is a direct result of random hardware failure and is independent of the scheduling algorithms under test. Hence, all plots display summary data from trials where the entire DC did not fail.

5.1. Resilience

Figure 6 shows simulation results for $J=10$ jobs using fixed DC size, equation (4). The proportion of successful job completions, S_j , is plotted against number of redundant task copies, R , for each algorithm: Random, Cluster, and Pack. In each graph, we see the intuitive result that success, S_j , increases with redundancy, R . However, whereas Random (blue circle) and Cluster (red square) reach 100% success under all conditions except bottom-right, Pack (green triangle) reaches a maximum in the range 80%-90% at approximately $R=5$ and then plateaus, with fluctuation, as R is increased further. As shown schematically in Figure 5, Pack schedules all tasks to fit on the smallest hardware set possible. However, this tactic of “putting all your eggs in one basket” is vulnerable to specific hardware failure events that may take out the entire set of tasks. Although such events are rare, across all runs

and all job sets they occur often enough to stop Pack from reaching $S_j=100\%$, regardless of R .

For all algorithms, we see that as the number of tasks per job, T , is increased from $T=10$, *left*, to $T=100$, *right*, more redundancy is needed to maintain a given level of resilience. This is intuitive. Since task failure results in job failure, the greater the number of tasks per job, the greater the chances of any one job failing; hence, the greater the number of redundant copies needed to counter this failure. Similarly, when the probability of hardware failure, f_{hw} , is increased from 0.05, *top*, to 0.10, *bottom*, to maintain resilience redundancy R must be increased. Once again, this is intuitive: as failure increases, so too does the likelihood of job non-completion.

Overall, across all conditions, Cluster is the most resilient. With low values of R , Cluster and Pack outperform Random. When $R \geq 7$, Cluster and Random outperform Pack. Further, there is no condition under which Cluster is significantly outperformed by either Random or Pack. Yet, there are several conditions under which Cluster significantly outperforms both alternatives. Thus, results suggest that Cluster is the most robust strategy. Interestingly, the default number of redundancies used in Hadoop's HDFS, $R=3$, appears to be a reasonable choice when $T=10$. As the number of tasks increases, $R=3$ does not suffice under our conditions.

5.2. DC Architecture

Figure 6 displays a clear relationship between increased redundancy, R , and increased resilience, S_j . However, the resilience graphs for Pack exhibit “dips”, for example at $R=8$ (*top-left* and *bottom-right*) and $R=7$

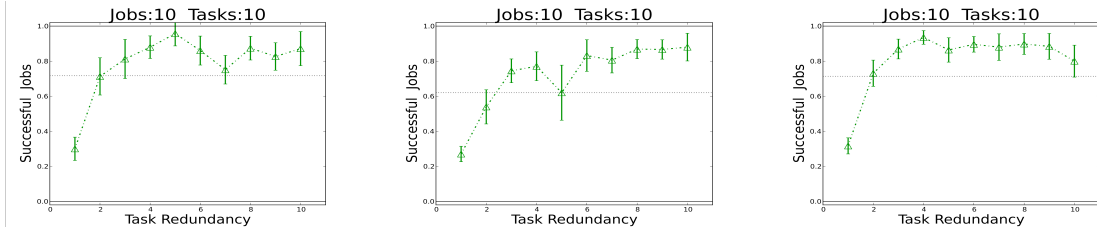


Figure 7: Pack scheduling using *h-8-4-16-16*, *h-5* and *h-10*, from left to right, respectively. Resilience dips when jobs exactly fit on underlying hardware.

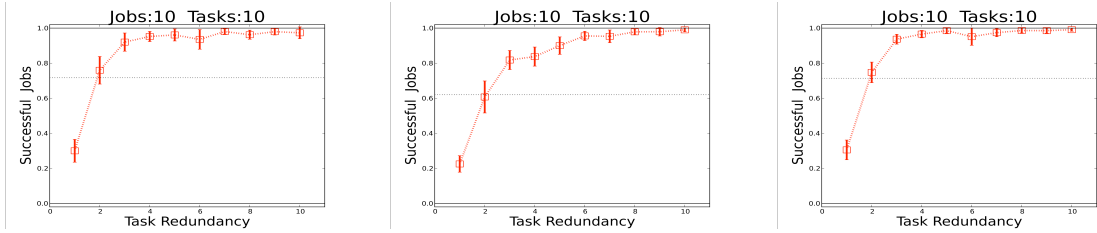


Figure 8: Cluster scheduling using *h-8-4-16-16*, *h-5* and *h-10*, from left to right, respectively. Cluster is largely insensitive to underlying hardware architecture.

(*bottom-left*), which raises some questions. Are these “dips” merely statistical aberrations, or are there underlying dynamics causing specific R values to result in lower S_J ?

To address this issue, a series of experiments were performed using alternative DC tree structures (Figure 2) to observe the effect that hardware hierarchy has on resilience. Three configurations were tested: *h-8-4-16-16*, *h-5* and *h-10*. In each case, data centre size $|\text{DC}|$ was variable; refer to equation (3).

For all hierarchy trees, results were qualitatively similar to those displayed in Figure 6, suggesting that the general behaviours of each scheduling algorithm are largely insensitive to the underlying hardware hierarchy configuration. However, Pack does display some idiosyncratic sensitivity to hierarchy. Figure 7 plots S_J against R for Pack under each tree hierarchy. Left: S_J increases with R until $R=5$, but then fluctuates around 80%, with a minimum at $R=7$. When DC hierarchy is changed to *h-5*, *centre*, Pack has a minimum at $R=5$. Finally, with hierarchy *h-10*, *right*, there is a minimum at $R=10$. This evidence suggests that Pack is sensitive to the underlying physical hierarchy of the DC. In particular, if all redundant copies of a job fit exactly onto one hardware unit—blade, chassis, rack, etc.—then a failure on that hardware will take out all copies of the entire job. Hence, with an *h-5* hierarchy, for instance, $R=5$ results in poor resilience for Pack. Under these circumstances, each job, $J_{T,R}$, contains 50 tasks, which exactly fit onto 2 chassis. Thus, two neighbouring chassis failure events will take out the entire job. In

comparison, when $R=4$ or $R=6$, task group copies will be more unevenly distributed over hardware, giving greater resilience to failure.

Figure 8 plots results for Cluster under the same conditions as Figure 7. Here, we see that Cluster is largely unaffected by the underlying tree structure of the data centre.

5.3. Jobs

To see how results scale with an increase in jobs, we ran experiments with $J=100$ jobs, using hierarchy *h-8-4-16-16* with variable data centre size, $|\text{DC}|$. Results are qualitatively similar to those of Figure 6. Pack outperforms Random with low R , but plateaus around $S_J=80\%$. Random has poor resilience when R is low, but outperforms Pack as R approaches 10. Finally, Cluster has best resilience overall. Results for other failure rates and number of tasks, T , are also qualitatively similar, indicating that resilience is insensitive to J .

5.4. Communication Costs

While it is important for scheduling algorithms to enable job completion, it is also important they do not induce prohibitive cost through wasteful communication. In this section, we explore the mean communication cost, C_J , for each successfully completing job. Results suggest that the relationships between algorithms are largely robust to variations in hierarchy-tree, number of jobs, J , number of tasks, T , and hardware failure rate, f_{hw} . Thus, in this section we consider only two conditions: variable $|\text{DC}|$; and fixed $|\text{DC}|$. Each time an *h-8-4-16-16* architecture is used.

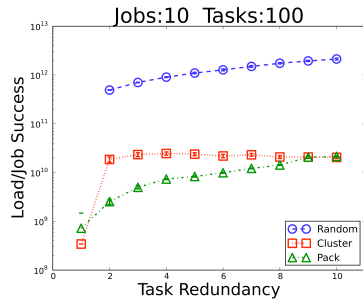


Figure 9: Costs per successful job, C_J , using fixed |DC| with $h=8-4-16-16$.

Figure 9 displays communication costs per successful job, C_J , using fixed |DC|. 95% confidence intervals are drawn, but are generally very small. With tasks uniformly distributed throughout the DC, Random produces the greatest communication costs per job. Conversely, with tasks placed as close together as possible, Pack has the smallest communication costs per job. For Pack and Random, an increase in redundancy, R , leads to a proportional increase in cost, C_J . When $R=10$, C_J is approximately 10 times greater than the value at $R=1$.

For Cluster, however, the story is different. When $R=1$, Cluster produces smaller C_J than Pack, since Cluster guarantees all job copies are placed on the lowest branch of the hardware tree that they fit; with Pack, however, depending upon number of tasks, T , and the underlying tree-hierarchy, some jobs will occasionally be split across hardware (see schematic Figure 3, for example), thus incurring greater communication costs. When redundancy is increased to $R=2$, communication costs, C_J , becomes a magnitude greater than Pack. As Cluster distributes job groups across the network, when an individual task fails, new communication links to alternative copies are likely to be long-range and costly. In contrast, Pack places all clones near each other, so communication with alternatives does not radically increase costs in most cases (refer to Figure 5). Interestingly, with fixed |DC| mean communication cost per successful job, C_J , remains constant when $R \geq 2$. Since Cluster distributes job copies uniformly across the data centre, the mean distance or cost for communication between tasks in different redundancy clusters is inversely proportional to R . Hence, additional redundancy reduces the mean communication cost a task must pay to communicate with an alternative clone, thus making C_J invariant under changes to R . It should be noted that the same is not true for Random, however. Unlike Cluster, since Random distributes all tasks uniformly independent of redundancy group, the majority of communication paths are inter-hardware and costly. Hence, doubling R will approximately double C_J .

When using a variable-sized data centre—equation (4)—results for C_J against R are similar to Figure 9 for Random and Pack. For Cluster, however, C_J is no

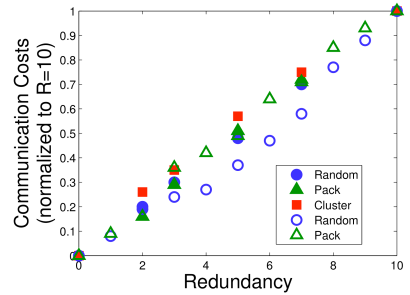


Figure 10: Communication costs per successful job, normalized to $R=10$. Clear-faced markers represent variable-sized DC; filled markers show fixed-size DC. We see that communication costs scale linearly with redundancy for all algorithms.

longer invariant to R and instead increases proportionally as R increases. As |DC| increases with each increase in R , the mean length between communicating tasks remains stable. Thus, as the number of tasks increases so too does overall communication costs.

Figure 10 plots normalized communication cost for each scheduling algorithm against redundancy, R . Clear faces show fixed |DC| (not including Cluster) re-plotted from Figure 9. Coloured faces show data from the equivalent set of runs using variable |DC|. In all cases, with all algorithms, there is clearly a linear relationship, suggesting that communication costs rise in direct proportion to R . Note, however that this is not the case for Cluster under fixed |DC| (not plotted): here, communication costs are invariant in R .

5.5. Summary of Findings

The main findings can be summarized as follows:

1. The network hierarchy tree has little effect on the resilience of scheduling algorithms (except in the case of Pack, where particular tree configurations have negative impact on particular levels of redundancy).
2. Cluster is the most resilient scheduling algorithm from the selection modelled. In contrast, Pack is a non-resilient high-risk algorithm.
3. Pack is the most efficient algorithm, Random the least. Cluster generates intermediate costs, but scales well under fixed data centre size.
4. Overall, Cluster is the most practical algorithm, effectively combining the efficiencies of Pack with the resilience of Random.

6. DISCUSSION

The aim of this work is to build an understanding of the general relationships between scheduling, resilience and costs (rather than perform a detailed analysis of any particular algorithm), the results presented support our basic endeavour to use simulation models as a methodological framework to design and test tools for elastic cloud-computing infrastructures. We do not

suggest that Random, Pack or Cluster are practical job-schedulers that should (or would) be used in industry, but rather that these purposely naïve algorithms provide a simple base-line set of strategies that enable us to tease out fundamental relationships between density, clustering and spread of jobs; and the impact each has on resilience and communication cost. By using simulation to better understand how these concepts interact, we gain access to a powerful off-line test-bed for algorithm development that provides a design route towards more robust and efficient cloud-computing services. The simulation model we have used makes some simplifying assumptions that should ideally be relaxed. However, despite this, the model is powerful enough to highlight how underlying complex interactions, such as between scheduling and the shape of the hierarchy-tree, can affect resilience. This is a promising indication of the value of pursuing the goal of creating an extensible simulation framework for cloud computing.

7. FUTURE WORK

Here, we outline potential future extensions:

1. More realistic modelling assumptions: the introduction of sequential task inter-dependencies, heterogeneous jobs and services, full-DC utilisation, etc.
2. Model verification and validation using real-world data. Retroactive validation of results through testing on real-world DCs.
3. Introduction of other scheduling algorithms from industry and the development of novel algorithms using evolutionary computation as an automated design tool.
4. Monitor individual failure events rather than failure over time, to observe how the system changes when failure occurs, and what exactly takes place at this lower level of description.
5. Compare the effects of scale-up versus scale-out: If the resource usage increases, what does it mean for the resilience if more services are used, rather than more powerful ones?
6. Introduce migration of services to the scheduling algorithm. This allows a task to be cloned when a parallel instantiation fails, and the clone can then be migrated towards the other tasks belonging to that job.

8. CONCLUSIONS

We have presented a simulation model for testing the effects that different scheduling algorithms have on the resilience and communication costs of jobs running “in the cloud” of a large scale data centre (DC). Modelling the data centre as a tree-hierarchy of physical machines and jobs as a collection of parallel tasks that can be cloned, we have demonstrated the effects that different job-scheduling algorithms have on network resilience and communication cost. As intuition would expect, Packing all tasks together in a small area of the DC greatly reduces communication cost but increases risk

of failure. Conversely, a Random distribution of tasks throughout the DC leads to greater resilience, but with a much elevated cost. Clustering tasks together in cohesive job-groups that are then distributed throughout the DC, however, results in a beneficial trade-off that assures resilience without prohibitive costs. This work provides a teasing glimpse into the powerful insights a cloud simulator can provide. Given the grand scale of the challenge, this work has naturally raised many open questions and introduced scope for future extensions.

ACKNOWLEDGMENTS

Financial support for this work came from the EPSRC grant:⁶ EP/H042644/17 (for J. Carlidge) and from Hewlett-Packard's Automated Infrastructure Lab, HP Labs Bristol (for I. Sriram). The authors would like to thank Prof. Dave Cliff and the sponsors for their support and interest in this topic.

REFERENCES

- Barroso, L. A. and Hölzle, U., 2009. The datacenter as a computer: An introduction to the design of warehouse-scale machines, *Synthesis Lectures on Computer Architecture*, no. 1, pp. 1–108, 2009.
- Buyya, R., Ranjan, R., and Calheiros, R. N., 2009. Modeling and simulation of scalable cloud computing environments and the cloudsim toolkit: Challenges and opportunities, in *International Conference on High Performance Computing & Simulation, HPCS '09*, pp. 1–11, June.
- Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A., and Gruber, R. E., 2008. Bigtable: A distributed storage system for structured data, *ACM Transactions on Computer Systems*, vol. 26, no. 2, article 4, pp. 1–26, Jun.
- Dean, J. and Ghemawat, S., 2008. MapReduce: Simplified data processing on large clusters, *Communications ACM*, vol. 51, no. 1, pp. 107–113.
- DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Voshall, P., and Vogels, W., 2007. Dynamo: Amazon's highly available key-value store, *SIGOPS Oper. Syst. Rev.*, vol. 41, no. 6, pp. 205-220.
- Hollnagel, E., Woods, D. D., and Levson, N., 2006. *Resilience Engineering: Concepts and Precepts*. Ashgate.
- Liu, D., Deters, R., and Zhang, W. J., 2010. Architectural design for resilience, *Enterprise Information Systems, 1751-7583*, vol. 4, no. 2, pp. 137–152, May.
- Nagel, L., 1975. *Spice2: A computer program to simulate semiconductor circuits*, University of California, Berkeley, Tech. Rep. UCB/ERL-M520.
- Sriram, I., and Cliff, D., 2011. SPEC12 - simulation program for elastic cloud infrastructures, to appear in *SummerSim '11*.

AUTHORS BIOGRAPHY

Dr **John Carlidge** is a Research Associate in cloud computing. His research interests include simulation modelling, automated trading and electronic markets, and evolutionary computing. **Ilango Sriram** is a final year PhD student soon to defend his thesis on simulation modelling for cloud-scale data centres.

⁶ <http://gow.epsrc.ac.uk/ViewGrant.aspx?GrantRef=EP/H042644/1>

Advanced Container Transportation Equipment using Transfer Robot and Alignment System

Young Jin Lee*, Dong Seop Han**, Dae Woo Kang***, Duk Kyun Lee****, Geun Ghoi****, and Kwon Soon Lee†

Korea Aviation Polytechnic College*, Dong A University**, Bumchang Eng. Co., Ltd. ****, Dong A University†

airlee011@kopo.ac.kr*, imdshan@gmail.com**, dwkang@dau.ac.kr***, dklee@bumchang.co.kr****, kslee@dau.ac.kr†

Abstract. In this paper, we proposed a horizontal-transfer robot system was designed using a hydraulic system and some electronic sensors, and a synchronized control method was applied to control each robot. An alignment system was also designed to align between the rail and a truck. The proposed system will be useful as an intermodal transportation system that is significantly considered an enhanced technique or future railroad logistics. The system can simplify the complicated job process in railroad-based transportation and, from an economic viewpoint, can reduce relevant logistic costs.

I. INTRODUCTION

The traditional railroad-based transport has difficulty directly moving freight in door-to-door service compared to the road method in the highway. This is the main reason for the hesitation in using the railroad for such purposes. Moreover, most locomotives were recently changed into electric-driven systems with catenaries for the electrification of the railroad, instead of gasoline engines. Unfortunately, as these catenary systems built over trains interfere with crane works, new parallel loading and unloading systems should be developed to enhance the classical vertical systems [1, 2].

In addition, the job process of the railroad-based transport system is largely complicated because auxiliary works such as operation of shuttle cars and stacker cranes are unavoidable, particularly at the starting and destination points, to connect with the roadway. Such additional processes increase the unnecessary logistic cost and time loss. As such processes also demand a wider storage space, more equipment, and more operators, the continued use of the traditional methods may bring about significant problems. In recent years, in many advanced countries, new transport systems have been developed, including intermodal transfer systems such as Piggy-back, CargoBeamer, Cargo Domino, Flexiwaggon, and Modalohr [3-11]. As there have been few discussions about Piggyback or Roadrailer in South Korea, however, a structural-design study of a horizontally moving system, such as the parallel-type intermodal transportation system, has been developed by Dong-A University [11, 12]. This system can transfer a container box horizontally between a trailer and a freight wagon. Moreover, when the robot systems travel, an alignment system of the two cars is necessary for safe loading. In this paper, a horizontally moving equipment is suggested as a container transport robot.

This system is controlled via synchronization and has an alignment system to align the railroad and trailer using electronic sensors (e.g., ultrasonic sensor, etc.).

II. DUAL-MODE TRAILER SYSTEM

A. Intermodal Transportation System

Because the problems of train only logistics complicated cargo handling and catenary, door-to-door service is impossible. In case of truck only logistics, the road traffic jam is severe because of increasing cars.

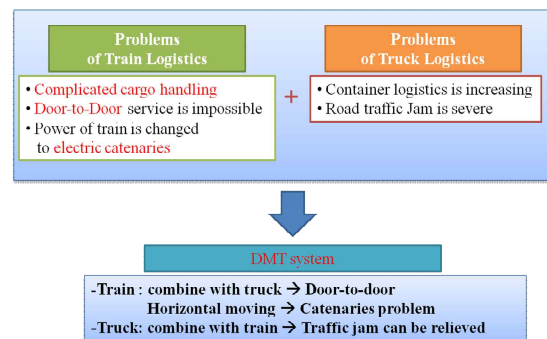


Figure 1. The problem of single mode transport system.

The intermodal transportation system (ITS) involves the transport of freight in an intermodal container or vehicle, using multiple modes of transportation (rail, ships, and trucks), without any handling of the freight itself when changing modes. This method can reduce the cargo handling time, damages, and losses, can improve the security, and allows freight to be transported faster. In relation to railroad transportation, in many countries, diverse studies on ITS have been conducted of late. Typically, this system can be classified into four types, as shown in Fig. 2: the parallel, cargo turning, piggyback, and Bogie changing types. The cargo-turning-type methodology involves transferring the container box by turning the wagon of the train. Modalohr is one of the best examples and is used in France. Even though this system has many advantages, it cannot be applied in South Korea because the height constraint does not allow it [11]. The parallel type is divided into CargoBeamer and Cargo Domino [13-17]. In the analysis based on reference [11], the horizontal loading and unloading system (e.g., Cargo Domino) showed better performance than the others.

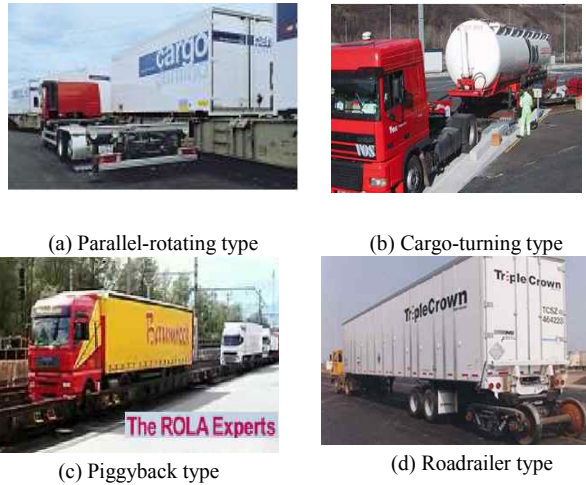


Figure 2. Intermodal transportation systems.

III. DESIGN OF THE DMT SYSTEM

As it is expected that the proposed horizontal-type system can avoid the use of catenaries and can provide higher advantages from a logistic-cost viewpoint, this research was focused on the parallel-type ITS, and proposed an improved transport system. The DMT system consists mainly of two parts: a horizontal-transfer robot (HTR) and an alignment system (AS). The key of the DMT system is the horizontal-transfer robot installed on the wagon or trailer. The working procedure of the proposed ITS is shown in Fig. 3.

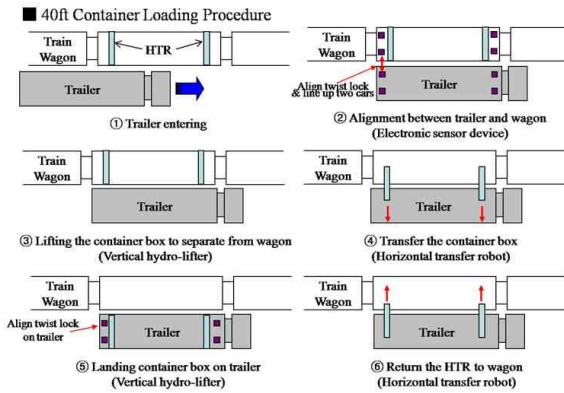


Figure 3. Operation procedure of the proposed ITS.

A container box is moved horizontally by the HTR from the wagon to the trailer or from the trailer to the wagon, without any overhead crane or reach stacker. In processes ① and ② in Fig. 3, the trailer should be aligned with the train wagon within the permitted clearance when a trailer is entering. Even if two cars are aligned within the permitted clearance (e.g., 320 or 350 mm), however, if the HTR loads a container from the wagon to the trailer or from the trailer to the wagon, the HTR will fall to the ground. This notwithstanding, the alignment within the clearance value and tolerance is very important because the HTR will autonomously trip on cars with a 40-ton container. This chapter mainly deals with HTR and AS.

A. HTR

HTR consists of a lifting part that separates a container from a cone on the trailer or wagon, and a horizontal-driving part for the loading and unloading of a container between a wagon and a trailer. This system is designed to deal with 20- and 40-ft containers. Therefore, the self-weight of the container is 40 ton, which is the maximum loading weight of the cargo in a wagon. This constraint would be adopted as the load condition for the design of the HTR. As a couple of HTRs are used for the loading and unloading of a container, one robot takes charge of the 20 ton weight. Accordingly, considering the dynamic-effect factor of 1.2, the container weight to be applied on each system should be 24 ton.

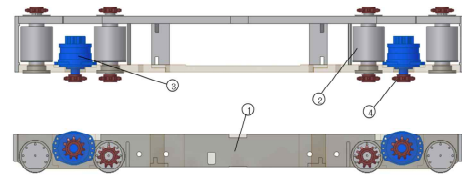
As the HTR suggested in this investigation lifts the under surface of the container and transfers it to a trailer or wagon, the height of the system should be below 300 mm due to the limitations in ROK. As a 300-mm space exists between a wagon and a trailer, this factor should be considered when designing the horizontal-driving part. The speed of the HTR and the lifting speed should be over 0.1 and 0.01 m/sec, respectively, because the total operating time for the loading or unloading of a container has to be less than 3 min to minimize the logistic costs and to reduce the time loss. These design conditions are shown in Table I.

TABLE I. DESIGN CONDITIONS OF THE HTR

Parts	Design Conditions
Horizontal-driving part	<ul style="list-style-type: none"> Capacity: 24 ton Moving distance: 2,800 mm Maximum speed: 0.1 m/sec
Lifting part	<ul style="list-style-type: none"> Capacity: 24 ton Moving distance: 100 mm Maximum speed: 0.01 m/sec

1. Horizontal driving part

The horizontal-driving part consists of an underframe and four wheels because this system moves a 300-mm space between a wagon and a trailer, and is driven by two hydraulic motors due to its heavy weight and the constrained space in the HTR.



1 : Frame 2 : Driving wheels
3 : Hydraulic motors 4 : Chain & Sprocket

Figure 4. Horizontal-driving part of the HTR system.

The power of the motor is transferred by the chain and sprocket to the wheels due to the large distance between the motor and the wheel shaft. When the rolling friction coefficient is 0.05, the required thrust force of the system for horizontal driving is a minimum of 2 ton. Hence, the wheel diameter was made 180 mm considering the outer diameter of the sprocket. Therefore, when the maximum driving speed is 0.1 m/s, the number of revolutions of the wheel is 10.6 rpm. The normal force (N_w) applied to each wheel is as follows:

$$N_w = \frac{\left(\frac{W_c}{2} + W_{TS}\right)}{4} \quad (1)$$

Figure 5. Forces applied to the HTR.

The horizontal thrust force (F_T) generated from the torque of the wheel (T_w) should be larger than the rolling friction force ($F_{f,R}$). Thus, the torque that is needed to drive the wheel is as follows:

$$T_M \geq r_w (\mu_R N_w) = r_w \left[\frac{\mu_R}{4} \left(\frac{W_c}{2} + W_{TS} \right) \right], \quad (2)$$

where r_w is the radius of the wheel, μ_R is the rolling friction coefficient, W_c is the self-weight of the container, and W_{TS} is the self-weight of the HTR. As one hydraulic motor rolls two wheels, the minimum required torque of the motor (T_M) should be two times the torque so that the wheel could be driven. Using equation (2), the minimum torque of the motor can be calculated as follows:

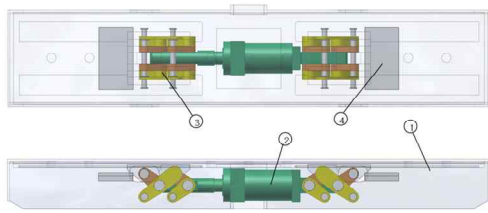
$$T_M = r_w \left[\frac{\mu_R}{4} \left(\frac{W_c}{2} + W_{TS} \right) \right], \quad (3)$$

$$= 2 \times 0.09 \left[\frac{0.05}{4} \left(\frac{40 \times 10^3}{2} + 1 \times 10^3 \right) \right] = 47.25 \text{ kg} \cdot \text{m}$$

where $r_w = 90 \text{ mm}$, $\mu_R = 0.05$, $W_c = 40 \text{ ton}$, and $W_{TS} = 1 \text{ ton}$.

2. Lifting part

To stably separate the container from the cone, the bucket has to be kept at a level where the bucket can lift a 40-ton container.



1 : Bucket 2 : Hydraulic cylinders
3 : Linkage 4 : Stopper

Figure 6. Lifting part of the HTR.

Accordingly, the “y>” combination-type linkage was adopted to keep the balance, and this linkage is driven by a hydraulic cylinder. The force (F_C) of the hydraulic cylinder that is needed for it to lift the container is as follows:

$$F_C \geq \frac{4V}{\tan \theta} \quad (4)$$

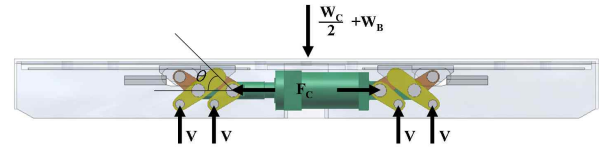


Figure 7. Forces applied to the hydraulic cylinder.

When the angle of linkage on the horizontal axis is 45° , the maximum force is applied to the cylinder. Using equation (4), the force (F_C) can be calculated as follows:

$$F_C \geq \frac{4V}{\tan \theta} = \frac{4 \left(\frac{W_c}{2} + W_b \right)}{4 \times \tan 45^\circ} = \frac{4 \left(\frac{40}{2} + 0.5 \right)}{4 \times \tan 45^\circ} = 20.5 \text{ ton}, \quad (5)$$

where $W_c = 40 \text{ ton}$ and $W_b = 0.5 \text{ ton}$. At this time, when the driving pressure is 280 bar, the required minimum force of the cylinder to lift the container is 20.5 ton.

3. Control algorithm of HTR

Fig. 8 shows the algorithm of the unloading procedure to convey the container from the wagon to the trailer. Information data about the alignment of the twist lock (cone) on the trailer and wagon should be sent to the HTR controller before moving the HTRs between the wagon and the trailer.

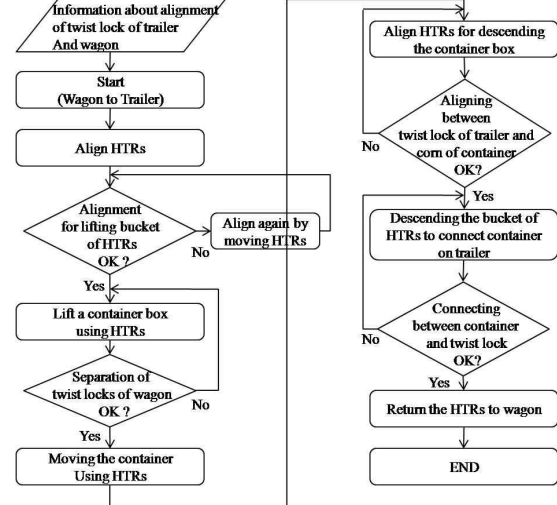


Figure 8. Control algorithm of HTR (unloading procedure).

4. HTR design

The main components of the HTR are a lifting hydraulic cylinder, a bucket, and two horizontal-driving hydraulic motors. The lifting cylinder can lift a 40-ton container box at 0.01 m/s to remove the twist lock. The bucket was designed to support the under frame of the container while the HTRs are moving. The horizontally driving motors carry a container box horizontally with the proportional hydraulic-valve controller. Fig. 9 shows the assembled HTR.

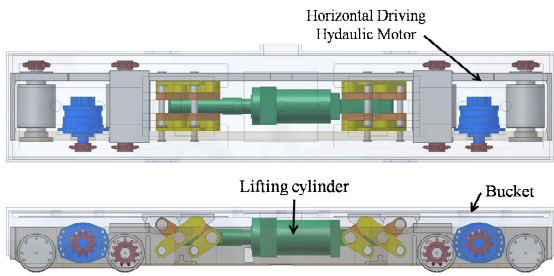


Figure 9. Assembled HTR.

Four HTRs are installed on the wagon or train and convey a container between the wagon and trailer, as shown in Fig. 10.



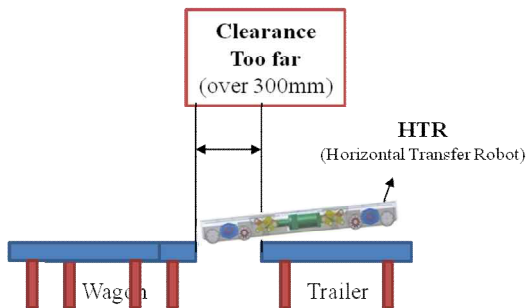
Figure 10. Installation position of the HTR.

As two HTRs are used when a container is carried, encoder sensors are used to synchronize the speed of the two robots, and to measure the distance of the trajectory for docking the container on the cone of the trailer or wagon after the arrival in the destination. Photoelectric sensors are used for driving control, such as acceleration and deceleration.

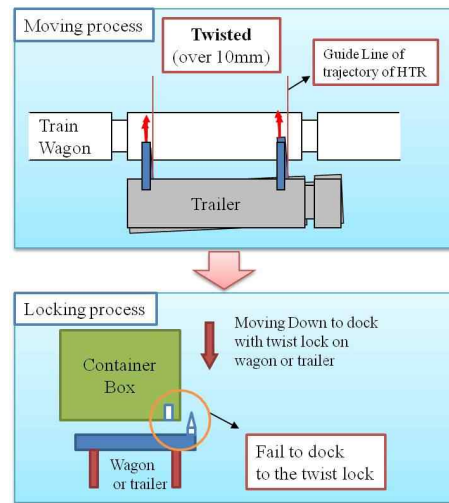
B. Design of the Alignment System

As mentioned in the previous chapter, when loading a container using HTR, as shown in Fig. 3, the wagon and trailer should be aligned under a permitted condition, as shown in Fig. 12. The following are the clearance and tolerance conditions required for the alignment of the DMT system:

- (a) Alignment of two cars: repositioning the wagon and the trailer using two ultrasonic sensors and CCD cameras
 - Clearance: 300 ± 10 mm
- (b) Alignment of the twist lock: aligning the wagon and trailer using 16 photosensors
 - Tolerance: under 20 mm



(a) Alignment of two cars



(b) Alignment of the twist lock
Figure 12. Alignment of two cars.

The alignment system that was designed measures the distance and distortion ratio between the trailer and the wagon using an electronic sensor device (e.g., ultrasonic sensor, photosensor). This gives the necessary information to the driver of the trailer to help him stop his car under the permitted conditions.

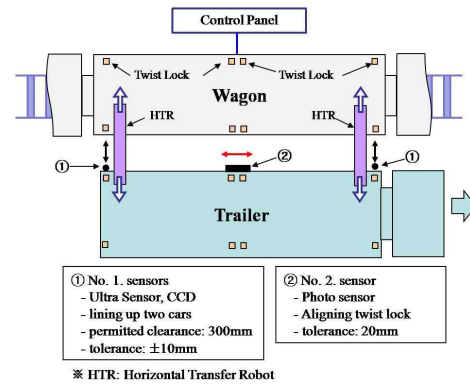


Figure 13. System layout of the positioning system between a trailer and a wagon.

As shown in Fig. 14, two ultrasonic sensors and two CCD cameras were applied to each trailer to measure the distance between the wagon and the trailer, and the driver can see the measured value at the LCD monitor in the driving room.



Figure 14. Display of the alignment process.

In Fig. 14, ① shows the photo sensor's output, ② the distance between the wagon and the trailer as detected by an ultrasonic sensor, and ③ the CCD images. From the above information, the driver of the trailer gets the alignment information and can stop the vehicle beside the wagon, with a tolerable distance. The deviation between the rear and fore sensor should be under 20 mm. All the lamps of ① in Fig. 16 are indicated by the red color, and the twist locks on the wagon and trailer are aligned. Fig. 15 shows the final alignment status of the real trailer and wagon. The twist locks on the wagon and trailer are aligned, and the clearance between the trailer and the wagon is set at the permitted value.



(a) Aligned twist lock (b) Two aligned cars

Figure 15. The two alignments between the trailer and the wagon.

Fig. 16 shows the algorithm of the alignment of the DMT system.

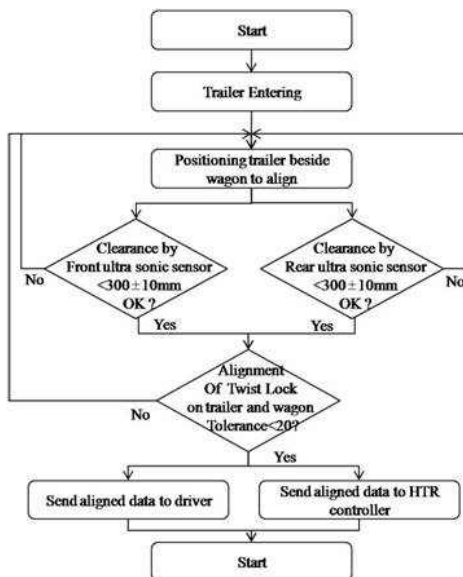


Figure 16. Alignment algorithm of the DMT system.

IV. EXPERIMENT RESULTS

A. Pilot System

To evaluate the performance of the alignment system, a real-scale pilot trailer and a real-scale wagon were manufactured. The developed alignment equipment was installed in the pilot trailer, as shown in Fig. 17.

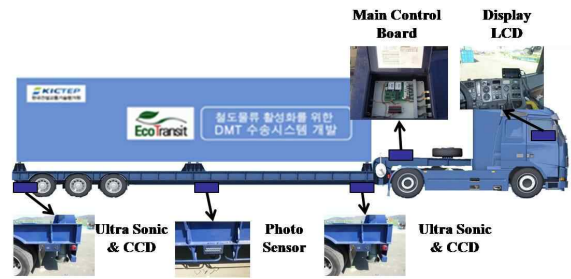


Figure 17. Installation position of the alignment system.

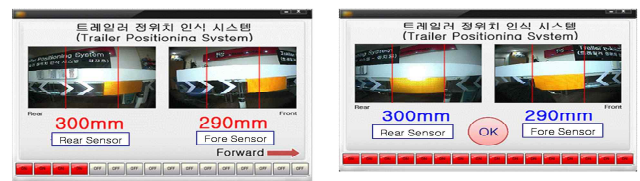
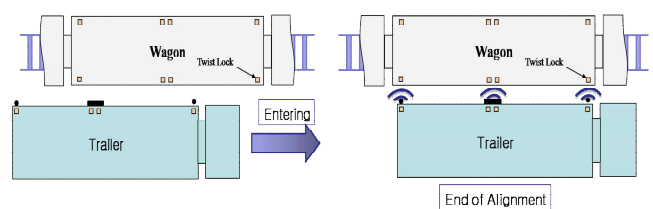
B. Test of the Alignment System

Table II shows the specifications of the alignment system. Three atmega128 microprocessors were used for the main controller and other interface boards. The RS485 protocol was mainly used for the communication system.

Table II. Specifications of the position control system

Sensor	Specifications	Function
Ultrasonic	<ul style="list-style-type: none"> •Vendor: Sontec/ROK, STMA-701ND •Operating range: 30-6,000 mm •Frequency: 40 khz •Used no.: 2/trailer 	Alignment of two cars
CCD camera	<ul style="list-style-type: none"> •Sensor: 1/3 CCD •Pixel: 768*494 •IP grade: 68 •Used no.: 2/trailer 	Alignment of two cars
Photosensor	<ul style="list-style-type: none"> •Vendor: D51-AP series, Italy •Detection method: polarized light reflect •IP grade: 67 	Twist lock alignment
Display panel	<ul style="list-style-type: none"> •CPU: AMD Geode LX800 •LCD: 7" TFT (800*480) •Communication: RS232, RS485 	For the driver
Controller	<ul style="list-style-type: none"> •Main controller: CPU - Atmega128 •Ultrasonic board: CPU - Atmega128 •Communication board: <ul style="list-style-type: none"> - CPU: Atmega128 - Communication: RS485 & RF - RF chip: A3007B, 424-447 Mhz 	System control

Fig. 18 shows the alignment process: (a) status of trailer entry; and (b) achieved alignment status.



(a) Trailer entering (b) Achieved alignment status

Figure 18. Alignment process.

C. Test of HTR

Two HTRs were manufactured by real scale to evaluate their performance in carrying a 40-ton container box.

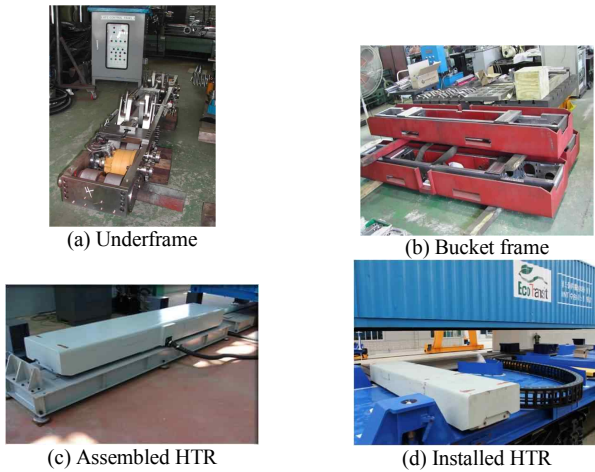


Figure 19. HTRs.

The specifications of the HTR are shown in Table III.

TABLE III. HTR SPECIFICATIONS

Parts	Specifications	Function
HTR	<ul style="list-style-type: none"> ▪ Vertical-lifting part <ul style="list-style-type: none"> - Hydromotor: 58 kg·m - Minimum torque: 47.25 kg·m - Lifting speed: 0.01 m/s - Lifting stroke: 120 mm - Solenoid valve: proportional valve ▪ Horizontal-transfer part <ul style="list-style-type: none"> - Hydromotor: 58 kg·m - Force: 20.5 ton - Speed: 0.1 m/s - Diameter of wheel: 180 mm - Position sensor: encoder (Sumtak IRS5) - Solenoid valve: proportional valve ▪ Total weight: about 1.5 ton 	Lifting and carrying a container
Controller	<ul style="list-style-type: none"> ▪ Main controller <ul style="list-style-type: none"> - CPU: Melsec-Q 7 series - Communication: RS485 and wireless RF - Data sampling: 20 msec 	System control

Fig. 20 shows the experimental testing process of the proposed DMT system with HTR and AS.

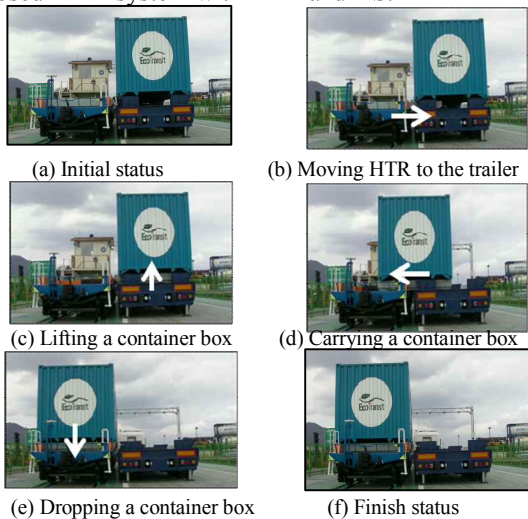


Figure 20. Experimental tests using HTR (unloading procedure).

Fig. 20(b) shows the HTRs being moved to under the container, (c) the lifting procedure by HTR, (d) the conveyance of the container, and (e) the docking process between the cone on the wagon and the container. With 30 containers, the switching time of DMT was only 45 min, as opposed to 78 min in the conventional system. This means that the proposed system provides a 42% switching time reduction.

V. CONCLUSIONS

In this paper, an innovative intermodal transportation system (ITS) named “dual-mode trailer (DMT)” was proposed, based on a horizontally moving robot. This system is a combined intermodal system linking the road and the railway transport system. It is believed that this system can enable the railway system to provide just-in-time service and can enable the on-the-road system to provide door-to-door delivery service. The proposed system is very useful under the high-voltage catenary system by virtue of its horizontal-transfer mechanism. It has been reported that the operation and maintenance costs of DMT will be lower by as much as 59% (KRW0.9 billion for one year) compared with the existing system. Finally, the most important contribution of the proposed system is its 90% CO₂ emission reduction compared with the cargo trucks on the road. It is believed that this system will be applied as an ITS in the near future due to its numerous merits.

ACKNOWLEDGEMENT

This work was supported by Korea Institute of Construction & Transportation Technology Evaluation and Planning.

REFERENCES

- [1] A Development of Core Technology of Railroad System, Korea Railroad Research Institute, 2002.
- [2] H. W. Kim, D. S. Moon, “A Primary Study on High Speed Intermodal Rail Freight Transportation,” Proceedings of the Korea Society for Railway Semiannual Spring Conference, 2003.
- [3] <http://www.modalohr.com>
- [4] <http://www.wabashnational.com/RoadRailer.htm>
- [6] http://sbbcargo.ch/index/magzin/02_domino.htm
- [8] <http://www.flexiwagon.se>
- [9] <http://www.cargospeed.ro>
- [10] <http://www.cargobeamer.com>
- [11] Y. J. Lee, H. C. Cho, J. W. Lee, D. S. Han, G. J. Han, and K. S. Lee, “Technical Trends and Analysis of Development of Dual Mode Trailer Systems for Enhancing Railroad Logistics,” 2009 IEEE International Conference on Automation and Logistics, 2009.
- [12] D. S. Han, J. M. Ha, Y. J. Lee, etc, “A Study on the Structural Design of a Horizontal Transfer System for DMT system,” Conference on Korean Society for Precision Engineering, 2009.
- [13] Thomas Pietsch, “Virtuelles Bahngleis,” Eurocargo, 6, 2003.
- [14] R. Rudel, “A Radical Innovation in the Intermodal Freight Market,” Alp-net Workshop Annecy, 2003.
- [15] <http://www.cargobeamer.com>
- [16] <http://www.cargobeamer.com>
- [17] M. Ruesch, Rapp Trans AG, “Research Initiatives and other Developments related to Intermodal Terminals in Switzerland,” UTP Steering Committee Meeting 9th March 2004.

Global Context Influences Local Decisions

Terry Bossomaier
CRiCS (Centre for Research in Complex Systems)
Charles Sturt University
Australia.
Email: tbossomaier@csu.edu.au

Michael Harré
CRiCS, Charles Sturt University
& Centre for the Mind, University of Sydney
Australia.
Email: mike.harre@gmail.com

Abstract—This paper studies the development of human expertise in the game of Go. Although superficially a simple game, Go is the most difficult of all established games for artificial intelligence, no computer program yet achieving top international level on a full 19x19 board. On smaller boards, such as 9x9 computers *are* competitive, implying that the understanding of the complex global interactions is the key to human superiority.

By mining thousands of positions online, we show that at some player levels the sequence of plays leading up to a *local* position is a stronger determinant of the next move than the position alone. This suggests that the sequence of plays is an indicator of global strategic factors and thus provides a context for the next move in addition to the local position itself.

Keywords-game of go; decision making; entropy; online data mining

I. INTRODUCTION

The big picture often influence or override local factors in many areas of human expertise, from board games to politics. Challenging games, such as Chess and Go, provide an excellent framework for studying expertise [1], [2], [3], [4] since they are both strategically deep but tightly constrained. This paper presents a striking demonstration of this, mined from thousands of decisions online. In recent work we have demonstrated transitions in the acquisition of expertise in the game of Go [5]. This game is interesting because it is currently the most difficult of all established games for computational intelligence. Unlike Chess, where the IBM computer Deep Blue [6], [7] triumphed over world champion Kasparov.

We also demonstrated therein, from calculation of mutual information between moves, that one of these transitions has the character of a *phase transition* [8]. The idea of a phase transition comes originally physics, such as the melting of ice to give water. When such a physical phase transition occurs there is a dramatic reorganisation of the system. In this case water molecules which were fixed rigidly in place in ice become free to move around and perhaps travel long distances. During a phase transition, systems exhibit long range order, where there are correlations in activity or structure over large distances and system parameters often exhibit power law behaviour, or fat-tailed distributions. Another example of a phase transition is in the formation of

random graphs. At the transition the average path length, in other words the number of steps from one node in the graph to another rises to a peak, and then drops back down again.

A dynamical system examples is the Vicsek model developed for studying magnetic transitions in solid state physics [9]. In this model particles travel around a two dimensional grid, and, when they come within some specified distance of each other, their directions of movement get slightly closer together. Phase transitions occur in this system as particles flow around in groups, like flocks of birds, but these groups are dynamic, continually forming and dissolving.

Mutual information is a system property which measures the extent to which the structure or behaviour of one part of a system predicts the behaviour of another. In the Vicsek model of above, at the transition the direction and velocity of one particle provides some information about the direction of all the other particles. The mutual information peaks during the phase transition [9], [10], and this is thought to be a general property of phase transitions along with the other characteristics, notably long range order and power law characteristics. We found a peak in mutual information as a function of rank amongst Go players from 1 Dan Amateur through to the very top players, 9 Dan Professional [8]. The previous work [8] has demonstrated phase transitions in collective human decisions in Go. This paper presents evidence that there is global influence on local decisions and that the influence is greatest during the phase transition.

A. State of the Art in Game Expertise

Much of the work on human expertise has been based on games, especially Chess, as in Gobet's extensive work [1], [11]. One of the key ideas, essentially from Nobel Laureate Herbert Simon, is that human expertise involves building a huge library of patterns [12], [13], although the application of these ideas in artificial intelligence for games is relatively new [14].

These patterns build up through the formation of *chunks*, and psychological observables, such as memory for Chess positions are well predicted by models such as CHREST [3]. The way the cognitive structures in the brain might change as expertise develops, and in particular the appearance of

phase transitions, is relatively new introduced by Harré and Bossomaier [8], [5].

Further recent advances have been limited, particularly in Go where a combination of the game space complexity of Go [15] and a lack of genuinely human like heuristics such as an evaluation function make progress difficult. However with the development of ever more effective random sampling techniques, such as the UCT-Monte Carlo approach currently favoured by AI system developers [16], some progress has been made in achieving strong amateur play. However these techniques do not address the inherent complexity of the game and the techniques that humans have developed in order to address such issues, almost completely because such techniques are not subject to easy investigation.

We argue that the sources of information players use in order to make good decisions are of two types: *local* and *global*. While every level of player in our study has learned a great deal about the game of Go over the course of their lives, it is how this information is implemented via the choices they make that is of interest to us. This relevance of the division of the problem space in to these two parts can be seen in the work of Stern et al [17]. They were able to produce ‘best in class’ move prediction for professional players in Go, achieving a 34% success rate. This was achieved by training their system on 181,000 expert game records and using the most modern techniques available for their analysis. The level of success achieved in this work highlights one of the principal difficulties of good performance in complex tasks: exact pattern matching is not enough; AI systems need to be able to model how non-local aspects i.e. information that cannot be derived by exactly matching board configurations, influence decisions. Loosely interpreted this is what is called *influence* in Go and before our recent work it had not been reported in the research literature.

II. METHODS

Harré and Bossomaier [8] examined the game trees 6 moves (i.e. 3 black and three white) deep for around 8,000 games across a range of Go expertise. At the low end were 2 kyu Amateurs, a rank reached by serious players after a couple of years club play through the highest amateur rank of 6 Dan Amateur (6A) to the top professional rank of 9 Dan Professions (9P). Game data was obtained from the *pandant* Go server. Full details of experimental procedures are given in Harré et al. [8]. The game trees were computed from 7x7 board sections in the corner, from games played between players of the same rank. No symmetry was exploited, apart from rotations to align each of the four corners (used to maximise data yield per game). Note that although these are the first 6 moves played in the region, they are not necessarily, and usually are not, the first 6 moves of the game.

For each possible move, m_i , three probability distributions were computed

- 1) the probability of the move, m_i , occurring, $P(m)$
- 2) the conditional probability, $P(m_I|q_i)$, of the move, m_i occurring from a given position, $q - I$
- 3) the conditional probability, $P(m_I|s_i)$ of the move occurring from a given position, *reached by a particular order of moves*, s_i

From these results the entropy and mutual information were calculated, but this paper addresses findings from the entropies alone. The move entropy, $H(M)$, is taken over all moves which can arise at each level in the game tree (i.e. for the 6 moves in the sequence). (eqn. 1).

$$H(M) = - \sum_i p(m_i) \log_2[p(m_i)] \quad (1)$$

Entropy is a measure of disorder or randomness and is maximal when the probabilities of all events are the same. When the first move in the region there are 49 possible positions and after 5 moves, 44, giving a maximal entropy of $\log_2 44 = 5.5$ bits. But since the moves are far from random the measured entropies are much lower than this.

The conditional entropy, $C(M|q_i)$, is the move entropy calculated from the moves which can arise in a given context, such as a position, q_j , or sequence of moves, s_j , leading to a position.

$$C(M|Q) = - \sum_i p(m_i|q_j) \log[p(m_i|q_j)] \quad (2)$$

with the same expression used for an ordered sequence of moves with s_j replacing q_j .

These entropic quantities are now calculated across all ranks from amateur 2q, denoted *am2q* through, the amateur dan ranks to *am6d*, and then to the highest rank of all, professional 9 dan, *pr9d* and are shown in figures 1–3.

III. RESULTS

Figure 1 summarises the key findings of the paper. It shows the conditional entropy as a function of move in the sequence of 6 averaged across all ranks, both amateur and professional. Error bars are calculated as in Harré et al. [5]. Up to move 3 the entropy for both the ordered and unordered cases are the same. At move three they fall dramatically, but the ordered average falls about a third more.

Figure 2 shows the entropy at each move from a given position. For purely random moves the entropy at each move in the sequence would be between 5 and 6 bits. The entropies observed are, of course, much lower, usually less than 2 bits, reflecting the structure inherent in the game.

The entropy for the third move is slightly less than for the first two, but the entropy falls a little for the fourth move and a lot more for the fifth and sixth moves. This is not surprising given the reduced options available as the number

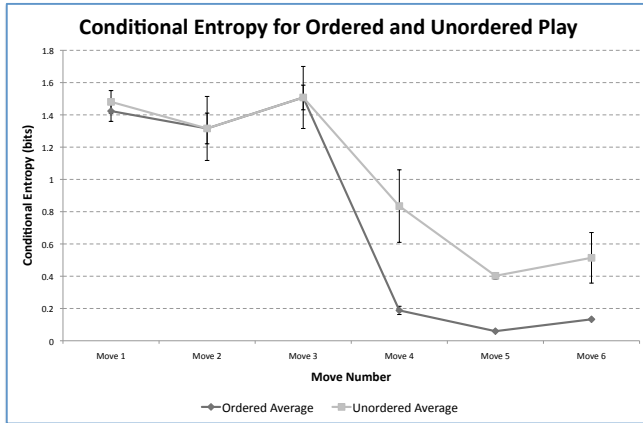


Figure 1. Conditional entropy (eqn. 2) as a function of move averaged over all ranks

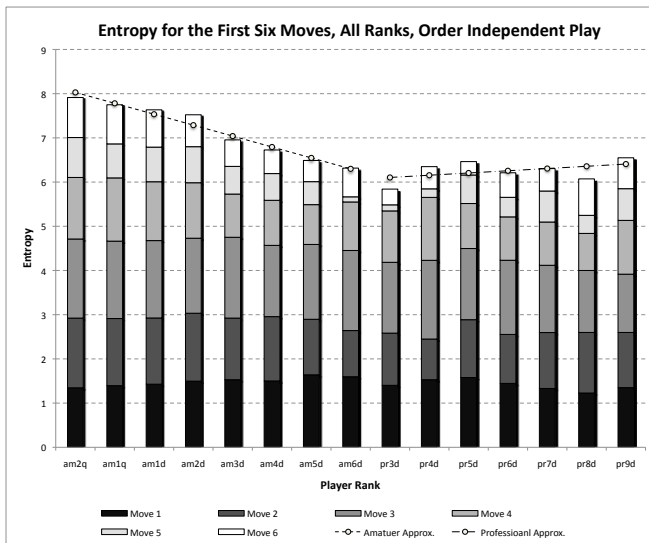


Figure 2. Entropies for moves from a given board position (reprinted from [8])

of stones on the board increases. The entropy summed over all moves declines linearly to the maximum amateur rank and then increases slightly from from the first professional rank.

Figure 3 shows the entropies which result from positions which arose from a particular sequence of play. These entropies are around 3 bits smaller, than for the unordered case. The slope of the regression line for the amateur levels is not so large, but the trend for the professionals displays a different pattern: the summed entropy jumps near the start of the professional ranks and then *decreases* with rank up to 9P with a slope very similar to that for the amateur ranks on the left of the figure.

The most interesting thing about this figure, though, is the way the entropy for the last three moves shrinks and vanishes as the amateur rank increases from 4A to 6A. In

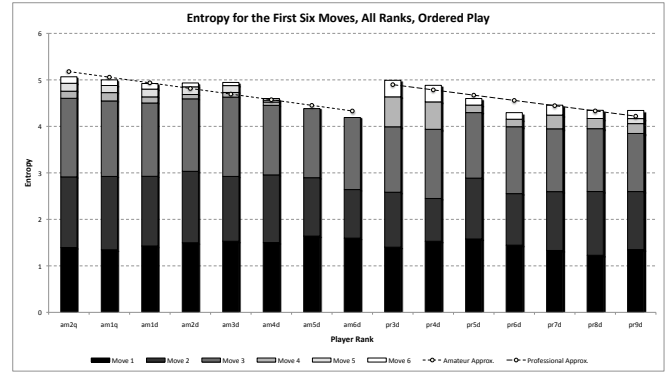


Figure 3. Entropy for the first six moves shown as a stacked bar chart. The black bars represent the entropy at move 1, the dark grey at move 2 and so on for all six moves. The dashed regression lines show the total entropy for the amateur and professional sequences.

fact the summed entropy for the first three moves is quite similar to the unordered case, so the three bit loss is almost all in the last three moves.

IV. DISCUSSION

There are two very interesting features of these results, which we consider in turn: the difference between ordered and unordered play; and the way the conditional entropy varies with rank.

That the ordered and unordered play differ, implies that the position at each move is *not* the sole determinant of the opponent response. The much lower conditional entropy after the first three moves for the ordered case strongly suggests that the sequence of moves has revealed something of the global context which has in turn fed back into move selection. To see this, imagine that black is strong in one area of the board and white in another. Since communication lines are of great strategic importance in Go, the locations of these areas will strongly influence the order of moves made in the local area we examine. The first three moves implicitly contain some of this information, which subsequently reduces the range of options in the second three moves.

The gradual decline in entropy with rank for amateur and professional reflects a gradual reduction in the space of range of options, which we could see as the elimination of poor moves in established situations, similar to the mastering of the opening in Chess.

Our data and results are explicitly based on an analysis of the local information, but by implication they also say a great deal about the global context that influences these localised decisions. The first three moves in our study have a reasonably similar conditional entropy of about 1.4-1.6 bits of information. This is the amount of information that is common between each successive move within the local region. Such measures of information are the best estimate of how much one stochastic variable can tell us about another [18]. The only other source of information available

to the players are the pieces on the board that were not included within our local region. We exclude the possibility of being able to read the other opponent. While it is a debated issue as to the importance of such skills, we believe that it is much less significant than all the other pieces on the board that were not within the local area of study. The changing influence that non-local information has on decisions during a game, is evident in the the significant drop-off in the conditional entropy after move 3 in Figure 1, a drop-off of shared information from one move to the next of nearly an order of magnitude for the ordered play and about half that for unordered play.

This change in the levels of conditional entropy as the game progresses in the corner region of the board might be due to the reducing size of the move space as the board fills up. Will this might have some minor influence on our results, we should also expect such changes to be almost linear as the number of available positions only drops by a total of 1/49 per move. It is also possible, but exceptionally unlikely, that after move 3 players choose much more randomly, i.e. without concern for the pieces on board either local or non-local, than they did for the first three moves. Considering the vast training literature available to players that readily teach them the many different variations of the first 6 moves within the corner, and then how to contextualise these decisions by considering what pieces occupy nearby areas, we consider this to be an unlikely strategy.

Instead we argue that it is just this external influence, the influence of the stones arrayed on the rest of the board that is having such a striking influence on the condition entropy. This is perhaps not so surprising when considered in the light of the state of the game itself after 3 moves have been played in the corner. These first moves can be thought of as establishing the game board layout in terms of an ‘opening book’, highly stylised moves of local pieces where the local pattern can be thought of as essentially uncoupled from the rest of the board, or at least coupled to the same extent for these first moves. This coupling then changes significantly from the 4th move onwards where greater consideration needs to be afforded to the other pieces on the board. This change in focus of the information effectively reduces significantly the information coupling between the local moves and the local stones on the board.

The complete disappearance of entropy at the high amateur ranks is very interesting. It suggests that the at this level play has become somewhat stereotyped, and a major change in thinking is needed to advance, which indeed seems to happen on turning professional. Thus this loss of entropy is consistent with the long range order found in phase transitions [8]. This accords with the findings in Harré et al [8], wherein a peak in mutual information is found at the transition to professional, indicating some sort of major cognitive reorganisation. At present we do not know how to quantify such a reorganisation and this remains an exciting

open question. Ongoing work is attempting to apply the CHREST models to Go [3] and to determine how phase transitions might be predicted.

ACKNOWLEDGEMENTS

This work was supported by the Australian Research Council under Discovery Project DP0881829 and the US Airforce under grant 104116.

REFERENCES

- [1] F. Gobet, A. d. Voogt, and J. Retschitzki, *Moves in Mind*. Psychology Press, 2004.
- [2] A. D. Groot and F. Gobet, *Perception and memory in chess: Heuristics of the professional eye*. Assen: Van Gorcum, 1996.
- [3] F. Gobet, P. Lane, S. Croker, P. Cheng, G. Jones, I. Oliver, and J. Pine, “Chunking mechanisms in human learning,” *Trends in Cognitive Sciences*, vol. 5, pp. 236–243, 2001.
- [4] K. Ericsson and N. Charness, “Expert performance: its structure and acquisition,” *American Psychologist*, vol. 49, pp. 725–7247, 1994.
- [5] M. Harré, T. Bossomaier, C. Ranqing, and A. Snyder, “The development of expertise in a complex environment,” *Minds and Machines*, vol. in press, 2011.
- [6] X. Cai and D. Wunsch, “Computer Go: A grand challenge to AI,” in *Challenges for Computational Intelligence*. Springer Berlin, 2007, pp. 443–465.
- [7] M. Campbell, A. Hoane, and F. Hsu, “Deep blue,” *Artificial Intelligence*, vol. 134, no. 1-2, pp. 57–83, 2002.
- [8] M. Harré, T. Bossomaier, A. Gillett, and A. Snyder, “The aggregate complexity of decisions in the game of go,” *European Physical Journal B ERA A*, vol. in press, 2011.
- [9] R. Wicks, S. Chapman, and R. Dendy, “Mutual information as a tool for identifying phase transitions in dynamical complex systems with limited data,” *Phys. Rev. E*, vol. 75, 2007.
- [10] S.-J. Gu, C.-P. Sun, and H.-Q. Lin, “Universal role of correlation entropy in critical phenomena,” *journal of physics A*, 5 2006.
- [11] F. Gobet and P. Chassy, “Expertise and intuition: a tale of three theories,” *Minds and Machines*, vol. 19, pp. 151–180, 2009.
- [12] W. Chase and H. Simon, “The mind’s eye in chess,” in *Visual Information Processing*, C. W.G., Ed. Academic Press, NY, 1973, pp. 215–281.
- [13] F. Gobet and H. Simon, “Five seconds or sixty? presentation time in expert memory,” *Cognitive Science*, vol. 24, pp. 651–682, 2000.
- [14] J. Rubin and I. Watson, “A memory-based approach to two-player texas hold’em,” in *AI 2009: Advances in Artificial Intelligence, Proceedings*, ser. Lecture Notes in Artificial Intelligence, A. Nicholson and X. Li, Eds. Springer, 2009, vol. 5866, pp. 465–474, 22nd Australian Joint Conference on Artificial Intelligence DEC 01-04, 2009 Melbourne, AUSTRALIA.

- [15] J. Tromp and G. Farneböck, “Combinatorics of go,” *Computers and Games*, pp. 84–99, 2007.
- [16] S. Gelly and Y. Wang, “Exploration exploitation in go: Uct for monte-carlo go,” in *Twentieth Annual Conference on Neural Information Processing Systems (NIPS 2006)*. Citeseer, 2006.
- [17] D. Stern, R. Herbrich, and T. Graepel, “Bayesian pattern ranking for move prediction in the game of go,” in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 873–880.
- [18] C. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Univ. Ill. Press, Urbana, 1949.

MODELING AND SIMULATION OF PETRI NETS FOR COMPLEX SCHEDULING RULES OF AUTOMATED MANUFACTURING SYSTEMS

Chulhan Kim and Tae-Eog Lee

Department of Industrial and Systems Engineering
KAIST
291 Daehak-ro, Yuseong-gu, Daejeon 305-701, South Korea

E-mail: chulhan.kim@kaist.ac.kr, telee@kaist.ac.kr

ABSTRACT

Discrete event systems such as automated manufacturing systems and engineering systems can be modeled and simulated by Petri nets. Precedence relations between activities or events, concurrent processes, synchronization, resource sharing, mutual exclusion, etc can be well modeled by Petri nets. However, discrete event systems, especially discrete event manufacturing systems, tend to have diverse complex scheduling rules to optimally utilize the resources, meet scheduling requirements and constraints, and optimize the performance measures such as makespan or cycle time. In this paper, we propose ways of modeling such complex scheduling rules by controlling the firing sequences and timing of the associated transitions in the Petri net model. We also present a Petri net model for scheduling a robotized indexer cell for flat panel display manufacturing.

Keywords: Petri net, scheduling, simulation, indexer

1. INTRODUCTION

A Petri net is a graphical and mathematical modeling framework for discrete event systems (Murata 1989). Discrete event systems such as automated manufacturing systems and engineering systems can be modeled and simulated by Petri nets. Transitions, places, arcs, and tokens represent activities or events, conditions or resources, precedence relations between transitions and places, and availability of resources or conditions, respectively. They are graphically represented by rectangles, circles, arrows, and dots, respectively. A more formal definition including transition enabling and firing rules can be found in Murata (1989). Precedence relations between activities or events, concurrent processes, synchronization, resource sharing, mutual exclusion, etc can be well modeled by Petri nets.

Scheduling is to determine the order and timings of processing jobs at a resource. The processing order may be fixed and independent of the system state such as the number of jobs at each resource, or can be dynamically changed depending on the system state. When there are multiple resources that can process a job,

a resource should be selected to process the job. Discrete event systems, especially discrete event manufacturing systems, tend to have diverse complex scheduling rules to optimally utilize the resources, meet scheduling requirements and constraints, and optimize the performance measures such as makespan or cycle time.

In a Petri net model, resources and activities that contend for using a resource can be modeled by a place and its output transitions, respectively. The tokens at such a conflict place are properly routed to the output transitions. Therefore, we expect that complex scheduling rules can be modeled by token routing rules, which can depend on the system state information such as the number of tokens and the token sojourn times at the places. However, it is not so straightforward to represent complex scheduling rules by token routing rules. For instance, even a simple state-independent cyclic scheduling rule that processes jobs at a resource in a fixed cyclic order is not simply modeled by routing tokens at a conflict place cyclically. It is because a transition firing for processing a job may be delayed due to delayed arrivals of tokens at the other input places of the transition. Furthermore, there are not enough studies on modeling and simulating scheduling rules in Petri nets. Most Petri net simulators can model decision-free nets that have no scheduling decisions or conflict places, or nets that have probabilistic token routing rules. There are few works on Petri net models or simulators that can model complex token routing rules or scheduling rules effectively.

In this paper, we propose ways of modeling such complex scheduling rules by controlling the firing sequences and timings of the associated transitions in the Petri net model. As an application, we also present a Petri net model for scheduling a robotized indexer cell for flat panel display manufacturing.

2. FIRING POLICIES OF PETRI NETS

We explain policies for controlling firings of transitions in a Petri net. Consider an example of Petri net in Figure 1. Places p_1 and p_4 are conflict places, each of which has two output transitions. A token routing rule basically determines an output transition to be fired at

each conflict place. A token that entered a conflict place and has stayed there as long as the token holding time of the place, if any, can be released to one of the output transitions of the conflict place. The token released to the output transition can join enabling the transition. The release can be made as soon as the token is ready to be released or delayed. An enabled transition can be fired immediately or after a prescribed firing delay of the transition, if any. It also can be delayed. Therefore, the token routing, delays of token releases and transition firings can be controlled by modeling and simulating scheduling rules. A *token routing rule* indicates such token route and release timing. A token routing rule eventually determines the order and timings in which the output transitions are fired. In the sense, it can be told that the scheduling decisions are made by a *transition firing policy* that determines the order and timings of firing the output transitions of each conflict place. When some intentional delays are made in scheduling decisions, but most practical scheduling rules for manufacturing systems start an activity or job as soon as possible. In other words, the token release and transition firings are not intentionally delayed. Therefore, in this paper, we focus on controlling only the order of transition firings. Of course, job release control rules such as kanban or CONWIP(Constant Work in Progress) intentionally delay jobs or activities. Therefore, such scheduling rules might be modeled by token routing rules or transition firing policies that make intentional delays in token release or transition firing appropriately. However, even such scheduling rules for timing regulation can be incorporated into Petri net model as an appropriate subnet that provides feedback of tokens from some transitions to other transitions (Lee and Park 2005; Mascolo, Frein, Dallery and David 1991). We therefore define four types of transition firing policies that determine only the order of transition firings.

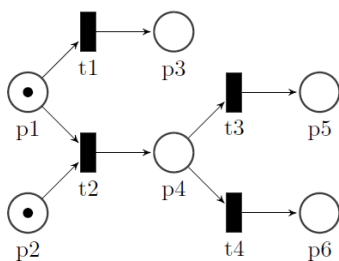


Figure 1: A Petri net which has two conflict places.

2.1. Probabilistic sequence

In a probabilistic sequence, the output transitions of a conflict place are fired in a random order. For example, in Figure 1, output transitions t_1 and t_2 of the place p_1 are fired with the same probability, 0.5. Of course, the probabilities may be unequal. Probabilistic sequence can be used for describing systems of which the order of activities is not important and systems with decision makers that behave randomly.

2.2. Cyclic sequence

A cyclic sequence is a fixed sequence which repeats firing the transitions based on a specified cycle. We use the following expression for a cycle:

$$C(S_1, S_2, \dots, S_n)$$

where S_i is i th firing transition and n is the length of the order list. For example, a cycle $C(t_a, t_b, t_b)$ repeats transition firing by an order of $(t_a \rightarrow t_b \rightarrow t_b \rightarrow t_a \rightarrow t_b \rightarrow t_b \rightarrow t_a \rightarrow t_b \rightarrow t_b \rightarrow \dots)$.

A cycle is *feasible* if every transition can be fired continuously by the order of it. The Petri net in Figure 1 has $4! = 24$ possible cycles but there is no feasible cycle. On the other hand, the Petri net in Figure 2 has $2! = 2$ cycles and $C(t_1, t_2)$ is feasible.

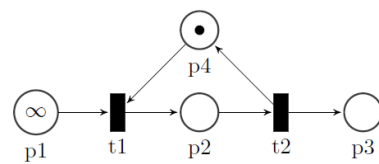


Figure 2: A Petri net which has a feasible cycle, $C(t_1, t_2)$.

2.2.1. K-cyclic sequence

A k -cyclic sequence is a special case of the cyclic sequence. It selects a firing transition based on a specified cycle which contains every transition of the net. Transitions are fired by the order of cycle and every transition is fired k times in a cycle. For example, 2-cyclic sequence $C(t_a, t_c, t_b, t_b, t_c, t_a)$ means the transitions are fired by an order of $(t_a \rightarrow t_c \rightarrow t_b \rightarrow t_b \rightarrow t_c \rightarrow t_a \rightarrow t_a \rightarrow t_c \rightarrow \dots)$.

Cyclic sequence is useful to describe systems in which the several events are repeated by a cyclic order. In case of the k -cyclic sequence, since every transition should be fired k times in a cycle and it is independent of the marking of Petri nets, the problem complexity is much more reduced. Hence, much research has been done for finding optimal cycles to maximize/minimize the value/cost using mathematical programming techniques.

2.3. Non-cyclic sequence

Some decision makers such as distribution machines in manufacturing cells may behave without any pattern. A non-cyclic sequence can be used to describe this kind of non-repeating firing orders. A list is described as

$$N(T_1, T_2, \dots, T_m)$$

The only difference from the order list of cyclic sequence is that it does not fire any transition after firing the last one in the list, regardless of the existence of enabled transitions. In most cases it leads to the end of simulation.

2.4. Rule

Probabilistic, cyclic, and non-cyclic sequences select a firing transition based on random numbers or a pre-specified order list which means that they do not care about the marking of Petri nets. In contrast, rule policy is marking-dependent. It means that we can use rule policy when we describe systems with decision makers that depend on the state of the systems.

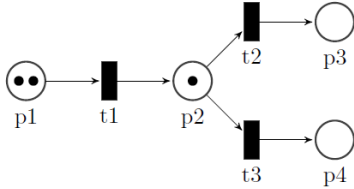


Figure 3: A Petri net.

Table 1: Two rules for a Petri net in Figure 3.

Rule	Statements
#1	If $M(p2) = 0$, then fire $t1$. Else if $M(p3) \geq 1$, then fire $t3$. Else, fire $t2$.
#2	If $M(p3) + M(p4) \leq 1$, then fire $t1$. Else if $M(p2) \geq 2$, then fire $t2$. Else, fire $t3$.

Basically, a *rule* is a series of *if-then* statements. Each statement follows a format of “if condition C is satisfied, then fire transition T”. Every time a transition is fired, the rule is considered to select a next firing transition. Table 1 shows two sample rules for a Petri net in Figure 3 where $M(p)$ is the number of tokens in place p . We do not state the enabling condition of transition in a rule for simplicity. For example, the first statement of Rule #1 in Table 1 actually means “If $t1$ is enabled and $M(p2) = 0$, fire $t1$.”. The firing order based on Rule #1 and Rule #2 in Table 1 are $(t2 \rightarrow t1 \rightarrow t3 \rightarrow t1 \rightarrow t3)$ and $(t1 \rightarrow t1 \rightarrow t2 \rightarrow t2 \rightarrow t3)$, respectively.

Using probabilistic, cyclic, non-cyclic sequences and rule policies properly, we can model diverse types of discrete event systems and simulate them more effectively.

3. GLOBAL AND LOCAL FIRING POLICY

In section 2, we defined four firing policies to describe various discrete event systems with Petri nets. In this section, we discuss the meaning of global firing policy, local policy, and the difference between adapting a single global firing policy and several local firing policies to a Petri net.

3.1. Global firing policy

A *global firing policy* is a firing policy that deals with every transition of a Petri net. For example, the global cycle policy should include all transitions of a Petri net in its firing order list. We use a single global firing policy for a Petri net to model a system with a single decision maker that controls the whole system.

3.2. Local firing policy

Some systems may have several independent decision makers like a manufacturing cell with two independent transportation robots. They act without considering other decision makers. Describing this kind of system with a global firing policy may not be easy. In this case, we use several *local firing policies* for each *conflict set* in a Petri net.

Definition: A *conflict set* of a Petri net is a subnet which includes two nonempty sets P_c and T_c , where

1. P_c is the smallest set of places such that $\bigcup_{p_i \in P_c} (p_i \bullet) = T_c$ where $p_i \bullet$ is the set of output transitions of p_i .
2. T_c is the smallest set of transitions such that $\bigcup_{t_i \in T_c} (\bullet t_i) = P_c$ where $\bullet t_i$ is the set of input places of t_i .

For example, in Figure 1, there are two conflict sets. The first one is a subnet including $p1$, $p2$, $t1$, and $t2$. Second one consists of $p4$, $t3$, and $t4$. There is no conflict between every pair of conflict sets. It means that firing of a conflict set does not disable transitions in the other conflict sets. That is, every conflict set works *concurrently*. In case of the Petri net in Figure 1, we need to specify two firing policies for each conflict set to describe the system with two independent decision makers.

4. SIMPN

There have been already a lot of free and commercial Petri net tools/simulators available. However, they are not well suited for modeling and simulation of discrete event systems which behave based on complex operating schedules with them because most of them deal with conflict situations randomly or by asking the users to select one of enabled firing transitions. Based on firing policies discussed in section 2 and 3, we developed a java-based p+-time simulator, *SIMP*N. Figure 4 illustrates the main frame of *SIMP*N.

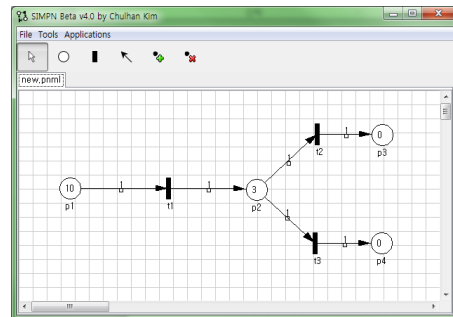


Figure 4: SIMPN

P+-time Petri net is an extension of a timed places Petri net (Wang 1998) whose places have random token holding time and maximum token delay. It is formally defined by Kim and Lee (2008) to be used for modeling cluster tools with non-deterministic processing time and time window constraints. P+-time Petri net is a useful

extension of the Petri net that can describe various types of discrete event systems especially manufacturing systems.

4.1. Simulator

The simulator of SIMPN supports Petri net simulations with a global firing policy and local policies. Basically, a simulation engine works based on the discrete event simulation framework. It controls an event list which prioritizes the events according to the times of event occurrences.

The users should specify firing policies for the Petri net to simulate it. In case of simulation with local firing policies, SIMPN automatically finds all conflict sets of the Petri net.

4.2. Statistical analysis

The users can make the simulator gather data from simulation and analyze the results. The following is a list of possible statistical analysis from SIMPN:

- Average token inter-arrival time of a place,
- Time interval between two firing epochs of transitions, and
- Gantt chart of token holding time of a place.

For example, in a manufacturing line, first option can be used when we want to know the average time interval between two consecutive finished products (cycle time). Second option can be used to measure the whole processing time of products (turnaround time). Lastly, we can make SIMPN draw a Gantt chart for a place to see overall status of processing chamber and to check patterns.

5. APPLICATION: LCD INDEXERS

An LCD indexer is a machine for sorting, loading and unloading glasses in the LCD panel manufacturing. Robot arms are responsible for transportation of glasses between two processing chambers. Therefore, the behavior of the robot arms is closely related to the overall performance or productivity of the LCD indexer.

LCD indexers are quite similar with cluster tools for the semiconductor manufacturing. A cluster tool combines several single-wafer processing modules with wafer handling robots in a closed environment (Lee 2008). There is much research for evaluating the performance of cluster tools (Chan, Yi and Ding 2010; Dwande, Geismar, Sethi and Sriskandarajah 2007) and finding optimal robot arm schedules, especially marking-independent cyclic schedules using timed Petri nets and mathematical programming techniques (Jung 2010).

However, finding optimal robot behavior of an LCD indexer is not simple due to some differences from cluster tools. First, we cannot assume processing time as deterministic parameters because the variance is too large to ignore. Second, some chambers may have time window constraints. That is, if a glass stays too long in a chamber with time window constraint, it becomes

defective due to heat or hazardous gases. Similar cases for cluster tools are introduced in (Kim, Lee, Lee and Park 2003; Lee and Park 2005; Kim and Lee 2008). However, they are about analyzing schedulability analysis of timed event graphs which are special cases of timed Petri nets. Third, some modules can process several glasses at once. Batch process tends to make the problem much complicated. Last main difference is that processing order of some abnormal glasses may not be the same as that of the other glasses.

For LCD indexers, we cannot get an optimal operating schedule using mathematical programming techniques because of their stochastic behavior and high complexity. Simulation approach can be used to find good schedules for them. We introduce a case of finding robust and efficient *dispatching rules* of an LCD indexer, annealing oven line.

5.1. Annealing oven line

An annealing oven line is a type of LCD indexer which anneals glasses with heat and cools them. It consists of one or more ovens, coolers and a robot arm. A single-arm robot is responsible for transportation of glasses. Figure 5 illustrates the annealing oven line that we deal with.

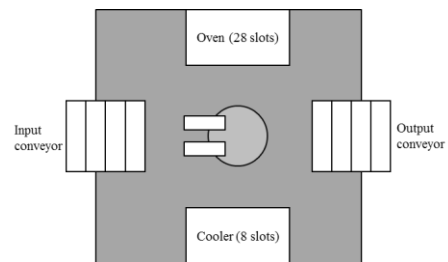


Figure 5: Annealing oven line

Normal glasses are unloaded from the input conveyor, visit oven, cooler and output conveyor, sequentially. On the other hand, abnormal glasses are not processed in the oven and cooler. They just move from the input conveyor to the output conveyor.

We assume no batch process in the system and 1% of total glasses are abnormal. Minimum and maximum processing times of the oven and cooler are 200, 300, 150 and 200 seconds, respectively. Due to hot environment, a glass should be unloaded from the oven within 100 seconds after it is processed. If a finished glass is not unloaded within the specified time window constraint, severe quality problems occur. Times for unloading/loading and transportation are all taken to be 1.

Figure 6 shows a p+-time Petri net model of the system and the meaning of each place and transition is explained in Table 2. In addition, Table 3 shows the time information of each place.

5.2. Deadlock prevention

There are two types of deadlock in this problem. First, deadlock occurs when the robot tries to load a glass into

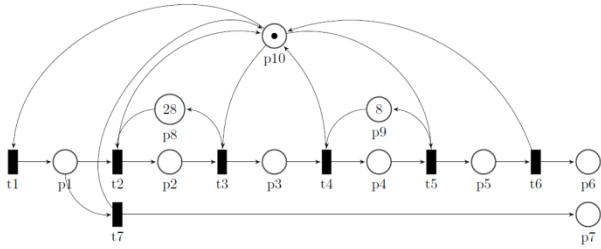


Figure 6: A p+-time Petri net model of annealing oven line in Figure 5.

Table 2: Legend for Figure 6.

Name	Meaning
p1	Robot moves a glass from input conveyor to oven.
p2	Oven is processing.
p3	Robot moves a glass from oven to cooler.
p4	Cooler is processing.
p5	Robot moves a glass from cooler to output conveyor.
p6	Finished normal glasses.
p7	Finished abnormal glasses.
p8	Availability of oven.
p9	Availability of cooler.
p10	Availability of robot arm.
t1	Unloading a glass from the input conveyor.
t2	Loading a normal glass to the oven.
t3	Unloading a normal glass from the oven.
t4	Loading a normal glass to the cooler.
t5	Unloading a normal glass from the cooler
t6	Loading a normal glass to the output conveyor.
t7	Loading an abnormal glass to the output conveyor.

Table 3: Time information of each place in Figure 6.

Place	Holding time	Time window constraint
p1	2	∞
p2	(200, 300)	100
p3	2	∞
p4	(150, 200)	∞
p5	2	∞
p6	0	∞
p7	0	∞
p8	0	∞
p9	0	∞
p10	2	∞

the oven or the cooler without an empty slot. We can prevent this kind of deadlock by defining conditions in firing rules as follows:

1. If $M(p8) \geq 1$, fire t1, and
2. If $M(p9) \geq 1$, fire t3.

Second possible deadlock is due to the time constraint of the oven. If a glass is not unloaded within 100 seconds after processing, the oven stops working.

Since we cannot directly control the processing times of the oven and cooler, this type of deadlock is almost impossible to be prevented by defining conditions. A good way to avoid the problem is changing the availabilities of the oven and the cooler on purpose. In this case, there are too many slots in the oven, so the cooler may not be able to handle coming glasses fast enough. Therefore, reducing the number of usable slots in the oven can help prevent deadlock.

5.3. Defining firing rules

Now we define firing rules for two conflict sets which have more than one output transition. For the first conflict set, we have to define firing rules for selecting a firing transition between t2 and t7. Since 1% of glasses are abnormal, rules can be set up as follows:

If $RND \leq 0.01$, fire t7;
Otherwise, fire t2,

where RND is a random number between 0 and 1.

The other conflict set has transitions t1, t3, and t5. This conflict set represents the behavior of the robot. Firing policies for the robot are directly related to the performance of this anneal oven line. For a simple experiment, we define two firing rules in Table 4. As the rules in Table 1, we omit the conditions for the enablement of transitions. For example, first statement in Rule #1 means "If t5 is enabled, fire t5."

Table 4: Two rules for a Petri net in Figure 6.

Rule	Statements
#1	Fire t5. Else if $M(p9) \geq 1$, then fire t3. Else if $M(p8) \geq 1$, fire t1.
#2	If $M(p8) \geq 1$, fire t1. Else if $M(p9) \geq 1$, then fire t3. Else, fire t5.

Briefly, a robot with Rule #1 always unloads a glass from the cooler whenever its cooling process is done, while a robot with Rule #2 always loads a glass to the empty oven slot right after it becomes available. The only difference between two rules is the priority of statements.

5.4. Simulation

Simulation experiments are conducted with SIMPN. First, we find the maximum number of available oven slots while avoiding deadlock. It can be simply found by doing simulation reducing the availability of oven one by one until no deadlock occurs during the simulation.

After finding the availability, we obtain the average cycle time of normal glasses. In this case, we can get this result by analyzing the average inter-arrival time of tokens in p6.

Simulation results with simulation time 100,000 are in Table 5. In case of Rule #1, even though all the oven slots are available, the average cycle time of

glasses is much greater than the case of Rule #2 which has 9 available slots. That is because it uses only one oven slot and cooler slot at once. In summary, Rule #1 guarantees no deadlock, but the efficiency is limited. On the other hand, Rule #2 ensures the good average cycle time even though it may have to disable some oven slots on purpose.

Table 5: Simulation results

Rule	Max availability	Average cycle time
#1	28	434.51
#2	9	28.86

6. CONCLUSION

We defined firing policies for Petri nets: probabilistic, cyclic, non-cyclic sequences and rule policy. These firing policies can be used to model complex and dynamic schedules of the system. One global firing policy is defined for a Petri net if there is only one decision maker. Several local firing policies can be also adapted to a Petri net in order to describe systems with multiple decision makers working independently. Petri nets with dynamic firing policies cannot be modeled mathematically in most cases. Simulation approach is valuable for efficient operating schedules for this kind of systems. We developed SIMPN which is a java-based p+-time Petri net simulator with firing policies. We conducted simulation for an annealing oven line for a case study. Experiment results showed the productivity of the system to be highly dependent upon its operating rules.

For further study, firing policies can be more generalized and well-organized for more general schedule. In addition, we have to specify a general framework for finding good firing policies especially marking-dependent firing rules.

ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0011438).

REFERENCES

- Chan, W.K.V., Yi, J. and Ding, S., 2010. Optimal scheduling of multicluster tools with constant robot moving times, Part I: Two-cluster analysis. *IEEE transactions on automation science and engineering*, 8 (1), 5–16.
- Dawande, M.W., Geismar, H.N., Sethi, S.P. and Sriskandarajah, C., 2007. *Throughput optimization in robotic cells*. USA: Springer Science + Business Media, LLC.
- Jung, C., 2010. *Cyclic scheduling of timed Petri nets: Behavior, optimization, and application to cluster tools*. Thesis (Ph. D.). KAIST.
- Kim, J.H., Lee, T.E., Lee, H.W and Park, D.B, 2003. Scheduling analysis of time-constrained dual-

armed cluster tools. *IEEE transactions on semiconductor manufacturing*, 16 (3), 521–534.

- Kim, J.H. and Lee, T.E., 2008. Schedulability analysis of time-constrained cluster tools with bounded time variation by an extended Petri net. *IEEE transactions on automation science and engineering*, 5 (3), 490–503.
- Lee, T.E. and Park, S.H., 2005. An extended event graph with negative places and tokens for time window constraints. *IEEE transactions on automation science and engineering*, 2 (4), 319–332.
- Lee, T.E., 2008. A review of scheduling theory and methods for semiconductor manufacturing cluster tools. *Proceedings of the 2008 winter simulation conference*, 2127-2135. December 7-10, Miami, F.L., USA.
- Mascolo, M., Frein, Y., Dallery, Y. and David, R., 1991. A unified modeling of Kanban systems using Petri nets. *International journal of flexible manufacturing systems*, 3 (3-4), 275–307.
- Murata, T., 1989. Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77 (4), 541–580.
- Wang, J., 1998. *Timed Petri nets: Theory and application*. USA: Kluwer academic publishers.

AUTHORS BIOGRAPHIES

CHULHAN KIM received the B.S. degree in Korea Advanced Institute of Science and Technology (KAIST) in 2010. He is in integrated master's and Ph. D. program in the same university. His research interests focus on simulation of Petri nets, and cyclic scheduling with Petri nets using mathematical programming techniques.

TAE-EOG LEE is a Professor with the Department of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology (KAIST). He is also the head of the department. His research interests include cyclic scheduling theory, scheduling and control theory of timed discrete-event dynamic systems, and their application to scheduling and control of automated manufacturing systems.

A BUSINESS PROCESS MODELING APPROACH TO SUPPORT PRODUCTION SYSTEMS ANALYSIS AND SIMULATION

Claudia Battista^(a), Giulia Dello Stritto^(a), Francesco Giordano^(a), Raffaele Iannone^(b), Massimiliano M. Schiraldi^(a)

^(a) “Tor Vergata” University of Rome, Department of Enterprise Engineering,
Via del Politecnico 00133 Roma (ITALY)

^(b) University of Salerno, Department of Industrial Engineering,
Via Ponte Don Melillo, 84084 Fisciano, Salerno (ITALY)

^(a)claudia.battista@uniroma2.it, ^(a)giulia.dello.stritto@uniroma2.it, ^(a)francesco.giordano@uniroma2.it,
^(b)riannone@unisa.it, ^(a)schiraldi@uniroma2.it

ABSTRACT

In this paper we propose a reference model conceived to simplify the development of production simulation paradigms as well as to support software houses in formalizing the main functions and properties of manufacturing systems simulation software. The proposed model results from a research project aiming to the design of a new manufacturing systems simulation tool, embedding the main production and logistics processes archetypes. Indeed, the designed tool natively entrenches several well-known production and inventory control policies on top of the greatest part of the typical processes and work methods in a manufacturing plant; the model is formally represented in Business Process Modelling Notation, which increases its clearness and the related benefits for industrial users.

The proposed reference model’s architecture and working logic have been validated through a manufacturing company case study.

Keywords: manufacturing systems, reference model, simulation, software architecture

1. INTRODUCTION AND LITERATURE REVIEW

Simulation is considered as a useful tool to study and optimize production processes. Several authors agree on simulation potentialities in analysing dynamic and stochastic behaviour of manufacturing system, predicting its operational performance and pointing out its critical factors (Smith 2003; Hlupic 1999; Law 1991). However, a lack of a commercial software that merges the critical functions for modelling different manufacturing phases in a user-friendly way is reported in literature; despite simulation is always described as one of the best approach for improving system efficiency, these limits seem to prevent the diffusion of these tools in manufacturing enterprises (Rogers 1993, 2002).

At present, most of the simulation software available on the market implement a graphical model-building approach - the so called “development environment” - where experienced users can model

almost any type of process using basic function blocks; then, some user-defined statistical functions evaluate the whole system behaviour. Occasionally, formal meta-languages are used to describe the relationships among the components, such as resources, entities, etc. (Van Beek et al. 2008). The usage of multi-purpose simulation software requires, on top on advanced modelling and simulation knowledge and skills, great effort in translating the real industrial processes logic into the modelling scheme. However, strong competences in operations research or statistics have never been the traditional background of the analysts in industrial companies (Davis 1994). Several authors (Bodner and Mc Ginnis 2002; Narayanan et al. 1998; Mujtabi 1994) underline the need of a standard reference framework to model production and logistics processes as a success factor for wide spreading simulation software in manufacturing industry.

The definition of conceptual model and, specifically, the model requirements, the development methodology, the model representation and communication rules are important issues that need to be addressed at first (Robinson 2006). Thus, literature suggests to concentrate on a new reference model development for simulating systems that implements a structure and logic much closer to real production systems, and which may effectively support different kind of analysis of industrial processes.

Since the sixties, business process modelling has emerged as a practical solution for obtaining a better understanding of business processes with an approach similar to that used for representing physical control systems (Williams 1967). Nowadays, Object Group’s Business Process Modelling Notation (BPMN) has become the de-facto business process modelling standard: BPMN is recognized to be an effective approach to model generic workflows in the companies (Chinosi and Trombetta 2011) either inside or outside production or supply chain management context, and to translate the result of the employees interviews on processes and procedures in a formal representation. It allows representing processes putting in evidence the differences among a present state (“as is”) and an improved future state (“to be”), so it is particularly

useful to support the what-if analysis typical of simulation approaches. Keramati and al. (2011) have applied BPMN to model and simulate as-is and to-be situations of sale and distribution process for some Iranian companies. For simulation purposes, BPMN can be put at work in conjunction with XPD (WfMC's XPD), while WS-BPEL (OASIS's WS-BPEL) can be considered a possible choice for translating BPMN diagrams into directly executable code.

In this specific application, an Italian manufacturing company has been selected and BPMN has been used to represent its functions inside a reference model with the aim of defining a standard to represent resources interactions and their relationships in manufacturing systems. The reference model is useful to define how manufacturing systems resources relate to, one another, and how these can be modelled and which roles can be played by each one. The model has been used to support the design of a software simulation tool, called O.P.U.S. (the acronym stands for Optimizing Production Using Simulation), that allows to build and to simulate manufacturing process models in accordance to the main operations management theories, thus natively embedding various production and inventory control policies inherent the typical processes and operating methods in a manufacturing plant.

The choice of BPMN, which is strictly linked to business process but is directly translatable in XML format, has helped to simplify the implementation/coding of the simulation software, as well as to reach a further standardization level. The software could be easily written thanks to the fact that the java-objects could behave according to the modelled rules, e.g. a machine may send an item picking request to a stock buffer or may send a confirmation message for an item release to downstream resources. Thus the shop floor functions were directly translated from BPMN into the Java classes and methods of the related objects, in order to obtain a complete compliance among company business processes, real production/logistics procedures and tool simulation logic.

The remainder of this paper is organized as follows: section 2 presents the architecture framework. In section 3 the reference model working logic on which the kernel of simulation tool is based is described. In section 4 the company case study is reported, in order to highlight to software working procedure and to show the validation of the proposed reference model.

2. THE ARCHITECTURE FRAMEWORK

An architecture framework here indicates the general solutions for the design of models inside a given domain; in this case, production systems simulation software development. The reference model, build in accordance to the requirements of the architecture

framework, includes standard design patterns to describe the operations. In this case, all the different functions, rules and procedures underneath the working logic in a manufacturing system are included.

The main problems in using multi-purpose simulation software to improve company processes reside in the contrast between the specificity of the application contexts and the generic high-level approach of these kind of tools. Significant approximations are always required to adapt the simulated model to the real context and this requires a lot of time on top of specific modelling competencies, which are seldom present in manufacturing companies. In order to solve this issue, the authors proposed an architecture framework dedicated to simulate only manufacturing production process. Despite the restriction of the application field, this approach results to be effective thanks to its clearness and the high intelligibility of the model for an operations expert rather than for an IT specialist. Thus, the pillars of the proposed architecture are:

- compliance between real-world objects and programming objects;
- separation of communication and production/logistics events;
- use of a discreet events simulation approach;
- existence of a single event handler.

Being dedicated to manufacturing systems, this approach helps in minimizing the use of artificial paradigms to model reality.

The dual-layer architecture, as well as the difference between production events and communication events, allows user to easily transfer to the model the most commonly used production algorithms and standards (e.g. MRP, lot sizing techniques, re-order level inventory management systems, etc.), relying on the fact that the interaction among the involved entities and resources will be automatically defined by the simulation engine.

The centralized event handler allows the simulation tool to build a single database, that can be queried to calculate the performance indicators and production cycles, namely the ones linked to the objects (machinery, resources, stocks) involved in the process. This solution drastically simplifies software development. The event handler manages all events thanks to the Future Event List (FEL), a sort of calendar that is progressively generated and scanned; this ensures the dynamic execution of the simulation.

A reference model based on these architecture pillars can easily:

- be used and understood from non-IT experts;
- embed operations logics and algorithms;
- embed performance indicators used in specific industries.

As a consequence, the simulation tools developed in accordance to this reference model will comply without difficulties to the main operations management theories:

- the only input comes from the typical manufacturing systems data structures: Bills Of

Materials, Master Production Schedules, Process Charts, etc.;

- basic production processes archetypes (e.g. set-ups, machine failures, etc.) are natively supported and no abstraction effort is required to the analysts;
- main production and inventory management policies (e.g. look-back and look-ahead material management policies, etc.) are natively supported;
- the distinction between information and physical layer is clear - considering that the data structure will provide all the required information to complete the physical flows (i.e. the items processing sequence information are already defined into the process charts).

3. THE REFERENCE MODEL

The architecture features presented in the previous section guarantees “ease of use”, short development time and highly reliable simulation models: the user does not need to describe the basic functions logic because the conceptual archetypes of industrial production systems are embedded into the reference model.

The reference model structure replicates the exact manufacturing system dynamic: Master Production Scheduling or buffer replenishment requests initialize material flows. Depending on the Bill of Material (BOM) and process chart (item paths), “Picking request” and “Production order” flow upstream the production process in order to satisfy MPS orders or buffer replenishment requests. Indeed this structure evidences how the explicit definition of logic relationships among the objects of the model is not required, nor to define the process constraints.

The reference model basic elements are the machine object and the stock buffer object. Each of these objects has specific lists for managing the physical and information flows progress. Specifically, in the following table, a “communication function list” is reported both for the machine and the buffer objects; here, the coherence between the real manufacturing system dynamic and the proposed framework is put in evidence.

Table 1: Resources communication function list

Resource	List name	List functions
Buffer	Inventory on hand list	Controls items inventory stock: on its base, the simulation engine satisfies a picking request or sets out a replenishment request.
	Picking list	Records picking requests from downstream resources. Pending picking request are fulfilled at the time of required materials are available.
	Storage request list	Records storage requests from upstream resources. Pending storage request are satisfied when enough storage space becomes available in the storage destination.
	Item release list	Records items ready to be transported to downstream resources. Depending

		on production resources status – idle, busy, ecc – physical material flow is generated.
Machine	Production orders list	Records production orders requests from downstream resources.
	Items order list	Records items required to fulfil production orders.
	WIP list	Records working progress material on the specific machine.

The model works on an event based logic: both machine and buffer functions are triggered by an event occurrence. Ten main events have been identified to represent the typical manufacturing production process (**Errore. L'origine riferimento non è stata trovata.**).

Table 2: Events list

ID	Event	Description
Ev1	Picking request	An item is requested to a buffer stock by some entity downstream (i.e. by another buffer, by a machine or by the Master Production Schedule in case of finite products)
Ev2	Available item alert	An item, which has been previously requested, becomes available in a buffer or a in a machine
Ev3	Supply order	An item is requested to a supplier outside the companies boundaries
Ev4	Item release	An item is transferred downstream to a buffer stock
Ev5	Production order	An item is requested to be produced by a machine in the process
Ev6	Setup end	A setup is completed and the machine is ready to process another kind of item
Ev7	Failure occurrence	A failure occurs in a machine and the machining phase is stopped
Ev8	Reparation end	A machine is restored after the occurrence of a failure
Ev9	Production end w/scrap	A machining phase is completed but the result is not compliant to quality requirements
Ev10	Production end	A machining phase is completed and the result is compliant to quality requirements

Thus the reference model is event-driven; each simulation cycle is performed in five phases which are:

- 1) advance the simulation time (clock);
- 2) identify the events scheduled to the current time;
- 3) identify the elements to be activated (objects) along with the related functions (methods);
- 4) execute the selected functions and update the system values (variables);
- 5) schedule the future events.

With reference to phase 3), activated resources manage information and physical flow through specific methods and this properly represents the simulation working logic. Each of the 10 previously presented events triggers the activation of certain objects, according to the following rules (Tab 3 - \oplus and \vee symbols stand for XOR and OR Boolean operators):

Table 3: Triggered resources by event

Event ID	Triggered resources
Ev1	Buffer
Ev2	(Buffer \oplus Machine) \vee Buffer
Ev3	No resource is triggered by Ev3
Ev4	Buffer \vee Machine
Ev5	Buffer \oplus Machine
Ev6 to Ev10	Machine

Then the FEL generation process is a consequence of each event. For instance, if at time t_{now} a *failure occurrence* event is recorded on the FEL for a certain machine, the simulation engine reads, on the input tables, the mean-time-to-repair (MTTR) data for the specific failure, on the specific machine. Then the engine returns a random number according to a pre-specified distribution probability function with a pre-specified standard deviation and MTTR as average. This number (Δt) represents a single random occurrence of that time-to-repair. Thus, the simulation engine will write a *reparation end* event at time $t_{now} + \Delta t$ on the FEL.

From FEL process generation BPMN diagram (see annex A, fig. 1) it should be clear that the proposed reference model embeds a look-back logic: Master Production Scheduling or buffer replenishment requests set off production process. Specifically, *picking request* event manages information flow propagation to the resources upstream in the production path: the buffer stock, once fulfilled the picking request on the base of the items inventory level, would propagate the replenishment requests. Thus, *production order* or *picking request* are triggered as a consequence of replenishment needs (this is the reason for the expression “look-back”). Buffer working logic, triggered by *picking request* event, is highlighted in the figure 2 that represent picking function.

In the next section a case study is presented to describe OPUS modeling process and to verify and validate the reference model proposed.

4. THE CASE STUDY

The proposed approach here described was conceived as a result of a public funded research project carried on from 2007 to 2011 by the Italian universities of Rome “Tor Vergata” and of Salerno, related to the design and development of the prototype of a production/logistics processes simulation tool dedicated to manufacturing SMEs. Thus, the proposed architecture framework and the reference model have been validated on the case of an Italian manufacturing company that was selected to participate in the design and testing phases of the research project. Note that each resource method has been modelled with BPMN diagrams in order to facilitate communication both with the Italian manufacturing firm – that has been able to verify the reference model compliance to their real business and production processes – and with the software house that

was encharged to the software OPUS coding, translating each resource function into java-method of the related objects.

The case study aims to show some of the features on which the OPUS architecture is based. In particular, the typical workflow of the development of a simulator starting from a real company is shown. In order to verify and validate information transfer mechanisms, UnisaGest management software (prototype of an ERP software developed by the Operations Management research group at the Department of Industrial Engineering, University of Salerno) was used which was connected to OPUS environment through a JDBC-ODBC connection that provides data transfer necessary for the operation of the simulator. The objective of the simulation is to verify the saturation of production resources for a given production plan for a period of 6 months. Therefore, for the validation of the model, media saturation data will be used for the machinery in manufacturing the product. The Execution Control is performed through a MES system continuously fed with data coming from the production environment. The company produces components for the automotive industry and it has several manufacturing plants in Italy.

For the construction of the physical level of the production environment to simulate, the OPUS architecture includes the ability to import the layout DWG and the placement of virtual resources on it. The metric environment makes it possible to drag the necessary objects, and once opportune scale adjustment operations are performed, to customize the objects so that they reproduce the machines actually present in the processing departments. The transport times of materials between resources will be proportional to their distance.

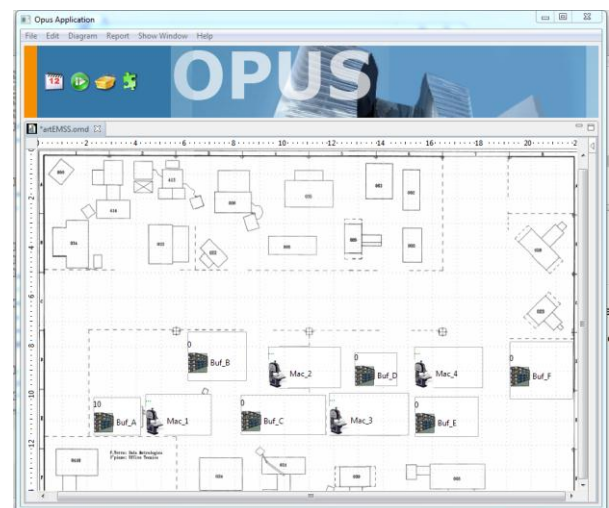


Figure 1: Layout Design

The data transfer will affect the components, the cycles and the Master Production Schedule for the production of the "brake pedal" (the subject of this

example). The Opus architecture makes it possible, as mentioned previously, to connect directly to the ERP system and collect data on:

- Bill of materials. This is done by selecting the finished products, possibly components and raw materials (in case you want to delete components of little interest to the analysis). The transfer allows you to automatically fill in the database of the simulator (Tab 4).

Table 4: "brake pedal" BOM

Lev	Item	Description	UM	Q.ty	WS
1	64786693GR	ASS.PEDALE FRENO DX VER	PZ	1,00	W
2	64786693GR1	ASS.PEDALE FRENO SALDATO	PZ	1,00	W
...3	6478700	PEDALE FRENO GUIDA DX	PZ	1,00	W
...4	64786709GR	PEDALE FRENO SCARICO MAT.	PZ	1,00	W
...5	CUBC80X620	FE 430C F.TO 620X450X8	KG	5,592	P
...3	64787909	FORCELLA	PZ	1,00	W
...4	64787909GR	FORCELLA GREZZA	PZ	1,00	W
...3	64786009	TUBO DISTANZIALE	PZ	1,00	W
...4	CIRC35X25	FE 360 CRUDO TRASF S/SAL	KG	0,15	P
...3	64786109	CIABATTA	PZ	1,00	W
...4	FEEC40X56	NASTRO FE360D 56X4	KG	0,13	P
...3	64786209	PERNO MOLLA	PZ	1,00	W
2	V420N00	VERNICE ARSONSIS	KG	0,025	P
1	64899009	COPRIPEDALE	PZ	1,00	P
1	SH655X800	SEPARATORI IN CARTA 655X800	PZ	0,0625	P
1	SH450X1450	SEPARATORI IN CARTA 450X1450	PZ	0,0250	P
1	BAG60	BUSTA PLURIBALL 270X470X60	PZ	1,00	P
1	64786409	BOCCOLA 34	PZ	2,00	P
1	13709003	SFERA-RB 11,906 TN 2481	PZ	1,00	P
1	GR0990G1	GRASSO AL LITIA JOTA 2/S	GR	0,0005	P

- Processing cycles. The operation takes place in an assisted manner making it possible to select the correspondences between the resources in cycles and those modelled in the previous phase. The figure 2 shows the processing cycle with integrated multi-level BOM of the 64786693GR1 component, as shown by UnisaGest.

Prodotto Finale	Livello I	Cl	Livello II	Cl	Livello III	Cl
64786693GR1	10 SALDATURA 6478700	46 1	10 RITRANCIATURA FORI 64786709GR	150 1	10 TRANCIATURA A BLOCCO CUB80X620X450	105 3,59
	64787909	1	10 FRESATURA 64787909GR	88 1		
	64786009	1	20 BURATTATURA	88		
	64786109	1	10 INTASTARE CIRC35X25	1 0,15		
			20 SMUSSARE	1		
	64786209	1	10 STAMPAGGIO PROGRESSIVO FEEC40X56	1 0,127		
			20 BURATTATURA	1		
	64786209	1				
	20 SALDATURA	130				
	30 PUNTATURA FORCELLA	50				
	40 ARROTONDAMENTO	400				
	50 GRANIGLIATURA	450				
	60 RIQUALIFICA	75				
	70 RIBADITURA A CALDO	300				
	80 CONTROLLO	300				

Figure 2: "brake pedal" process path

- MPS. The transfer operation makes it possible to select the orders whose production you want to simulate in the virtual environment. In this case, the selection applies only to orders for the "brake pedal" part number (Fig 3).

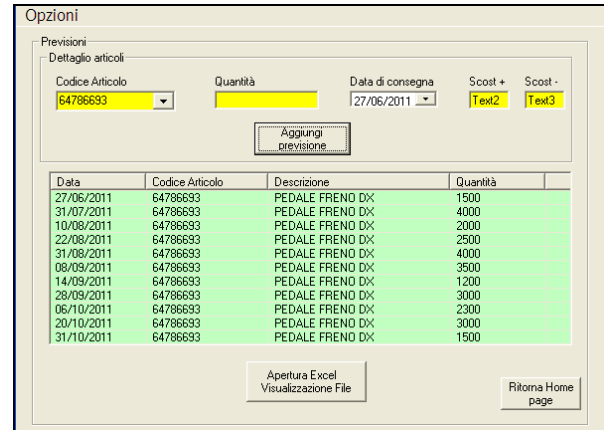


Figure 3: MPS

Lastly, the configuration operations make it possible to change the standard settings of the buffer management policies, initial inventory and resource parameters. In the example buffer management policy was changed in order to simulate a recovery operation, an infinite level of stocks of raw materials was also set and lastly, a distribution of random type processing times was implemented.

The verification of the proper operation of the simulation model was carried out through an analysis at "step by step" mode. In this manner it was possible to display the arrows showing the transfer of information and materials between objects and events generated during simulation in the "Log" side window. This operation made it possible to verify that the model accurately reproduced the modelled work environment.

Instead, the validation phase was conducted by comparing (in the simulation environment) historical data (extracted from the MES) of saturation, obtained by creating the same production used as input for the virtual model. The results of the comparison are shown in the figure 4.

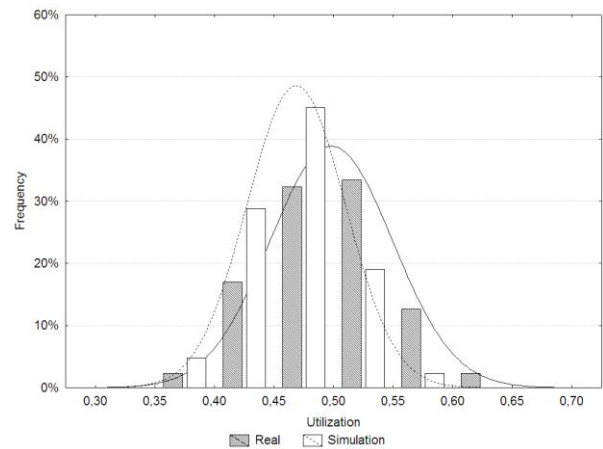


Figure 4: Validation analysis

The average saturation data, linked to all the resources used in the production (measured day by day), showed a mean value lower than the corresponding figure for the real plant. This difference is due mainly to the simultaneous presence of other components on the real plant that during the period of observation used the resources and were not considered in this example. The validation process can be achieved quickly by configuring the connection to the MES and comparing the results automatically.

The designing phases, execution of RUNs and analysis of the results are similar to those of traditional simulation environments. Even if the research group is designing an environment for result analysis integrated with the decision support system, to make the environment more suitable for supporting the daily choices of planners.

5. CONCLUSIONS

The authors proposed an approach that may be looked at as a reference for manufacturing system simulation tool development. The aim was to create a standard for both production system objects (entities, functions, items, data, etc.) and their relationships to one another. An architecture framework and a reference model were described: the latter was presented through the usage of Business Process Modelling Notation, in order to allow a better understanding to Companies, which – differently from software developers - tend to think in term of processes instead of functions and procedures. On top of this, the basic functions of manufacturing systems have been embedded into the properties of the modelled objects, so that any simulation software adopting the proposed reference model could automatically inherit all the typical processes, procedures and operating methods of a manufacturing system: basic production processes archetypes (e.g. set-ups, machine failures, etc.), main production and inventory management policies (e.g. MRP, ROC&ROL policies, etc.), input data format (BOM, MPS, Process Charts, etc.). As a consequence, creating objects that behave according to the proposed design patterns, the work of engineers and developers who need to develop manufacturing systems simulation tools is made easier.

6. REFERENCES

- Bodner, D. A. and L. F. McGinnis (2002). A structured approach to simulation modelling of manufacturing systems. Proceedings of the 2002 Industrial Engineering Research Conference, Georgia.
- Chinosi, M. and Trombetta, A. (2011). BPMN: An introduction to the standard. Computer Standards and Interfaces.
- Cull, R. and Eldabi, T. (2010). A hybrid approach to workflow modelling. Journal of Enterprise Information Management, Vol. 23, No. 3, pp. 268–281.
- Davis, L. and Williams, G. (1994). Evaluating and Selecting Simulation Software Using the Analytic Hierarchy Process. Integrated Manufacturing Systems, Vol. 5, pp. 23 – 32.
- Hlupic, V. A. (1999a). Guidelines for selection of manufacturing simulation software. IIE Transactions, Vol. 31, No. 1, pp. 21-29.
- Hlupic, V.; Irani, Z. and Paul, R. J. (1999b). Evaluation Framework for Simulation Software. International Journal of Advanced Manufacturing Technology, Vol. 15, pp. 366-382.
- Hopp, W. and Spearman, M. (1996). Factory Physics. Boston: McGraw Hill Education.
- Keramati, A.; Golian, H. R.; Afshari-Mofrad, M. (2011). Improving business processes with business process modelling notation and business process execution language: an action research approach. International Journal of Business Information Systems, Vol.7, No. 4, June 2011 , pp. 458-476(19).
- Law, A. a. (1991). Simulation modelling and analysis. Singapore: McGraw-Hill.
- Mujtabi, M. S. (1994). Simulation Modelling of Manufacturing Enterprise with Complex Material. Information and Control Flows. International journal of Computer Integrated Manufacturing, Vol. 7, No. 1, pp. 29-46.
- Narayanan, S.; Bodner, D.A.; Sreekanth, U.; Govindaraj, T.; McGinnis, L.F. and Mitchell, C.M. (1998). Research in object-oriented manufacturing simulations: an assessment of the state of the art. IIE Transactions, Vol. 30, No. 9.
- Robinson S. (2006). Conceptual modelling for simulation: issues and research requirements. Proceedings of the 2006 Winter Simulation Conference, IEEE, pp. 792-800, Piscataway, NJ.
- Rodriguez, A.; Fernandez-Medina, E. and Piattini, M. (2007). A BPMN extension for the modeling of security requirements in business processes. The Institute of Electronics, Information and Communication Engineering TRANS. INF. & SYST., Vol. E90-D, No. 4, pp. 745–752.
- Rogers, P. and Gordon R. J. (1993). Simulation for real time decision making in manufacturing systems. Proceedings of the 25th conference on winter simulation, Los Angeles, California, pp. 866-874, ACM New York, United States.
- Rogers, P. (2002). Simulation of manufacturing operations: optimum-seeking simulation in the design and control of manufacturing systems experience with optquest for arena. Proceedings of the 34th conference on Winter simulation:

exploring new frontiers, pp. 1142-1150, San Diego, California, United States.

Smith, J. (2003). Survey of the use of simulation for manufacturing system design and operation. *Journal of manufacturing systems*, Vol. 22, No. 2, pp. 157-171.

Van Beek D.A.; Hofkamp, A.T.; Reniers, M.A.; Rooda J.E. and Schiffelers R.R.H. (2008). Syntax and Formal Semantics of Chi 2.0. Available from: <http://se.wtb.tue.nl/sereports>, [Accessed: 2010-05-17].

Williams, S. (1967). Business Process Modeling Improves Administrative Control. *Automation*. December, 1967, pp. 44 - 50.

*** (2010) <http://www.omg.org/spec/BPMN/2.0/> - Object Management Group, Business Process Model And Notation (BPMN 2.0), [Accessed on: 2010-06-13]

*** (2008) <http://www.wfmc.org/xpdl.html> - WfMC, XML Process Definition Language (XPDL 2.1), WfMC-TC- 1025, WfMC, [Accessed on: 2011-05]

*** (2007) <http://docs.oasis-open.org/wsbpel/2.0/wsbpel-v2.0.html> - OASIS, Business Process Execution Language (WS-BPEL 2.0), wsbpel-v2.0-OS, OASIS, [Accessed on: 2011-05]

7. ANNEX A

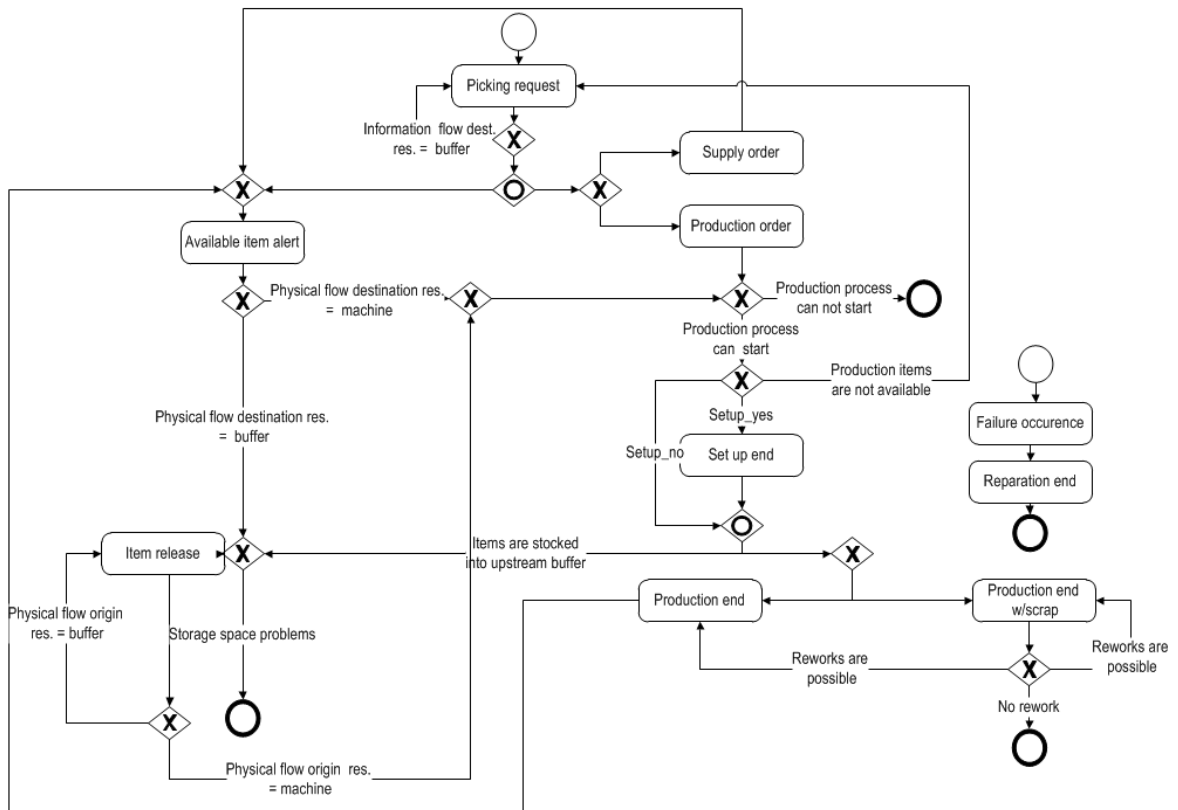


Figure 1: FEL generation process

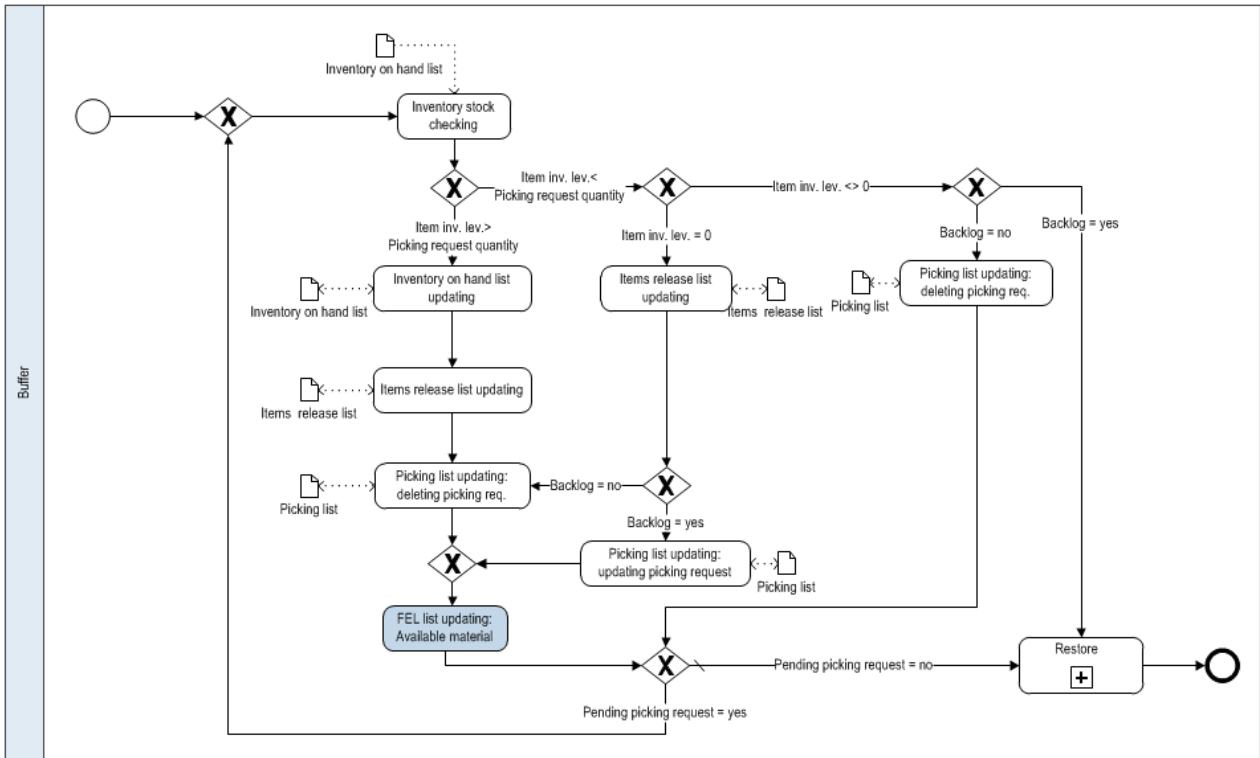


Figure 2: Buffer picking function

ONLINE COLLABORATIVE SIMULATION CONCEPTUAL MODEL DEVELOPMENT

Bhakti S. S. Onggo^(a), Suchismita Hoare^(b)

^(a)Department of Management Science, Lancaster University Management School, Lancaster, United Kingdom

^(b)Digital Modelling Research Group (Faculty of S & T), Anglia Ruskin University, United Kingdom

^(a)s.onggo@lancaster.ac.uk, ^(b)suchismita.hoare@student.anglia.ac.uk

ABSTRACT

Despite significant advancements in Web-based simulation research, its commercial applicability and adoption by users has not widened to the desired extent. We believe that by providing tools to support interaction, collaboration and formation of social networks, we may stimulate more interest in, and the wider adoption of Web-based simulation. This paper presents our work in developing an online collaborative simulation modeling tool that is aimed at supporting simulation conceptual model development, allowing users to form relationships and self-organize into virtual communities around common interests.

Keywords: simulation, web-based simulation, conceptual modelling, collaboration

1. INTRODUCTION

Web-based simulation is a term used to capture a cross-section of simulation methodologies and the World Wide Web (the Web, thereafter), specifically the utilization of the Web for supporting model design, model execution and analysis of generated simulation results. There are a number of potential benefits that can be reaped from harnessing the web infrastructure and technologies. First, it allows the sharing of simulation resources such as computers, software, storage and models. Second, it may improve the accessibility and availability of the model and the generated simulation results. Third, the Web may enable the reuse of existing simulation models by utilizing model repositories on the Web. Fourth, the Web can prove to be advantageous in enabling users on disparate sites to collaborate on model design and analysis of generated simulation results by leveraging the communication capability provided by the Internet. Finally, the Web provides an alternative platform for the use of simulation in education.

On searching papers using 'Web-based simulation' as the keyword from ACM, IEEE, and SAGE (include the SCS publications) digital libraries, we found that there was an explosive growth in the number of publications between 1996 and 2002. However, after the peak in 2002, the number of publications dropped very quickly. Kuljis and Paul (2003) argued that the problem with this stems from the fact that there was a mismatch evident between the main characteristics of

the Web and the approach taken by the domain of Web-based simulation, which failed to take full advantage of the features of the Web including common standards, interoperability, ease of navigation and use, etc. In other words, the focus of many Web-based simulation endeavours was on the re-implementation of existing standalone and distributed simulation software, utilizing Web-related technologies.

In this paper, we emphasize on the importance of supporting collaborative activities in a simulation project. Pidd (2004, chapter 3) suggests that simulation projects require analysts to operate in two parallel domains: technical and organizational. In the technical domain, the team must abstract and simplify the systems of interest so as to develop and use a computer simulation model. In the organisational domain, the team must manage the project properly so as to gain the required insights in an appropriate timescale and within the budget. At each stage, the team members need to interact and collaborate with other team members and other stakeholders in the simulation project. A good interaction and collaboration may lead to good relationship between the people involved in the project and hence a social network could eventually be formed. Robinson and Pidd (1998) found out that good communication and relationship between the people involved in a simulation project is critical. Therefore, a tool to support the interactive and collaborative nature of conducting a simulation project may provide a strong incentive for users to adopt Web-based simulation more readily. In this paper, we present a tool that we have developed that supports collaboration at the early stage of a simulation project, i.e., conceptual model development.

The remainder of this paper is organized as follows. Section 2 provides the literature study of the research in web-based simulation. Section 3 discusses the importance of supporting interaction and collaboration in simulation projects especially at the conceptual model development. Section 4 explains the design and implementation of the tool. Finally, we present our conclusion and highlight some avenues for future work in Section 5.

2. LITERATURE STUDY

One of the objectives of computer networks (including the Internet) is to share resources. In the context of Web-based simulation, these resources include computing power, simulation software, data, models, simulation results and storage. It is fair to say that most, if not all, Web-based simulation endeavours have resource sharing as one of their objectives. Specific examples that emphasize on resource sharing include project PUNCH (Purdue University Network-Computing Hubs) that allowed users from more than fifty universities to share simulation tools via the Web and the Web-based simulation project conducted by Wainer et al. (2008) that supported sharing of computing power, data, models and experiments on a global scale.

The Web could make simulation models and the information they generate more accessible than is the case with traditional simulation platform, owing to the distributed nature of the Web. Guru et al. (2000) developed a Web-based simulation system that utilized a database server for model storage and an application server running the simulation engine. The objective was to allow users to store the simulation model via the Web so that they can be easily located, retrieved, and updated using a web interface, and to execute the models using the same web interface. The simulation results were returned as HTML documents; hence they could be accessed easily using any web-browser. Another example is the construction process simulation project done by Halpin et al. (2003). In this project, a Web-based simulation was developed to provide an easy-to-access environment for studying and analyzing construction processes via the Web. General Motor (GM) Enterprise Systems Laboratory developed a Web-based simulation for Order-to-Delivery evaluation and prediction. The objective was to enable GM staff to conduct the simulation analysis anywhere at any time through the Internet.

The fact that simulation models can be shared and be accessible from the Internet leads to the concept of model reusability. In an ideal case, simulation analysts can search existing simulation models from various model repositories on the Web. Subsequently, these models can be re-used as is or they can be assembled to form a bigger model. The reuse of validated models aims at reducing the cost of simulation model development and the time required for model development while increasing the simulation's accuracy through more robust simulation programs. However, this turns out to be easier said than done.

The first issue is that providing a way to efficiently locate and organize simulation models is challenging. Most web search engines are not designed for this purpose. Even for a specialized search engine, brute-force techniques are not adequate and we need to rely on good heuristic techniques, for formal NP-completeness proof see Page and Oppen (1999). The effectiveness of the specialized search engines can be

improved through standardized model representation and/or the use of ontologies for simulation.

One of the standards for simulation model components is the Base Object Models (BOM) published by the Simulation Interoperability Standards Organization (SISO). BOM defines the syntax and the semantics needed to represent a simulation conceptual model and an interface of a simulation component (SISO 2006). The utilization of the repository of BOM-based components in the component-based simulation model development can be found in Moradi et al (2006).

Ontology is a formal descriptions used to describe and categorize concepts and the relationships among concepts within a particular knowledge domain (Gruber 1995). Web-based simulation can benefit from research in ontology for simulation, since it allows simulation models to be represented as a collection of ontology instances. Effective techniques to search for models based on a set of criteria can be done using specialized query languages. There have been a number of initiatives that explored the ontology-based model representation. Benjamin et al. (2006) outlined the architecture of the Ontology-driven Simulation Modeling Framework (OSMF), in which an ontology-based repository of simulation models was utilized. The University of Georgia's DeMO (Discrete-event Modeling Ontology) was developed to provide a comprehensive ontology for discrete-event simulation that covered the three world views: process interaction, activity scanning and event scheduling (Miller et al. 2004). PIMODES (Process Interaction Modeling Ontology for Discrete Event Simulations) used the same approach as DeMO but focused more on the process interaction world view (Silver et al. 2006).

The second issue is related to the concept of composability. Composability is the capability to select and assemble reusable simulation components in various combinations into simulation systems to meet user requirements (Weisel et al. 2003). Naturally, composability lends itself to a plethora of other issues such as: portability, interoperability (multiple resolution, communication protocols, etc.) and validity. Hence, composability, in itself is an important research topic both within the Web-based simulation research and the distributed simulation research in general. In general, we noticed at least three major approaches to composability: web-service (for example Chandrasekaran et al. 2002), BOM (for example, Gustavson and Chase 2004, Moradi et al. 2007) and ontology (for example, Silver et al. 2007).

The fact that simulation models can be shared might also be advantageous in enabling users on disparate sites to collaborate on model design by leveraging the communication capability provided by the Internet (Kuljis and Paul 2003). However, Henriksen et al. (2002) noticed that there was an apparent paucity of research into the collaboration aspect of web-based simulation. They developed a prototype that combined several existing software tools

(such as: project management, animation and simulation) to support the collaborative aspect of a simulation project. Wang and Liao (2003) developed a Web-based simulation environment that provided facility for group communication and collaborative model design. The collaboration was achieved by providing a facility for team members to view the same model at the same time. Any proposal for changes had to be sent electronically to a coordinator. The coordinator would then edit the model so that the amended model could be viewed by the team members immediately. Araújo-Filho et al. (2004) developed a synchronous groupware environment that supported the conceptual modelling, computer implementation and experimentation processes of a simulation project. Unlike Wang and Liao (2003), the role of 'coordinator' was replaced by a software locking mechanism where at any point in time, only one member of the team could modify the model.

3. GROUP MODEL DEVELOPMENT

The area of Web-based simulation has witnessed some success in novel methods for executing models through both client and server-side applications, besides addressing the issue of model reuse. However, even with these advancements and the critical mass of research knowledge available, the applicability of Web-based simulation has not widened to the desired extent, and the potential benefits to be derived have remained unrealized from a commercial perspective (Miller et al 2001). On the basis of information gleaned from the literature, there seems to a disproportionate level of interest in the technology of Web-based simulation in the academic community and potential users who are likely to influence its commercial use; with the area of Web-based simulation remaining more of a scholarly endeavour, and thus principally academic in nature.

Fishwick (2002) noted that Web-based simulation can be an effective problem solving technique and decision support tool that has the potential of stimulating a paradigm shift in simulation; with the shift being one from a single simulation analyst running experiments and analyzing results on his computer to one of global proportions involving multiple interacting simulation analysts.

Simulation projects involve interactions and collaborations among different stakeholders (clients at different management levels, domain experts, simulation analysts, statisticians, etc.). Different stages in a simulation project involve different types of stakeholders and require different degrees of interactions and collaboration. For example, the conceptual model development stage may require heavy interactions between clients, domain experts and simulation analysts, while the computer implementation stage may require regular interactions between programmers and simulation analysts. Therefore, we believe that by providing facilities that support the interaction and collaboration in simulation modelling

process, we may stimulate more interest in, and the wider adoption of Web-based simulation.

As far as we know, the formation of social networks through Web-based simulation has not been investigated. Engeström (cited in Breslin and Decker 2007) has argued that social networking sites' longevity is proportional to the degree to which people are connecting via items of interest related to their jobs, workplaces, hobbies, and so on. We believe that Web-based simulation should provide a facility that supports the formation of social network communities. We make a conjecture that this may further attract people to adopt Web-based simulation. People tend to form relationships and self-organize into communities around common interests. Therefore, people may form communities related to the simulation software they use, the type of domain they are involved in (such as: public sector, healthcare and manufacturing), the type of modelling paradigm they frequently use (such as: discrete event, system dynamics and agent-based), and so on. To a certain degree, different simulation communities exist in the physical world (such as: discrete event, system dynamics, agent-based simulation and simulation application in healthcare). Hence, it is plausible that the virtual social networks through the use of Web-based simulation may work. Further, the ontology-based Web based simulation may provide a facility for them not only for sharing their conceptual knowledge but also their simulation models (as long as it is legal to do so). Web-based simulation that supports the formation of social networks may also draw the attention of the new type of users who are referred to as the 'natural-born Webbers' in Kuljis and Paul (2003).

4. CSM WEB: A PROTOTYPE

To prove our concept, we have developed a Web-based simulation tool called Collaborative Simulation Modelling Web Application (CSM Web) that supports the interaction, collaboration and the formation of social networks among simulation enthusiasts either for specific simulation projects (professional), for leisure (perhaps in the form of 'Grab-and-Glue, run, reject, retry' approach as described in Kuljis and Paul (2003)) or for altruistic reasons (such as correcting errors in models posted by other people). At this stage, we focus on the conceptual model development because it is arguably the stage that requires high degree of interactions and collaborations (the same sentiment is also shared by Araújo-Filho et al. (2004)). Onggo (2009) used a number of diagrams to communicate different components in simulation conceptual models. One of the diagrams that can be used in simulation conceptual modeling is the Event Relationship Diagram (ERD) (Schruben 2008). ERD is independent of any software implementation and it can be converted into a target simulator automatically.

There are three types of user: model owner, collaborator, and visitor. A user can use all functionalities on her models (she is the owner of the

models). A user cannot delete another user's models, even if she is a collaborator for the models (but as a collaborator, she can modify the models). A visitor has the least privilege. Figure 1 shows the Use Case diagram for the users and Figure 2 shows the Use Case diagram for invitation processing.

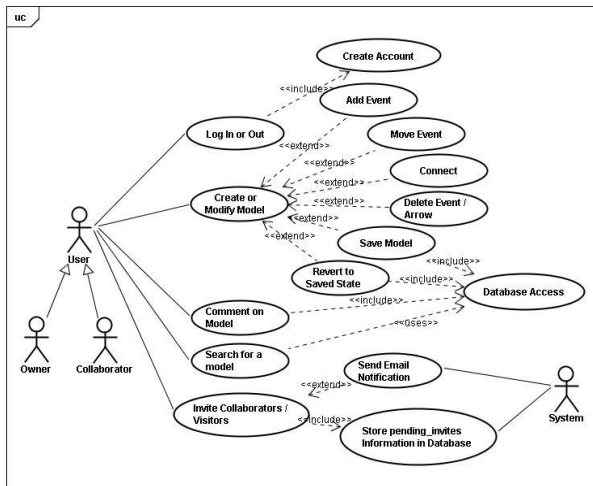


Figure 1: Use Case Diagram – Owner and Collaborator

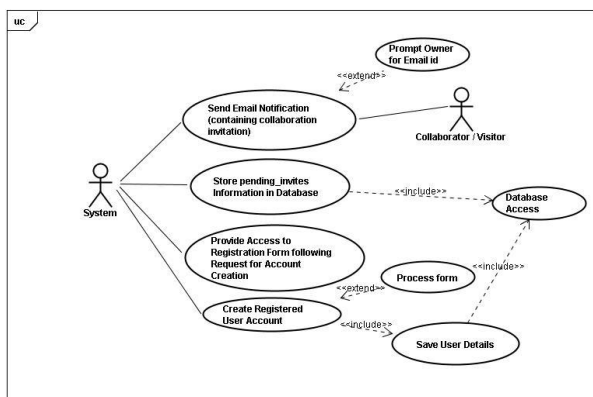


Figure 2: Use Case Diagram – Collaboration System

The CSM Web is implemented using in-browser VML/SVG for creation of models (drawing of events and arrows on canvas), and AJAX for the client-server communication. VML (Vector Markup Language), a standard proposed to the W3C by Microsoft (and others, including Visio Corporation, Macromedia, Inc., etc.) for Web vector graphics, is an application of XML 1.0 for supporting the markup of vector graphic information that is supported by Internet Explorer. SVG (Scalable Vector Graphics), on the other hand, is a W3C Recommendation open standard for describing two-dimensional graphics in XML, and supporting vector graphics in other standards-compliant browsers like Firefox, Opera, etc. We use a small JavaScript library called Raphaël which uses VML and the SVG W3C recommendation as a base for creating graphics, and provides an adapter to make drawing vector graphics for the Web cross-browser compatible and easy. As for the back end, the application utilizes PHP with a MySQL database to store model states and for the other features.

PHP (Hypertext Preprocessor) is a commonly used server-side scripting language especially suited for creating dynamic Web applications (through access of data from databases and execution of back-end services in an intuitive manner). It is widely used with database management system such as MySQL and PostgreSQL. The combination of PHP and AJAX, is widely used to provide a powerful platform for the creation of Web-based applications. The architecture of the software is shown in Figure 3.

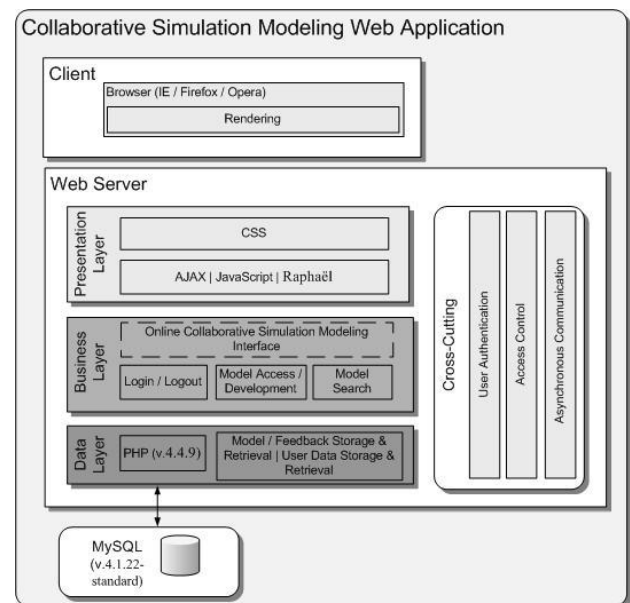


Figure 3: CSM Web – Architecture [Based on a template from the Application Architecture Guide 2.0 by Meier et al. (2009)]

The architecture is structured in three layers. The Data Layer creates a central point for data access (from the MySQL database) through storage, update and retrieval of necessary data. The Business Layer implements the business logic of the prototype, such as login and logout, search etc. While the Presentation Layer performs the task of interacting with the user, such as accepting username and password (for granting access to the system), search criteria (for performing a search on the relevant database), etc. At the client side, the Client Layer (which is the simplest one), composed of a Web browser in which the application is rendered, allows users to view the results of their requests for the available features, and that of the interactions with the features.

The CSM home page (Figure 4) allows a user to log-in or to search simulation models (without logging-in). Once a user has logged-in, the customized, main user interface is shown (Figure 5). The user interface comprises two parts: model creation frame and feedback frame. The model creation frame is used to create, modify, share, and delete models (Figure 6). The feedback frame is used to write comments on a model or to read comments from collaborators/visitors (Figure 7). Finally, Figure 8 shows how the same model is seen by the owner and a visitor. The new added event is

shown in red in the owner screen. The new addition will be shown to the visitor once the screen is refreshed. CSM Web prevents entire page refreshes; thus ensuring that small, targeted traffic is passed between the client and server. In addition, the prototype breaks the traditional page update model by performing immediate, asynchronous part-page updates (by restricting updates to specific components) in response to changes made to a model, logging in of a user, or arrival of a new comment on a model; leading to significant reduction in wait time for screen updates.

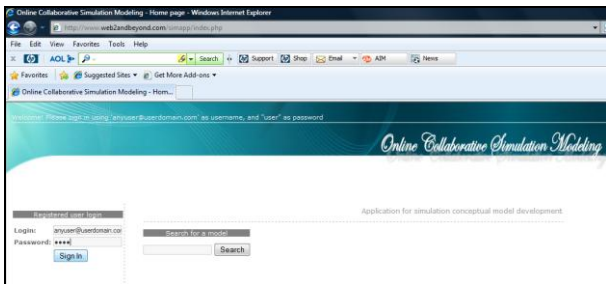


Figure 4: CSM Web – Home Page

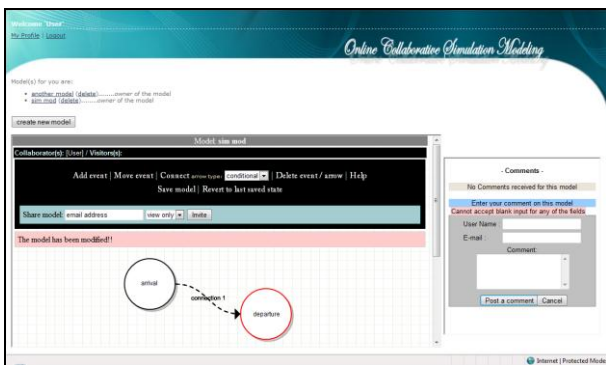


Figure 5: CSM Web – User Interface

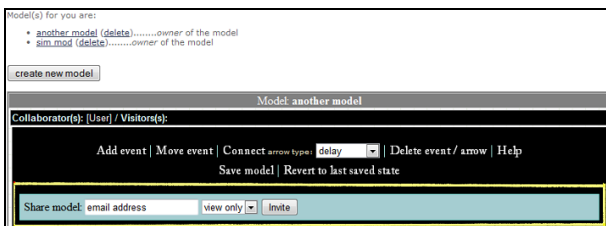


Figure 6: Model Creation Frame

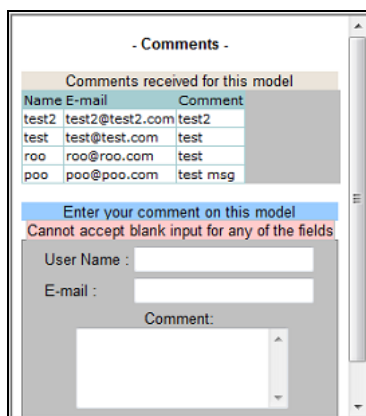


Figure 7: Feedback Frame

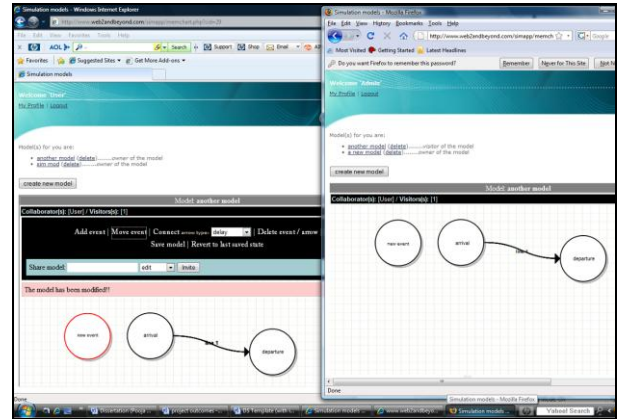


Figure 8: Collaboration View

5. CONCLUSIONS AND FUTURE WORK

We have shown from the literature that more efforts are needed for Web-based simulation to exploit the interactive and collaborative nature of simulation modelling process. We believe that by providing tools to support interaction, collaboration and formation of social networks, we may stimulate more interest in, and the wider adoption of Web-based simulation. We have shown through CSM Web that it is possible to build such Web-based simulation tool using the current Web technologies, such as: AJAX, Raphaël (JavaScript library), PHP and MySQL.

The CSM Web is able to demonstrate basic functionalities needed for our study. However, in order to conduct proper social experiments, we need to improve the functionalities and the user interface. This includes adding the functionality to run the simulation model and adding more social components such as friend list. We need to investigate whether open platforms such as: Drupal, Moodle, Wikis, etc. can be used as an effective collaboration platform for simulation modeling before investing valuable efforts in taking our prototype into a production version.

ACKNOWLEDGMENTS

CSM Web is developed by Suchismita Hoare as part of her Master of Science degree with the University of Liverpool and Laureate Online Education.

REFERENCES

- Araújo-Filho, W., Hirata, C.M. and Yano, E.T. 2004. GroupSim: A Collaborative Environment for Discrete Event Simulation Software Development for the World Wide Web. *Simulation: The Society for Modeling and Simulation International*, 80 (6), 257–272.
- Benjamin, P., Patki, M. and Mayer, R. 2006. Using ontologies for simulation modelling. *Proceedings of the 38th Conference on Winter Simulation*, pp. 1151–1159. December 03–06, Monterey (California, USA).
- BOM Product Development Group. 2006. *Base Object Model (BOM) Template Specification - SISO-STD-003-2006*. Available from:

- <http://www.sisostds.org/> [accessed: 17 March 2009].
- Breslin, J. and Decker, S. 2007. The future of social networks on the Internet: The need for semantics. *IEEE Internet Computing*, 11 (6), 86–90.
- Chandrasekaran, S., Silver, G., Miller, J.A., Cardoso, J. and Sheth, A.P. 2002. XML-based modeling and simulation: Web services technologies and their synergy with simulation. *Proceedings of the 34th Conference on Winter Simulation*, pp. 606–615. December 08–11, San Diego (California, USA).
- Fishwick, P.A. 2002. XML-based modeling and simulation: using XML for simulation modeling. *Proceedings of the 34th Conference on Winter Simulation*, pp. 616–622. December 08–11, San Diego (California, USA).
- Guru, A., Savory, P. and Williams, R. 2000. A Web-based interface for storing and executing simulation models. *Proceedings of the 32nd Conference on Winter Simulation*, pp. 1810–1814. December 10–13, Orlando (Florida, USA).
- Gustavson, P. and Chase, T. 2004. Using XML and BOMs to rapidly compose simulations and simulation environments. *Proceedings of the 36th Conference on Winter Simulation*, pp. 1467–1475. December 05–08, Washington (DC, USA).
- Gruber, G. R. 1995. Toward principles for the design of ontologies used in knowledge sharing. *International Journal of Human Computer Studies*, 43 (5-6), 907–928 .
- Halpin, D.W., Jen, H-y. and Kim, J-w. 2003. A construction process simulation web service. *Proceedings of the 35th Conference on Winter Simulation*, pp. 1503–1509. December 07–10, New Orleans (Louisiana, USA).
- Henriksen, J.O., Hanisch, A., Osterburg, S., Lorenz, P. and Schriber, T.J. 2002. Web based simulation center: Professional support for simulation projects. *Proceedings of the 34th Conference on Winter Simulation*, pp. 807–815. December 08–11, San Diego (California, USA).
- Kuljis, J. and Paul, R. J. 2003. Web-based discrete event simulation models: Current states and possible futures. *Simulation & Gaming*, 34 (1), 39–53.
- Meier, J.D., Homer, A., Hill, D., Taylor, J., Bansode, P., Wall, L., Boucher Jr, R. & Bogawat, A. (2009) patterns & practices: App Arch Guide 2.0 Knowledge Base, Microsoft® pat-terns & practices [Online image]. Available from: <http://www.codeplex.com/AppArch/Wiki/View.aspx?title=Visio%20Index&referringTitle=Home> (Accessed: 21 June 2009).
- Miller, J.A., Fishwick, P.A., Taylor, S.J.E., Benjamin, P. and Szymanski, B. 2001. Research and commercial opportunities in Web-based simulation. *Simulation Practice and Theory*, 9 (1–2), 55–72 .
- Miller, J.A., Baramidze, G.T., Sheth, A.P. and Fishwick, P.A. 2004. Investigating Ontologies for Simulation Modeling. *Proceedings of the 37th Annual Symposium on Simulation*, pp. 55. April 18–22, Arlington (Virginia, USA).
- Moradi, F., Nordvaller, P. and Ayani, R. 2006. Simulation Model Composition using BOMs. *Proceedings of the 10th International Symposium on Distributed Simulation and Real-Time Applications*, pp. 242–252. October 02–04, Malaga (Spain).
- Moradi, F., Ayani, R., Mokarizadeh, S., Shahmirzadi, G.H.A. and Tan, G. 2007. A Rule-based Approach to Syntactic and Semantic Composition of BOMs. *Proceedings of the 11th International Symposium on Distributed Simulation and Real-Time Applications*, pp. 145–155. October 22–24, Chania (Crete Island, Greece).
- Onggo, B.S.S. 2009. Towards a unified conceptual model representation: A case study in healthcare. *Journal of Simulation*, 3 (1), 40–49 .
- Page, E.H. and Opper, J.M. 1999. Observations on the complexity of composable simulation. *Proceedings of the 31st conference on Winter simulation*, pp. 553–560. December 05–08, Phoenix (Arizona, USA).
- Pidd, M. 2004. *Computer simulation in management science*, 5th edition. Chichester, UK: Wiley.
- Robinson, S. and Pidd, M. 1998. Provider and customer expectations of successful simulation projects. *Journal of the Operational Research Society*, 49 (3), 200–209 .
- Schruben, L. 2008. Analytical simulation modeling. *Proceedings of the 40th Conference on Winter simulation*, pp. 113–121. December 07–10, Miami (Florida, USA).
- Silver, G.A., Lacy, L.W. and Miller, J.A. 2006. Ontology based representations of simulation models following the process interaction world view. *Proceedings of the 38th Conference on Winter Simulation*, pp. 1168–1176. December 03–06, Monterey (California, USA).
- Silver, G.A., Hassan, O.H. and Miller, J.A. 2007. From domain ontologies to modeling ontologies to executable simulation models. *Proceedings of the 39th Conference on Winter Simulation*, pp. 1108–1117. December 09–12, Washington (DC, USA).
- Wang, Y-h. and Liao Y-c. 2003. Implementation of a Collaborative Web-based Simulation Modeling Environment. *Proceedings of the 7th International Symposium on Distributed Simulation and Real-Time Applications*, pp. 150–157. October 23–25, Delft (The Netherlands).
- Wainer, G., Liu, Q., Chazal, J., Quinet, L. and Traoré, M.K. 2008. Performance analysis of Web-based distributed simulation in DCD++: A case Study across the Atlantic ocean. *Proceedings of the 2008 Spring Simulation Multiconference*, pp. 413–420. April 14–17, Ottawa (Canada).
- Weisel, E.W., Petty, M.D. and Mielke, R.R. 2003. Validity of models and classes of models in semantic composability. *Proceedings of the 2003*

Fall Simulation Interoperability Workshop.
September 14–19, Orlando (Florida, USA).

AUTHORS BIOGRAPHY

Bhakti Satyabudhi Stephan Onggo is a lecturer in Business Process Modelling and Simulation at the Department of Management Science at the Lancaster University Management School. He completed his PhD in Computer Science from the National University of Singapore and his MSc in Management Science from the Lancaster University. His research interests are in the areas of simulation methodology (modelling paradigms and conceptual modelling), simulation technology (parallel and distributed simulation) and business process modeling and simulation applications.

Suchismita Hoare is a PhD student at Anglia Ruskin University, United Kingdom. She completed an MSc in IT (specialization in Software Engineering) from the University of Liverpool (Laureate Online Education), United Kingdom. Her research interests include web-based simulation and cloud computing.

dSPACE based direct-driven permanent magnet synchronous wind power system modeling and simulation

Yan-xia Shen^(a), Xiang-xia Liu^(b), Zhi-cheng Ji^(c), Ting-long Pan^(d)

^(a)Institute of Electrical Automation, Jiangnan University, Wuxi,China 214122

^(b)Institute of Electrical Automation, Jiangnan University, Wuxi,China 214122

^(c)Institute of Electrical Automation, Jiangnan University, Wuxi,China 214122

^(d)Institute of Electrical Automation, Jiangnan University, Wuxi,China 214122

^(a)shenyx@jiangnan.edu.cn, ^(b)xiangxia321@126.com, ^(c)zcji@jiangnan.edu.cn, ^(d)tlpan@jiangnan.edu.cn

ABSTRACT

When the wind speed is below rated value, the efficiency of captured wind energy must be maximized and the mechanical oscillation is guaranteed to be small. This essay deals with these problems, not only describes the advantages of the direct-driven permanent magnet synchronous wind power system, but also introduces two frequency loop model based on frequency separation principle, and suggests LPV model and LPV control method for high-frequency part of the system, the output of the high-frequency part is used to compensate the mechanical torque. The mathematical model is built with MATLAB, and the online test is carried out by dSPACE. The simulation results show that the controller reduces mechanical oscillation effectively, and enhances the system reliability.

Keywords: LPV, direct-driven permanent magnet synchronous wind turbine, Wind Power Conversion System, dSPACE

1. INTRODUCTION

Wind power is a clean renewable resource. After 20 years development, the cost of wind power generation is greatly reduced, and the system performance is gradually improved. Wind power has become an important strategy for national economic development (Akpinar, E.K and Akpinar, S 2006; Yazhou Lei and Gordon Lightbody 2005).

The wind power system mostly adopts DFIG at present, and it has gearbox between generator and wind turbine, gearbox not only consumes a part of energy, but also has big noises, high failure rate and high maintenance costs. Permanent magnet synchronous generator is a newer type motor without brush and commutator, so it has big power factor. Direct-driven permanent magnet synchronous wind turbine has no gearbox, namely that the wind turbine and generator is connected directly, so it can promote the efficiency, improve the reliability of the system and decrease failure rate and maintenance costs.

In the wind power system, the control objective under rated wind speed is to maximize the efficiency of captured wind energy, but the traditional control methods usually cause high mechanical oscillation. Since the wind energy conversion system (WECS) is a typical strong non-linear system, many control means need to transform non-linear model to linear model, such as PI and PID control(Tapia A, Tapia G and Ostolaza Jx 2003), and some advanced control ways, like LQ and LQG control(Muhando Be, Senjyu T and Urasaki N 2007), however, the robustness of all these control methods is low, this becomes an applying restrictions for these control methods. The randomness of wind often results in low resolution linear model, document(Shamma,J. and Athans,H. 1991) proposed LPV(Linear Parameter Varying) control which can solve above problems efficiently.

Firstly, this paper introduces double frequency loop model of the system(Inlian Munteanu, Nicolaos Antonio Cutululis and Antoneta Iuliana Bratcu 2005;2008), and then the low-frequency part uses PI control, and build LPV model for the high-frequency part, based on this model, adopts LPV control method, it can promote the accuracy of the model, also, can maximize the efficiency of captured wind energy and restrain oscillation. This paper built simulation model based on MATLAB, then loaded onto dSPACE to make on-line experiment, the experiment result shows LPV control can improve the performance of the system, and it proved the feasibility and superiority of the control method.

2. WIND ENERGY CONVERSION SYSTEM MODEL

The dynamic process of generator electromagnetic response is ignored in this paper, since the electromagnetic time constant is much smaller than the mechanical time constant. The system model established in this paper is ideal. The structure of variable speed constant frequency wind power energy conversion system is shown in Figure1.

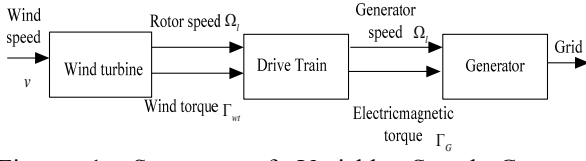


Figure 1: Structure of Variable Speed Constant Frequency WECS

In Figure 1, the wind turbine and generator are connected directly, Drive Train represents the shaft, and there is no gearbox between them. For this reason, they have the same speed Ω_g .

2.1. Wind Speed Model

Wind speed $v(t)$ which is a non-statistical random process is decomposed into two components in many references (Bianchi F, De Battista H and Mantz RJ 2006; Nichita C, Luca D, Dakyo B and Ceang E 2002), that is

$$v(t) = \bar{v}(t) + \Delta v(t) \quad (1)$$

Where $\bar{v}(t)$ is low-frequency component, which describes long-time scale and low-frequency changes, it is usual to assume to be a Weibull distribution; and $\Delta v(t)$ is high-frequency component, and the high-frequency part $\Delta v(t)$ is made up of Gaussian white noise $e(t)$ as a disturbance signal composed of a first-order filter.

$$\Delta \dot{v}(t) = -\frac{1}{T_w} \Delta v(t) + \frac{1}{T_w} e(t) \quad (2)$$

Where T_w is the time constant of the filter, and $T_w = L_t / \bar{v}$. L_t is the length of wind speed turbulence.

2.2. Wind Turbine Model

According to the Bates theory, the mechanical power captured by turbine is:

$$P_{wr} = 0.5\pi\rho R^2 C_p(\lambda) v^3 \quad (3)$$

Where ρ is air density, R is the radius of wind turbine, v is wind speed, λ is tip speed ratio, and $\lambda = R \cdot \Omega_r / v$, Ω_r is the angular velocity of wind turbine rotor, $C_p(\lambda)$ is the power factor of wind turbine, which is defined as:

$$C_p(\lambda, \beta) = 0.22 \left(\frac{116}{\lambda} - 0.4\beta - 5 \right) e^{-\frac{12.5}{\lambda}}$$

Where β is the pitch angle of variable-pitch control, for the fixed-pitch control, $\beta = 0$.

The torque of wind turbine is:

$$\Gamma_{wr} = \frac{P_{wr}}{\Omega_r} = 0.5\pi\rho R^3 v^2 C_T(\lambda) \quad (4)$$

Where $C_T(\lambda)$ is torque coefficient which is defined as $C_T(\lambda) = C_p(\lambda) / \lambda$.

2.3. Drive Train Model

Neglecting the transients, the rigid drive train is expressed as:

$$J_t \dot{\Omega}_t = \Gamma_{wr} - \frac{i}{\eta} \Gamma_G \quad (5)$$

Where J_t is the total inertia of drive train, i is the gear box ratio, for the direct-driven case, its value is 1, η is the efficiency of the transmission shaft, Γ_G is the electromagnetic torque of the generator. As mentioned above, the formula (4) and (5) constitute the basic low-frequency model of the wind power conversion system.

2.4. Model of PMSG

To suppose the permanent magnet synchronous motor is ideal, when we analyze its basic electromagnetic relations. So it fulfills:

1. Ignore the influence of the core magnetic saturation, excluding the eddy current and hysteresis loss.
2. The conductivity of permanent magnetic materials is zero.
3. There is no damper winding in rotor.
4. The three-phase of stator is symmetrical and the induced EMF (electromotive force) is sinusoidal

The electromagnetic torque of permanent magnet generator in d,q coordinate system is

$$\Gamma_G = p(\Phi_d i_q - \Phi_q i_d) = p[\Phi_m i_q + (L_d - L_q) i_d i_q] \quad (6)$$

It is assumed that the load of the generator R_l is independent and symmetric three-segment, and the states and input of the system are defined as:

$$x = [x_1(t) \quad x_2(t)]^T \equiv [i_d(t) \quad i_q(t)]^T$$

$$u \equiv R_l$$

The state model of the generator can be expressed:

$$\begin{cases} \dot{x} = \begin{bmatrix} \frac{1}{L_d + L_s}(-Rx_1 + p(L_q - L_s)x_2\Omega_l) \\ -\frac{1}{L_q + L_s}(-Rx_2 - p(L_d + L_s)x_1\Omega_h + p\Phi_m\Omega_h) \end{bmatrix} + \\ \begin{bmatrix} -\frac{1}{L_d + L_s} & 0 \\ 0 & -\frac{1}{L_q + L_s} \end{bmatrix} \\ y \equiv \Gamma_G = p\Phi_m x_2 \end{cases} \quad (7)$$

Where R is the stator resistance, L_d and L_q are the inductance of the stator in d,q coordinate system, i_d and i_q are the stator current, L_s is the equivalent inductance of grid and converter, Φ_m is the flux, and is a constant for the permanent material, p is the pole pairs of the generator.

3. LPV CONTROLLER DESIGN

The LPV theory was firstly proposed by Professor Shamma, its dynamic characteristics depend on the adjustable parameters which are measured in real time(Shamma,J. and Athans,H. 1991). Since these parameters can reflect the nonlinearity of the system, LPV system can be applied to describe non-linear system. Design gain scheduling controller using linearization method to make controller gain change with the parameters.

Refer to the references (Inlian Munteanu, Nicolaos Antonio Cutululis and Antoneta Iuliana Bratcu 2005; 2008), the LPV model of the system can be expressed:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) + Le(t) \\ y(t) = Cx(t) \end{cases} \quad (8)$$

The formula (8) shows that external interference $e(t)$ exists in the LPV model. In order to effectively suppress wind disturbance and improve the system dynamic performance, the controller is designed based on the LPV dynamic model, making the H_∞ norm of the closed-loop transfer function $T_{ez}(s)$ from the disturbance input $e(t)$ to the control output $y(s)$ less than a given performance index, that is

$$\|T_{ez\infty}(s)\|_\infty < \gamma_\infty$$

then the designed state feedback controller is

$$u(t) = K(\rho(t)) \cdot x(t)$$

the closed-loop system is obtained:

$$\begin{cases} \dot{x}(t) = (A(r(t)) + B(r(t))K(r(t)))x(t) + Le(t) \\ y(t) = C(\rho(t))x(t) \end{cases} \quad (9)$$

The affined parameters of this LPV model depend on $\rho(t)$, and the controller matrix is solved according to theorem 1.

Theorem 1(Junling Wang 2008)

For the LPV model described by formula(8) and a given positive constant, if there are continuously differentiable symmetric positive definite matrix $X(\rho(t))$, symmetric positive definite matrix Y , matrix V and $R(\rho(t))$ satisfying formula(10) for all the parameters, then the parameters of closed-loop (9) are quadratic stability and meet the given H_∞ performance.

$$\begin{bmatrix} -(V+V^T) & * & * & * & * & * \\ M & -X(\rho(t))+Y & * & * & * & * \\ 0 & 0 & -Y & * & * & * \\ L^T(\rho(t)) & 0 & 0 & -\gamma_\infty I & * & * \\ 0 & C(\rho(t))V & 0 & 0 & -\gamma_\infty I & * \\ V^T & 0 & 0 & 0 & 0 & -X(\rho(t)) \end{bmatrix} < 0 \quad (10)$$

where

$M = V^T A^T(\rho(t)) + R^T(\rho(t))B^T(\rho(t)) + X(\rho(t))$. If the inequality (10) has feasible solution, then the state feedback controller gain matrix which dependent on the parameters is

$$K(\rho(t)) = R(\rho(t))V^{-1}$$

according to the formula (5),

$$J_T \overline{\Delta\Omega_l} = \overline{\Delta\Gamma_{wt}} - \frac{1}{\eta} \frac{\overline{\Gamma_G}}{\overline{\Gamma_{wt}}} \overline{\Delta\Gamma_G} \quad (11)$$

Where $J_T = J_l \overline{\Omega_l} / \overline{\Gamma_{wt}}$, $\overline{\Delta\Omega_l} = \frac{\Delta\dot{\Omega}_l}{\Omega_l}$, $\Delta\dot{\Omega}_l = \dot{\Omega}_l - \overline{\dot{\Omega}_l}$,

$\overline{\Omega_l}$ is the stable value of the Ω_l , also for the $\overline{\dot{\Omega}_l}$, $\overline{\Delta\Omega_l}$ and $\overline{\Delta\Gamma_{wt}}$.

According to the high-frequency pulsation wind speed modeling method in the reference [9] and formula (2), there is

$$\overline{\Delta v} = \frac{1}{T_w} (e - \overline{\Delta v}) \quad (12)$$

and according to formula (4) and low-frequency sub-model,

$$\overline{\Delta\Gamma_{wr}} = \gamma \cdot \overline{\Delta\Omega_i} + (2-\gamma) \overline{\Delta v} \quad (13)$$

γ depends on the low-frequency operating point of the system, its value is $\gamma = \frac{\bar{\lambda} C'_p(\bar{\lambda})}{C_p(\bar{\lambda})} - 1$, and

$$C'_p(\bar{\lambda}) = \frac{dC_p(\bar{\lambda})}{d\bar{\lambda}}$$

Put the formula (11) and (12) into (13), there is

$$\begin{aligned} \overline{\Delta\Gamma_{wr}}(t) &= \left(\frac{\gamma}{J_T} - \frac{1}{T_w} \right) \overline{\Delta\Gamma_{wr}}(t) + \frac{\gamma}{T_w} \overline{\Delta\Omega_i}(t) \\ &- \frac{\gamma}{J_T \eta} \frac{\overline{\Gamma_G}}{\overline{\Gamma_{wr}}} \overline{\Delta\Gamma_G}(t) + \frac{2-\gamma}{T_w} e(t) \end{aligned} \quad (14)$$

Formula (11) and (14) constitute the high-frequency sub-model of the conversion system, for this model and according to the theory above, $u(t) = \overline{\Delta\Gamma_G}$ is chose as the control input, $x(t) = [\overline{\Delta\Omega_i} \quad \overline{\Delta\Gamma_{wr}}]^T$ is the state vector, $y(t) = \overline{\Delta\lambda}(t) = C(\rho(t))x(t)$ is defined to be the output vector, and the matrixes obtained by formula (14) are

$$\begin{aligned} A &= \begin{bmatrix} 0 & 1/J_T \\ \gamma/T_w & \gamma/J_T - 1/T_w \end{bmatrix}, & B &= \begin{bmatrix} -1 \overline{\Gamma_G} \\ J_T \overline{\Gamma_{wr}} \\ -\gamma \overline{\Gamma_G} \\ J_T \overline{\Gamma_{wr}} \end{bmatrix}, \\ L &= \begin{bmatrix} 0 & (2-\gamma)/T_w \end{bmatrix}^T, & C &= \begin{bmatrix} 2 & 1 \\ 2-\gamma & -2-\gamma \end{bmatrix} \end{aligned}$$

Use the LMI toolbox to describe the matrix of the theorem 1 to obtain the controller K.

4. ONLINE SIMULATION AND RESULTS ANALYSIS

According to the Figure 1 and the analysis above, the general structure diagram of the direct-driven permanent magnet synchronous wind power system based on LPV is shown in Figure2, and the online experiment is done by dSPACE. dSPACE system is a development and testing working platform based on MATLAB/Simulink in real-time environment, it can be connected with MATLAB/Simulink seamlessly, and can realize real-time control and modify for the control system, so it is much easier and overcomes some inconvenience of out-line simulation.

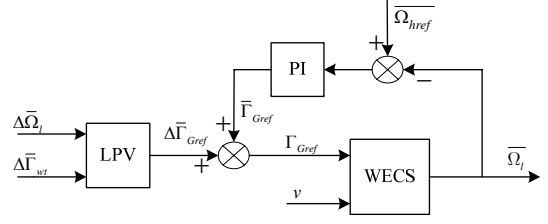


Figure 2: Gain Scheduling Control Structure Based on LPV

In MATLAB simulation environment, find the solution of matrix in theorem 1 with LMI and Simulink toolbox, and build the general simulation diagram of the system, then download the simulation to the dSPACE to do the real-time simulation experiment. The parameters of the experiment are shown in Table 1.

Table 1: Experiment Parameters

Name	Value	Name	Value
R	2.5 m	ρ	1.25kg/m ³
T_w	21.4286s	η	0.95
J_T	0.5632 kg*m ²	C_{pmax}	0.476
Γ_{Gmax}	40Nm	λ_{opt}	7

The experiment results of power factor of the wind turbine C_p and tip speed ratio λ (lam) are shown in Figure3 and Figure4.

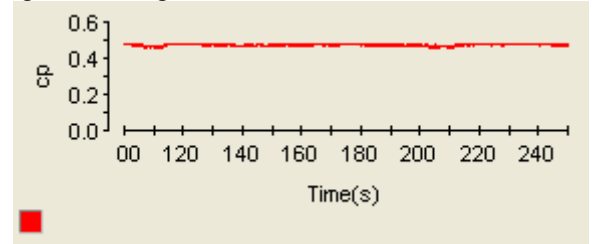


Figure 3: Power Factor

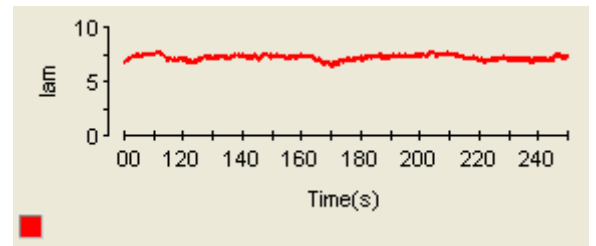


Figure 4: Tip Ratio

From Figure 3, it is easy to see that the value of C_p is stable and very closed to the maximum value 0.476; and the Figure 4 shows that the accuracy of tip speed ratio tracking the best value is high, and the robustness of the system is good, so the controller can capture the wind energy as much as possible, and the oscillation is small, solve the contradictions between the wind energy capturing and big oscillation. The experiment results show the effectiveness and advantages of this control method.

5. CONCLUSION

The situation discussed in this paper is below the rated wind speed of the direct-driven permanent magnetic synchronous wind power system, we need to capture the maximum wind energy and to ensure the smaller mechanical vibration, while increasing the model precision, firstly we build the basic model of the system, then for the high-frequency part, according to the LPV theory to linear the part and design the LPV controller, with PI control of the low-frequency part, the torque can be controlled well. The experiment results based on the dSPACE show that the control method of this paper can fulfill the control objective and efficiently improve the performance. The research of this paper has broad application prospects in direct-driven permanent magnetic synchronous wind power system.

ACKNOWLEDGMENTS

This work is supported by Program for New Century Excellent Talents in University NCET-10-0437 and Special Foundation of China Postdoctoral Science (201003553).

REFERENCES

- Akpinar, E.K, Akpinar, S., 2006. An investigation for wind power potential required in installation of wind energy conversion system. *Proceedings of the Institution of Mechanical Engineers*, 220(1): 1-13.
- Junling Wang, 2008. *Delay Linear parameter varying system stability analysis and gain scheduling control*. Beijing: Science Press.
- Shamma,J., Athans,H., 1991. Guaranteed properties of gain scheduled control for linear parameter-varying plants. *Automatica*, 27(3), 559-564.
- Muhando Be, Senjyu T, Urasaki N, et al, 2007. Gain scheduling control of variable speed WTG under widely varying turbulence loading. *Renewable Energy*, 32(14): 2407-2423.
- Nichita C, Luca D, Dakyo B, Ceang E., 2002. Large band simulation of the wind speed for real time wind turbine simulators. *IEEE Transactions on Energy Conversion*, 17(4): 523-529.
- Tapia A, Tapia G, Ostolaza Jx, et al. 2003. Modeling and control of a wind turbine driven doubly fed induction generator. *IEEE Transactions on Energy Conversion*, 18(2): 194-204.
- Inlian Munteanu, Antoneta Iuliana Brarcu, Nicolaos Antonic Cutululis et al. 2008. *Optimal Control of Wind Energy Systems*. London: Springer.
- Inlian Munteanu, Nicolaos Antonio Cutululis, Antoneta Iuliana Brarcu, et al. 2005. Optimization of Variable Speed Wind Power Systems Based on a LQG Approach. *Control Engineering Practice*, 13(7): 903-912.
- Yazhou Lei, Gordon Lightbody. 2005. wind power and electricity market. *Automation of Electric Power Systems*, 29(10): 1-5.
- Bianchi F, De Battista H, Mantz RJ, 2008. *Wind turbine control systems principles, modelling and gain scheduling design*. London: Springer.

AUTHORS BIOGRAPHY

Yan-xia Shen received the Ph.D. degree in power electronics and motor drives from China University of Mining and Technology, Xuzhou, China, in 2004. Her main research interests are in the fields of multi-objective optimization of high-power wind turbine control system, optimal control of high performance servo system.

She is a member of the IEEE Transactions on Power Electronic, the deputy secretary general of Wuxi Municipal Association of Automation.

Xiang-xia Liu Master candidate in control theory and application of Jiangnan University ,Wuxi, China. Her research interest is the optimal control of direct-driven permanent synchronous wind power system.

Zhi-cheng Ji received the Ph.D. degree in power electronics and motor drives from China University of mining and technology, Xuzhou, China, in 2003. His main research interests are multi-objective optimization of high-power wind turbine control system, the complex nonlinear control, network control.

Prof. Ji is a member of Professional Committee of Chinese Association of Automation Applications, and the director of Chinese Association for System Simulation.

Ting-long Pan received the Ph.D. degree in power electronics and motor drives from China University of mining and technology, Xuzhou, China, in 2004. His main research interest is the optimal control of variable pitch high-power wind turbine control system.

From 2008 to 2009, he was with the Intelligent Power Electronics and Energy System Laboratory, Electrical and Computer Engineering Department, University of Miami, America.

MISSING DATA ESTIMATION FOR CANCER DIAGNOSIS SUPPORT

Witold Jacak^(a), Karin Proell^(b)

^(a)Department of Software Engineering

Upper Austria University of Applied Sciences Hagenberg, Softwarepark 11, Austria

^(b)Department of Medical Informatics and Bioinformatics

Upper Austria University of Applied Sciences Hagenberg, Softwarepark 11, Austria

^(a) Witold.Jacak@fh-hagenberg.at ^(b) Karin.Proell@fh-hagenberg.at

ABSTRACT

The paper presents a heterogeneous neural network based system that can be used for estimation of missing tumor marker values in patient data and in a second step for calculating the possibility of a cancerous disease. For estimation of missing values we use different approaches: neural network based estimation of a specific marker depending on existing values of a related marker and neural network based estimation of missing tumor markers depending on standard blood parameter measurements. Finally we compare the results for calculating the possibility of a cancerous disease using different methods for missing value estimation of patient data.

Keywords: neural network, tumor marker prediction, missing values in biomedical data

1. INTRODUCTION

Tumor markers are substances produced by cells of the body in response to cancer but also to noncancerous conditions. They can be found in body liquids like blood or in tissues and can be used for detection, diagnosis of some types of cancer. For different types of cancer different tumor markers can show abnormal values and the levels of the same tumor marker can be altered in more than one type of cancer.

Blood examination tests only a few tumor marker values and for this reason the usage of such incomplete data for cancer diagnosis support needs estimation of missing marker values. Neural networks are proven tools for prediction tasks. For example neural networks were applied to differentiate benign from malignant breast conditions bases on blood parameters (Astion et Wilding 1992), for diagnosis of different types of liver disease (Reibnegger et al. 1991), for early detection of prostate cancer (Djavan et al. 2002; Matsui et al. 2004), for studies on blood plasma (Liparini et al. 2005) or for prediction of acute coronary syndromes (Harrison et al. 2005).

In this work we present a heterogeneous neural network based system that can be used for tumor marker value estimation and for prediction of cancer possibility.

We combine different neural networks, which calculate the possibility of a cancerous disease in different ways,

- without estimation of missing marker values,
- with neural network based estimation of missing marker values depending on existing values of other markers, and
- with neural network based estimation of missing marker values depending on standard blood parameters.

2. GENERAL CANCER DIAGNOSIS SUPPORT SYSTEM

We focus our considerations on the design of a complex decision support system for the calculation of the possibility of cancerous diseases.

The cancer prediction system is based on data coming from vector $\mathbf{C} = (C_1, \dots, C_m)$ of tumor marker values. We use several parallelly coupled neural networks to calculate the possibility of general cancer occurrence.

The basic problem in such approaches is data incompleteness of training data, which leads to problems in training neural networks. Data completeness can be achieved in two ways.

First we can use the existing values of markers in vector $\mathbf{C} = (C_1, \dots, C_m)$ of tumor marker values to estimate the missing values of other markers in the same vector.

Second we can estimate the missing values of markers by using supporting data - in this case a blood parameter vector $\mathbf{P} = (P_1, \dots, P_n)$ of each patient is used. Frequently also this vector is incomplete too. For that reason it is necessary to develop a system using complete or partially complete blood parameters for directly estimating values for missing tumor markers or for a classification of them. The structure of the system is presented in Figure 1.

3. TUMOR MARKER VALUES BASED CANCER DIAGNOSIS SUPPORT SYSTEM

Cancer diagnosis support uses parallel working systems ($Cancer^k$), with the same structure of networks trained for different types of cancer. The input of each $Cancer^k$ system is the complete or incomplete vector \mathbf{C} of tumor marker specific for the chosen type of cancer, and the

output represents the possibility (values between 0 and 1) of a cancerous disease. Output values of the network system greater than 0,5 are treated as cancer occurrence.

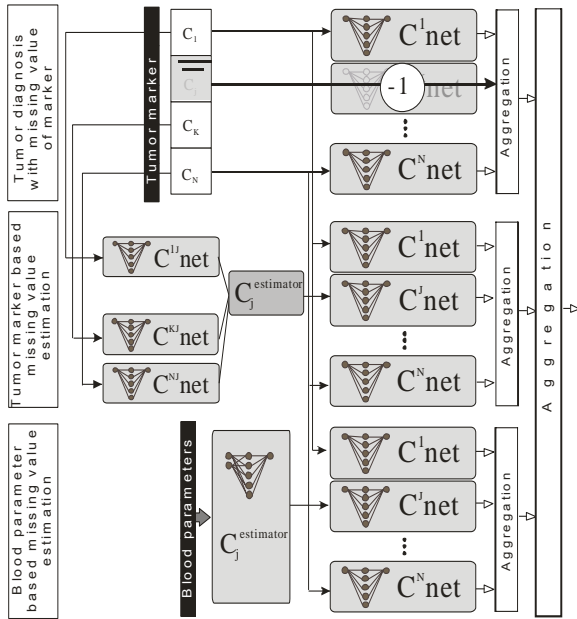


Figure 1. Architecture of Data Driven Cancer Diagnosis Support System

Each $Cancer^k$ system consists of many different groups of neural networks (see Figure 1.).

- Group of neural networks (C_{net}^k) for individual marker C_i ; $i=1, \dots, m$.
- Feed forward neural network (C_{net}^{Group}) for a whole vector of marker C , with complete or incomplete values.
- Group of neural networks (C_{net}^{kj}) for estimation of marker value of marker C_j based on marker C_k
- Feed forward neural network (P_{net}^{Group}) for estimation of maker value C_j , depending on blood parameters
- Cascaded coupled aggregation method for final calculation of cancer plausibility.

4. GROUP OF SEPARATE NEURAL NETWORKS FOR INDIVIDUAL MARKER (C_{NET}^K)

The first group of neural networks contains parallel coupled neural networks, which are individually trained for different tumor markers. Each neural network is of type feed forward with one hidden layer having 6 -10 neurons, activation functions tan/sigmoid, further one input (normalized tumor marker value) and one output (diagnosis: 0 – no cancer (healthy) and 1– cancer (ill)). The networks were trained independently of type of cancer disease (i.e. for all types of cancer diseases).

The values of markers are further categorized to four intervals (Classes). The first interval includes all values less than a *Normal Value* of marker, the second interval includes all values between the *Normal Value* and an *Extreme Normal Value* of marker, third interval

includes values between the *Extreme Normal Value* and a still *Plausible Value* of marker and fourth interval includes all values greater than *Plausible Value*.

The input values of each network for each training and testing process are normalized using the respective upper bound of *Plausible Value*. Each value of marker, which extends that upper bound, obtains the value 1. The individually trained networks represent a generalized cancer occurrence prediction, disregarding specific type of cancer and based only on one specific tumor marker.

4.1. Case study: Breast Cancer – C125, C153, C199 and CEA marker Group

One example of a trained neural network (with 6 neurons in hidden layer) for tumor marker C125 is presented in Figure 2. The x-axis represents the normalized values of tumor marker C125 and the y-axis represents cancer possibility. Network-output values greater than 0,5 (middle line) are interpreted as cancer occurrence.

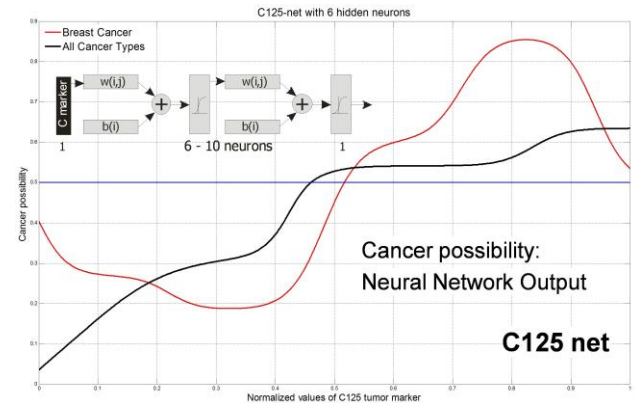


Figure 2. Output of individual trained neural network for tumor markers C125.

The networks were trained with 2598 datasets for C125 marker, with 2442 datasets for C153 marker, with 4519 dataset for C199 marker and with 7153 dataset for CEA marker. The datasets contain data with different cancer types from C00 to C96 ICD 10 code (44%) and data without cancer occurrence (56%).

Additionally, we trained these networks with smaller sets of data for one specific type of cancer disease. Figure 2 presents an example of trained C125 representing breast cancer.

Generally the network trained for all cancer types is more pessimistic, it means this network predicts cancer possibility greater than 0,5 for smaller values of tumor marker as the network trained for one special cancer type. In the example in Figure 2 the threshold points of networks trained for all cancer types is 0,46 (75,9 U/ml) for C125 marker. For marker C153 the threshold was 0,32 (34,8 U/ml), for C199 0,55 (73,1 U/ml) and or CEA 0,25 (14,1 ng/ml). The threshold points of networks trained for breast cancer are 0,52 (85,5 U/ml) for C125, 0,44 (48,2 U/ml) for C153, 0,42 (55,8 U/ml) for C199 and 0,37 (20,4 ng/ml) for CEA.

(Values for Markers C153, C199 and CEA are not shown in Figure 2).

The regressions between network outputs trained for all cancer types and breast cancer type are 0.69, 0.88, 0.88, and 0.91 for C125, C153, C199 and CEA, respectively. The influence of networks trained in this way on the final diagnosis prediction will be discussed in the next section.

The input of parallel coupled C_{nets} is the vector of tumor marker $C = (C_1, \dots, C_m)$, where some C_i are missing. When the tumor marker value in vector C is available, then the adequate C_{net} calculates the predicted cancer possibility. When a marker value in vector C is not available, then the output of C_{net} is set to -1. The individually calculated output values of C_{nets} can be aggregated in many different ways. We compare three methods of aggregation:

1. Maximum value of all individual network outputs: $C_{net}(C) = \max\{C_{net}^i(C_i) | i=1, \dots, m\}$
2. Average value of all individual network outputs, without missing values: $C_{net}(C) = \text{avg}\{C_{net}^i(C_i) | i=1, \dots, m \& C_i \neq -1\}$
3. $Net_{aggregation}$ - neural network trained on individual networks outputs (this neural network can be trained with data of only one chosen cancer type $Cancer^k$). $C_{net}(C) = net_{aggregation}(C_{net}^i | i=1, \dots, m)$

Other interesting possibility is using the thresholded values of outputs of individual networks (i.e. if $C_{net}(C) < 0,5$ then $C_{net}(C)=0$, else $C_{net}(C) =1$) as input for the perceptron type network.

We use one aggregation type in the full system. In case of max aggregation: If only one marker of the marker group shows a greater value than the aggregation has yielded this value is taken.

The diagnosis prediction based on aggregation of separately cancer predictions of individual marker networks C_{net} is not sufficient for generalization of cancer occurrence. It is necessary to reinforce the information coming from data of whole group of markers. Therefore two neural networks with cumulative marker groups are added. These networks will be trained only for a specific cancer type. When the tumor marker value in vector C is not available, then this value is set to -1. Based on this assumption we can generate training sets for a specific cancer type ($Cancer^k$) and train the neural network: A feed forward neural network with 16-20 hidden neurons and tansig/linear activation functions (C_{net}^{group}) with a vector C on input and diagnosis of tumor occurrence on output. This network can be used additionally to the individually trained networks for diagnosis prediction without and with estimated missing values of vector C .

5. SYSTEM FOR PREDICTION OF MISSING TUMOR MARKER VALUE

5.1. Estimation of missing tumor marker values based on other tumor marker

Estimation of missing values of tumor markers can be done by values of other tumor markers from tested marker group C . It can be accomplished by preparation of pair wise trained neural networks C_{net}^{ij} where the pattern set includes values of C_i marker as input and values of C_j marker as output in case both marker values are available. Not all tumor marker values can be predicted with sufficient level of plausibility.

In our case study we use feed forward neural networks with 5-10 hidden neurons for the estimation of a missing value of marker C_j based on a known value of marker C_i .

Figure 3 presents the prediction of C125 based on marker C153 and the prediction of marker value of CEA based on marker C199.

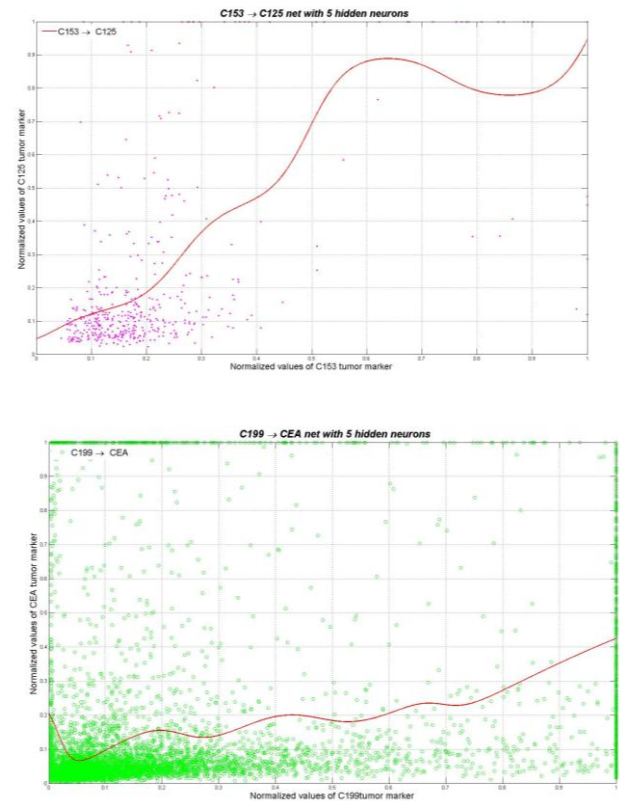


Figure 3. Outputs of neural networks for tumor markers C125 based on C153, and CEA based on C199.

In the first case we can observe clear dependency of marker C125 on marker C153 and in the second case a dependency does not exist.

In cases more than one value exists in vector C , then the estimator for a missing value C_k can be computed as:

$$C_k = \max\{C_{net}^{ik}(C_j) | i \neq k\} \quad \text{or} \quad C_k = \text{avg}\{C_{net}^{ik}(C_j) | i \neq k\}$$

where C_i represent the existing values markers in vector C . The quality of the estimation of a missing value will highly depend on existing values in vector $C = (C_1, \dots, C_m)$.

5.2. Estimation of missing tumor marker values based on blood parameters.

Typically in labor blood examination 27 blood parameters such as HB, WBC, HKT, MCV, RBC, PLT, KREA, BUN, GT37, ALT, AST, TBIL, CRP, LD37, HS, CNEA, CMOA, CLYA, CEOA, CBAA, CHOL, HDL, CH37, FER, FE, BSG1, TF and tumor markers such as AFP, C125, C153, C199, C724, CEA, CYFRA, NSE, PSA, S100, SCC, TPS etc. are measured. For each parameter and marker are experimentally established upper and lower bounds of values. We divide the values range of marker C and blood parameter P into k non-overlapping intervals, called classes.

In our case study we define four classes ($k = 4$). *Class 1* includes all values less than *Normal Value* of marker or blood parameter, *Class 2* includes all values between *Normal Value* and *Extreme Normal Value* of marker or blood parameter, *Class 3* includes values between *Extreme Normal Value* and *Plausible Value* of marker or blood parameter and *Class 4* include all values greater than *Plausible Value*. These classes and their limits are used in normalizing process of parameter and marker values. In normalizing process, we replace the missing value with the value -1.

The system consists of three heterogeneous parallel-coupled artificial neural networks and a decision-making system based on aggregation rules (Jacak et al., 2010a).

The input and output values of each network for training and testing are normalized using the respective upper bound of *Plausible Value*. Each value of parameter or marker, which is greater than this upper bound, obtains the normalized value 1.

The general marker value estimation system contains three neural networks.

- Feed forward neural network (*FF*) with p inputs (normalized values of blood parameter vectors P) and one output, normalized values of marker C_i
- Pattern recognition neural network (*PR*) with p inputs (normalized values of blood parameter vectors P) and k outputs, k -dimensional binary vector coding classes of marker C_i
- Combined feed forward neural network (*FC*) with p inputs (normalized values of blood parameter vectors P) and two outputs: normalized values of marker C_i (as in network *FF*), and normalized classes of marker C_i :

All neural networks have one hidden layer and tan-sigmoid or log-sigmoid transfer function. The output values of neural networks belong usually to interval $[0, 1]$.

Based on the neural networks calculated estimation of marker value we can establish four hypotheses $x_1, x_2,$

x_3, x_4 concerning the class of marker. For each hypothesis x_1, x_2, x_3, x_4 the possibility value is calculated too. These hypotheses are to be verified for finding the maximal possible prediction. (Jacak et al., 2010b)

The empirical test shows that the best results are achieved with networks having 40-60 neurons of hidden layer.

It can be expected that not all markers can be predicted with good quality. The examples of regression between blood parameters test data and estimation system output for tumor markers C153 (regression 0,71) and CEA (regression 0,53) are presented in Figure 4.

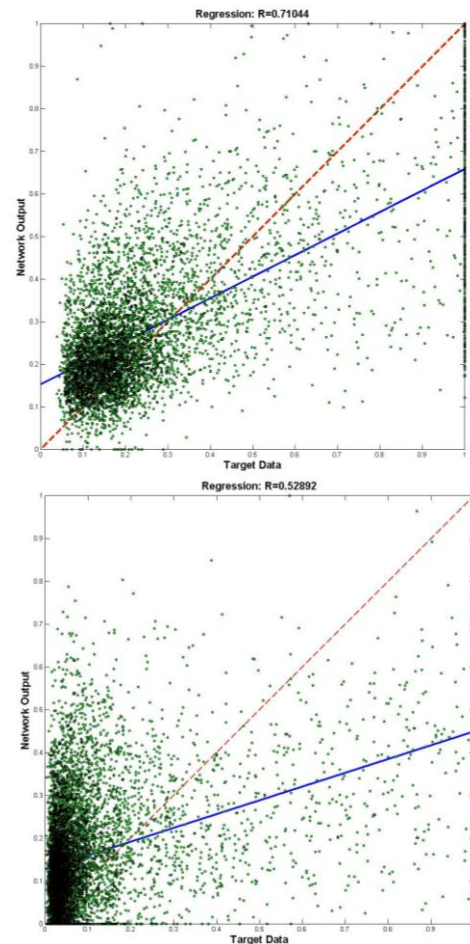


Figure 4: Regression between test data and predicted tumor marker values for markers C153 (first) and CEA (second).

In this method of value estimation, the quality of estimation will highly be dependent on missing values of marker in vector $C = (C_1, \dots, C_m)$.

The two methods presented have different properties and it will be necessary to combine both estimations results for a final prediction of cancer occurrence. We aggregate the outputs of the three diagnosis prediction networks (without estimation of missing values, with estimation of missing values based on other existing markers values and with estimation of missing values based on blood parameters) by the

applying the maximum function. The result of prediction quality is presented in the next section.

6. RESULTS AND COMPARISON OF PREDICTION QUALITY OF DIFFERENT APPROACHES

For comparison between previously described methods of breast cancer prediction based on marker group $C = (C_{125}, C_{153}, C_{199}, CEA)$ we have prepared a test data set containing 695 positive cases (diagnose coded as 1) and 765 negative cases (diagnose coded as 0). By assumption that the general probability of positive and negative cancer occurrence is 0,5. We can estimate the probability $P(1/1)$ (true positives) and $P(0/0)$ (true negatives) for the various systems.

The results of prediction are presented in table 1.

Table 1: Prediction of cancer occurrence

System	P (correct)	P (1/1)	P (0/0)
Individually trained networks without estimation of missing values	0,63	0,29	0,93
FF network with C vector, no estimation of missing values	0,67	0,45	0,89
Individually trained networks with estimation of missing values, based on existing additional values in vector C (max as aggregation function)	0,63	0,35	0,90
FF network with C vector, with estimation of missing values based on existing additional values in vector C (max as aggregation function)	0,66	0,57	0,74
Individual trained networks with estimation of missing values based on blood parameters values	0,66	0,57	0,73
FF network with C vector, estimation of missing values based on blood parameters	0,63	0,42	0,80
Aggregated prediction based on three network outputs (maximum aggregation function applied)	0,70	0,77	0,63

Diagnosis prediction performed without estimation of missing values of marker values works increases the probability $P(0/1)$ of false negatives (see Figure 5). Prediction based on missing value estimation decreases false negatives rate but increases false positives rate. All confusion matrices of chosen experiments are presented in Figures 5-8. The whole system increases the probability of correct cancer diagnosis and decreases the false positives rate. The confusion matrix of the overall system for breast cancer diagnosis is presented in Figure 8.

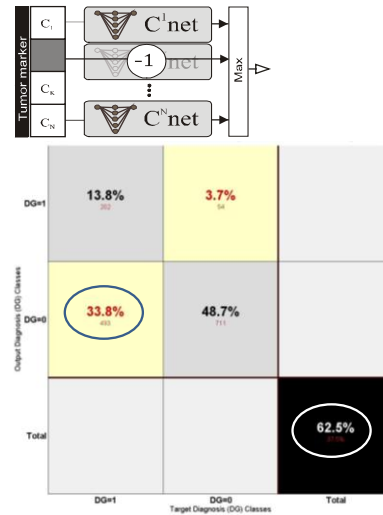


Figure 5. Confusion matrix of breast cancer diagnosis based on C125, C199, C153 and CEA marker group without estimation of missing values. False positives rate is 34%.

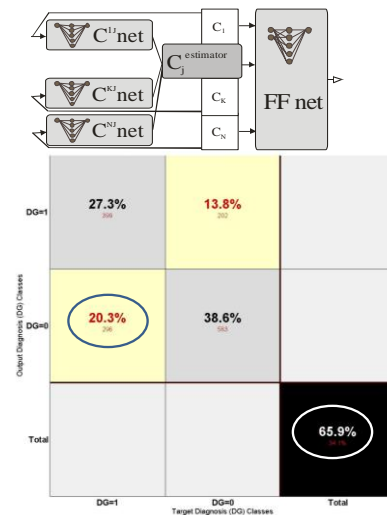


Figure 6. Confusion matrix of breast cancer diagnosis based on C125, C199, C153 and CEA marker group with marker based estimation of missing values of markers. The false positives rate is 20%.

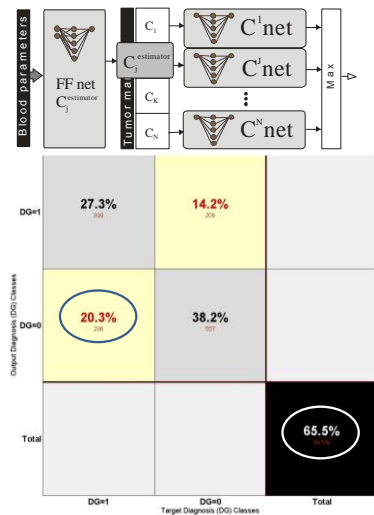


Figure 7. Confusion matrix of breast cancer diagnosis based on C125, C199, C153 and CEA marker group with blood parameters based estimation of missing values of markers. False positives rate is 20 %.

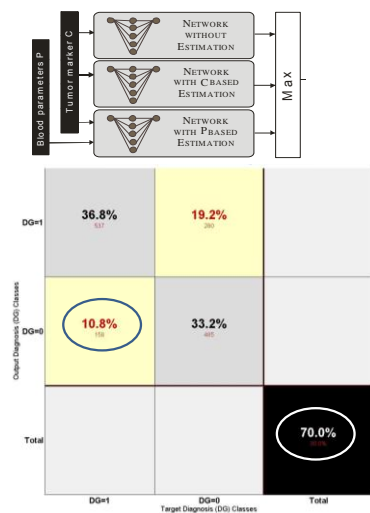


Figure 8. Confusion matrix of breast cancer diagnosis based on aggregated outputs of 3 networks with different estimation of missing values. The false positive rate is 11%.

REFERENCES

Astion, M.L., Wilding P., 1992, Application of neural networks to the interpretation of laboratory data in cancer diagnosis. *Clinical Chemistry*, Vol 38, 34-38.

Djavan, et al., 2002. Novel Artificial Neural Network for Early Detection of Prostate Cancer. *Journal of Clinical Oncology*, Vol 20, No 4, 921-929

Djavan, B., Remzi, M., Zlotta, A., Seitz, C., Snow, P., Marberger, M., 2002, Novel Artificial Neural Network for Early Detection of Prostate Cancer.

Journal of Clinical Oncology, Vol 20, No 4, 921-929

Harrison, R.F., Kennedy, R.L., 2005, Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation. *Ann Emerg Med.*; 46(5):431-9.

Jacak W., Proell K., 2010a, Data Driven Tumor Marker Prediction System, *Proceedings of EMSS 2010*, Fes, Marokko

Jacak W., Proell K., 2010b, Neural Network Based Tumor Marker Prediction, *Proceedings of BroadCom 2010*, Malaga, Spain

Liparini, A., Carvalho, S., Belchior, J.C., 2005, Analysis of the applicability of artificial neural networks for studying blood plasma: determination of magnesium on concentration as a case study. *Clin Chem Lab Med.*; 43(9):939-46

A NEW DEVS-BASED GENERIC ARTIFICIAL NEURAL NETWORK MODELING APPROACH

S. TOMA^(a), L. CAPOCCHI^(b), D. FEDERICI^(c)

^{(a)(b)(c)} SPE UMR CNRS 6134 Laboratory, University of Corsica, Quartier Grimaldi, 20250, Corte, France

^(a)toma@univ-corse.fr, ^(b)capocchi@univ-corse.fr, ^(c)federici@univ-corse.fr

ABSTRACT

The Artificial Neural Network (ANN) is a black box model capable of resolving paradigms that linear computing cannot. Therefore, the configuration of ANN is a hard task for modeler since it depends on the application complexity. The Discrete Event system Specification (DEVS) is a formalism to describe discrete event system in a hierarchical and modular way. DEVS is mainly used to defragment a system or a model in an easy way allowing the interaction with the architecture and behavior of the system. This paper presents a new artificial neural network modeling approach using DEVS formalism in order to facilitate the network configuration by introducing a new scheme of the training phase. We validate our approach with a simple not linearly separable data set example provided by two-dimensional XOR problem.

Keywords: artificial intelligence, discrete event systems, artificial neural networks, learning systems, modeling, simulation.

1. INTRODUCTION

Throughout the years, the computational changes have brought growth to new technologies. Such is the case of ANNs; they have given various solutions to the industry. Designing and implementing intelligent systems has become a crucial factor for the innovation and development of better products for society. Such is the case of the implementation of artificial life as well as giving solution to interrogatives that linear systems are not able resolve (Bishop 1995, Mas and Flores 2008, Agatonovic-Kustrin and Beresford 2000). In the world of engineering, neural networks have two main functions: Pattern classifiers and non linear adaptive filters (Bishop 1995). As its biological predecessor, an ANN is an adaptive system where each parameter is changed during its operation and it is deployed for solving the problem in matter (Drew and Monson 2003).

ANN is a system capable of resolving paradigms that linear computing cannot. It is a system based on the operation of biological neural networks, in other words, it is an emulation of biological neural system. Another aspect of the ANN is that there are different architectures, which consequently requires different types of algorithms, so it might look like a complex system (Agatonovic-Kustrin and Beresford 2000).

Always said that the ANN is a black box system and we can never interact with its structure. Sometimes depending on the architecture or the algorithm used some parameters must be initialized. Some of these parameters are a function of the complexity of the system that we will try to solve. For certain types of problem try and error are able to get the best network configuration, but it will be better to find some algorithms to automate this process (Bishop 1995, Agatonovic-Kustrin and Beresford 2000).

DEVS is a formalism which allows the behavior modeling of a non linear system (Zeigler and Praehofer and kim 2000). This formalism provides a model (atomic) in order to define the behavior of a system (Concepcion and Zeigler 1988). DEVS and ANN are two concepts that are able to simulate complex systems and problems. Combining DEVS and ANNs could make a perfect match because of the nature of each of these concepts. In (Choi and Kim 2002) we can see this combination was the extraction of the DEVS from a trained ANN. An another interesting approach has been presented in (Filippi and Bisgambilia and Delhom 2001) where the ANN behavior is encapsulated into only one atomic DEVS model making a hybrid system that offers a better simulation.

In order to go much further with hybrid systems this paper presents a new modeling approach of the ANN using DEVS aspects. This new model will concern presenting an ANN into certain number of atomic and coupled models. This approach will be able to facilitate the network configuration that depends a lot on the application. In other words the new model will be able to give the space to implement algorithms and plug-ins to automate the network configuration as the network efficiency.

The remainder of the paper is organized as follows. In section 1, we introduce the DEVS formalism and the ANN concepts, showing their usage, their structure and how could they be implemented. Section 2 describes the new hybrid system that transforms the ANN into several DEVS models. Section 3 describes a test comparison of the new model using an XOR problem in order to present our new design. Finally, we conclude and present future works.

2. BACKGROUND

2.1. DEVS Formalism

DEVS is a formalism introduced by Zeigler (1976) to describe discrete event system in a hierarchical and modular manner. A manner way means that the system has input and output ports that allow it to interact with the external environment (Concepcion and Zeigler 1988). This formalism is distincter between the simulation approach and the modeling one. The DEVS modeling approach captures dynamic behavior with atomic models. The simulation approach is responsible for the automatic generation of the simulation behaviors into two models: atomic and coupled model. Each model type could be considered like a black box with some behaviors and interactions with the external environment through input and output ports. The atomic models can be linked together in a well-defined way to produce more complex coupled models whose behaviors are described by their atomic models and a set of a relation between those models. Any real system can be modeled using DEVS into a collection of coupled and atomic models (Barros and Zeigler and Fishwick 1998).

2.1.1. Atomic and Coupled Models

An atomic model is a model for a system that has a set of inputs, outputs, states, transition functions, a time advance function and an output function. Any atomic model can be defined by the following structure.

$$M_{atomic} = \langle X, Y, S, \delta_{ext}, \delta_{int}, \lambda, t_a \rangle$$

- X is a set of inputs
- Y is a set of outputs
- S is a set of states
- $\delta_{int} : S \rightarrow S$ is the internal transition function.
- $\delta_{ext} : Q \times X \rightarrow S$ is the external transition function, Where $Q = \{(s, e) \mid s \in S, 0 \leq e \leq t_a(s)\}$ is the set of total states, e is the time elapsed since the last transition.
- $\lambda : S \rightarrow Y$ is the output function
- $t_a : S \rightarrow R^+_{0,\infty}$ is the time advance function.

The internal transition function describes state changes that occur in the absence of input over time. The external transition function responds to an input with a certain state change. Based on the current state the output function produces and output event. The time advance function calculates the amount of time before the next internal state transition takes place (assuming no inputs arrive in the interim).

A coupled model is a confirmation on the hierarchical notation of DEVS. It is consisted of a set of sub-models. Sub-models could be either atomic or coupled models. The behavior of such models is defined by the behavior of its models components and the relations between them.

The coupled models consist of a set of inputs, outputs, states, a set of sub-models with the influences between them and three types of coupling between models (Figure 1).

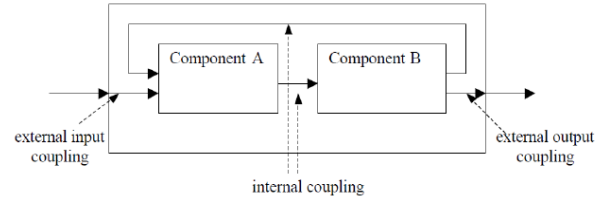


Figure 1: DEVS coupling in coupled models

The coupled model has inputs and outputs defined the same way as the atomic model. The couplings between the sub-models are defined under three categories: (i) External input coupling, (ii) Internal coupling, (iii) External output coupling. With the atomic and the coupled models it is easy to model and simulate discrete event systems.

2.1.2. DEVS Softwares

Nowadays the number of tools implementing the DEVS formalism is growing too quickly. In (Lara and Vangheluwe 2002) the authors present a General User Interface (GUI) allowing multi-paradigm (including DEVS) modeling. PowerDEVS (Kofman and Lapadula and Pagliero 2003) is an excellent GUI for the DEVS modeling and simulation focused on hybrid systems. Another interesting tool is described in (Baati and Frydman and Giambiasi 2007) that is often used for pedagogical aspects. In the case of ANN modeling we need a GUI only based on DEVS formalism allowing us to design and implement any architecture and ANN algorithm.

The main benefit using this kind of software is the simplicity of modeling the ANN algorithms and the possibility of creating libraries of reusable and "viewable" components. The GUI software allows creating, deleting, handling or switching the models in a simply way using toolbar or shortcut. Moreover, the simulation process is automatic and performed by clicking on a simple button.

DEVSImPy (Python Simulator for DEVS models) (Capocchi and Santucci and Poggi And Nicolai 2011) is a user-friendly interface for collaborative modeling and simulation of DEVS systems implemented in Python. Python is a programming language known for its simple syntax and its capacity to allow modelers to implement quickly their ideas (Sanner 1999). The DEVSImPy project uses the python language and provides a GUI based on PyDEVS (Bolduc and Vangheluwe 2001) Application Program Interface (API) in order to facilitate both the coupling and the reusability of PyDEVS models. This API is used in the excellent multi-modeling GUI software named ATOM3 (Lara and Vangheluwe 2002) which allows the usage of several formalisms without focusing on DEVS. DEVSImPy is an open source project under GPL V3 license and its development is supported by the SPE

research laboratory team. It uses the wxPython graphic library and it can be downloaded from <http://code.google.com/p/devsimpy/>.

The main goal of this environment is to facilitate the modeling of DEVS systems using the GUI dynamic library and the drag and drop functionality. With DEVSimPy, models can be stored in a dynamic library in order to be reused and shared. The creation of dynamic libraries composed by DEVS components is easy since the user is coached by dialogs and wizard during the building process. We propose in this paper the DEVS modeling of ANNs algorithm through DEVSimPy in order to implement a generic ANN library. Thereby, the DEVSimPy developer will be able to use this library when the ANN is needed in the modeling of complex systems at all times.

2.2. Artificial Neural Network

Basically, an ANN is a system that receives input, process the data, and provides an output. This system can be used for two main functions: Pattern classifiers and as non linear adaptive system. By adaptive, it means that the system parameters can be changed during operation to solve the faced problem. This is called the training phase. During this phase a classifier or a non linear adaptive system ANN tries to adapt its parameters to solve the problem in matter. The ANN has many different architectures and for everyone there is some personalized algorithms. In this section the Feed-Forward Neural architecture with Backpropagation algorithm is presented.

2.2.1. Feed-Forward Neural Architecture

Any ANN has a certain number of entities called neuron. The power of the network comes from the weighted connections between different neurons. When neuron receives weighted inputs it calculates the sum and then passes the data through a transfer function. The transfer function is the element that introduces the non-linearity aspect into the neural network. Many types of function can be used: hyperbolic, threshold, piecewise-linear, and the sigmoid functions.

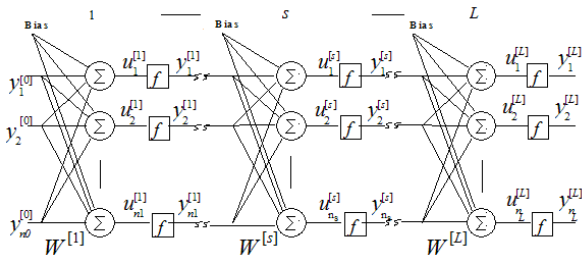


Figure 2: Artificial Neural Network Architecture.

The multilayer perceptrons (MPLs) is the most used class of ANN in all applied fields. A MPL consists of a set of input units (input layer), one or more computation layers (hidden layers), and one output layer (computation/output). In feed-forward architecture (Figure 2) a neuron on layer s always connects to a neuron on layer $s+1$. A fully connect network means that all neurons on layer s are connected to each neuron

on layer $s+1$. Every neural network must be trained before we can use it. So a training algorithm is chosen to adapt and change the connection weights between each layer.

2.2.2. Backpropagation Learning Algorithm

The training algorithm that is used to adjust the network weights is a principal factor of the network accuracy and performance. Two major types of algorithms can be found to train an ANN: supervised and unsupervised (Omatu and Khalid and Yusof 1996, Bishop 1995). Unsupervised learning is used when no output is desired for the ANN. Supervised training is used when we have a well defined desired output. First the input propagate forwardly through layer and an output is calculated. The supervised algorithm calculates the error between desired and calculated output. The layer connection weights are modified trying to minimize this error. This cycle is repeated many times until the network is trained (Agatonovic-Kustrin and Beresford 2000).

$$\mu_{pj}^{[s]} = \sum_{i=1}^{n_{s-1}} w_{ij}^s y_{pi}^{[s-1]} \quad (1)$$

$$y_{pj}^{[s]} = f(\mu_{pj}^{[s]}) \quad (2)$$

$$\delta_{pj}^{[s]} = (d_{pj} - y_{pj}^{[L]}) f'(u_{pj}^{[L]}) \quad (3)$$

$$\delta_{pj}^{[s]} = f'(u_{pj}^{[L]}) \sum_{r=1}^{n_{s+1}} (\delta_{pj}^{[s]} w_{rj}^{s+1}) \quad s = L - 1, \dots, 1 \quad (4)$$

$$w_{ji}^{[s]}(t) = w_{ji}^{[s]}(t) + N(\delta_{pj}^{[s]}(t) y_{pji}^{[s-1]}) + M(\delta_{pj}^{[s]}(t-1) y_{pji}^{[s-1]}) \quad (5)$$

$$E = \frac{1}{N} \sum_{p=1}^m \sum_{k=1}^{n_L} (d_{pk} - y_{pk}^{[L]})^2 \quad (6)$$

The backpropagation (BP) is the most common supervised learning algorithm to a feed-forward neural network used as classifiers (Figure 2). A BP network learns by example; in other word it learns by training sets. At the beginning of the training phase, all weights are initialized by random values - say between -1 and +1. Next the input patterns (p) propagate through the network layers (s) calculating the output value (Eq.1, Eq.2). After the first propagation of data the calculated output normally is different than the desired output. At that point the BP algorithm comes to calculate the error of each output neuron. After that a reverse procedure to the forward propagation takes place trying to calculate new weights as a function of the calculated error. This back propagation of values is calculated using the derivative of the activation function, the inputs, outputs of each layer, a momentum factor (M) and a learning factor (N). Equation 3 is used for the output layer error calculation and for all other layer the equation 4 is applied. Equations 5 and 6 are used for the weight adaption and to calculate the quadratic error.

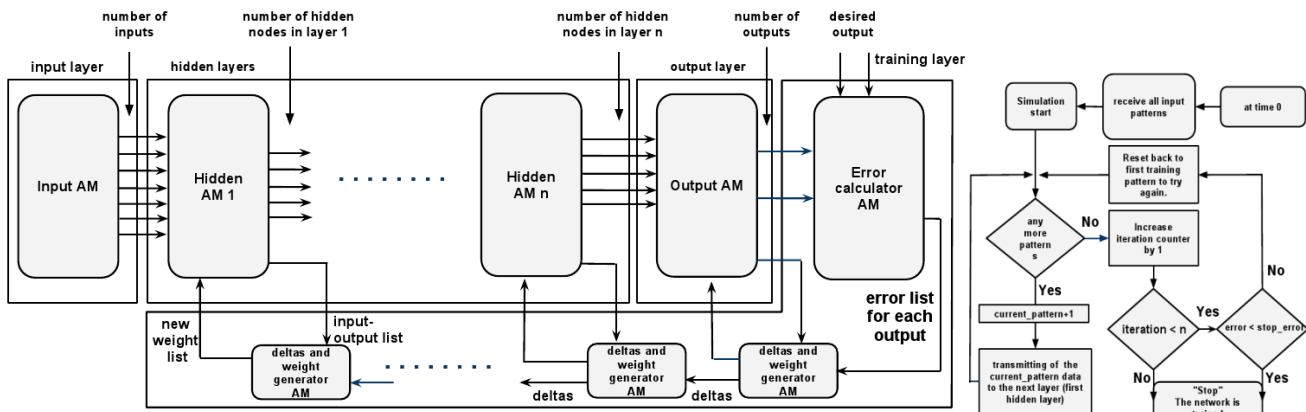


Figure 3: (a) ANN/DEVS model transformation

(b) Input Atomic Model block diagram.

3. PROPOSED ANN MODELING APPROACH

3.1. ANN/DEVS Compatibility

Since its birth the ANN is inspired from the human brain neurons and known as a black box. When an ANN is build some configuration must take place. First the number of layers that we are going to use and this depends on the complexity of the problem that we are trying to solve. Usually we use a two layer neural network (two calculation layer + input layer) that is sufficient to solve most non-linear problems. Second the number of neuron in each layer: it is fixed for the input and output layers by the input and output data length; for the hidden layer it is one of the hardest choices. The choice of the number of neuron in the hidden layer is different for each application but some recommendations may take place; taking half the number of input neurons plus one is one of them. Also the bias value in each layer can be a difficulty too. Third the algorithms that can learn or train the network can have a great effect on the network performance. The BP algorithm is one of the most common ones. As a result for the BP choice, learning and momentum factors (N,M) must be chosen. Forth the stop condition for the training phase. Too much training could be considered as over-training, which means that the network will learn the data noise and not the desired pattern. Too short training could lead to an untrained network. So an iteration number depending on the application complexity must be chosen or define a stop condition.

As shown in the previous paragraph too many parameters must be calibrated before using the neural network. Creating a generic ANN library that can be automatic configured and even new aspect and algorithms can be plug-in directly to the network without re-building it. Trying to break the idea that the ANNs are always black boxes and they are not easy to calibrate, a DEVS model is proposed to simulate an ANNs. The ANNs are by default using discrete event; the network is always waiting to an input event to generate an output one. Even inside the network itself, every layer is waiting to receive data from the previous

layer to start calculations. Also during the BP algorithm the recalculation of weights starts when an output is calculated; than the error is calculated; than modification of weight is done layer after layer waiting the changes in the previous ones. So ANNs have a natural compatibility with DEVS formalism, which makes it obvious to create the generic neural library using this formalism.

3.2. ANN/DEVS Mapping Approach

Usually when we represent the neural network we see it as in Figure 2, but what we can see in Figure 3 (a) could be a little bit different. In this paper we propose a DEVS model per layer, which means that for the input, hidden and output layers in the neural network will be presented as a standalone atomic model. And a new training layer will be presented into multiple atomic models.

The input atomic model (input layer) could be considered as the leader of the network. It is called leader because it controls the data propagation throw the network. Controlling the data propagation means that it controls when to propagate the learning, testing, or the real data patterns. First this input model receives with all pattern types and the stop learning condition (iteration number). After initialization it starts to push learning patterns into the network to start calculations (Figure 3 (b)). An iteration number can be determined but also a minimum error to reach to not to get the over training problem. In this design the hidden and the output models are almost the same model. The unique difference is the number of neurons in the output model is fixed by the number of outputs as it is the rule of any neural network. Then the calculations made in the hidden model are the same as in the output model. Both of them are multiplying the inputs by the weight list for each neuron inside this model and then go through the transfer function that must be chosen before the calculation starts (Eq.1, Eq.2).

This is the first time to see something called training layer. All models that help only to train the network will be considered as the training layer (Figure 3 (a)). The idea of having a neural DEVS network came

while trying to enhance and automate the training of any neural network and make it as generic as possible. In this layer the learning algorithm takes place, so any learning algorithm can be implemented with its own design. One of these algorithms is the BP shown in the previous section. So the training layer will be composed of an error calculator and deltas and weights generators. The error calculator model has two functions; First is to calculate the error of each output, which means the difference between the calculated output and the desired one for each single output; Second is to calculate the global quadratic error (Eq.6). The deltas and weight generator is the model where the learning algorithm appears. Algorithm 1 shows the external function of this generator model and how the BP algorithm can be implemented.

Algorithm 1:

```

errors = msg_received
for i in range(len(outputs)):
    deltas[i] = self.dactivation(outputs[i]) * errors[i]
    for j in range(len(inputs)):
        for k in range(len(outputs)):
            change = deltas[k] * nputs[j]
            weights[j][k] = weights[j][k] + N * change + M * C[j][k]
            C[j][k] = change
        for i in range(len(inputs)):
            for j in range(len(outputs)):
                GError[i] = GError[i] + (deltas[j] * weights[i][j])

```

As shown in Figure 3(a) the first deltas and weights generator receives the output error list from the error generator and then starts to calculate the deltas and the new weights for the output model, then the next deltas and weights generator does the same for all hidden models. Trying to prevent the overtraining phenomenon an error stop condition will be implemented. After the training phase ends either by a fixed iteration number or another stop condition the test phase begins automatically by the input model. During the test phase the error generator still works but only to calculate the global error and send nothing to the deltas and weight generator so no more weight modification could be done during this phase.

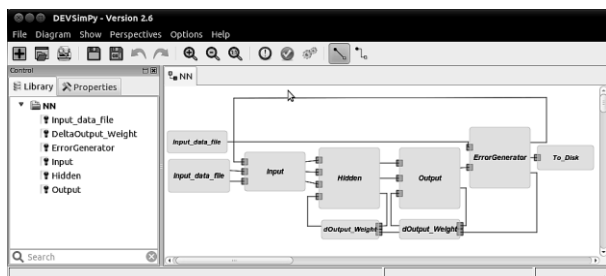


Figure 4: DEVSimPy Dynamic library.

Also sometimes we need to use some validation patterns during the training phase. Using DEVS and so DEVSimPy we can send during the training phase in

parallel to the training patterns and the validation patterns too; which can help to have at the end of the training and the validation error data after each iteration. The validation pattern can help us to see if the network is going through an over training or not.

To implement this modeling approach into DEVSimPy an ANN library is created that contains six atomic models that will help to construct a full ANN: input, output, hidden, error calculator, deltas and weights generator, input file (Figure 4). The input file is a model that extracts data from a file and gives it to the network in form of patterns. Some validation tests using this library will be presented in the next section.

4. VALIDATION AND ANALYSIS

The ANN can solve many non-linear problems and depending on the problem complexity the configuration of the network changes. The complexity of a network depends on two parameters: First the problem dimension; second is if the problem data is linearly separable or not. A simple example of a data set which is not linearly separable is provided by two-dimensional X-OR problem (Mas and Flores 2008) In this example we can show the problem of learning to classify a given data set, where each input vector has been labeled as belonging to one of two classes C1 and C2. The input vectors $x = (0,0)$ and $(1,1)$ belongs to class C1, while the input vectors $(0,1)$ and $(1,0)$ belongs to class C2. It is clear that there is no linear decision boundary which can classify all four points correctly. The neural network using the bias value and the weights values in each layer which is a linear discriminant that will lead to perfect classification. A comparison is made between the designed DEVSimPy neural network models (Figure 4)

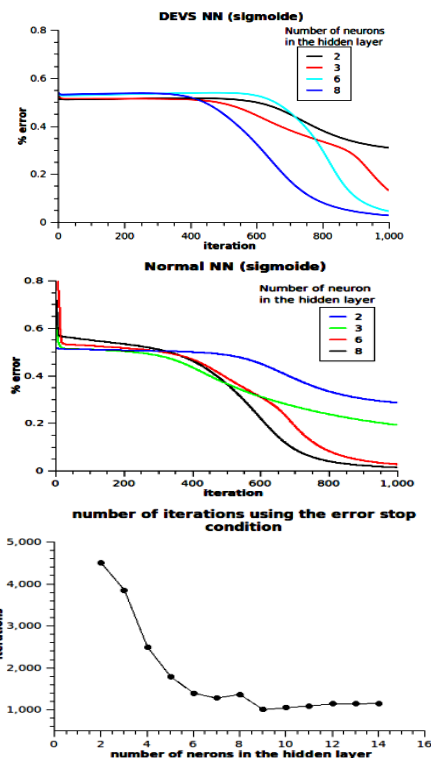


Figure 5: Test and Validation

and a mono-thread ordinary neural network written with python programming language. The tests represent the effect of the incrementation of neurons number inside the hidden layer; also the implementation of the stop learning condition is presented with the different number of neurons to show the effect of increasing neurons number and the iteration number.

Figure 5 represent the comparison between the proposed neural DEVS network and the ordinary neural network when the transfer function is sigmoid with different number of neurons in the hidden layer. These results shows that is the Neural DEVS network is capable to solve the same problem as the normal ANN but using DEVS has a big facility to add or remove hidden layers using the Drag and drop ability that DEVSimPy offer. Also using DEVS formalism gives more automatic configuration that can be added and many plug-ins and enhancements could be developed too. One of those implemented enhancements is the automatization of the stop condition. In other word the learning phase stops only when the network has been learned with minimum accepted error. Figure 5 shows the number of iteration as a function of number of neurons in the hidden layer when an XOR problem is solved using sigmoid transfer function.

5. CONCLUSION

In this paper a new DEVS-model for artificial neural network is presented. This model shows a technique during the training phase by implementing an additional layer named as the training layer. The training layer is composed of several small atomic DEVS models that control the weights adaption independently for each layer. This approach was tested and compared to the standard network. With this technique a great facility to change the training algorithm or to add additional algorithms to make it more efficient is available. The stop error condition is added to the training layer as a test of additional algorithms that can be added as a new model and that could be removed at any time. With the same idea more algorithms to enhance the network performance could simply be added. Moreover there is a work in progress to implement the pruning algorithm. The pruning is a very interesting algorithm that minimizes the number of neurons in the hidden layers to get the smallest network to solve the problem in question. With this defragmentation of the training layer into several models the pruning algorithm can intervenes as an atomic DEVS model just before the entry of the new weights into the hidden layer to make its elimination decision. On the other hand the decision of using DEVS formalism to implement this defragmentation of the neural network opens a new dimension to implement new individual small models or plug-ins to enhance the performance of the network.

REFERENCES

Agatonovic-Kustrin, S., Beresford, R., 2000. Basic Concepts of Artificial Neural Network (ANN) Modeling and its Application in Pharmaceutical

- Research. *Journal of Pharmaceutical and Biomedical Analysis*, vol. 22, no. 5, pp. 717-727.
- Baati, L., Frydman, C., Giambiasi, N., 2007. LSIS DME M&S Environment Extended by Dynamic Hierarchical Structure DEVS Modeling Approach, in *Proceedings of the 2007 spring simulation multiconference*, Vol. 2, pp. 227-234. San Diego, CA, USA.
- Barros, F. J., Zeigler, B. P., Fishwick, P. A., 1998. Multimodels and Dynamic Structure Models: an Integration of DSDE/DEVS and OPMM, *Proceedings of the 30th conference on Winter simulation*, pp. 413-420. Los Alamitos, CA, USA.
- Bishop, C. M., 1995. *Neural Networks for Pattern Recognition, 1st ed. Oxford University Press*, USA.
- Bolduc, J. S., Vangheluwe, H., 2001. The Modelling and Simulation Package PythonDEVS for Classical Hierarchical DEVS, *MSDL Technical Report MSDL-TR--01*. Montreal, Quebec, Canada.
- Capocchi, L., Santucci, J.F., Poggi, B., Nicolai, C., 2011. DEVSimPy: A Collaborative Python Software for Modeling and Simulation of DEVS Systems, Accepted in 20th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, Paris.
- Choi, S. J., Kim, T. G., 2002. Identification of Discrete Event Systems Using the Compound Recurrent Neural Network: Extracting DEVS from Trained Network, *Simulation*, vol. 78, no. 2, p. 90.
- Concepcion, A.I., Zeigler, B.P., 1988. DEVS Formalism: A Framework for Hierarchical Model Development, *IEEE Transactions on Software Engineering*, vol. 14, no. 2, pp. 228-241.
- Drew, P. J. J., Monson, R. T., 2003. Artificial Neural Networks, *Surgery*, vol. 127, no. 1, pp. 3-11, jan.
- Kofman, E., Lapadula, M., Pagliero, E., 2003. PowerDEVS: A DEVS-based Environment for Hybrid System Modeling and Simulation, Tech. Rep., Rosario National University.
- Filippi, J., Bisgambiglia, P., Delhom, M., 2001. Neuro-DEVS, an Hybrid Methodology to Describe Complex Systems, in *Actes of SCS ESS 2001 conference on simulation in industry*, vol. 1, pp. 647-652.
- Lara, J., Vangheluwe, H., 2002. AToM 3: A Tool for Multi-Formalism and Meta-Modelling, *Fundamental Approaches to Software Engineering*, pp. 174-188.
- Mas, J. F., Flores, J. J., 2008. The application of Artificial Neural Networks to the Analysis of Remotely Sensed Data, *International Journal of Remote Sensing*, vol. 29, no. 3, p. 617.
- Omatu, S., Khalid, M., Yusof, R., 1996. *Neuro-Control and its Applications*, Springer.
- Sanner, M. F., 1999. Python: A Programming Language For Software Integration and Development, *J. Mol. Graphics Mod.*, vol. 17, pp. 57-61.
- Zeigler, B.P., Praehofer, H., Kim, T.G., 2000. Theory of Modeling and Simulation, *Academic press*, 2nd Edition.

AN IMPROVED TIME-LINE SEARCH ALGORITHM TO OPTIMIZE INDUSTRIAL SYSTEMS

Miguel Mujica^(a), Miquel Angel Piera^(b)

^(a,b)Autonomous University of Barcelona, Faculty of Telecommunications and Systems Engineering,
08193, Bellaterra, Barcelona

^(a)miguelantonio.mujica@uab.es, ^(b)miquelangel.piera@uab.es

ABSTRACT

The coloured Petri net formalism has been used recently to analyze and optimize industrial systems making use of the state space analysis. This approach has great potential to give very good results when it is properly implemented. In this article an improved version of the algorithm known as the *time line search* for optimizing the makespan of manufacturing models is presented. The algorithm uses a compact state space of coloured Petri net models in order to analyze the highest possible number of configurations.

Keywords: timed Petri nets, state space, optimization, simulation, manufacturing.

1. INTRODUCTION

In this article an improved version of the time line search algorithm is presented. The initial version of the TLS (Mujica and Piera 2010) was implemented as a heuristic to generate the state space using an algorithm in two phases (Mujica et al. 2010). The performance of the implementation yielded very good results when it was implemented in models that had state spaces small enough so they could be stored in the computer memory. Based on those results an algorithm that analyses and generates the compact timed state space (CTSS) in a better way was devised. The original algorithm, (which will be called from now on as the *old* algorithm) presented some shortages that caused some time penalties due to the node evaluation activity. The new implementations are devised with the purpose of overcoming those drawbacks giving as a result a better version of the time line search algorithm. The new algorithm has been tested with the industrial model of a CNC eye-glass machine (Mujica and Piera 2009b) which generates big state spaces when small workloads are simulated.

2. TIMED COLOURED PETRI NETS

Coloured Petri Nets (CPN) is a simple yet powerful modelling formalism which allows to properly model discrete-event dynamic systems which present a concurrent, asynchronous and parallel behaviour (Moore et al. 1996, Jensen 1997). CPN is a bipartite graph which is composed of two types of nodes: the place nodes and the transition nodes. Place nodes are

commonly used to model system resources or logic conditions, and transition nodes are associated to activities of the real system. The entities that flow in the model are known as tokens and they have attributes known as colours. The characteristics of the formalism allow modelling not only the dynamic behaviour of systems but also the information flow which is a key attribute in decision making.

In order to evaluate systems performance it is necessary to make an extension to the formalism attaching time stamps that determine the availability of tokens, a global clock that represents the model time and a time delay to transitions that model the time consumed by the activities. Formally they can be defined as follows.

Definition 1. Timed Coloured Petri Nets (TCPN)

$TCPN = (P, T, A, \Sigma, V, C, G, E, D, I)$ where

1. P is a finite set of places.
2. T is a finite set of transitions T such that $P \cap T = \emptyset$
3. $A \subseteq P \times T \cup T \times P$ is a set of directed arcs
4. Σ is a finite set of non-empty colour sets.
5. V is a finite set of typed variables such that $Type [V] \in \Sigma$ for all variables $V \in V$.
6. $C: P \rightarrow \Sigma$ is a colour set function assigning a colour set to each place.
7. $G: T \rightarrow EXPR$ is a guard function assigning a guard to each transition T such that $Type [G(T)] = Boolean$.
8. $E: A \rightarrow EXPR$ is an arc expression function assigning an arc expression to each arc a , such that:
 $Type [E(a)] = C(p)$
Where p is the place connected to the arc a
9. $D: T \rightarrow EXPR$ is a transition expression which assigns a delay to each transition (this delay is commonly represented with a '@' sign).
10. I is an initialization function assigning an initial timed marking to each place p such that:
 $Type [I(p)] = C(p)$

$EXPR$ denotes the expressions used by the inscription language, and $TYPE[e]$ denotes the type of an

expression $e \in \text{EXPR}$, i.e. the type of values obtained when evaluating e . The set of free variables in an expression e is denoted $\text{VAR}[e]$ and the type of a variable v is denoted $\text{TYPE}[v]$.

Type $[[p]] = C(p)$

In TCPN context the state of every model is also called the *timed marking* which is composed by the expressions together with their time stamps associated to each place p and they must be closed expressions i.e. they cannot have any free variables.

The markings are defined as follows:

Definition 2. The *timed marking* of a TCPN is a function $M^T: P \rightarrow \text{EXPR}$ such that $M^T(p) \in C(p)$. It maps each place p into a multi set of values $M^T(p)$ representing the timed marking of place p . The individual elements of the multi set are called *timed tokens* and the expressions contain also time stamps.

Definition 3. The *untimed marking* M^U of a TCPN model is a function $M^U: P \rightarrow \text{EXPR}$ that maps each place p into a multi set of values $M^U(p) \in C(p)$ representing the untimed marking of place p . In this case the expressions do not contain any time information.

In order to fire a transition, the number of tokens in the input place nodes (the directed arcs go from the places to the transition) must satisfy not only the arc inscriptions but also the restrictions imposed by the guard expressions. Only the tokens that have time stamp values less than or equal to the global clock participate in the transition enabling procedure.

When a transition occurs, the output tokens will have a time stamp Δt time units larger than the current global clock Gc which simulates time delay due to the execution of an activity. The time stamp calculated by formula (1) represents the earliest model time when the output tokens can be used again for a new transition firing, i.e. the token will not be available for Δt time units.

$$t_o = Gc + \Delta t \quad (1)$$

Where t_o is the time stamp value that must be attached to the output tokens when the transition firing takes place, Gc is the global clock of the model when the firing occurs and Δt is the time associated with the transition.

3. THE COMPACT TIMED STATE SPACE

The reachability graph is a directed graph which has been traditionally used by scientific community for the verification and analysis of behavioural properties of timed and non-timed CPN models (Christensen et al. 2001, Kristensen and Mailund 2002, Wolf 2007). The reachability graph is also known as the state space (SS)

because it generates and stores all the different reachable states from an initial one. The SS analysis can be performed with timed or untimed models to evaluate properties, such as liveness, boundedness and reachability of states among others (Jensen et al. 2001) to determine the behaviour of the modelled system.

In timed models each node of the SS represents a timed marking of the TCPN model. Some authors have developed different ways of representing the timed state space (TSS) basing their representations on different structural characteristics of the model (Chiola et al. 1997, Jensen et al. 2001). Those representations have been developed in order to reduce or delay as much as possible the state explosion problem (Valmari 1998) without losing the necessary analysis capabilities to verify model properties.

The following definitions are common to almost every state space representation.

Definition 4. Let \mathfrak{M}^T be the set of timed markings of a state space, and $M_i^T, M_k^T \in \mathfrak{M}^T$ be timed markings. A state M_k^T will be called *old node* if it is exactly the same (together with its time values) as one that have been previously generated in any other level of the SS, i.e. $M_k^T = M_i^T$

It is possible to reduce the amount of states to be analyzed when the symmetry of the colours in the tokens is exploited without taking into account the time stamps of the markings. Therefore it is possible to define a special kind of "repeated state", the symmetric old node or *S-old node*.

Definition 5. Let \mathfrak{M}^T be the set of timed markings of a state space. Let M_i^T and M_k^T be timed markings with their correspondent untimed markings M_i^U and M_k^U .

A marking M_i^T is an *S-old node* to another M_k^T marking when the following condition holds:

$$M_i^T, M_k^T \in \mathfrak{M}^T \wedge M_i^U = M_k^U$$

The representation using the S-old nodes is called the *compact timed state space* (CTSS) since it is possible to reduce the amount of space needed to store all the information generated in the state space (Mujica et al. 2010).

Other characteristics common to the CTSS and the TSS are:

- The root node represents the initial marking of the system.
- The successor or children nodes correspond to the new states or markings obtained once the enabled transitions have been fired.
- For each node in the tree as many successor nodes as enabling combination of tokens in the marking must be generated.

- Each node is connected with its successor nodes through directed arcs.
- The connecting arcs represent transition firings and they also contain the information concerning the transition fired and the tokens used in the firing.

4. NEW ALGORITHM TO GENERATE THE CTSS

The old *time line search* algorithm has been developed based on an algorithm in two steps (Mujica and Piera 2009a). It generates the CTSS making use of an incremental variable and the firing time of every marking is used as a key that matches the value of the incremental variable. The main elements needed for the algorithm are:

- A key that is assigned to each node in the CTSS. It will be used to determine the sequence of evaluation. The key takes into account the global clock when the firing takes place and the time stamps of the marking using the following formula:

Being $M_i^T \in \mathcal{M}^T$ a timed marking, T_i its correspondent time stamp list and Gc the global clock of the timed marking. The function $tline_k$ assigns the key in the following way.

$$tline_k : \mathbb{N}^k \times \mathbb{N} \rightarrow \mathbb{N}$$

$$(T_i, Gc) \mapsto \text{Max}\{\text{Min}\{T_{i_1}, \dots, T_{i_k}\}, Gc\} \quad (2)$$

- An incremental variable which determines the group of nodes that must be evaluated next. The nodes whose keys match the value of the variable will be the ones to be evaluated next.
- A list of node numbers with the sequence of evaluation based on the progress measure.

Making use of these elements the state space is generated and analyzed in two phases. The first phase generates the CTSS following the mentioned sequence and the second phase is used to improve the feasible path that has been found during the first phase.

Since the constructed CTSS is event-driven (Mujica and Piera 2009a), the markings to be evaluated are those that occur closest to the current global clock. Based upon this characteristic, the states with a firing time value greater than the current global value will not be evaluated until the progress value reaches the value of the firing time of the group. The latter situation caused in the old version of the TLS algorithm that the evaluation of some S-old nodes with good potential was delayed even if they would be fired with a global time earlier than the one from the already generated node. Figure 1 illustrates the situation when an S-old node with good potential was forced to a later evaluation.

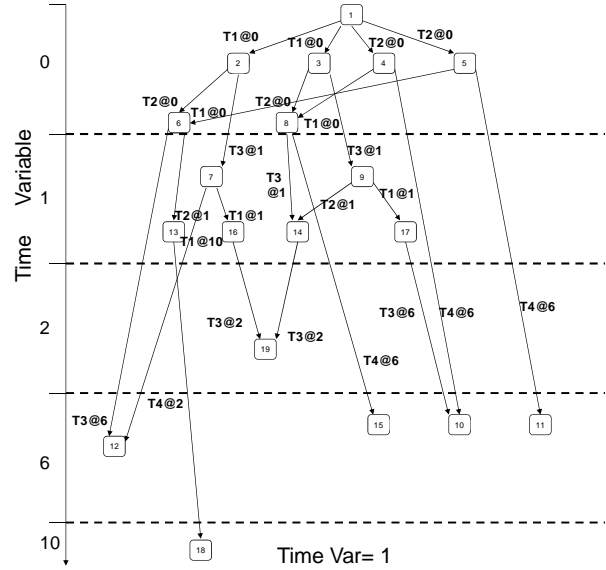


Figure 1: Generation of the SS (drawback)

This figure represents a compact timed state space. The nodes in the figure are being evaluated in order of appearance but the sequence of evaluation depends on the value of the time variable. The first group of nodes (nodes #1,#2,#3,#4,#5,#6 and #8) have been evaluated when the time variable had the value of 0 units. If we put focus on node #6, during its evaluation it generated nodes #12 and #13. In the case of node #12 the firing time was 6 time units while the node #13 was 1 time unit. In the case of node #2 it generated node #7 at a firing time of 1 unit. When all the nodes of the 0-time group were evaluated, the time variable (Time Var) headed to the next value (1 time unit). The second group of nodes was evaluated starting in node #7 which in this case generated the S-old node #12 and the node #16. If we put focus on the S-old node #12 it can be appreciated that it had been already generated by node #6 which produced it at a firing time of 6 units while the same S-old node can be generated by node #7 with a time value of 2 time units!

The latter situation illustrates the shortage that was incurred by the old algorithm. In order to follow the principle of the time line, node #12 should have been evaluated when the leading variable reached 2 time units instead of 6 which happened with the old approach. The first improvement to the new algorithm aims to overcome this shortage, and it is explained in the following subsection.

4.1. IMPROVEMENT A: UPDATING THE TIME LINE

A procedure which analyzes on-the-fly whether a node is better or not has been developed in order to avoid the shortage discussed in the previous sub-section.

The improved algorithm will use the same elements of the old TLS algorithm but during the evaluation it will

verify that the firing time of the correspondent nodes is certainly the smallest one.

The verification is performed every time a group of nodes has been evaluated. Therefore it is assured that the firing times used so far have the earliest values.

Figure 2 illustrates graphically the procedure performed by the algorithm to overcome the shortage.

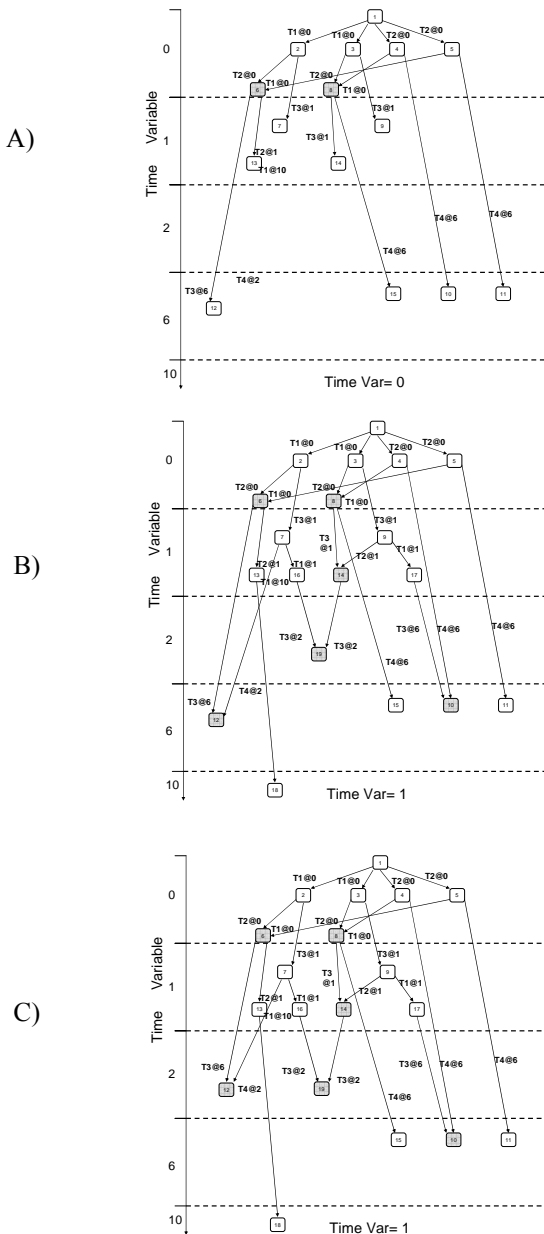


Figure 2: Generation of SS (overcoming the shortage)

Figure 2A represents the state space when the time variable has the value 0. During this instant of time the first nodes to be evaluated are nodes #1, #2, #3 #4 and #5 which generate nodes #6 to #15. During these evaluations nodes #6 and #8 are generated with a firing time of 0 units. The time variable value will not change until all the nodes that fall in the group of 0-firing-time value have been evaluated including nodes #6 and #8 (which were generated during the first evaluation). The

rest of the nodes take their place in the list waiting for their evaluation time to come.

When all the nodes of the 0-firing-time value group have been evaluated, the S-old nodes generated so far (the gray-shaded nodes in the figure) must be analyzed in order to determine if they have the smallest firing-time value for the successive evaluations. In Figure 2A the S-old nodes #6 and #8 are generated with the 0-firing-time value from their father nodes.

Figure 2B represents the next step in the evaluation procedure, the time variable heads to the new value (1 time unit). The group that matches the new time value is evaluated following the sequence #7,#9,#13 and node #14. The resulting state space from that evaluation sequence is illustrated in Figure 2B. It can be appreciated that four new S-old nodes have been generated during this evaluation, nodes #10, #12, #14 and #19.

In this figure, node #12 represents the S-old node that has been generated when node #6 was evaluated.

When all the nodes from the group of 1-firing-time value have been evaluated, the firing times of the new S-old nodes are verified in order to ensure that the firing time value is the smallest possible:

- The algorithm takes the list of the S-old nodes and makes a comparison between the firing times of the nodes that generate an S-old node.
- If it is found a firing time that is less than the original firing time, it will update the time values of the found node (time stamps and global time) and the information related to the father node that generated the node will be replaced with the one that produce the best time values. If the node has not been evaluated yet it will be removed from the group that originally belonged to and it will take its position into the new group. In the case of node #12 its father node will be changed from node #6 to node #7, Figure 2C.
- If the node has been already processed then it will update its time values and afterwards it will update the time values of the branch that hangs from the node.

Implementing this analysis strategy, it is ensured that the evaluated nodes in the generation phase of the CTSS use the smallest time values.

It can be argued that the branch-updating operation would need a lot of operations in order to perform the updating; but it is expected that the number of nodes to be updated are few since they have been evaluated accordingly to the progress measure, therefore the branch is not big compared to the total amount of nodes to be generated.

4.2. IMPROVEMENT B: Differentiating similar markings

A problem arises when there are groups of nodes that result difficult to distinguish based upon their

potential to improve the final node. Due to this condition it is not an easy task to decide which father node must be maintained for the subsequent evaluations. The latter situation shows up when two states are fired at the same global time and the operations take similar time; in that case the resulting markings will represent different logistic states but the time stamps of both markings will have similar values. In order to establish a difference between the conflicting nodes the following implementation was developed:

- If two states have the same firing time, the one with the earliest time stamp will be selected. This action assumes that this token will be ready earlier than the one from the other state.
- If the two characteristics hold (the firing time is the same and the earliest time stamp is the same) then it will be selected the second earliest time stamp and so on until it is found one time stamp that differentiate both markings.

The previous implementation assumes that all the tokens that compose the markings have the same probability of enabling a transition of the model. Figure 3 gives an example of the developed procedure to distinguish between two nodes.

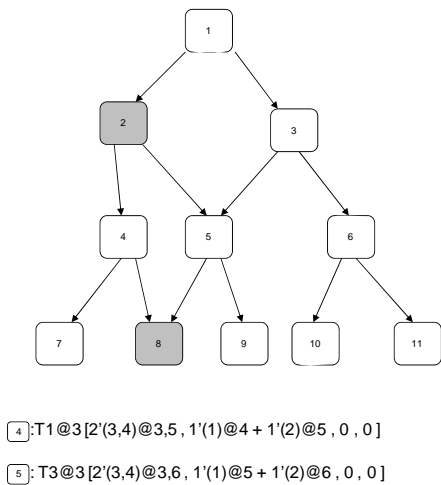


Figure 3: Differentiating two markings

In this example the S-old node #8 can be generated from nodes #4 and #5. Both nodes have the same firing time (3 time units) therefore they can only be differentiated evaluating their time stamps. Using the previous approach, the time stamp values that are taken into account for the decision have the values 3,4,5 for the S-old node that can be generated from node #4 and 3,5,6 for the S-old node that can be generated from node #5. The time stamp that differentiates both nodes is the second one thus in this example node #8 would be evaluated using the time values generated from node #4. If the S-old node #8 were originally generated from another node than node #4 then node #8 would be

updated with the new time values. If the node has not been evaluated yet then the updating process ends in that node; if there were a sub branch that hanged from the node then the time values of the complete branch must have be updated.

4.3. Algorithm of the improved time line algorithm

In this section the flowchart of the new time line algorithm is presented.

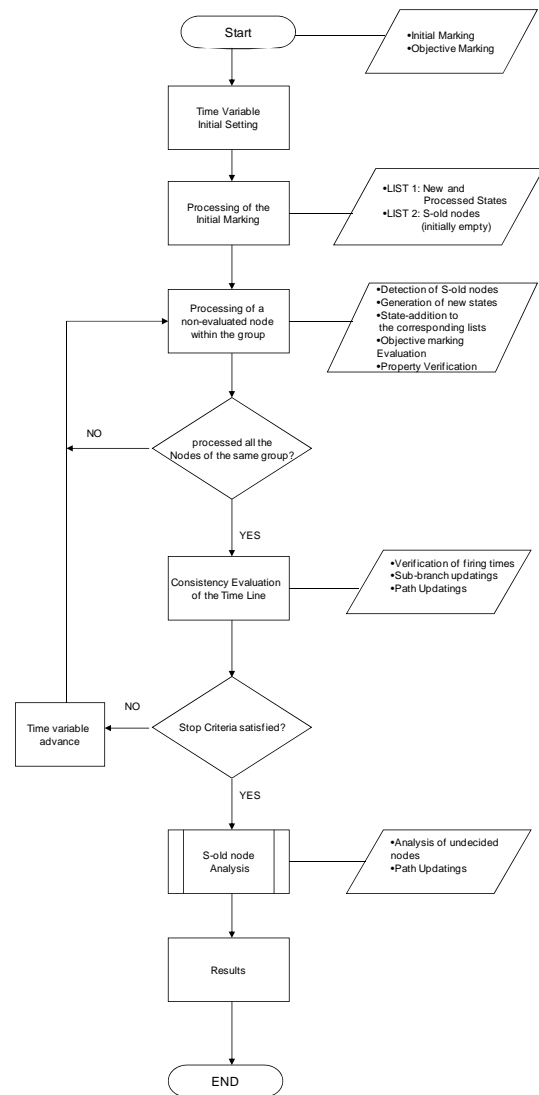


Figure 4: Improved time line search algorithm

This algorithm presents some advantages that come from the old algorithm and from the one in two-phases (Mujica and Piera 2009a):

- Good information management to store all the generated information
- Property verification can be performed prior to the optimization phase
- Good initial feasible path

The *consistency evaluation* step of the time line verifies if the S-old nodes to be evaluated certainly have the smallest possible firing time. If that is not the case the correspondent time value switching is performed and the sub-branches are updated if there are any.

The *S-old node analysis step* of Figure 4 evaluates in a second phase the generated S-old node list in order to optimize the feasible path whenever is needed. In this case it is fair to mention that the initial path can be used if a fast response of the algorithm is needed (Mujica and Piera 2010).

4.4. Experimental Results

In order to verify the improvements achieved with the new algorithm, it has been tested with small workloads of the CNC machine (Mujica and Piera 2009b). The proposed workloads are small enough to be stored in the computer memory therefore the achievements can be appreciated when a comparison is made testing three different algorithms: the two phase algorithm, the old TLS algorithm and the new one. Table 1 presents the results obtained from the performed optimizations.

Table 1: Obtained results from the benchmark

Final Marking of the Buckets Place node	Obtained Makespan DFS Approach	Obtained Makespan Old/TLS Approach	New TLS Implementation	Structural Information of the CTSS
Buckets: 2'(1,1,1,2,6,6,215,220)	1,039 sec.	635 sec.	590 sec.	No. Nodes: 19,232 No. Arcs: 59,610 No. OLD Nodes: 15,431 Levels: 53
Buckets: 3'(1,1,1,2,6,6,215,220)	1,335 sec.	960 sec.	833 sec.	No. Nodes: 172,242 No. Arcs: 765,177 No. OLD Nodes: 145,911 Levels: 84
Buckets: 2'(1,1,1,2,6,6,215,220) +1'(2,1,3,4,6,6,120,120)	955 sec.	821 sec.	720 sec.	No. Nodes: 562,799 No. Arcs: 1,800,951 No. OLD Nodes: 471,939 Levels: 81
Buckets: 2'(1,1,1,2,6,6,215,220) +1'(2,1,3,4,6,6,120,120) +1'(3,1,5,6,6,6,540,540)	1,913 sec. (500,000 nodes explored)	Unable to reach the Objective State	Unable to reach the Objective State	No. Nodes: --- No. Arcs: >2,541,330 No. OLD Nodes: >676,223 Levels: 105

It can be appreciated that making the implementations presented in this article the new TLS algorithm gives a better makespan. It can also be appreciated that in all the presented cases the new algorithm outperforms the previous ones.

5. CONCLUSIONS AND FUTURE WORK

The scheduling of industrial systems is a challenging problem due to the combinatorial nature present in most of them. The experiments with a timed coloured Petri net model of a CNC machine shows that the makespan is improved when the nodes are evaluated on a smallest-firing-time basis. The implementations presented in this article are crucial in order to develop an approach such as the time line search algorithm which uses the CTSS as the search space for performing the optimization of the makespan of industrial models.

In order to have the capacity to evaluate big state spaces a garbage collection algorithm for the CTSS is needed. This algorithm is being part of the current research of the authors.

REFERENCES

- Chiola, G., Dutheillet, C., Franceschinis, G., Haddad, S., 1997. A Symbolic Reachability Graph for Coloured Petri Nets. *Journal of Theoretical Computer Science*, vol.176 (1-2), pp. 39-65.
- Christensen, S., Jensen, K., Mailund, T., Kristensen, L.M., 2001. State Space Methods for Timed Coloured Petri Nets. *Proc. of 2nd International Colloquium on Petri Net Technologies for Modelling Communication Based Systems*, pp. 33-42, Berlin.
- Jensen, K., 1997. Coloured Petri Nets: Basic Concepts, Analysis Methods and Practical Use. vol. 1 Springer-Verlag, Berlin.
- Jensen K., T. Mailund, L.M. Kristensen, 2001. State Space Methods for Timed Coloured Petri Nets. *Proceedings of 2nd International Colloquium on Petri Net Technologies for Modelling Communication Based Systems*, Berlin
- Kristensen, L.M., Mailund, T., 2002. A Generalized Sweep-Line Method for Safety Properties. *FME*, Springer-Verlag, pp. 549-567, Berlin- Heidelberg.
- Moore, K.E., Gupta, S.M., 1996. Petri Net Models of Flexible and Automated Manufacturing Systems: A Survey. *International Journal of Production Research*, Vol. 34(11), pp. 3001-3035.
- Mujica, M.A., Piera M.A., (2009a). A Two Step Algorithm to Improve Systems Optimization based on the State Space Exploration for Timed Coloured Petri Net Models. *Proc. of the TiStoWorkshop*, Paris, France, pp.47-61.
- Mujica, M.A., Piera M.A., (2009b). Performance Optimization of a CNC Machine through exploration of Timed State Space. *Proc. of the International Modelling Multiconference (I3M)*, Tenerife, Spain, pp.20-25, 23-25 Sept.
- Mujica M.A., Piera M.A., Narciso M., 2010. Revisiting state space exploration of timed coloured petri net models to optimize manufacturing system's performance. *Simulation Modelling Practice and Theory*, Vol.18, 9, p.p. 1225-1241.
- Mujica M.A., Piera M.A., 2010. Time Line Search for the State Space-based Optimization Algorithm for Timed Coloured Petri Nets. *Proc. of MCPL IFAC'10*, Coimbra, Portugal, 8-10 Sep 2010.
- Valmari, A., 1998. The State Explosion Problem. *Lecture Notes in Computer Science*, vol. 1491, Springer-Verlag, London, pp. 429-528.
- Wolf, K., 2007. Generating Petri Net State Spaces. *Proc. of the 28th int. conf. on applications and theory of Petri nets and other models of concurrency*, pp.29-42, Springer.

SIMULATION AND MODEL CALIBRATION WITH SENSITIVITY ANALYSIS FOR THREAT DETECTION IN THE BRAIN

Keegan Lowenstein^(a), Brian Leventhal^(b), Kylie Drouin^(c), Robert Dowman^(c), Katie Fowler^(b), Sumona Mondal^(b)

^(a)Clarkson University, Department of Computer Science

^(b)Clarkson University, Department of Mathematics

^(c)Clarkson University, Department of Psychology

^(a) keegan.lowenstein@gmail.com, ^(b) leventhbc@clarkson.edu, kfowler@clarkson.edu, smondal@clarkson.edu,
^(c) rdowman@clarkson.edu, drouinke@clarkson.edu

ABSTRACT

In this study we use optimization techniques and sensitivity analyses to provide more rigorous, quantitative connectionist models of the functional interactions between the brain areas involved in detecting and orienting attention towards threat. A toolkit has been developed using flexible neural network modeling with automated parameter estimation. A sensitivity analysis is provided within the framework to identify significant model parameters and better understand model dependencies. These studies emphasize the importance of fitting the models to behavioral reaction time and brain activation data. They also show that the specific architecture of the model, and the numerical precision in the model parameters is important in determining an acceptable fit of experimental data, and that it is not the case that any model will work given the appropriate set of connection strength parameters.

Keywords: Neural Network, Least-Square Minimization, Analysis of Variance

1. INTRODUCTION

Our survival depends in part on being able to detect a threatening stimulus that occurs outside the focus of attention and to redirect attention towards the threat so it can be dealt with (Bishop 2008; Corbetta & Shulman 2008; Norman & Shallice 1989). The ability of threats to capture and hold attention can also have an impact on mental health. For example, hypervigilance towards threat is thought to play an important role in the etiology and maintenance of many anxiety disorders (Bishop 2008; MacLeod et al. 2004; Mogg & Bradley 2004). Yet despite its importance to survival and mental health there remain significant gaps in our understanding of this fundamental cognitive process (Corbetta & Shulman 2002; Öhman 2000; Phelps & LeDoux 2005).

Several studies have shown that threats are better at capturing and holding attention than non-threatening stimuli (Bar-Haim et al. 2007; Bishop 2008; Cisler et al. 2009; Öhman 2000, 2005). The enhanced ability of threats to capture attention is evidenced by faster

behavioral reaction times for threatening than non-threatening stimuli. Interestingly, the bias in attentional capture has been more difficult to demonstrate than the ability of threats to hold attention (see Bar-Haim et al. 2007; Bishop 2008; Cisler et al. 2009; Wyble et al. 2008). The neurophysiological underpinning of the attentional bias towards threat has also been the focus of a number of studies. Although several key brain areas have been implicated, such as the amygdala and insula, (Öhman 2000, 2005; Phelps & LeDoux 2005), a detailed understanding of how these areas interact with the brain areas involved in perception, response generation, and attention is lacking.

One approach to investigating the interactions between these brain areas is to combine experimental work with computational modeling. In this approach the computational models provide rigorous tests of hypotheses generated by the experimental work, and importantly, should provide novel predictions that can be tested in future work (e.g., Yeung et al. 2004). Little work has been done applying this approach to studying the attentional bias towards threats, and the few that have relied on qualitative fits of the experimental data rather than quantitative fits (Armony & LeDoux 2000; Dowman & ben-Avraham 2008; Wyble et al. 2008). In this study we describe our efforts at applying a connectionist model to quantitatively fit behavioral and brain activation data obtained in our studies of the attentional bias towards threats to the body (somatic threats).

An important focus of this work involved comparing model architectures simulating the different functional interactions between the brain areas thought to be involved in threat detection and orienting. To accomplish this we applied optimization techniques and sensitivity analyses to allow more rigorous, quantitative comparison of the different architectures and to explore the properties of the parameter space.

We explain the experimental setting in section 2 and follow in section 3 with the modeling. We present two different brain architectures and the model calibration results in section 4 with the follow-up sensitivity analysis in section 5. We end with a discussion of future directions.

2. EXPERIMENTATION

In our somatic threat studies (Dowman 2007a, 2007b) subjects performed two tasks: a visual color discrimination task and a somatic intensity discrimination task alternating in random order within the same session. The visual discrimination task consisted of indicating whether a red or a yellow LED was lit, and the somatosensory discrimination task consisted of indicating whether a high or low intensity electrical stimulus was delivered to the sural nerve at the ankle. A symbolic cue given at the beginning of each trial signaled which of the two tasks was forthcoming. The target stimulus was correctly cued on a randomly determined 75% of the trials (validly cued condition) and incorrectly cued on the remaining 25% of the trials (invalidly cued). The subject was instructed to focus his/her attention on the cued target stimulus, but to respond to the target regardless of whether or not it was correctly cued. Note that in the validly cued condition the target stimulus was presented within the subject's focus of attention, and in the invalidly cued condition the target was presented outside the focus of attention.

Two different sural nerve electrical stimulus intensities were used. In Dowman (2007a) both were strong and threatening (one at pain threshold and the other moderately painful), and in Dowman (2007b) both were weak and non-threatening. The attentional bias towards the somatic threat was evidenced in our experimental studies by the reaction time difference between the validly and invalidly cued conditions (validity effect) being smaller for the threatening somatic than the non-threatening somatic or visual target stimuli (Dowman & ben-Avraham 2008). (Reaction time differences due to stimulus intensity and sensory modality precluded a direct comparison between the threatening and non-threatening target stimuli). The smaller validity effect is consistent with the idea that threat is better able to capture and shift attention than non-threatening stimuli.

Electrophysiological measurements obtained during these experiments revealed three brain areas that appear to play an important role in detecting and orienting attention towards somatic threats. That is, for the threatening sural nerve target stimuli these brain areas exhibited greater activation when they were presented outside the focus of attention (invalidly cued) than when they were presented within the focus of attention (validly cued) (Dowman, 2007a, 2007b; Dowman & ben-Avraham 2008).

The electrophysiological data suggest that somatic threats are detected by somatic threat detectors located in the dorsal posterior insula. The threat detector activity is in turn monitored by the medial prefrontal cortex, which then signals the lateral prefrontal cortex to shift attention towards the threat (Dowman & ben-Avraham 2008). The greater activation of the somatic threat detectors in the invalidly than the validly cued condition suggests that the ability of somatic threats

to capture attention is greater when they are presented outside the focus of attention, and is consistent with the smaller reaction time validity effect for threatening targets observed in the reaction time data.

3. MODELING

We further examined the threat detection and orienting hypothesis using artificial neural network modeling (Dowman & ben-Avraham, 2008). The model was based on the work of J.D. Cohen and co-workers on response conflict (Botvinick et al. 2001; Yeung et al. 2004). The response conflict modeling studies, in conjunction with behavioral and functional imaging measurements, have provided convincing evidence that the medial prefrontal cortex is involved in monitoring situations that require a change in attentional control (e.g., response errors, response conflict, unattended threats) and signals the lateral prefrontal cortex to make the change. We modified the Yeung et al. (2004) model by replacing the response conflict component with threat detectors.

We compared several different model architectures in order to test the different physiologically feasible functional interactions between the brain areas responsible for detecting and orienting attention towards somatic threats. The architecture that provided the best qualitative fits of the reaction time and brain activation data is shown in Figure 1.

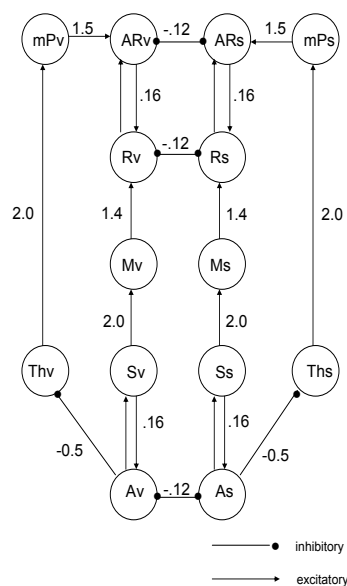


Figure 1: Artificial Neural Network Model of the Threat Detection and Orienting Hypothesis.

The model includes two stimulus-response pathways corresponding to the two tasks, where the Ss-Ms-Rs nodes and their connections simulate the somatosensory intensity discrimination task, and the Sv-Mv-Rv nodes and connections simulate the visual color discrimination task. The S and M nodes correspond to brain areas involved in early and late sensory processing, respectively, and the R nodes correspond to brain areas involved in the response. Note that the

model does not attempt to simulate the discrimination task performance in each stimulus modality, but rather only the reaction time differences for threatening vs. non-threatening target stimuli. The lateral prefrontal cortex areas controlling attention are simulated by attention nodes, one for each of the visual and somatosensory sensory nodes (Av and As, respectively), and one for each of the visual and somatosensory response nodes (ARv and ARs, respectively). The threat detectors for the visual and somatosensory systems are simulated by the Thv and Ths nodes, respectively, and the medial prefrontal cortex is simulated by the mPs and mPv nodes for the visual and somatosensory systems respectively.

The activation for each node was computed over 55 cycles. During the first 5 cycles external inputs were added to the As or Av nodes to simulate the allocation of attention as directed by the cue. The validly cued condition was simulated by adding an external input of 1.0 to the somatosensory attention node (As) and 0.0 to the visual sensory attention node (Av). The invalidly cued condition was simulated by adding an external input of 0.0 to the somatosensory attention node (As) and 1.0 to the visual sensory attention node (Av). During the stimulus cycles (cycles 6-10) external inputs were added to the somatosensory sensory node (Ss) to simulate the presentation of the somatosensory target stimulus. A threatening somatosensory stimulus was simulated by also adding external input to the somatosensory threat detector node (Ths) during the stimulus cycles. Due to the symmetry in the model, targets were only presented on the somatosensory side.

In the remaining 45 cycles, activation was allowed to spread through the model. For a model with M nodes, the activation levels of the nodes were computed using the following activation function:

$$A_i = \frac{1}{1+e^{(4-N_i)}}, \quad (1)$$

where A is a column vector containing M elements which represent the activation levels for all nodes in the model for the i th cycle. N_i was defined as:

$$N_i = N_{i-1} + (W * A_{i-1}) - (N_{i-1} * \delta), \quad (2)$$

N is also a column vector of size M , δ is a scalar decay constant, and W is an $M \times M$ weighted connection matrix. N_{i-1} represents the value of N during the cycle prior to i . Similarly, A_{i-1} represents the value of A during the preceding cycle. The product of the weighted connection matrix and the last known activation values ($W * A_{i-1}$) accounts for the input to each node due to its incoming connections. A_0 and N_0 were both null vectors initially.

The reaction time was defined as the cycle where the response node activation equaled 0.2. Reaction time was converted to milliseconds using the following function:

$$Reaction\ Time = 20c + 500, \quad (3)$$

where c is the cycle at which the activation level of the response node equals 0.20, 20 is an estimate of the number of milliseconds per cycle (based on the brain activation data), and 500 is a constant that accounts for perceptual and decision processes that are not accounted for by the neural network model (Dowman & ben-Avraham, 2008).

Dowman & ben-Avraham (2008) tested a number of different model architectures simulating different functional interactions between the brain areas thought to be involved in detecting and orienting attention towards somatic threats. As noted above, the architecture shown in Figure 1 demonstrated the best qualitative fits with the experimental reaction times and brain activations. Interestingly, this architecture led to the prediction that the attentional bias towards somatic threats will only be observed when the threat is presented outside the focus of attention (invalidly cued) and not when it is presented within the focus of attention (validly cued). As noted earlier, we could not directly test this hypothesis using the sural nerve stimuli because of the stimulus intensity confound. However, this prediction was recently confirmed in a study using pictures of somatic threats (Dowman et al. 2010), where the neutral and somatic threat target stimuli were matched for stimulus intensity and hue.

Dowman & ben-Avraham (2008) used the same set of connection strengths for all of the architectures (see Figure 1). These connection strengths were based on those published by Yeung et al. (2004), and were modified manually to provide acceptable qualitative fits of the data. The comparisons were straightforward given that many of the architectures could not simulate the direction of change in both the reaction time and brain activation data. It is possible of course, that had we chosen a different set of connection strength parameters that another architecture would have fit the data better. Therefore, a much better approach for model architecture comparison would be to use optimization techniques to find the best fit connection strength parameters for each of the architectures. Optimization techniques have the added advantage of allowing us to search for the best quantitative fit of the experimental data, something that is not feasible when the connection strengths are adjusted manually.

It is also important to perform sensitivity analyses to explore the parameter set. Of particular interest is determining whether the fit is dependent on a small range of connection strength values, or whether a wide range of combinations produce a good fit. Together, the optimization techniques and sensitivity analyses will provide a more rigorous quantitative comparison of the different architectures. Importantly they will allow us to determine if the architecture is important in fitting the data, or whether any architecture can be made to fit the data given the right set of connection strength parameters. Clearly the former outcome is of much greater interest in using the models to help determine the functional interactions between these brain areas.

4. MODEL CALIBRATION

Our previous effort involved manually adjusting connection strength parameters to provide acceptable fits, and then using these parameters to compare the different model architectures. This process was slow, tedious and at best led to rough qualitative fits of the data. More recently we developed a MATLAB[®]-based toolkit that provides automated calibration of connection strengths using optimization techniques (Lowenstein 2010). The optimization involved minimizing

$$J(W) = \frac{1}{P} \sum_{i=1}^P \left(\frac{e_i - m_i}{n_i} \right)^2, \quad (4)$$

where W is the matrix of connection strengths, P corresponds to the number of statistics that are being fit, e_i is the experimental value of a given statistic, m_i is the corresponding modeled value (which depends on W), and n_i is a normalization factor which ensures that each statistic contributes equally to the cost function. For our purposes, this normalization can be accomplished by setting n_i equal to e_i .

Within the toolkit, the Nelder-Mead simplex method is used for the minimization of Eq. (4) (Nelder & Mead 1965). Nelder-Mead has previously been shown to be effective in parameterizing connectionist models (Bogacz & Cohen 2004). A benefit of Nelder-Mead is that no gradient information is needed and minimization is based solely on function evaluations using a simplex that changes at each iteration based on the best point found. The function to be minimized can be non-differentiable, non-convex, or even discontinuous. This is an attractive feature of the toolkit because it allows for a general framework for the model calibration. Thus changes made to the simulation tool itself will have little, if any impact on the calibration process.

For each model, the five optimal parameters sought were *asr*, *at*, *in*, *smr*, and *tmr* (see Figure 2). Nelder-Mead is well known to be a local optimization method that can be highly dependent on an initial simplex. Consequently, multiple optimizations are usually required to better search the design space. Here 20 optimization runs were obtained for each model. To find the starting values for each Nelder-Mead optimization run, 1000 connection strength parameter sets were randomly chosen and the fits calculated. The set with the lowest $J(W)$ value was used as the starting values.

4.1 Numerical Results

First we determined whether the original architecture (Figure 1) provides a good, quantitative fit of the reaction time and brain activation data. As noted above, the stimulus intensity confound prevented us from directly comparing behavioral reaction times obtained for the non-threatening and threatening

somatic target stimuli. For modeling purposes we approximated the stimulus intensity confound-free invalidly cued threatening somatic target reaction time by multiplying the increase in the invalidly cued reaction time relative to the validly cued condition for the threatening somatic target (i.e., [invalidly cued – validly cued]/validly cued) to the validly cued non-threatening somatic target reaction time. Also, owing to the uncertain relationships between the scalp potentials used to measure the brain activations, the underlying brain activity, and the activation function in the model, we used the percent change in the electrophysiological measurements in the invalidly cued condition relative to the validly cued condition (i.e., {[invalidly cued – validly cued]/validly cued} * 100). As described in detail in Dowman & ben-Avraham (2008) the electrophysiological measures of the threat detector activation also include activation of the adjacent sensory area. Hence, this activity was modeled by combining the activations of the Ss and Ths nodes. The scalp potential measurements do not provide acceptable isolation of the medial and lateral prefrontal cortex activities, hence they were not included in the modeling studies.

The original model was able to provide excellent fits of the reaction time data (least-square error $\sim 1.0e-9$), but could not also fit the activation data (least-square error = 1.0). The failure of the original model to fit reaction time and activation data was because it could not account for the lack of change in the Ss node across the validly and invalidly cued conditions for the non-threatening somatic targets, as was originally pointed out by Dowman & ben-Avraham (2008). Rather, in that model the Ss node was smaller in the invalidly than the validly cued conditions. Our experimental studies have reported that brain areas involved in later sensory processing do show this attention effect (Dowman 2007a). Hence, we altered the model architecture by connecting the sensory attention nodes (As) to the middle layer (Ms), where the latter simulates the later stage of sensory processing. This model is shown in Figure 2 below. The result was a much better fit of the reaction time and activation data. Specifically, of the 20 best-fit parameter sets obtained from the Nelder-Mead algorithm, three gave a least-square fit to within the measurement error (i.e., least-square error $\sim 1.0e-5$, modeled reaction times within 5 milliseconds).

Dowman & ben-Avraham (2008) also compared architectures where the threat signal from the medial prefrontal cortex (mPs) to the response attention node (ARs) to one that has the threat signal going to both the sensory (As) and response attention nodes. This architecture is shown in Figure 3. The latter architecture is more consistent with the known anatomical connections between the medial and lateral prefrontal cortices (Miller & Cohen 2001). Dowman & ben-Avraham (2008) could not find any difference between the two architectures. We re-ran this simulation using the best-fit connection strength parameters for each architecture to see if this would make a difference, and

indeed it did: the architecture sending the threat signal to both the sensory and response attention nodes provided a noticeably better quantitative fit of the reaction time and activation data.

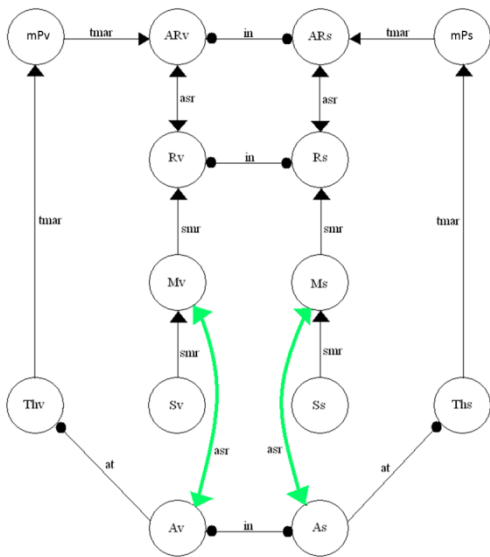


Figure 2: Modified Architecture to Sensory Attention Applied to the Late Sensory Processing Stage.

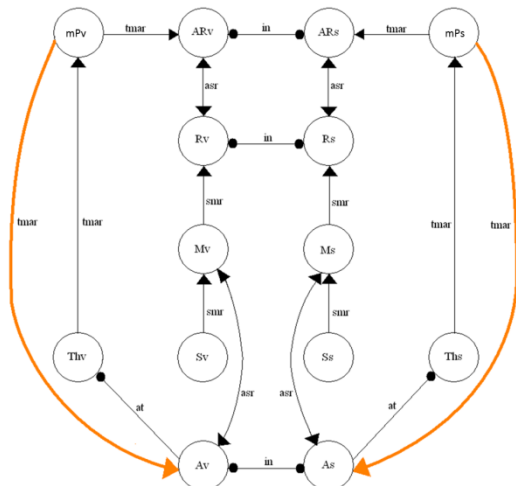


Figure 3: Architecture to Account for the Threat Going to Both the Sensory and Response Attention Nodes

For this model, 7 of the 20 optimization runs produced acceptable fits. The least-square errors for these parameters were an order of magnitude smaller than the previous model ($2.0702e-6$ vs. $6.8900e-5$, respectively) (Table 1), where the modeled reaction times were within 1 millisecond of the experimental data (Table 2) and the modeled percent change in brain activations equaled the experimental data.

We also examined one of the architectures that Dowman & ben-Avraham (2008) reported was unable to fit the reaction time and activation data. In this architecture the threat signal from the medial prefrontal cortex was sent to the sensory attention node instead of

the response attention node. The same result was obtained here when trying to quantitatively fit the reaction time and threat detector/sensory activation data (least-square error $\sim 1.0e1$).

Table 1: Best Fit Model Parameters and Least-Square Error for Architecture in Figure 3

Parameter	Optimal Value
<i>asr</i>	0.5640
<i>at</i>	-0.2119
<i>in</i>	-0.4327
<i>smr</i>	0.0256
<i>tmar</i>	0.2073
$J(w)$	$2.0702e-6$

Table 2: Comparison on Experimental and Model Reaction Times.

Experimental Condition	Reaction Time (milliseconds)
Valid Non-Threat Exp.	694.2
Valid Non-Threat Model	694.8
Invalid Non-Threat Exp.	826.1
Invalid Non-Threat Model	826.1
Valid Threat Exp.	694.2
Valid Threat Model	693.4
Invalid Threat Exp.	770.5
Invalid Threat Model	770.7

Interestingly, we consistently found that almost all of the architectures that we tested could provide excellent fits of the reaction time data when the best fit connection strengths were used (least-square errors $\leq 1.0e-8$). However, only the 2 architectures described here provided acceptable fits of the reaction time and brain activation data. Clearly, the brain activation data appears to provide critical constraints on the model.

5. SENSITIVITY ANALYSIS

The optimization results described above show that the model architecture is critical in obtaining fits of the reaction time and brain activation data. We next sought to determine the range of connection strength parameters that produced acceptable fits of the data. The mean \pm SD of the connection strength parameters for the architecture producing the best fit of the data (see Figure 3) is shown in Figure 4. The 7 optimization runs that produced acceptable fits (least-square errors = $2.1-7.5e-6$) were all tightly clustered around the same values. This was not the case for the 13 runs that

resulted in unacceptable fits (least-square error = 6.6-9.1e-3). Recall that the starting values for the Nelder-Mead optimization were determined by 1000 iterations of randomly selecting parameter values and computing the least-square error, and using the values that produced the best fit as the starting point in the optimization. This strategy reduces the probability that the optimization will always converge on the same local minimum. Hence, the tight coupling of acceptable fit parameters around the same values strongly suggests that range of best-fit connection strengths is very narrow.

This result was confirmed with a sensitivity analysis. When developing and studying mathematical models, it is common in practice to perform a sensitivity analysis to gain a deeper understanding of the model behavior, regardless of whether optimization is part of the design process. Analysis of variance (ANOVA) is one approach to studying the impact of changes in model parameters on model output. Specifically, ANOVA may reveal that some parameters have little effect on the overall model while others have a profound effect. In such cases, certain insignificant parameters can be set to a reasonable value while optimization can be done to fit the sensitive parameters and thereby reduce the problem size for the least-squares problem. This approach can also determine the specificity of the connection strength parameters. That is, does each connection have to be within a tight range for the model to work, or can changes in one connection be offset by a change(s) elsewhere in the model. This analysis can have a significant impact on interpreting the functional significance of the connection strength values. A benefit of ANOVA is that only sets of parameters and output are required as opposed to needing any derivative information.

ANOVA compares the ratio of the variation between sample means to the variation within each sample. The starting point for the procedure is to sort each parameter into groups. Analysis is done by considering changes in a response (here the least-squares error) as the group changes. Specifically, ANOVA is a hypothesis test with null hypothesis, $H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$, where k is the number of experimental groups. Each μ represents the mean of the single parameter, often called a factor, that is being found by the values in each experimental group. When rejecting the null hypothesis, the alternative hypothesis states that at least one mean is different from another, however it does not specify which one. The experimental groups are different equally spaced intervals for a single variable. The ANOVA examines the source of variation by finding the sum of squares of deviation from the mean for each of these groups. Using a statistical F test, the procedure is able to determine whether or not at least one mean is deviating from the others. The F test will produce a p-value; If this value is below a significance of 0.05 then the null hypothesis is rejected.

The model calibration experiments described above revealed that small changes in even the third decimal place of the connection strengths could strongly impact the overall least-square error and result in a poor model fit. For the sensitivity study presented here, tight bounds were placed on each parameter based on the best point found. We provide the details of the sensitivity analysis for the architecture in Figure 3, since it provided the overall best fit to the experimental data. For the sensitivity analysis, we chose the best of the 7 parameter sets that provided acceptable fits. The bounds are shown in the second and third columns of Table 3 below. For the analysis, each parameter was divided into 8 equally spaced groups and 500 values of each parameter were chosen via a Latin hypercube sampling, giving 2,500 parameter sets.

The Kolmogorov-Smirnov normality test was applied to the response variable (here the least-squares error) and we found the data was not generated from a normally distributed population. Thus, the non-parametric ANOVA method, the Kruskal-Wallis test, was used to calculate the corresponding p-values, shown in the last column of Table 3. Three parameters had values close to zero (*asr*, *at*, and *in*) indicating that they are significant in the modeling process. However, *smr* has a p-value of 0.052, which is very close to our level of significance 0.05, and we still can consider it to be a sensitive parameter. The parameter *tmar* was identified as insensitive and this was also evident in the values identified by the optimizer for the seven best parameter sets found during optimization. For the significant parameters, the standard deviation was always less than 0.03 but for *tmar* it was 0.07, indicating that a range of values would still lead to reasonable fitting to the data. These results intuitively make sense because *tmar* and *smr* are feed-forward connections, the rest are bidirectional. Clearly the positive feedback associated with a bidirectional connection will make it much more sensitive to change than a feed-forward connection. Furthermore, *at* is particularly sensitive since it is largely and only responsible for the brain activation fit.

Table 3: Lower and Upper Bounds for Parameter Study on Architecture in Figure 3

Parameter	Lower Bound	Upper Bound	p-value
<i>asr</i>	0.5	0.6	≈ 0
<i>at</i>	-0.3	-0.2	≈ 0
<i>in</i>	-0.5	-0.3	≈ 0
<i>smr</i>	0.02	0.03	0.052
<i>tmar</i>	0.2	0.29	0.493

To this end, an interval plot can provide a deeper insight into how the response values are distributed. We show these for *smr* and *in* in Figures 5 and 6. Here, the shaded, red dots correspond to points values of $J(W)$

while the average for each group is shown with a blue \oplus . For *smr*, the average values are relatively constant, with small fluctuations across the groups, but there is actually a broad range of response values within a group. For *in*, the average values change significantly across the groups, which is expected since *in* was identified as a sensitive parameter. It is important to note that of the 2,500 parameter sets randomly chosen from within the bounds given in Table 3 the least-square errors were greater than $1.0e-2$, which is four orders of magnitude greater than the best-fit values.

The sensitivity analysis for *in* produced an unexpected result. Sensitivity analysis is often used to guide the starting parameter set values for optimization. That is, the factor (parameter range) showing the best fit values is chosen in the optimization. However, sensitivity analysis suggests that the best fit point(s) would lie within factor 7 whereas the optimal value for *in* actually falls within factor 2 (-0.4327). The interval plots (Figure 6) reveal that the neighborhood around this point is considerably small thus requiring high accuracy in the optimization process. This implies that caution must be applied when interpreting the sensitivity analysis results for models with a very narrow range of best-fit values. An important clue that the sensitivity analysis results may not provide meaningful information on selecting starting points and/or bound constraints for the optimization was that even the best $J(W)$ values were four orders of magnitude greater than the optimal value.

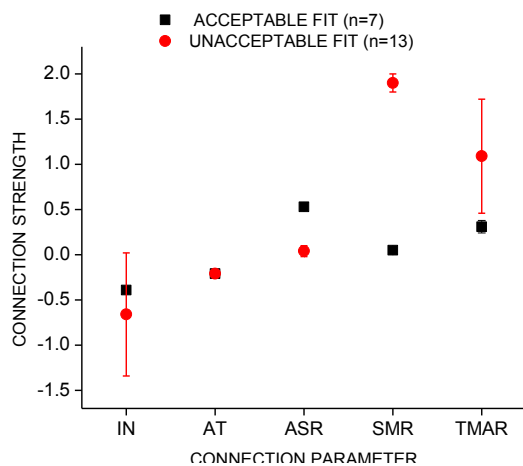


Figure 4: Mean \pm SD Connection Strength Parameters For the Best-Fit Architecture.

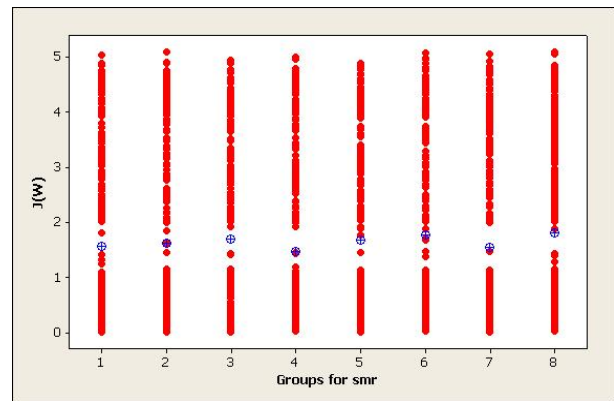


Figure 5: Range of Response Values Across Groups for *smr*

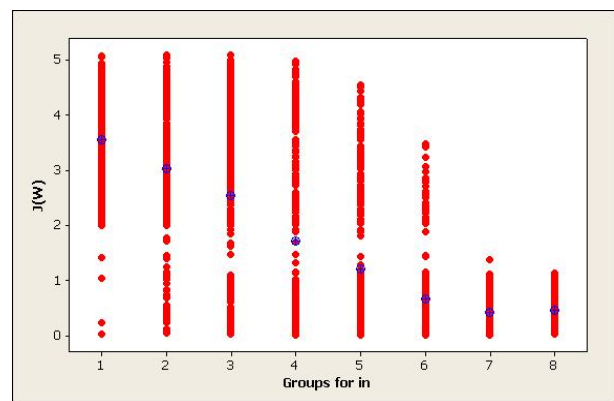


Figure 6: Range of Response Values across Groups for *in*

6. DISCUSSION

We have developed a flexible toolkit to develop and test artificial network models of the brain mechanisms for detecting and orienting attention towards threats to the body. Using optimization techniques were able to provide excellent quantitative fits of behavioral reaction time and brain activation data. These studies demonstrate that the model architecture is critical in producing good quantitative fits of the reaction time and brain activation data. Indeed, of the several models examined by Dowman & ben-Avraham (2008), only 2 provided acceptable fits. However, essentially all of the architectures could fit the reaction time data given optimal set of connect strength parameters. Clearly, including the brain activation data is critical in obtaining meaningful results in this type of work.

The sensitivity analysis suggests that only a very narrow range of connection strength parameters will fit the data. This implies that for fits of reaction time and brain activation at least, it is not case that the acceptable fits are an artifact of having a large number of parameters to fit the data. These results strongly suggest that the sensitivity analysis should not be used to determine the starting parameter values and ranges when the range of optimal values is very narrow. Future

work will include understanding the interaction of parameters.

Future experimental studies are aimed at testing predictions derived from the model. Of most interest is the prediction that the attentional bias towards threats are only seen when the threat is presented outside the focus of attention (Dowman et al. 2010). We are also performing modeling studies to determine if the model can simulate the attentional bias towards threats that have been reported using other experimental paradigms (e.g., Koster et al. 2007).

ACKNOWLEDGMENTS

This research work was supported by National Science Foundation UBM (Undergraduate Biology-Mathematics) Grant DBI-0926568 and the Clarkson University Honors Program.

REFERENCES

- Armony, J.L., & LeDoux, J.E. (2000) How danger is encoded: Toward a systems, cellular and computational understanding of cognitive-emotional interactions in fear. In: M.S. Gazzaniga (Ed.) *The New Cognitive Neurosciences*, Cambridge Massachusetts: The MIT Press, pp. 1067-1079.
- Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M.J. & van IJzendoorn, M.H. (2007) Threat-related attentional bias in anxious and non-anxious individuals: A meta analytic study. *Psychological Bulletin*, 133, 1-24.
- Bishop, S.J. (2008) Neural mechanisms underlying selective attention to threat. *Annals of the New York Academy of Sciences*, 1129, 141-152.
- Bogacz, R. & Cohen, J.D. (2004) Parameterization of connectionist models. *Behavior Research Methods, Instruments, & Computers*. 36, 732-741.
- Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., & Cohen, J.D. (2001) Conflict monitoring and cognitive control. *Psychological Review*, 108, 624-652.
- Cisler, J.M., Bacon, A.K. & Williams, N.L. (2009) Phenomenological characteristics of attentional biases towards threat: A critical review. *Cognitive Therapy and Research*, 33, 221-234.
- Corbetta, M. & Shulman, G.L. (2002) Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 31, 201-215.
- Corbetta, M., Patel, G., & Shulman, G.L. (2008) The reorienting system of the human brain: From environment to theory of mind. *Neuron*, 58, 306-324.
- Dowman, R. (2007a). Neural Mechanisms Of Detecting and Orienting Attention Towards Unattended Threatening Somatosensory Targets. I. Modality Effects. *Psychophysiology*, 44, 407-419.
- Dowman, R. (2007b). Neural Mechanisms Of Detecting and Orienting Attention Towards Unattended Threatening Somatosensory Targets. II. Intensity Effects. *Psychophysiology*, 44, 420-430.
- Dowman, R., Quin, J., & Sieg, E. (2010) Mechanisms Underlying the Capture of Attention by Somatic Threats. American Psychological Society 22nd Annual Meeting, Boston MA May 27-30.
- Dowman, R., & ben-Avraham, D. (2008) An artificial neural network model of orienting attention towards threatening somatosensory stimuli. *Psychophysiology*, 45, 229-239.
- Koster, E.H.W., Crombez, G., Verschuere, B., Vanvolsem, P., & De Houwer, J. (2007) A time-course analysis of attentional cueing by threatening scenes. *Experimental Psychology*, 54, 161-171.
- Lowenstein, K. (2010) *A Toolkit for Developing Neural Network Models of How the Brain Detects Threat*. Honors Thesis. Clarkson University.
- MacLeod, C., Campbell, L., Rutherford, E., & Wilson, E. (2004) The causal status of an anxiety-linked attentional and interpretive bias. In: *Cognition, Emotion and Psychopathology. Theoretical, Empirical and Clinical Directions* (J. Yiend, Ed.) New York: Cambridge University Press, 172-189.
- Miller, E.K., & Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. *Annual Reviews of Neuroscience*, 24, 167-202.
- Mogg, K., & Bradley, B.P. (2004). A cognitive-motivational perspective on the processing of threat information in anxiety. In: *Cognition, Emotion and Psychopathology. Theoretical, Empirical and Clinical Directions* (J. Yiend, Ed.) New York: Cambridge University Press, 68-85.
- Nelder, J.A., & Mead, R. (1965). A Simple Method for Function Minimization. *Computer Journal*, 7,308-313.
- Norman, D.A., & Shallice, T. (1986) Attention to action. In: R.J. Davison, G.E. Schwartz, & D. Shapiro (Eds.) *Consciousness and Self-Regulation. Advances in Research and Therapy*. Plenum Press: New York, pp. 1-18.
- Öhman, A. (2000) Fear and anxiety: evolutionary, cognitive, and clinical perspectives. In: M. Lewis & J.M. Haviland-Jones (Eds.) *Handbook of Emotions*, New York: Guilford Press, pp. 573-593.
- Ohman, A. (2005) The role of the amygdala in human fear: automatic detection of threat. *Psychoneuroendocrinology*, 30, 953-958.
- Phelps, E.A., & LeDoux, J.E. (2005) Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron*, 48, 175-187.
- Wyble, B., Sharma, D., & Bowman, H. (2008) Strategic regulation of cognitive control by emotional salience: A neural network model. *Cognition and Emotion*, 22, 1019-1051
- Yeung, N., Botvinick, M.W. & Cohen, J.D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review*, 111, 931-959.

SIMULATION AND MODELLING OF THE FLAT-BAND VOLTAGE FOR BELOW 200nm SOI DEVICES

C. Ravariu¹, F. Babarada¹

¹ Politehnica University of Bucharest, Faculty of Electronics and Telecommunications, Splaiul Independentei 313, 060042, Bucharest, Romania,
E-mail: cristir@mcma.pub.ro
and babflorin@yahoo.com

ABSTRACT

The nowadays SOI technologies frequently offer below 200nm, even up to tens of nanometre, for film on insulator. The flat-band voltage is one of main parameter in the electrical characterization of the SOI devices. The conventional models for this voltage were established for thicker structures, with 0.5...2 μm Si-film thickness and 2-3 μm buried oxide thickness. The electric charge from the buried oxide was ignored because the interesting conduction occurs in the vicinity with the front oxide. The pseudo-MOS transistor is a dedicated device for the electrical characterization of SOI wafers and works with a buried channel. The downscaling consequences of the SOI sizes on the flat-band voltage modelling were studied in this paper, with applications on the pseudo-MOS device.

KEY WORDS

Modelling, simulation, SOI, buried interface, devices

1. Introduction

The algorithm of modelling and simulations of physical processes spread over a large spectrum of applications, [1], [2]. The SOI structures represent a promising candidate for the nanodevice implementation, as classical [3] or novel architecture, [4]. This miniaturization is possible only if it is accompanied by proper models for the developed electronics devices. Some device parameters lost or change their classical meaning for new technologies. For example, the threshold voltage cannot be defined in a SON Transistor, [5], but it still arises in the transfer characteristics of a pseudo-MOS device, [6].

An excellent device for the electrical characterization of the SOI wafers is the pseudo-MOS transistor. This device represents an up-side-down SOI-MOSFET, with a back-gate command and electrical conduction through the film bottom, [7]. This paper comparatively presents some new analytical models, versus some simulation results, regarding the flat-band voltage of a pseudo-MOS transistor.

A new reference model will be confronted with others, analytically deduced in the next paragraph. A comparative

analysis between the classical model, new model and the simulation results are presented, as novelty.

2. The analytical model

The definition of the flat band voltage is related to the compensation of the positive electric charges from the buried insulator in order to bring the film surface potential to zero volts.

Adapted from the SOI-MOSFET to the pseudo-MOS transistor, this parameter is classically expressed as, [8]:

$$V_{FB} = -\frac{Q_{it}}{2C_s} - \frac{Q_{it}^2}{2q\epsilon_{Si}N_A} + \Phi_{MS} \quad (1)$$

where N_A [cm^{-3}] is the doping concentration in substrate, Q_{it} [e/cm^2] is the surface electric charge density, Φ_{MS} [V] is the metal-semiconductor work function, $C_s = \epsilon_{ox}/x_{ox}$ [F/cm^2] is the specific oxide capacitance, $\epsilon_{Si/ox}$ is the dielectric permittivity of Silicon, respectively oxide, x_{ox} is the BOX (Buried Oxide) thickness and $q=1,6 \times 10^{-19}\text{C}$ is the elementary electric charge. Firstly, in the analytical model $\Phi_{MS} = 0\text{V}$ will be assumed.

At the interface Si/SiO₂ usually exists a positive electric charge, Q_{it} , due to the presence of two kinds of charges: the interface charge Q_t representing the electrons trapped on the fast surface states and the fixed charge Q_f representing an excess of the ionic silicon solved in oxide and frozen at the Si/SiO₂ interface during the end of the annealing. The global charge is noted in this paper by $Q_{it} = Q_t + Q_f$.

Its sub-components can be: $Q_{it1} = 10^9 \div 10^{10} \text{e}/\text{cm}^2$, $Q_{it2} = 10^{10} \div 10^{11} \text{e}/\text{cm}^2$, $Q_{f1} = 10^{10} \text{e}/\text{cm}^2$, $Q_{f2} = 10^{12} \text{e}/\text{cm}^2 = 10^2 \text{e}/\text{nm}^2$; where the index "1" is used for the upper SOI interface and "2" for the bottom SOI interface. Frequently, the effect of Q_t on V_{FB} is neglected. For example, the contribution of Q_t charge is just 0,01V in V_{FB} value for the density of states $10^{10} \text{eV}^{-1} \text{cm}^{-2}$ in a bulk MOSFET with $N_A = 10^{15} \text{cm}^{-3}$ and $x_{ox} = 100\text{nm}$.

Therefore, in this paper we will work with the total positive electric charge Q_{it1} , Q_{it2} were considered, fig. 1. The flat-band voltage represents that gate voltage, which reduce to zero volts the potential in the Si-film, equivalent

with $V(0)=0$ and $E(0)=0$ in fig.1. By integration of the Poisson's equation, yields:

$$V_{FB} = \left(-\frac{Q_{it1}}{C_s} \right) + \left(\frac{qN_A}{2\epsilon_{Si}} x_d^2 - \frac{Q_{ox1} + Q_{ox2}}{\epsilon_{Si}} x_d \right) \quad (2)$$

The first parenthesis represents the potential drop over the buried oxide and the second parenthesis is the potential drop over substrate. The notations correspond to the fig. 1, where x_d is the width of the depleted region in substrate.

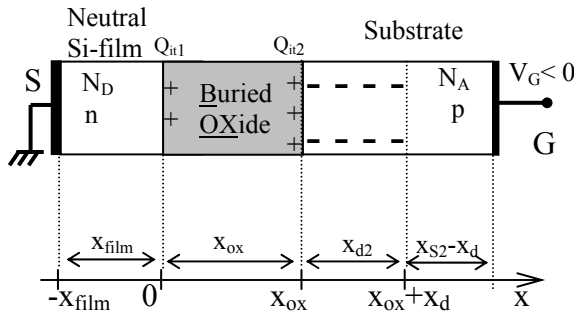


Fig. 1. The analyzed SOI structure with positive fixed charges in BOX and negative ions in substrate.

The limit conditions give the electric field:

$$\epsilon_{ox} E_{ox}(0) - \epsilon_{Si} E_{Si}(0) = Q_{it1} \Rightarrow E_{ox} = \frac{Q_{it1}}{\epsilon_{ox}} \quad (3)$$

$$\begin{aligned} \epsilon_{Si} E_{SB}(x_{ox}) - \epsilon_{ox} E_{ox}(x_{ox}) &= Q_{it2} \Rightarrow \\ E_{SB}(x_{ox}) &= \frac{Q_{it1} + Q_{it2}}{\epsilon_{Si}} \end{aligned} \quad (4)$$

From Gauss'law for $x \in (x_{ox}, x_{ox} + x_d)$ results:

$$E_{SB}(x) = -\frac{qN_A}{\epsilon_{Si}} \cdot (x - x_{ox}) + \frac{Q_{it1} + Q_{it2}}{\epsilon_{Si}} \quad (5)$$

From the limit conditions: $E_{SB}(x_{ox}+x_d)=0$ in (5), the x_d expression results:

$$x_d = \frac{Q_{it1} + Q_{it2}}{qN_A} \quad (6)$$

By replacing x_d from (6) in (2), the final expression of V_{FB} is obtained:

$$V_{FB} = -\frac{Q_{f1}}{C_s} - \frac{(Q_{it1} + Q_{it2})^2}{2\epsilon_{Si}qN_A} \quad (7)$$

The traditional Lim and Fossum model completely ignores the second interface, considering $Q_{it2}=0$ and also the depletion of the substrate, $x_{d2}=0$, [6]. Hence, V_{FB} is $-Q_{it1}/C_s$. This happened at the beginning of the SOI structures, with Micronics sizes. Obviously, “ Q_{it} ” is a model parameter in (1) and hasn't a physical meaning. It is named “the global charge from BOX”, but from eq. (1) it must be measured in $[C/cm^2]$, being a superficial electrical charge density.

A first disagreement between models (1) and (7) consists in different values of Q_{it} , and $Q_{it1}+Q_{it2}$. Considering additionally the electric charge from the second interface $Q_{it2} \neq 0$, from the limit conditions the accurate model is (7). The classical model (1) systematically under-evaluates the flat-band voltage value. Additionally, Q_{it} from the first and second ratio in eq. (1) hasn't quite the same values.

Another correction concerns the “2” factor that is missing in the model (7), first fraction at denominator, due to an average value assigned to Q_{it} .

In fact, either interface comprises fixed charges Q_f and interface trapped charges, Q_t . Consequently, the model (7) can be detailed as:

$$V_{FB} = -\frac{Q_{f1} + Q_{t1}}{C_s} - \frac{(Q_{f1} + Q_{t1} + Q_{f2} + Q_{t2})^2}{2\epsilon_{Si}qN_A} \quad (8)$$

where $Q_{t1, 2}$ respectively are the electric charge densities due to the electrons captured on the fast-states from the Si-film/BOX and BOX/Substrate interfaces. The correct value of the fixed charge density, $Q_{f1, 2}$ must be extracted from V_{FB} parameter after the $Q_{t1, 2}$ subtractions from V_{FB} in eq. (8). In the spirit of the classical model (1), model (8) could be corrected by averaging:

$$V_{FB} = -\frac{Q_{it1} / 2}{C_s} - \frac{((Q_{it1} + Q_{it2}) / 2)^2}{2q\epsilon_{Si}N_A} \quad (9)$$

In this way, two targets are reached: the problem of “2” missing at denominator of first ratio of model (7) is solved and a better agreement between simulations and the analytical model is obtained; the second ratio from model (7) overestimate the flat-band voltage, while the second ratio from model (9) brings the analytical values closer to the simulation results. The insight for Q_{it1} must be $Q_{f1}+Q_{t1}$ and for Q_{it2} must be $Q_{f2}+Q_{t2}$.

In the following simulations, a reverse way was investigated: the interface global charge densities were selected for different pseudo-MOS transistors and the flat-band voltage was extracted from definition. The scope was to accomplish the best fitting between V_{FB} simulated and V_{FB} analytical.

3. Simulations

The simulated SOI structure had: $\Phi_{MS1}=\Phi_{SM2}=-0,32V$, as is shown in fig. 2, $\Phi_{s-s}=0$, selecting the same p-type semiconductor as film and substrate with $N_A=5 \times 10^{15} \text{cm}^{-3}$.

The interface charge densities were chosen accordingly with some typical experimental results, [9]. The total front charge $Q_{it1}=5 \times 10^{10} \text{e/cm}^2$ placed at $x=0$ in fig.1 and the total bottom charge, $Q_{it2}=5 \times 10^{11} \text{e/cm}^2$ placed at $x=x_{ox}$ in fig.1, was selected. None charge in the front oxide was select, in order to be focused just on the buried oxide.

The simulations started with an SOI structure having $x_{film}=200\text{nm}$ and continue to 50nm , $x_{ox}=400\text{nm}$, $x_{SB}=750\text{nm}$. Figure 2 presents the simulation results for a pseudo-MOS with 200nm . In these conditions, a holes distribution still arises along the structure.

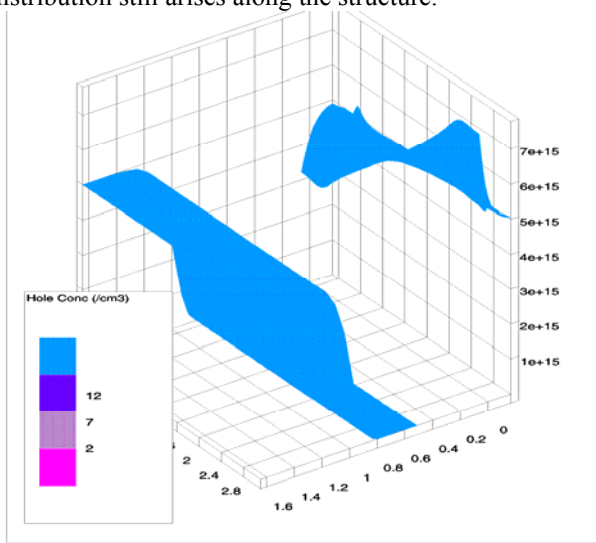


Fig.2. The holes concentration in 200nm Si-film structure.

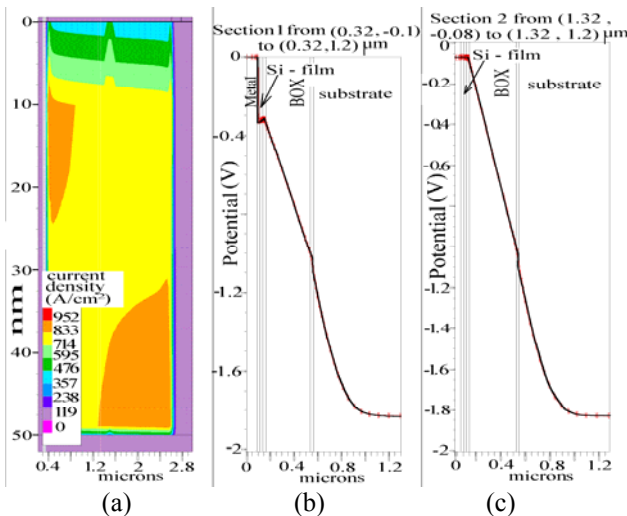


Fig. 3. (a) Detail of the current density in the 50nm SOI Si-film biased at $V_S=0V$, $V_D=0,5V$, $V_G=-1,5V$; (b) the potential distribution across the SOI near the source contact; (c) the potential distribution across the SOI through the middle.

In this case a global values $Q_{it1,2}$ values were established, as is modelled in (9).

Therefore, the discussion regarding the subtracting of the $Q_{it1,2}$ from the global density $Q_{it1,2}$ resting just at theoretical level. However the simulation can reveal some discrepancies between the classical model (1) and the proposed models (7) and (9).

Figure 3 a presents the current flow density through the Si-film in the case of biased structure at: $V_S=0V$, $V_D=0,5V$, $V_G=-1,5V$. The conduction prevails through the film bottom as is expected. Figures 3,b and c provide the adopted method for the extraction of the simulated flat-band voltage, V_{FBsim} . The gate voltage was increased in modulus till the film potential becomes zero. Then, the potential graph was translated with $-0,32V$ value, correcting the metal-semiconductor work function, in order to extract the simulated flat-band voltage, $V_{FBsim}=-1,92V$, affected just by the surface electric charges densities, $Q_{ox1,2}$.

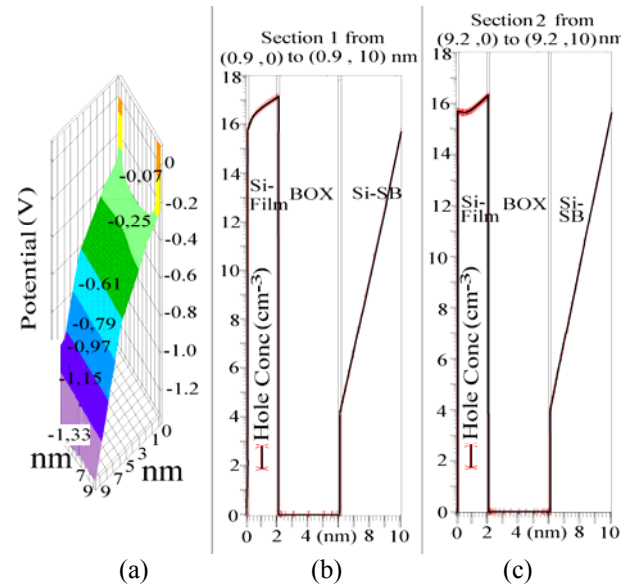


Fig.4. The 2nm Si-film structure at $V_S=0V$, $V_D=0.5V$, $V_G=-1.33V$: (a) potential distribution, (b) holes concentration near source, (c) holes concentration near drain.

Accordingly with fig. 4.a, at $V_G=-1,33V$ applied on the back gate, the hole concentration $p > 10^{16} \text{cm}^{-3} > N_A$. Hence a lower flat-band voltage is searching. Finally, $V_{FBsim} = -0.85V$ for previously mentioned $Q_{it1,2}$ values.

6. Discussions

For the investigated SOI structures, with $x_{film}=50\text{nm}$, $x_{ox}=400\text{nm}$ and SOI with $x_{film}=2\text{nm}$, $x_{ox}=4\text{nm}$, the same amount of positive interface charge density was used for both structures in order to provide a comparison. These

contributions were centralized in the table 1. Here can be compared some situations simulated and computed for different sizes. The notations are: $V_{FB\ sim}$ for the simulated value of V_{FB} , $V_{FB(1)}$ for the value deduced with the model (1), $V_{FB(7)}$ for the value deduced with the model (7), $V_{FB(11)}$ for the value deduced with the model (9).

x_{film} (nm)	x_{ox} (nm)	Q_{it1} (ecm^{-2})	Q_{it2} (ecm^{-2})	$V_{FB(1)}$ (V)	$V_{FB(7)}$ (V)	$V_{FB(11)}$ (V)	V_{FBsim} (V)
50	400	$2 \cdot 10^{10}$	$5 \cdot 10^{11}$	-0.191	-4.51	-1.31	-1.82
50	400	$5 \cdot 10^{10}$	$5 \cdot 10^{11}$	-0.501	-5.55	-1.84	-1.95
50	400	$2 \cdot 10^{10}$	10^{12}	-0.191	-16.3	-4.32	-3.92
2	4	$2 \cdot 10^{10}$	10^{12}	-0.007	-15.9	-4.14	-1.95
2	4	$5 \cdot 10^{10}$	10^{12}	-0.042	-16.9	-4.63	-2.11
2	4	$1 \cdot 10^{10}$	10^{11}	-0.002	-0.18	-0.056	-0.01

Table 1: Comparisons for structures with 50nm and 2nm film thickness.

Table 1 highlight that the analytical model (1) always underestimate the flat-band voltage, considering all the time just the first interface charge, Q_{it1} . The inclusion of the second interface charge Q_{it2} , with some correct limit conditions but in the depletion approximation, systematically overestimate the flat-band voltage accordingly the model (7). The best model at thick or thin sizes is the analytical model (9).

For thinner SOI films, a lower interface area results, within a lower quantity of negative ionic charge in substrate, in order to fulfill the flat-band conditions. In the ultra-thin SOI structures, the substrate isn't inverted, being in incipient depletion regime. Hence, the depletion approximation used in the deduction of the model (7), is more justified in ultra-thin SOI films than in thicker films. In SOI nanofilms the components Q_{f1} , Q_{f2} change the balance of importance on V_{FB} parameter. In thick BOX, some values like $Q_{it1}=10^{10}e/cm^2$, $Q_{it2}=10^{12}e/cm^2$, influence the potential of Si-film mainly via Q_{it1} parameter. In the case of some nanometres thickness of film and BOX and a device area= $10 \times 10 nm^2$ the prior charges densities are: $Q_{it1}=10^{12}e/cm^2=0.01$ electrons/device area – in probability terms quite negligible and $Q_{it2}=10^{12}e/cm^2=1$ electron/ device area – has a strong activity through a 2-5 nm thickness of buried oxide.

6. Conclusion

The classical model (1) induces high errors in thin SOI films because it entirely ignores the back charge interface that was true at thick BOX. The model (7) accurately deduced by Poisson equation integration systematically overestimated the flat-band voltage, because it use the depletion approximation in substrate and ignore the inversion layer arisen at the substrate surface. Simulator that proved accumulation of electrons at the substrate surface surprises the superposition. Therefore, the best model is (9), based on the averaging of the known interface charge, Q_{it1} and Q_{it2} .

In conclusion the charge placed at the bottom interface BOX/Substrate has a maximum influence on V_{FB} parameter extraction in the thin SOI films and it is partially annihilated by the negative inversion layer formed at the substrate surface during the device operating, in thicker SOI films.

Acknowledgment. This work is supported by the POSDRU /89/1.5/S/62557, PN2 no. **12095, 62063**.

References

- [1] Antonio Cimino, Francesco Longo, Giovanni Mirabelli, *A General Simulation Framework for Supply Chain Modeling: State of the Art and Case Study*, International Journal of Computer Science Issues, Volume 7, Issue 2, No 3, pp 1-9, March 2010.
- [2] Michael Affenzeller, Stefan Wagner, Stephan M. Winkler, *Effective allele preservation by offspring selection: an empirical study for the TSP*, International Journal of Simulation and Process Modelling 2010 - Vol. 6, No.1 pp. 29 - 39.
- [3] F. Babarada, et al., *MOSFET Modelling Including Second Order Effects for Distortion Analysis*, IASTED Proc., Applied Simulation and Modelling 2006, Rhodes, Greece, pp. 506-510.
- [4] C. Ravariu, et al., *Modelling and simulation of a nanostructure for a single electron technology implementation*, 5th International Mediterranean Modelling Multiconference, EMSS, Briatico, Italy, 16-19 Sept, 2008, ISBN 978-88-903724-0-7, pp.312-315.
- [5] J. Pretet, S. Monfray, S. Cristoloveanu and T. Skotnicki. Silicon-On-Nothing MOSFETs: performance, short channels effects and back gate coupling. IEEE Trans. Electron Devices, vol.51, no.2, pp. 240-245, 2004.
- [6] C. Ravariu, A. Rusu, *Parameters extraction from some experimental static characteristics of a pseudo-MOS transistor*, Bucharest, UPB Scientific Bulletin, ISSN 1454-234X, Series C, vol. 70, no. 1, pp. 29-34, 2008.
- [7] S. Sato, K. Komiya, N. Bresson, Y. Omura, S. Cristoloveanu, Possible influence of the Schottky contacts on the characteristics of ultrathin SOI pseudo-MOS transistors, IEEE Transactions on Electron Devices, vol 52, no.8, pp. 1807-1808, 2005.
- [8] H.K.Lim, J.G.Fossum, Threshold voltage of thin-film silicon-on-insulator (SOI) MOSFET's, *IEEE Trans. Electron. Devices*, vol. ed-30, no.10, October, 1993.
- [9] CEA-Leti R&D, 20nm Fully Depleted SOI process, EUROSIO, Newsletter, October, vol XXVI, 2010.

SIMULATION OF HUMAN BEHAVIOR IN SITUATION OF EMERGENCY

Samira Benkhedda, Fatima Bendella, Karima Belmabrouk

SIMPA Laboratory, Department of Computer Science,
University of Science and Technology of Oran, Algeria.

benkhedda.usto@gmail.com, bendella@univ-usto.dz, karimabelmabrouk@yahoo.fr

ABSTRACT

Multi-agent systems are used to study the complex natural and social phenomena. In this context, simulations are based on agent-oriented representation to describe the characteristics of the situation that all actions can be performed by actors. These tools have been used in situations of crises that are usually very difficult to manage, because of their complexity or the damage on help. In this paper. We propose a new approach for simulating multiple agents in a medical emergency based on practical reasoning of human and using the notion of simulation of complex systems.

Keywords: Multi agent systems, simulation, simulation of complex systems, medical emergency

1. INTRODUCTION

In a Multi-Agent System (MAS), agents interact and cooperate to perform a task or to achieve a common goal.

In this paper we present our simulation of a medical emergency that is 'heart attack', because if no action is taken immediately in emergency assistance and if do not act soon, the victim's life is in danger in a short term. We describe the features which we've featured in our simulation, by determining the scientific value of the project and identifying our agents, their roles and interactions between them.

2. CONCEPT OF MULTI-AGENTS SIMULATION

The multi-agent simulation is a simulation that employs the concept of multi-agents systems in the conceptualization, specification and implementation. A multi-agents system simulated living in a simulated environment; the multi-agents simulation directly represents the people, their behavior, their actions in the environment and their interactions. The multi-agents simulation is Interactions, Agents and Environment [1]. Ferber [2] notes that the multi-agents simulation allows the study of complex systems.

It represents the complexity of a phenomenon through the interaction of a single set of entities called agents. Each agent can:

- Communicate with other agents to exchange information.
 - Perceive and act on all or part of the environment
 - Apply knowledge, skills and other resources to perform their individual personal goals.
- The objective of the multi-agents simulation is :
- to Infer the nature of the functioning of the entities of a complex system.
 - System Analysis.

Simulator knows two phases of use :

- A research phase in which the simulator acts as an incubator model.
- An operational phase: once the model is validated, the simulator becomes a tool in the field.

2.1. Simulation of complex system

Simulate a complex system is to model its components, their behaviors and interactions between them and with their environment and then run the model obtained numerically. A feature of these systems is that one cannot predict the evolution of the modeled system without going through this phase of simulation. The approach "experimental" simulation makes it possible to reproduce and to observe complex phenomena (eg biological or social) in order to understand and anticipate their evolution [3]

3. MULTI AGENT SIMULATION SYSTEM FOR RAPIDLY DEVELOPING INFECTIOUS DISEASE MODELS IN DEVELOPING COUNTRIES (IDESS) [4]

3.1. Model Overview and context

IDESS (Infectious Disease Epidemic Simulation System) is a system able to build a simulation model to detect infectious diseases from the existing data in a geographical area.

IDESS is characterized by:

- The ability to create a simulation model for any location worldwide.

- Uses existing data to generate the simulation model.
- The ability to view the results in several ways.
 - From a software engineering perspective, it can change the behavior of agents and interactions with others.

3.2. IDESS Implementation

They have implemented the approach in a multi-agent system application-specific. At the heart of this simulation, they used the agent person (Person Agent PA) that acts like a normal person; the PA interacts with other PA in the model simulation and the environment in terms of Agent City (Town Agent TA). PA and TA have parameters and interactions that are associated with each agent. TA agents vary according to their consciences on the changes in the population of the city. The agent TA has connections with other agents TA and PA. The interaction between the TA may change if a containment strategy was invoked by isolating and TA officer concerned.

IDESS was used to quickly build a simulation model based on agents that can be used in the investigation of the spread of disease in a given geographic.

This approach is flexible in its ability to model any geographic location by processing the unpredictable nature of the spatial location where an outbreak of an infectious disease will occur.

This system is dynamic because you can edit and add new agents and information to the model.

4. AGENT-BASED SYSTEM FOR THE EVACUATION OF THE BUILDING IN CASE OF FIRE (IFI) [5]

3.1. Model Overview and context

This system is designed to simulate agent-based building evacuation in an emergency, and more specifically they simulated the building of the IFIs in the case of the fire.

In this work, they modeled and simulated fire. They built the model with the map of the IFI building (3 floors with 18 rooms, two staircases, and 3 outputs). The initial state is that all agents are in the rooms. When the program starts, an emergency occurs.

All agents try to leave the room. They determine the nearest door to get out of them. When moving they have to avoid obstacles. If they want to pass the door is full (many agents want to spend, there is more room) they have to go to another door if available. After leaving the room, the agents determine the direction to move. With agents who know the plan of the building, they measure the distance between them is the stairs. They take the stairs closest to you. The strategy is that all agents who know the plan always use the shortest path to the exit. In the event that there are many agents who pass the stairs, congestion occurs, some agents waiting at the tail end will choose another staircase. If agents do not know

the plan and around them there is no agent who knows the plan, they move according to two strategies: always running to the west or east still running, if they meet they descend the stairs, for if cons in their neighborhood, an agent knows the plan, then they follow it.

When agents encounter a fire, they change their direction, increasing their speed and find another way out. The strategy to choose another way is to take the stairs or the nearest fire is not yet declared. Other agents that meet these agents will learn the information lights and follow them.

4.2. IFI Implementation

This work aims to build a simulation that is closest to reality as possible so the authors used the BDI architecture (Belief, Desire and intention) because pedestrians must communicate and reason to find the way to the exit, and then Agents are getting smarter.

In this model, the environment is the building of the IFIs and the fires are considered an agent. There are two types of agents:

- 1) Agents who know the plan of the building.
- 2) Workers who do not know the plan of the building.

Each agent type has a different algorithm, but they have the same process to observe, to update the world (the state building and state of the stairs that wants to spend). If an agent is affected by the fire that minimal heat, it will be hurt, and the heat of fire is greatest, the agent will die.

The application is flexible; we can change the building plan by modifying the xml file. You can also change the number of agent in each room, the number and position of obstacles, fire, and the percentage of agent type and percentage rate of agent.

The reactions of the agents respond to environmental changes, they are reasonable, they can avoid congestion, avoid fire and teach others. The reasoning of the agents is following the reasoning of real people.

The application meets one important limitation: The information exchanged between agents are small, there is little communication. If agents communicate and exchange much more information, the model becomes more real.

5. MULTI-AGENTS SIMULATION IN THE CASE OF HEART ATTACK

5.1 The heart attack [6]

A heart attack is a serious problem caused by a blood clot in a coronary artery or one of the chambers of the heart. Cardiac arrest is the sudden stoppage of the heart that pumps more blood.

5.2. Signs of heart attack [6]

Defibrillation and cardiopulmonary resuscitation (CPR) quickly can save lives. Prompt treatment to break up clots can greatly increase the chances of survival of the person who suffers of heart attack. Since prompt

treatment can make a difference, it is important to know the early signs of heart attack.

In case of heart attack, you may experience one or more of the following:

- Discomfort in the center of the chest that lasts more than 5 minutes or comes and goes. It takes the form of an uncomfortable pressure, tightness, a feeling of heaviness or pain.
- Discomfort in other parts of the body, such as pain or discomfort in one or both arms, neck, at the jaw or stomach.
- Shortness of breath is often accompanied by chest discomfort but can occur before the discomfort. Among other signs, call the cold sweats, nausea and a sensation of floating.

Women who are a heart attack may not experience the usual symptoms, which can delay their care. Among the symptoms: an atypical or unusual pain in the chest, abdominal pain, nausea, shortness of breath and unexplained fatigue.

5.3. Mobile-Learning [7]

Mobile-learning is a logical extension of e-learning. In this sense it refers to the provision of courses or learning objects through mobile devices such as PocketPC, cell phones or the PalmPilot Users will have lessons in reduced format, but the main advantage of such a solution is accessible at any time of day and from any location with a network is nearby. Today the knowledge of students in mobile technology (accessibility, ease of use, speed of adaptation) makes the m-learning possible.

5.4. Complex SIMUL

The main objective of our project (Complex SIMUL) is to simulate an emergency based on the simulation of complex systems. Complex SIMUL includes reactive agents that will work in a complementary way and deliberative agents (BDI type) that will incorporate the concept of practical reasoning in humans. Our system includes three different types of agents (figure1):

5.4.1. Victim Agent

It is a passive actor, it can take three different cases (a patient with a blue face, red or a deceased victim), it is reactive, and it reacts to the actions of the agent rescuer (move, help him ...).

5.4.2. Environment Agent

it is the agent that identifies the location of the accident, it is reactive type and it is dynamic (the road, the greenery ...).

5.4.3. Succourer Agent

It is the active player of our emergency; there are three types of agents:

- The emergency agent and the doctor agent: aid workers are expert-like agent, they are reactive agents.
- The no-expert agent: an agent's type BDI (Belief, Desire and intention), the agent has beliefs about the world in which it operates, it must meet the desires by making intentions).

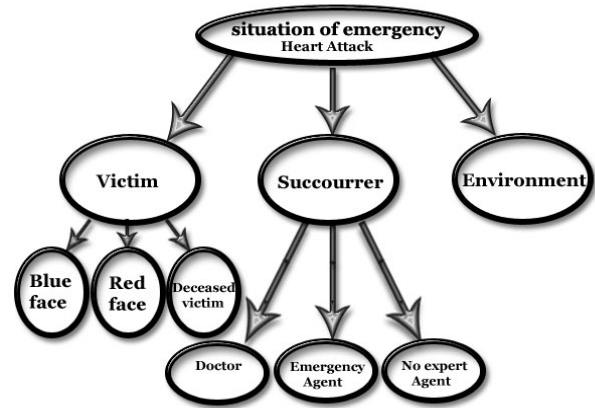


Figure1: Architecture of Complex SIMUL.

5.4.4. Why a BDI agent type?

This architecture is the most valued. In theoretical point of view, a BDI agent can perform any type of task. The architecture allows the agents to solve complex problems.

In this architecture, agents have a feature that allows evaluating the utility of each action. Contrary to the cognitive architecture of agents, the agents ask random actions when they were not able to achieve their goals. BDI agents can determine the action (or actions) to be performed to get as close as possible goal. This means that when an agent can achieve its goal by setting an action, then it will select the action as close as possible to the goal or action which will achieve this goal as quickly.

5.4.5. The operation of our Succourer Agent like a BDI agent (figure2)

Our agent includes an event queue (the actions of the agent to whom he rescued victim) by storing the internal events of the system, the beliefs (knowledge of the agent), a library of plans (know-how of the agent), a battery of desires (goals of the agent) and a stack of intentions (instantiated plans to achieve goals). The BDI interpreter cycle begins by updating the event queue and beliefs of the agent. It then activates new desires by selecting the plans of the library, so our agent has the opportunity to decide by running the first selected action stack of intentions, and so on.

This agent must:

1. Observe the environment agent and the victim agent for the possible risks.
2. Report a warning (phone call).
3. Select a plan (and it depends on his desires).

4. Select a plan by running the corresponding plan of action (the ABC of first aid).

Our no-expert agent is going to learn during the simulation using his mobile phone through the M-Learning. During the M-Learning, our BDI agent will learn along the ABC of first aid.

We used the concept of priority for emergency agents, the doctor agent has the highest priority then the emergency agent but no-expert agent will take the lowest value of priority zero boots (0) then after every successful in emergency and in the same case this value will increment to 1 and so on.

The priority of the no-expert agent is less than or equal to the priority of the emergency agent but the last two priors are always less than that of the doctor agent.

The BDI agent will protect the victim agent and his self from danger (the middle of a road) then it will use the ABC of first aid.



CARDIOPULMONARY RESUSCITATION (CPR)

The CPR is a combination of chest compressions with rescue breathing done on the victims is believed to be in cardiac arrest. When cardiac arrest occurs, the heart stops pumping blood. CPR can support a small amount of blood flow to the heart and brain to "buy time" to restore normal heart function [9].

8. CONCLUSION

In this paper, We have tried to show the methodology for the design of our simulation, we have implemented a multi-agent simulation and an application of M-Learning which is flexible and intuitive enough to allow the rescuer to the first aid to save the life of a victim of a heart attack because the speed of treatment is very important in this case.

We currently implement the agents of our simulation and integrate them into a learning application.

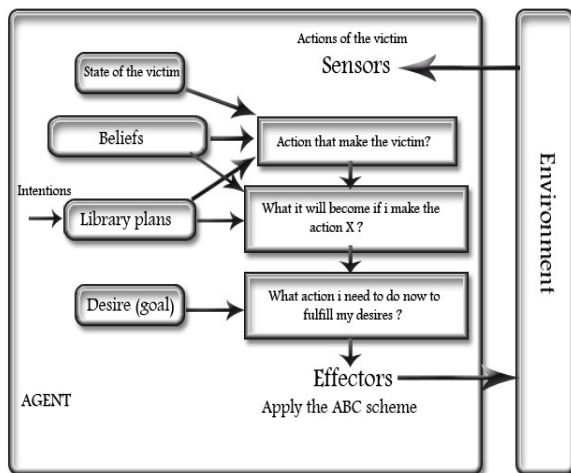


Figure2: The BDI agent of Complex SIMUL.

6. THE ABC OF FIRST AID

The priorities of first aid are...

A AIRWAY

B BREATHING

C CIRCULATION (and bleeding)

We will explain only the first operation and the Cardiopulmonary resuscitation CPR

Airway

The airway of an unconscious person may be narrowed or blocked, making breathing difficult and noisy or impossible. This happens when the tongue drops back and blocks the throat. Lifting the chin and tilting the head back lifts the tongue away from the entrance to the air passage. Place two fingers under the point of the person's chin and lift the jaw, while placing your other hand on the forehead and tilting the head well back. If you think the neck may be injured, tilt the head very carefully, just enough to open the airway [9].

REFERENCES

- [1] Wafa Ketata, Wided Lejouad and Chaari. 2007. *Une Ontologie pour la réutilisation des Interactions dans un Système Multi-Agents*. JFO, 18-20 octobre 2007, Sousse, Tunisie.
- [2] J. Ferber, 1999. *Multi-agent systems*. Reading MA : Addison-Wesley.
- [3] Alain Boucher, NGUYEN Nhu Van, 2007. *L'interaction dans simulation multi agents*. Hanoi.
- [4] Dean Yergens, Julie Hiner, Jörg Denzinger et Tom Noseworthy. *Multi Agent Simulation System for Rapidly Developing Infectious Disease Models in Developing Countries*. The Second International Workshop on Multi-Agent Systems for Medicine, Computational Biology, and Bioinformatics.
- [5] NGUYEN Thi THUY Nga et HO Tuong Vinh, Juillet 2009. *SMA pour la simulation à base d'agents d'évacuation de bâtiment dans les cas d'urgence*. Hanoi,.
- [6] McKesson Provider Technologies, 2006. *La crise cardiaque : Signes annonciateurs précoces*.
- [7] J, MUESSER., 2005. *Conception et réalisation d'un objet pédagogique pour périphérique mobile*. projet professionnel IUP PSM 3ème année.

- [8] B.David, juillet 2006. *Mobile-learning pour des activités professionnelles*. cours architecture informatique, école d'été du CNRS « EIAH ».
- [9] <http://tilz.tearfund.org/Publications/Footsteps+11-20/Footsteps+18/The+ABC+of+first+aid.htm>.

AUTHORS BIOGRAPHY

Dr. Fatima Bendella is a Senior Lecture in the Department of Computer science in USTO; she got an engineer diploma in computer science at the University of Oran in 1988, a magister of USTO in 1995 and a doctorate in 2005. She directs several theses of magister and doctorate in the application of multi-agent systems in software development. She is responsible of many research projects and a national research project (PNR) approved in May 2011.

Simulation, Optimisation and Design a Platform for in-vivo Electrophysiological Signals Processing

F. Babarada¹, C. Ravariu¹, J. Arhip²

¹ University Politehnica of Bucharest, Faculty Electronics Telecommunications and Information Technology, DCAE, ERG, Bucharest, Romania

² S.C. Seletron Software si Automatizari SRL, Bucharest, Romania

Abstract— The paper presents a hardware solution of the in vivo electrophysiological signals continuous processing and using a data vector acquisition on PC. The originality of the paper comes from some blocks proposal, which selective amplify the biosignals. One of the major problems in the electrophysiological monitoring is the difficulty to record the weak signals from deep organs that are covered by noise and the cardiac or muscular strong signals. An automatic gain control block is used, so that the high power skin signals are less amplified than the low components. The analog processing block is based on a dynamic range compressor, containing the automatic gain control block. The following block is a clipper since to capture all the transitions that escape from the dynamic range compressor. At clipper output a lowpass filter is connected since to abruptly cut the high frequencies. The data vector recording is performing by strong internal resources microcontroller including ten bits A/D conversion port. Design of analogical blocks is assisted by electronics circuit's simulation and optimization.

Keywords— **Simulation, Design, Health care, Compressor technique, Electrophysiological signal.**

I. INTRODUCTION

The common techniques from the human electrophysiology are non-invasive, with electrodes placed on the tissue (e.g. metallic electrodes in contact with gastric mucosa in electro-gastro-graphy [1] or at cutaneous level in the classical electro-cardio-graphy ECG [2]).

Therefore, a main problem arises when the electrodes are placed onto skin: the useful weak signals are buried in high level parasitic signals. The non-invasive electrophysiological methods suffer from noise, collected by the surface electrodes. There are many types of noise to be considered:

- *Inherent noise in electronics equipment*: It is generated by all electronics equipment and can't be eliminated. It is only reduced by high quality components using. It has a frequency range: 0 – several thousand Hz, [3].

- *Ambient noise*: The cause is the electromagnetic radiation, with possible sources: radio transmission, electrical wires, fluorescent lights. It has a dominant

frequency of 60Hz and amplitude of 1 – 3 x EMG signal, [4].

- *Motion artefact*: It has two main sources: electrode/skin interface and electrode/cable, having a frequency range of 0 – 20Hz. It is reducible by a proper circuitry and set-up.

- *Inherent instability of signal*: All electronics equipments generate noise and the amplitude is somewhat randomized, being in correlation with the discrete nature of the matter. This noise has a frequency range of 0 – 20Hz and cannot be removed.

In this situation, a gastric signal for instance, recorded at skin level, is hundreds times lower than the parasitic signals.

The most accurate solution is the invasive one, straight to the target organ, using microelectrodes, [5]. Unfortunately, the majority of organs are inaccessible without a surgical act which adds two great disadvantages: the health-state in danger and high costs.

This paper presents an analog processing and digital recording system for low power electrophysiological signals, with the possibility to use them in medical applications like ECG, EGG, EMG etc. For low contact electrodes area, the noise introduced by the electrodes begins to be most significant. As the results of modelling of the ensemble source-electrode, it is recommended for the amplifier to be implemented by a low noise and distortions, transimpedance amplifier stage followed by one low passing filter. For very low electrophysiological signals it is necessary a differential amplifier because it has a high common mode rejection of parasitic signals characteristic [6].

II. THE ELECTROPHYSIOLOGICAL SIGNALS PROCESSING

As in the case of many concepts from engineering, automatic gain control was also discovered by natural selection.

A. The automatic gain control

Automatic gain control (AGC) is an adaptive system found in many electronic devices. The average output signal level is feedback to adjust the gain to an appropriate level

for a range of input signal levels. AGC algorithms often use a proportional-integral-differential controller.

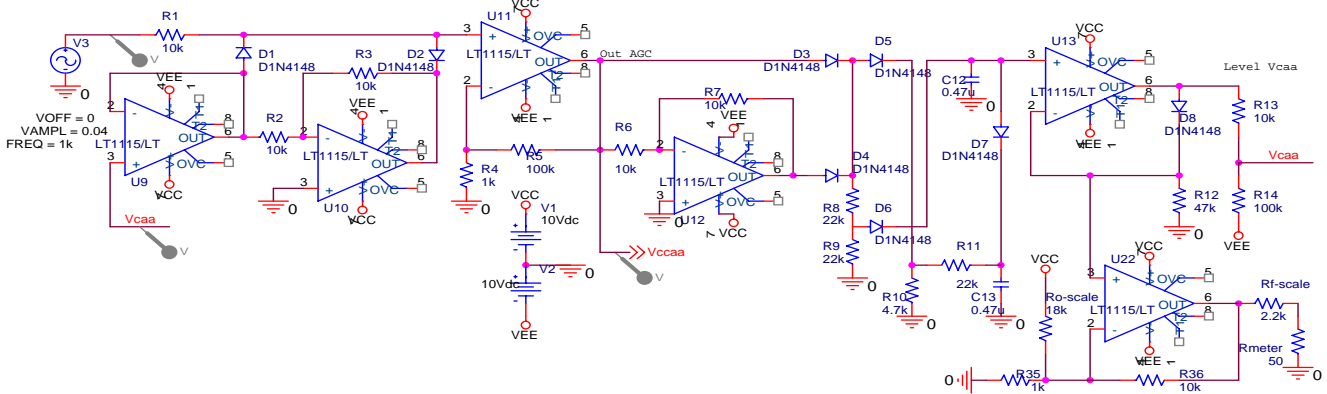


Fig. 1 The automatic gain control

The basic components of compressor are the U9, U10 integrated circuits, which biases the D1, D2 diodes, fig. 1, at their I-V curve knee. The input voltage is in the range of 10 to 300mVp and the output voltage is in the range of 5 to 10mVp. The voltage command of AGC is in the range of 300 to 600mVdc. The resistor R3 allows the circuit to be balanced and adjust the output voltage so it does not produce distortion in the output when gain reduction is active. In order to provide the voltage command of AGC (Vcaa) we choose a feedback configuration design. This design contain the amplifier, composed by U11, R4, R5 with the amplification around 101, the full wave rectifier, composed by D3, D4, R6, R7, U12 which bring the signal to the absolute value and the positive voltage detection realized with D6, C12 connected trough half voltage divider R8, R9.

The voltage over the condenser C12 is exactly the voltage command of the automatic control amplifier Vcaa. Discharging of the condenser C12 is made through the diode D7. This diode is opposite polarized by a voltage greater than Vcaa, respectively the voltage produced by diode D5, which is not reduced by half and loaded at the absolute value the condenser C13 through resistances R10 and R11. At reduction of the input signal amplitude the voltage Vcaa remains constant until C13 is discharged by R11 and R10. Thus at transient simulation at 1kHz, the amplification remains constant 5ms and then increase in time of 15ms. To reduce the temperature dependence of the automatic gain control we used the IC stage realized with U13, the diode D8 and the resistance R12, achieving the circuit from fig. 1. The voltage command Vcaa can be adjusted from resistive divider composed with resistances R13 and R14.

In order to display the output level of the dynamic range compressor signal we added the stage composed with operational amplifier U22 that adapt the connection and the adjusting of zero and end of scale for a linear display with LEDs bargraph.

The fig. 2 presents the beginning action of the automatic gain control for the input amplitude signal 20mV.

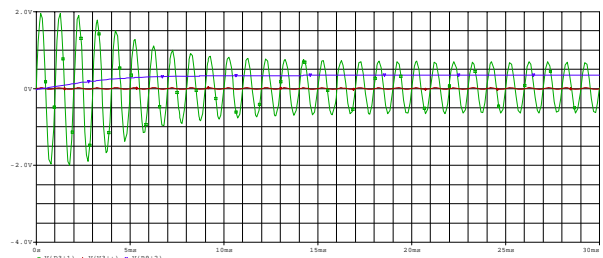


Fig. 2 Simulation of output voltage CCAA and the detected voltage VCAA for the input signal amplitude of 20mV.

B. The clipper

A clipper circuit was added to catch all the transitions that escape out of the dynamic range compressor. This is done with two diodes D10 and D11, connected in parallel, fig. 3. Each diode is reverse biased so that they do not drive until they reach a certain amount of tension. This voltage is set using a resistive divisor consisting of R20 and R21 to a value of approximately 1V and may be adjusted to set the threshold for clipper. It is applied to the uninverted input of the operational amplifier U18 and from the output of U18 to the inverted input of the second operational amplifier U19.

Between the AGC system and the clipper, an adapting gain block is installed. If this block is operating at unity

gain, it is practically transparent and the clipper is operating at threshold.

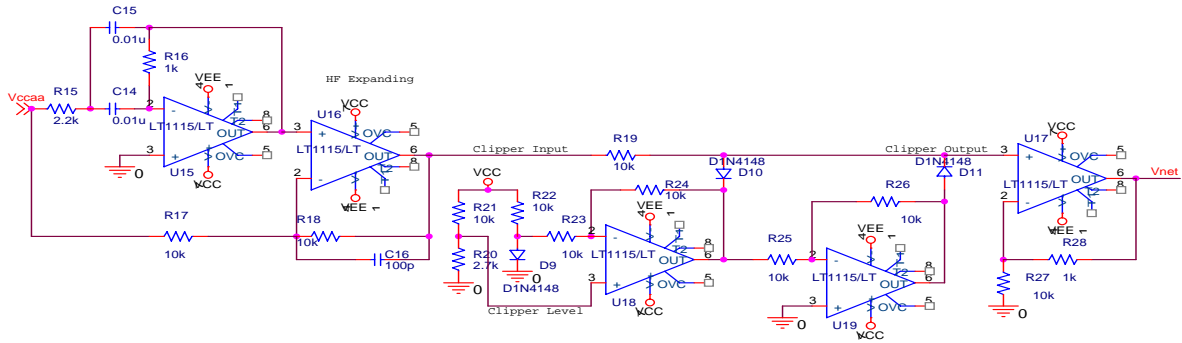


Fig. 3 The clipper

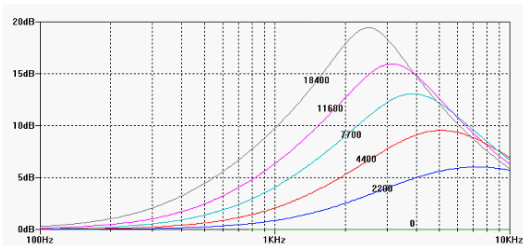


Fig. 4 The clipper driver frequency response

When the resistor R16 in the schematic fig. 3, is set to zero ohms, this circuit is simply a unity gain block. As R16 is

Adjusted, the upper audio frequencies are increased and resulting the family of curves from fig. 4. The reason for introducing the resistance R22 and diode D9 is for the clipper output signal temperature compensation.

C. The lowpass filter

Since the clipper circuit can create higher harmonic to the output, we add a filter to cut frequencies over 3KHz. Must be a lowpass filter, with a flat response and an abrupt shape from the maximum passing frequency. Therefore we choose a filter of order five Chebyshev with 0.2 dB wave amplitude in passing band, fig. 6. The low pass filter circuit was optimized in order to have the best lowpass filter transient response.

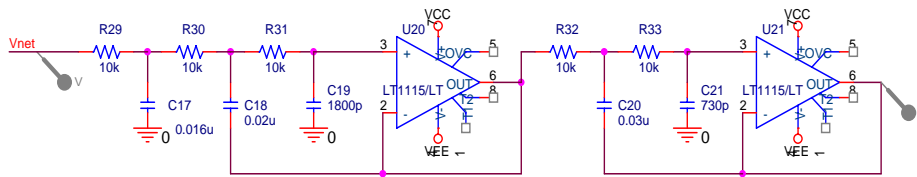


Fig. 6 The lowpass filter

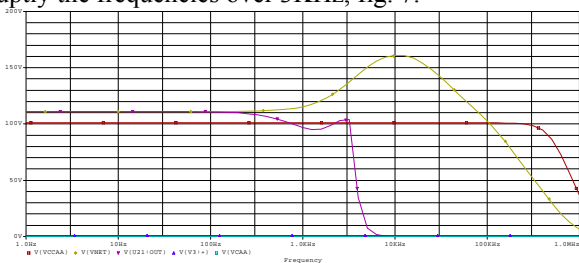
Frequency response of the entire chain corresponds with the block level simulations and makes the designed behaviour. Thus the compressor frequency response is smooth over 500KHz, the driver stage of clipper emphasizes high frequencies and low-pass filter cut abruptly the frequencies over 3KHz, fig. 7.

Fig. 7 Frequencies response of the whole chain of signal processing

III. THE PC INTERFACE

The electrophysiological signals acquiring begins from the source of the bioelectric signals coupled with the electrodes, amplification, processing, analog-digital conversion and data storage in some file format. For the electrophysiological studies, the data storage is necessary, for a long time, as vector data storage.

Acquisition and data storage are performed by an 8-bit microcontroller series AVR (Atmel), namely ATmega32 on a development board that has its own power source, a real



time clock circuit, an EEPROM memory, a LED display and a serial interface adapter (RS232 or RS485), fig. 8. This microcontroller has strong internal resources, allowing data acquisition and digital conversion through a 10-bit ADC, provided with eight inputs multiplexer and its own high accuracy reference voltage reference [6, 7].

Different interesting voltages are collected to internal DAC by means of microcontroller port A, ADC0 to ADC7. The conversion of analog data to a digital vector is synchronized by an internal clock which allows for choose different sampling rates. A conversion cycle starts by clearing the memory locations for the measured values. After that, every input is converted in a 10 Bits word and temporary stored into the internal RAM memory then the next input is also converted, and so on. At this moment we have an eight 10 Bytes words representing a sample of the analog entry signal. This word is now completed with the conversion time, extracted from the external "Real Time Clock" (RTC), the U10 chip.



Fig. 8 The digital recording module using the microcontroller with integrated Port-A analog-digital converter

The vector obtained looks like: 5AYYMMDDHH mmsshhV0V1.....V7, [8]. The whole record is now 24 bytes long and it is stored into the external flash memory U2. This memory has 65536 bytes allowing for over 2700 records. At a rate of 20 samples/second that means there is enough space for more of 2 minutes of records. The acquired data can be extracted by a serial link, the chip U3 providing for the RS232 specification, including hardware handshake by RTS-CTS pair. The communications parameters have been choose to meet the MODBUS specification, as is: 1 START bit, 8 data bits, 9600 Bauds, 1 Even parity bit, 1 STOP bit. During the recording process time information and recorded values are displayed cyclic.

IV. CONCLUSIONS

Usually, only non-invasive electrophysiological methods can be accepted, in respect with the tissue particularities. The paper was focused on signal processing because the electrophysiological signals have a high dynamic range and can be easily covered by the artefacts noise.

The presented vector data collection, processing and recording have the possibility to use many input channels that give the possibility to simultaneously test different versions of source-electrodes-amplifier blocks. Later this facility can be used to multipoint measuring or to increase the resolution. A specific data vector recording was presented, with the advantage of development for new remote methods in electrophysiology.

ACKNOWLEDGMENT

This work was supported by projects 62063, 12095 financed by Romanian National Authority for Sic. Research.

REFERENCES

1. Květina, J.; Varayil, J.E.; Ali, S.M.; Kuneš, M.; Bureš, J.; Tachecí, I.; Rejchrt, S. & Kopáčová, M. (2010). Preclinical electrogastrography in experimental pigs. *International Journal of Interdiscip. Toxicol.*, Vol.3, No.2, (June 2010), pp. 53-58.
2. Al. Rusu, N. Golescu, C. Ravariu, (2008) Manufacturing and tests of mobile ECG platform, IEEE Conf. Sinaia Romania, 2008, pp 433-436
3. Bogdan, D.; Craciun, M.; Dochia, R.I.; Ionescu, M.A. & Ravariu, C. (2009). Circuit design for noise rejection in electromyography, *Proceedings of INGIMED 2009 2nd National Conference on Biomedical Engineering*, pp. 76-81, Bucharest, Romania, ICPE-CA Publisher, November 12-14, 2009
4. Merletti, R. & Parker, A.P. (2004). *Electromyography: Physiology, Engineering, And Non-invasive Applications*, IEEE Computer Society Press, New York, USA
5. B. Firtat, R. Iosub, D. Necula, F. Babarada, E. Franti, C. Moldovan, (2008) Simulation, design and microfabrication of multichannel microprobe for bioelectrical signals recording, IEEE Int. Conf., Sinaia, Romania, 2008, pp 177-180
6. F. Babarada, J. Arhip, (2009) Electrophysiology Signal Data Vector Acquiring, Congress of Romanian Medical Association, Bucharest, 2009, pp 82
7. Rustem Popa, (2006) Medical Electronics. Matrix House, Bucharest
8. ATmega32 data sheet, http://www.atmel.com/dyn/resources/prod_documents/doc2503.pdf

A SIMULATION-BASED FRAMEWORK FOR INDUSTRIAL AUTOMATED WET-ETCH STATION SCHEDULING PROBLEMS IN THE SEMICONDUCTOR INDUSTRY

^{a)}Adrián M. Aguirre, ^{a)}Vanina G. Cafaro, ^{a)}Carlos A. Méndez*, ^{b)}Pedro M. Castro

a) INTEC (Universidad Nacional del Litoral - CONICET), Güemes 3450, 3000 Santa Fe, Argentina.

b) UMOSE, Laboratório Nacional de Energia e Geologia, 1649-038 Lisboa, Portugal

* cmendez@intec.unl.edu.ar

ABSTRACT

This work presents the development and application of an advanced modelling, simulation and optimization-based framework to the efficient operation of the Automated Wet-etch Station (AWS), a critical stage in Semiconductor Manufacturing Systems (SMS).

Lying on the main concepts of the process-interaction approach, principal components and tools available in the *Arena*[®] simulation software were used to achieve the best representation of this complex and highly-constrained manufacturing system. Furthermore, advanced *Arena* templates were utilized for modelling very specific operation features arising in the process under study.

The major aim of this work is to provide a novel computer-aided tool to systematically improve the dynamic operation of this critical manufacturing station by quickly generating efficient schedules for the shared processing and transportation devices.

Keywords: Discrete-event simulation, Semiconductor Manufacturing System (SMS), Automated Wet-Etch Station (AWS), Arena Software.

1. INTRODUCTION

Semiconductor wafer fabrication is perhaps one of the most complex manufacturing systems in the modern high-tech electronics industry. Wafer facilities typically involve many production stages with several machines, which daily perform hundreds of operations on wafer lots. Moreover, different product mixes, low volume of wafer lots and hot jobs are some of the typical issues arising in this type of system.

Wet-Etching represents an important and complex operation carried out in wafer fabrication processes. In this stage, wafer's lots are automatically transferred across a predefined sequence of chemical and water baths, where deterministic exposure times and stringent storage policies must be guaranteed. Hence, automated material-handling devices, like robots, are used as shared resources for transferring lots between consecutive baths.

An important process restriction is that each robot can only transport a single wafer lot at a time and it cannot hold a wafer lot more than the exact transfer time. Due to the lack of intermediate storage between consecutive baths, this condition can be considered as a non-intermediate storage (NIS) policy in every bath,

which must be respected by robots for all transfer movements.

Another constraint adding more complexity to the system operation is that baths must process wafer lots one by one, during a predefined period of time, avoiding the overexposure in the chemical ones, which can seriously damage or contaminate the wafer lot. In spite of this, wafers can stay longer than its processing time only in water baths. So, a zero wait (ZW) and local storage (LS) policy must be strictly satisfied in every chemical and water bath, respectively.

As a direct consequence, an effective schedule of material movement devices and baths along the entire processing sequence will provide a better utilization of critical shared-resources and, at the same time, an important reduction in the total processing time.

In the last years, different methods have been developed to achieve convenient solutions to this challenging problem. Main approaches to large-sized problems lie mainly on heuristic and meta-heuristic methodologies, such as the ones presented by Geiger et al. (1997) and Bhushan and Karimi (2004). In these works, tabu search (TS) and simulated annealing (SA) procedures, together with other different algorithms, were developed to provide a quick and good-quality solution to the job sequence problem and also, a feasible activity program for the robot.

A more recent approach under the concepts of Constraint Programming (CP) was developed by Zeballos, Castro and Méndez, (2011) to handle the sequencing problem of jobs and transfers in the AWS. This method could obtain better results than the ones reported by Bhushan and Karimi (2004) for industrial problem instances in a reasonable CPU time.

To the best of our knowledge, efficient systematic solution methods need to be developed to represent and evaluate the complex dynamic behaviour of the AWS. Thus, a discrete event simulation environment becomes a very attractive tool to analyze the impact of different solution schemes in the system.

In this work, a modelling, simulation and optimization-based tool is developed to validate, test and improve the daily operation of the AWS, allowing an easy evaluation of different operative schemes and possible alternative scenarios. To do this, a discrete event simulation model was developed by using most of the tools and capabilities that are available in the Arena simulation environment. The principal aim is to provide a highly dynamic and systematic methodology to reach the best feasible schedule of limited resources by testing

different measures of effectiveness and performance rates for the system.

Thus, the paper is organized as follows: Section 2 introduces the major features of the problem addressed. Then, Section 3 describes the proposed solution method, highlighting its advantages in comparison with other existing methods and tools as well as the main objectives of this work. Later, the simulation structure is explained in detail in Section 4. A brief description concerning the simulation tool is presented. Software integration and principal interfaces between different tools are discussed. A detailed analysis regarding external and internal logic of the model and the implementation of this solution in a discrete-event simulation environment is also presented.

In Section 5, an alternative solution strategy is tested using several examples, with the main idea of validating the model and, at the same time, comparing results of different solution methods.

Finally, the solutions generated and the comparative study results are reported in Section 6. Conclusions and future work are stated at the end.

2. PROBLEM STATEMENT

The AWS scheduling problem provides a complex interplay between material-handling limitations, processing constraints and stringent mixed intermediate storage (MIS) policies (Figure 1). We can summarize major features of the system in the following way:

- Material-handling devices (robot) can only move one wafer lot at a time. No intermediate storage is allowed between successive baths. So, NIS policy is applied between consecutive baths.

- Waiting times are not allowed during the transportation of a wafer lot.

- Robots and baths are failure-free.

- Setup times are not considered for robots.

- Every bath can only process one wafer lot at a time.

- A ZW storage policy must be ensured in chemical baths whereas LS policy is allowed in water baths.

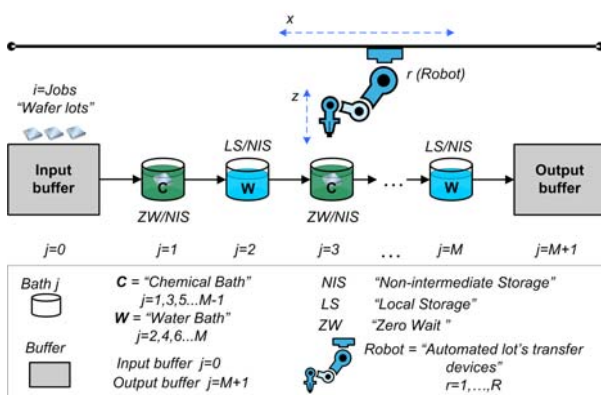


Figure 1: Automated Wet-etch Station (AWS) process scheme.

For this problem, it is assumed that each wafer lot, also called job, i ($i=1,2,\dots,N$) has to be processed in every bath j ($j=1,2,\dots,M$), by following a predefined processing sequence. In addition, it considers that a single robot ($r=1$) is available, which has to perform all the transportation activities in the system.

Consequently, the problem to be faced corresponds to the scheduling of N jobs in M baths, in a serial multiproduct flowshop, with ZW/LS/NIS policies. The use of a single shared robot with finite load capacity for the wafer movement between consecutive baths is explicitly considered in this work.

3. PROPOSED SOLUTION METHODOLOGY

This work introduces an efficient discrete-event simulation framework, which faithfully represents the actual operation of the automated Wet-etch Station (AWS) in the wafer fabrication process.

The main advantage of this computer-aided methodology is that it permits to systematically reproduce a highly complex manufacturing process in an abstract and integrated form, visualizing the dynamic behaviour of its constitutive elements over time (Banks et al. 2004).

The proposed simulation model represents the sequence of successive chemical and water baths, considering the automated transfer of jobs.

Based on a predefined job sequence, which is provided by an optimization-based formulation, the model structure allows the evaluation of many different criteria to generate alternative efficient schedules.

The major aim here is to efficiently synchronize the use of limited processing and transportation resources. This methodology allows also evaluating and improving the operation and reliability of baths and robot schedules. What is more, simulation runs permit addressing industrial-sized problems with low computational effort.

As a result, a basic model is generated to achieve an effective solution to the whole AWS scheduling problem. It becomes also very useful for making and testing alternative decisions to enhance the current process performance.

4. THE SIMULATION-BASED FRAMEWORK

In order to formulate a computer-aided representation to the real-world Automated Wet-Etch Station (AWS) described above, it was decided to make use of the simulation, visualization and analysis tool set provided by the *Arena* discrete-event simulation environment (Law et al., 2007, Kelton et al., 2007).

The simulation model developed in *Arena Software* provides an easy way to represent the AWS by dividing the entire process in specific sub-models (Initializing, Transfer, Process and Output). For each sub-model, the detailed operative rules and strategic decisions involved are modelled using the principal blocks of *Arena Simulation Tool* and, at the same time, a set of visual monitoring objects is used to measure the

utilization performance of all baths and resources in the system.

Additionally, the model allows working with a user-friendly interface with Microsoft Excel for simultaneously reading and writing different data. In next sections, we will describe these features in detail.

4.1. Software integration

The simulator allows an easy communication with Excel spreadsheets. Thus, this tool permits reading, writing and processing important data for the simulation model. Figure 2 illustrates the data flow between Excel and Arena. Both tools support Visual Basic for Applications (VBA) that can be used to move data between them. As shown in the figure, a hybrid solution framework is proposed on these tools. The Mixed integer linear programming (MILP) model provides an initial solution that is written in Excel as input data of the Arena's model. Using that input data, Arena simulation software runs the process model to generate many important statistics that are collected by Excel files as output data. The procedure of reading and writing data is used to dynamically generate a solution schedule by updating the start and finish times of every job in each bath and, simultaneously, determine the status of each job in every stage of the system.

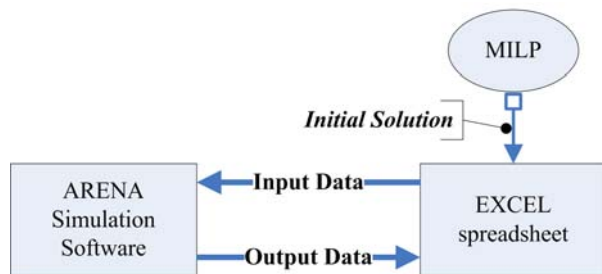


Figure 2: Information exchange between Excel – Arena – MILP Software

4.2. Proposed simulation model

As shown in Figure 3, the entire logic of the simulation model is divided into four main modules (input, transfer, process and output). The first module is the *Initializing* sub-model. The initializing process receives as input data the processing time of each job at each

chemical and water bath and a job sequence provided by a MILP model, which is considered as an initial alternative solution. Then, the discrete-event simulation model generates as many entities as wafer lots are to be scheduled. Here, the logic behind the automated transfer of jobs is performed in order to generate a feasible schedule for the robot activities.

The subsequent simulation module is the *Transfer* sub-model, which defines the needed delay time to transfer a wafer lot to the next bath. This module is used to explicitly simulate the time spent to transfer the jobs between the input buffer to the first bath, between successive baths, according to the predefined sequence, and also between the last bath to the output buffer. Only after the transfer is finished, the bath from where the wafer comes is released. It should be noted that a transfer can be only executed if the robot and the destination bath are both available.

In order to simulate the process itself, one *Process* sub-model for each bath is defined. There is a different logic depending on the type of bath (chemical or water). The wafer residence times in chemical baths must be controlled strictly (as soon as chemical bath finishes, the wafer must transferred to the succeeding water bath). While holding time in water baths is allowed. Thus, for every baths, the logic performs the following tasks: (i) reports the time at which the process begins and ends; (ii) seizes the following bath after the delay time finishes; (iii) performs the transfer to the following bath, only if the robot and the destination bath are empty.

It is important to notice that the logic driving in the *Process* sub-model permits to easily identify why and when a given wafer's lot is discarded. Basically, it may occur because the robot and/or next bath are not available. This allows making a detailed analysis about the behaviour of the system, executing, if necessary, the corresponding adjustments when unexpected events occur or when different strategies are tested in the way to improve the process performance. So, *Process* sub-models permit to evaluate and also validate the feasibility of the internal logic algorithm proposed in the *Initializing Process* of the system, identifying the possible causes of infeasibility to be corrected.

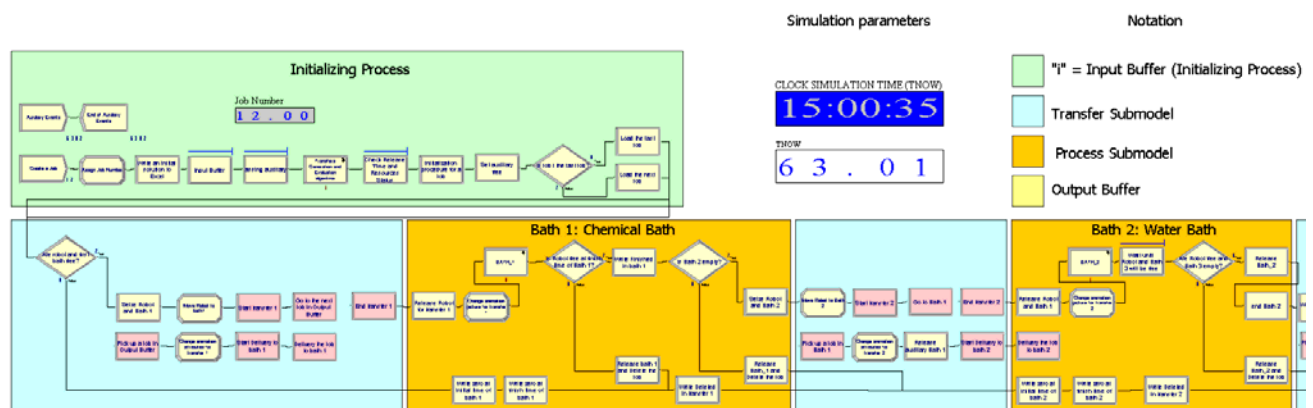


Figure 3: Partial size view of the in-progress simulation model generated in the *Arena* environment.

The last module is the *Output buffer*. The logic of this sub-model represents the final stage of each job. At this module the final processing time (*Makespan*) of each job is reported. It is the ending point for entities created at the input module. Here, the model reports if the current job has been successfully finished or has been discarded.

4.2.1. Advanced internal logic for the robot

The principal aim of modelling the internal robot logic is to explicitly represent the finite capacity of transportation resources for transferring jobs between consecutive baths. The sequence and timing of transfers will depend on the stringent storage restrictions to be satisfied in the baths (ZW / NIS / LS) as well as on the availability of a transportation resource to carry out the transfer.

Since there is only a single robot to do all the job movements, the sequence in which the transfers will be performed needs to be clearly defined. Transfers related to a particular job can never overlap because they are carried out after the corresponding processing stages finish. Consequently, no pair of transfers of the same job may be performed simultaneously.

Therefore, the sequencing problem of transfers must only be focused on the comparison of transfer activities of different jobs in order to determine a feasible robot schedule.

For that reason, a complex internal logic for the robot was embedded in the simulation model to compare and update the start ($ts_{(i,j)}$) and end times ($te_{(i,j)}$) of transfers ((i,j)). The aim is to define the earliest time at which each transfer can be executed. This logic permits to sequence the different transfers in a correct way, generating a feasible schedule for the robot and a near-optimal solution for the whole system, considering a predefined sequence of jobs.

By using this logic, the transfers related to a given job are sequentially inserted according to the order in which they will be processed at every different bath ($j=1,2,3,\dots,M+1$). Then, the transfers are compared successively with all the transfers that were previously inserted into the schedule (according to a predefined processing sequence).

The application of strict storage policies such as ZW and LS in the baths and the NIS rule in the robot significantly complicates the solution of the problem. Enforcing a ZW policy in the chemical baths j implies that the start time of the transfer to the water bath $j+1$ must strictly satisfy equation (1).

$$ts_{(i,j+1)} = ts_{(i,j)} + tp_{(i,j)} + \pi_{(j)} \quad j=1,3,5\dots M-1; \forall i=1\dots N \quad (1)$$

For that reason, the value of $ts_{(i,j)}$ allows directly determining the value of $ts_{(i,j+1)}$. Here $tp_{(i,j)}$ represents the processing time of job i in bath j while $\pi_{(j)}$ denotes the transfer time for every job from bath $j-1$ to j .

On the other hand, if the LS rule is applied to a water bath j , inequality (2) must be satisfied.

$$ts_{(i,j+1)} \geq ts_{(i,j)} + tp_{(i,j)} + \pi_{(j)} \quad j=2,4,6\dots M; \forall i=1\dots N \quad (2)$$

Let $p = \{p_1, p_2, p_3, \dots, p_N\}$ define a permutation processing sequence N different jobs. p_w represents the w -th position of a job i ($i=1\dots N$) in the processing sequence. It means that the job processed in the w -th position will be always before the job processed in the $w+1$ position in the sequence p .

Due to the NIS policy in the transfers and constrains on finite load capacity of the baths and the robot, the equation (3) is to be defined.

$$ts_{(w,j)} \geq ts_{(w-1,j+1)} + \pi_{(j+1)} \quad j=1,2,3\dots M+1; \forall w=1\dots N \quad (3)$$

So, any transfer of a job processed in the p_w position, at bath j , has to wait the ending of the transfer of the job located in the p_{w-1} position at the succeeding bath $j+1$ to be processed.

In the next section, we will explain the transfer comparison algorithm developed to solve the described problem. Only one robot is considered to be available for the execution of the transfers in the system.

4.2.2. Generation and evaluation algorithm for transfers

This algorithm is mainly based on the major ideas of the JAT (Job-at-a-time) algorithm, developed by Bhushan and Karimi (2004). The JAT algorithm always prioritizes the transfers related to jobs that were previously inserted in the system, following a predefined processing sequence. For transfers related to the same job, they are executed according to the fixed sequence of baths to be visited ($j=1\dots M+1$). So, based on processing constrains (1)-(3) and assuming that all the jobs follow the same processing stages, no job in the p_w position may leave the system before the one located in the p_{w-1} position. This means that all jobs will be processed in the different baths following the same p sequence, what is known as "flowshop permutation schedule".

Our algorithm, as the JAT algorithm, selects a job to be processed and then generates (Generation Process) and evaluates (Evaluation Process) all the transfers for this job, one at a time, before going to the next job of the sequence. The principal difference between the proposed algorithm and the JAT algorithm is the evaluation procedure used for the system transfers.

In the proposed evaluation process, every selected transfer is compared with all the transfers previously inserted into the system. Thus, a detailed schedule of the robot operations is defined.

The aim of this process is to avoid that any transfer previously inserted (w',j') can be performed (for $p_w \leq p_{w'}$

and for all j) between the starting time ($ts_{(w,j)}$) and the ending time ($te_{(w,j)}$) of the inserted transfer (w,j)

During this iterative evaluation process the transfer times are initialized (Initialized Process), then they are compared with all the other transfers times (Comparison Process) and finally, they are updated (Updating Process). This loop is repeated successively for a given transfer, until all the comparisons, with the previously inserted transfers, do not introduce new updates at the compared transfer times. So, the comparison and updating processes end. Then, the transfer is evaluated and loaded onto the system with its respectively times [$ts_{(w,j)}$, $te_{(w,j)}$], the counter number of iteration without

change ($iter$) and the number of transfers loaded onto the system ($transf$) are updated and the next transfer from the list (w,j) with $j=j+1$ if $j < M+1$; or $w=w+1$ and $j=1$ if $w < N$, is taken for the comparison. The algorithm ends when there are no more transfers to be compared in the system ($j=M+1$ and $w=N$).

The simplified logic proposed is summarized in Figure 4.

Next, the Generation Process is explained more in detail as well as the procedures of Initialization, Comparison and Update of the Evaluation Process, all of them generated by our algorithm.

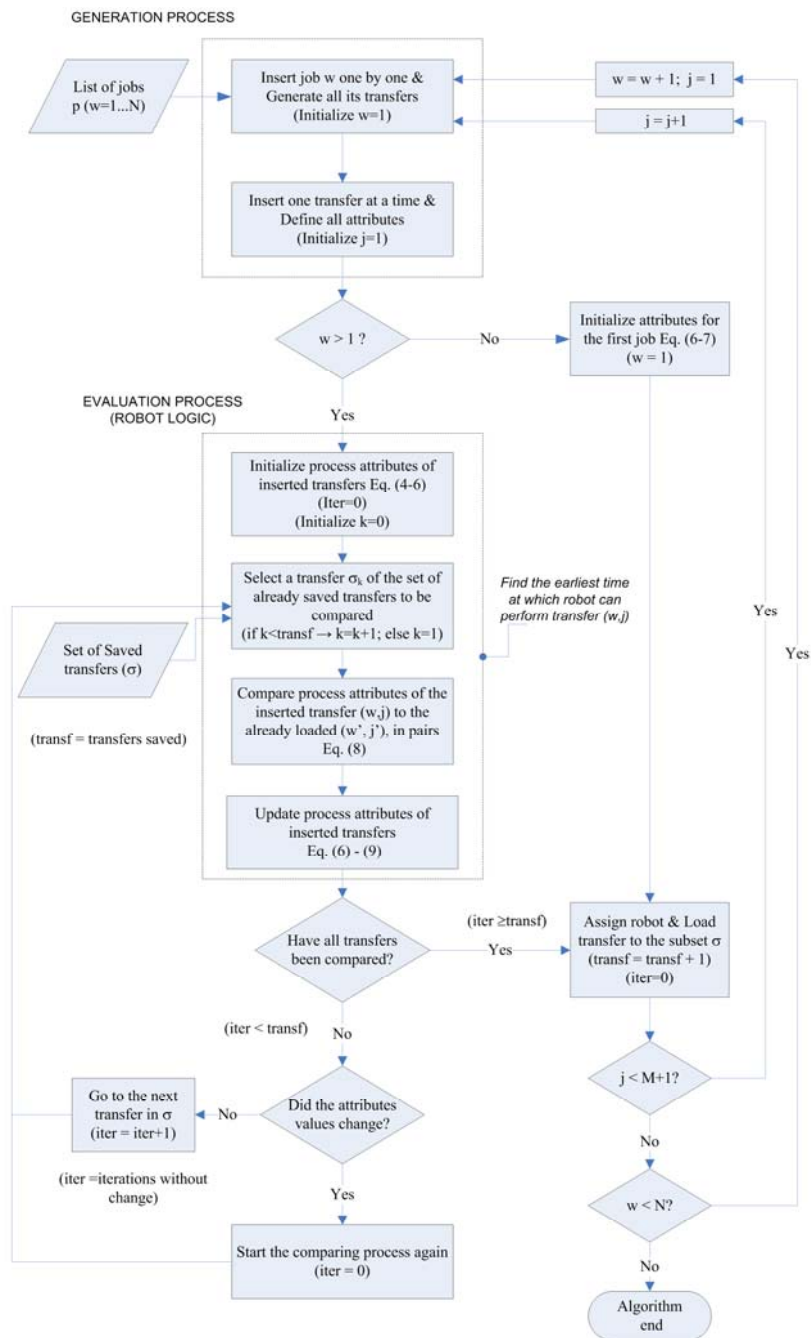


Figure 4: Pseudocode of the Generation and Evaluation Algorithm for transfers

Generation process

To apply the logic in the simulation model it was necessary to define each transfer as a particular new entity in the system, together with the entities associated to the jobs in the system. Consequently, a given job “ i ” will have associated a certain number of transfers and/or entities (i, j) corresponding to the quantity of baths into the system $j=1 \dots M+1$.

Therefore, to start the processing of a given job i , all its j transfers must be pre-loaded into the system. Going back to equations (1) and (2), we can notice that the treatment of the transfers must be done in successive pairs. In order to define the start and end time of the transfer, it is necessary to correctly arrange the successive transfers in the robot, without overlap with any other transfer in the system. So, infeasible schedules are avoided. For that, it is necessary to define a set of attributes $[ts_{(i,j)}, te_{(i,j)}]$ $[ts_{(i,j+1)}, te_{(i,j+1)}]$, for each transfer (i, j) in order to define a correct sequence of transfer over time, avoiding infeasible solutions for the future transfer at the same job i ($i, j+1$).

Evaluation process

After defining all the attributes of the inserted transfer, we proceed to determine an initial value.

Initialized Procedure: the initialization procedure consists on determining the lower value at which the transfer can be initialized, assuming that there are not limitations of resources. So, we can determine the initial value $ts_{(w,j)}$ for each transfer using the following equations (4)-(5).

For chemical baths (baths with odd number), equation (4) is applied:

$$ts_{(w,j)} = \text{Max} \begin{bmatrix} te_{(w-1,j+1)} \\ te_{(w-1,j+2)} - \pi_{(j)} - tp_{(w,j)} \\ te_{(w,j-1)} + tp_{(w,j-1)} \end{bmatrix} \quad j=1,3,5 \dots M+1; \forall w=2 \dots N \quad (4)$$

While for water baths (baths with even number), equation (5) is applied:

$$ts_{(w,j)} = \text{Max} \begin{bmatrix} ts_{(w,j-1)} + \pi_{(j-1)} + tp_{(w,j-1)} \\ te_{(w-1,j+1)} \end{bmatrix} \quad j=2,4,6 \dots M; \forall w=2 \dots N \quad (5)$$

There, $te_{(w,j)}$ is calculated for all baths j with the equation (6).

$$te_{(w,j)} = ts_{(w,j)} + \pi_{(j)} \quad j=1,2,3 \dots M+1; \forall w=1 \dots N \quad (6)$$

So, for any job $w > 1$ the initial state of the attributes in the system is determined: $[ts_{(w,j)}, te_{(w,j)}]$; $[ts_{(w,j+1)}, te_{(w,j+1)}]$.

Instead, for $w = 1$, the initial values of the attributes are defined following equation (6) and (7).

$$ts_{(w,j)} = \sum_{j'=1}^{j'-j-1} tp_{(w,j')} + \sum_{j'=1}^{j'-j-1} \pi_{(j')} \quad j=2,3 \dots M+1; \forall w=1 \quad (7)$$

For the first transfer in the system ($w=1$ and $j=1$), the initial value is equal to zero ($ts_{(1,1)} = 0$).

Comparison Procedure: Once transfers are initialized to $w=1$, they are loaded in the system by updating the subset of charged transfers σ . In σ there are all the transfers (w, j) that have been previously compared and assigned to the robot in a correct way. The value σ_k represents the k -th transfer analysed and initialized into the system according to the priorities described above.

The comparison procedure is applied to the p_w position with $w > 1$. During this iterative procedure, the inserted transfer (w, j) is compared in pairs with a transfer (w', j') of the subset σ , already assigned to the robot (being the $p_{w'}$ position $< p_w$, that means $w' < w$).

If analyzing the attributes (w, j) ($[ts_{(w,j)}, te_{(w,j)}]$ and $[ts_{(w,j+1)}, te_{(w,j+1)}]$) of the transfer with the ones already inserted (w', j') ($[ts_{(w',j')}, te_{(w',j')}]$) there is any overlap between the values of them, then the algorithm will update them for avoiding overlaps (see Equation (8)). Otherwise, the attributes will not be updated. That means that transfer (w, j) does not overlap with (w', j').

$$\begin{aligned} \text{If} \quad & (te_{(w,j)} \leq ts_{(w',j')}) \vee (ts_{(w,j)} \geq te_{(w',j')}) \\ \text{Then} \quad & [ts_{(w,j)} = ts_{(w,j)}] \wedge [te_{(w,j)} = te_{(w,j)}] \\ \text{Else_If} \quad & (ts_{(w,j)} < te_{(w',j')}) \wedge (te_{(w,j)} > ts_{(w',j')}) \\ \text{Then} \quad & [ts_{(w,j)} = te_{(w',j')}] \wedge [te_{(w,j)} = ts_{(w,j)} + \pi_{(j)}] \\ & \forall (w, j) \neq (w', j'); \forall w = 2 \dots N; \forall w' \leq w; \forall j, j' \quad (8) \end{aligned}$$

As can be seen, the updating process consists in delaying the start time of transfer (w, j) when overlapping with (w', j') are observed. Initially, it is necessary to compare the attributes $[ts_{(w,j)}, te_{(w,j)}]$ vs. $[ts_{(w',j')}, te_{(w',j')}]$ and then $[ts_{(w,j+1)}, te_{(w,j+1)}]$ vs. $[ts_{(w',j')}, te_{(w',j')}]$. Thus, we try to ensure that if $ts_{(w,j)} \geq te_{(w',j')}$, then by equation (1) and (2) $ts_{(w,j+1)} \geq te_{(w',j')}$, else if $te_{(w,j+1)} \leq ts_{(w',j')}$ then $te_{(w,j)} \leq ts_{(w',j')}$.

Update Procedure: This procedure is used to generate the earliest time at which the analyzed transfer (w, j) can be executed, in relation with the transfers previously inserted (w', j') and taking into account the resource constrains. As result, the efficient assignment and the detailed program of the robot is determined.

The procedure tries to recalculate the value of the attributes $[ts_{(w,j)}, te_{(w,j)}]$ and $[ts_{(w,j+1)}, te_{(w,j+1)}]$ from the (w, j) transfer fulfilling the equations (1) and (2). As result of the comparison process, the attributes $[ts_{(w,j)}, ts_{(w,j+1)}]$ will be updated according to equation (9).

$$\begin{aligned}
&\text{If } (ts_{(w,j)} + tp_{(w,j)} + \pi_{(j)}) \geq ts_{(w,j+1)} \\
&\text{Then } ts_{(w,j+1)} = ts_{(w,j)} + tp_{(w,j)} + \pi_{(j)} \\
&\text{Else_if } (ts_{(w,j)} + tp_{(w,j)} + \pi_{(j)}) < ts_{(w,j+1)} \\
&\text{Then } ts_{(w,j)} = ts_{(w,j+1)} - tp_{(w,j)} - \pi_{(j)} \\
&\quad j = 1,3,5 \dots M+1; \forall w = 2 \dots N \quad (9)
\end{aligned}$$

For $j=2,4,6 \dots M$ or if not met any of the conditions, only the values of $[te_{(w,j)}, te_{(w,j+1)}]$ are updated. Both of the values are recalculated according to equation (6).

It may be possible that after the end of the processes of comparison and updating, some of the analyzed transfer's attributes overlap again with the previously compared transfer or with some other in the system. If this occurs, the algorithm makes a loop in the comparison process selecting the next σ_k transfer, saved in the σ subset. It also updates the iterations counter to zero ($iter = 0$).

If there isn't any change in the attributes, the algorithm returns to the comparison process and evaluates the analyzed transfer with the next transfer in the σ sequence. Then, the iteration counter $iter$ is updated to $iter + 1$ ($iter = iter + 1$).

In both cases, the comparison is made with the transfer of job σ_k , where $k=k+1$ if $k < transf$, or otherwise: $k=1$; being $transf$ equal to the number of elements in the σ set ($transf = card(k)$).

This iterative process is performed for all the possible comparisons. While this method may be not efficient from the procedural standpoint, since there are unnecessary and redundant comparisons, it tries to avoid the generation of unwanted or erroneous results after the updating stage.

Since to the comparison process is simple and the additional number of events does not report high updating times, we can demonstrate that our algorithm is able to deal with industrial scale problems with modest computational cost.

Finally, when the analyzed transfer (w,j) has been compared dynamically with all the transfers (w',j') of

the σ sequence without updating attributes, that is that the algorithm iteration number ($iter$) is greater or equal than the σ set cardinality ($iter \geq transf$), then the last transfer is loaded into the system with its respectively times, and the number of elements of σ set are updated ($transf = transf + 1$). The iteration counter is initialized ($iter = 0$) and the robot is assigned to the (w,j) transfer during the time between the interval $[ts_{(w,j)}, te_{(w,j)}]$. The next transfer will be $(j = j+1)$ if $j < M+1$.

Otherwise, if j is the last bath of the sequence ($j=M+1$) and w is not the last job of the p sequence ($w < N$) then, the process continues with the next ($w = w+1$) job in the p sequence and $j=1$ is established.

The algorithm ends when there are not more transfers to be evaluated. In this case, $w=N$ and $j=M+1$, that means that all transfers have been loaded into the system ($transf = N * M + 1$).

As result, our algorithm ensures that no pair of transfers inserted into the system and assigned to the robot may overlap over time. Thus, a feasible schedule for both, the process and robot, is generated.

4.3. Implementation in the simulated model

Once the timing of transfers is defined, the model is able to emulate the real system behaviour while satisfying the job processing time, the mixed intermediate storage policies and the assignment of transfers to the limited shared resource.

The simulation is run by using the model resources (baths and robot) and the waiting modules (Queues/Hold/Match). The waiting modules hold the entities until a given condition is met.

While jobs are being processed in the system, according to the predefined job sequence given by p , the transfers' values are updated using specific writing and reading modules (Read/Write). Thus, a fast and simplified way of interacting with Microsoft Excel® is permitted (see Figure 5), defining dynamically the detailed schedule for the baths and robot, together with the generation of dynamic charts representing the evolution of the different works (operations and transfers) over time.

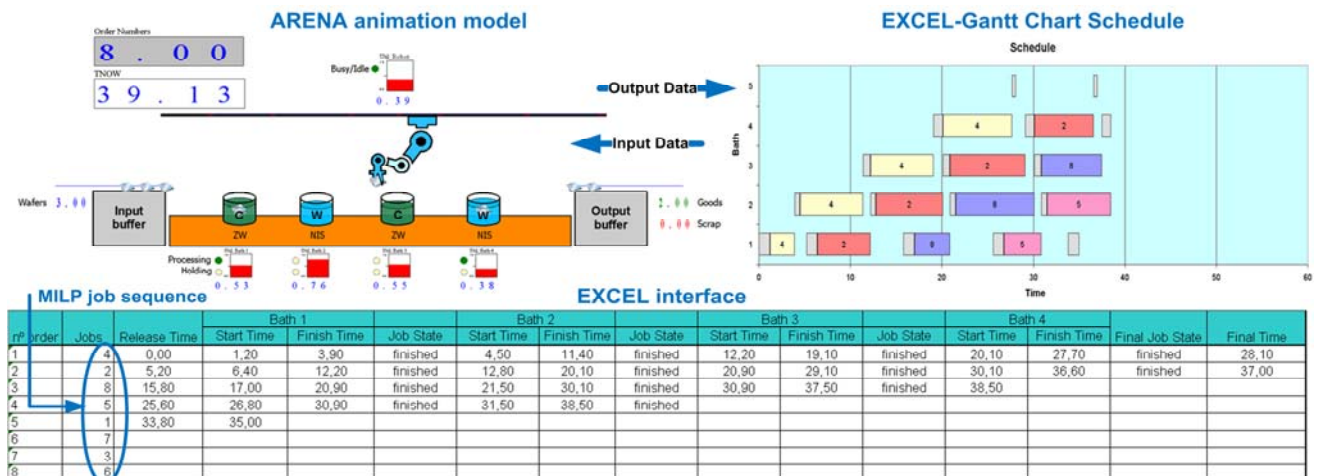


Figure 5: Dynamic Gantt Chart Schedule Generated by a User-friendly Excel Interface

As a result, the dynamic operation can be controlled and analyzed in a global perspective. Failures and/or possible improvement actions can be easily observed by analysing the graphical interface. For example, it can be easily identified how a change in the process sequence impacts over the processing time of each bath and in the availability of the shared resource.

Also, the simulated model progressively evaluates the utilization of the system resources (bath and robot), using monitors or animated screens (see Figure 6), which allow to execute a detailed control of the shared resources performance over time. Thus, it is possible to identify the critical points (resources and/or stages intensively used in the system) with the aim of evaluating alternative modifications in the process design (change the number of resources, or use parallel resources) and/or in the process operation (resources assignment and priority of processing)

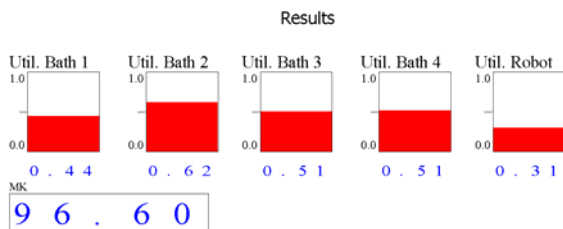


Figure 6: Monitoring the resource utilization

4.4. Animation module of the AWS station

Additionally, the model displays the dynamic behavior of the AWS station through the animation of main system components (entities, resources, performance indicators). Thus, the system operation, involving baths (chemicals and water) and robot activities can be easily evaluated (see Figure 7).

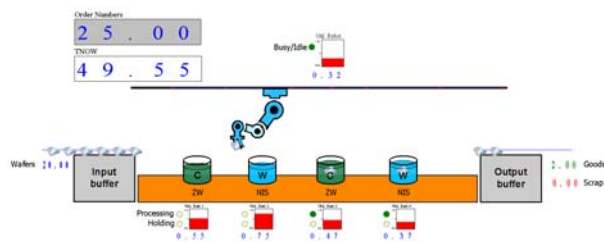


Figure 7: In-progress animation of the AWS station

4.5. Performance measures and termination criterion

The proposed algorithm looks for the best permutation sequence p of the different jobs to be scheduled. This is determined based on the timing of jobs at the consecutive stages and also by the detailed feasible schedule of the transfer robot activities.

Start and end times of activities are dynamically reported in Excel®, according to the different events that take place in the system, at each stage of the process through the simulation.

The main goal is to achieve the shortest completion time of all jobs in the system. So, the objective function

can be estimated with the final time of the last transfer of the robot in the AWS station ($te_{(w,j)}$ for $w=N$ and $j=M+1$).

For our model, the estimation of this time is determined by the MK (*Makespan*) variable. This variable analyses the simulation variable $TNOW$ every time a job is finished.

$TNOW$ is a global variable managed by the simulator that indicates the actual time at which the different events are happening throughout the simulation. In turn, the time for completing the last job in the system represents the stopping criterion of the simulation run (Termination Criterion).

Other performance measures are the utilization of baths and robot. In our particular case, they are used to compare alternative solutions in order to determine alternative policies and logic for the robot allocation.

5. ALTERNATIVE SOLUTION STRATEGIES

A natural way to get a good result of complex problems is to try to break the whole problem at different stages (Bhushan and Karimi, 2003). An iterative solution involves decomposing the whole problem into independent sub-problems, using the solution from one stage as input data for the next one, in order to obtain a global solution in a sequential manner.

In our particular case, generating a good initial p sequence for all the jobs to be processed in the system may notably reduce the complexity of sequencing robot decisions.

The use of meta-heuristics (Bhushan and Karimi, 2004) and mathematical programming models (MILP) (Bhushan and Karimi, 2003; Aguirre, Méndez and Castro, 2011; Zeballos, Castro and Méndez, 2011), are some of the existing tools used to obtain a good initial sequence p for large size AWS scheduling problems.

Here, we present an alternative solution to the robot sequencing problem, based on modern simulation techniques and tools. We also know that in these highly combinatorial problems there exist always a trade-off between computational times and optimal solutions.

For this reason, we have proposed an interesting alternative for obtaining an efficient solution. It is based on a MILP model that provides the best solution to the problem assuming unlimited robots, in order to obtain the optimal p sequence of the jobs in the system. Then, this information is taken as input data by the simulator in order to obtain a feasible and efficient solution to the whole problem, involving the sequencing robot activities. For this, we use the solution provided by a continuous-time formulation developed by Aguirre, Méndez and Castro (2011). Thus, we will initially solve different cases without considering the robot constraints, to subsequently incorporate the results of the sequence into the simulation model.

Finally, in order to validate the model developed, we compare the results with the ones obtained by a rigorous mathematical formulation (MILP), considering the same p sequence in both solutions and also with the results obtained by a full-space MILP model

considering all robots restrictions. Several examples of different sizes are efficiently solved by using this strategy. We will analyse the results obtained from the comparison of those techniques.

5.1. Cases Studies

To prove the applicability of the internal and external logic of the simulation model, different examples using the proposed method are tested. Also, the results generated are compared with optimal MILP solutions found by Aguirre, Méndez and Castro (2011) and the heuristic procedure RCURM from Bhushan and Karimi (2003), by using the previous mentioned MILP model.

The problem instances have been obtained from literature, for an specific $M \times N$ configuration for the first M baths and N jobs of the original problem presented by Bhushan and Karimi, 2004.

6. RESULTS AND COMPARISONS

The heuristic methodology RCURM ("Resource Constrained Unlimited Robot Mathematical Model") is based on a MILP model that can solve moderate size problems with reasonable computational effort in comparison with pure mathematical models. Two alternative models, URM ("Unlimited Robot Model") and ORM ("One Robot Model") were solved

sequentially in order to obtain a solution for the whole problem.

The first one, i.e. the URM, is used to generate an optimal job sequence that ignores the robot restrictions. The URM just only takes explicitly into account the predefined transfer times, assuming that a robot will be always available to perform the transfer operations.

The ORM, in turn, considers the impact of limited transfer resources in the objective function. This proposed model also takes into consideration the sequential use of the single transfer movement device, which enforces a proper synchronization of bath schedules and robot activities.

The idea of the RCURM is to first solve the problem using the URM model, to then fix the production sequence obtained by this model and solve the detailed robot schedule through the ORM formulation. Following this idea, the simulation model will receive as input data the sequence obtained by the URM to then simulate the whole process including the robot activities. As it is shown in Table 1, for the examples validated, the Simulation Model gives the same MK value than the RCURM-MILP for the first three problem instances. This is a good indicator to conclude that the simulation logic may generate results that are as effective as optimal MILP solution, that can be obtained with a modest computational effort.

Table 1: Model Statistics for a few $M \times N$ problem instances

$M \times N$	Statistics	Unlimited Robot Model (URM-MILP)	One Robot Model (ORM-MILP)	Resource Constrained Unlimited Robot Model (RCURM-MILP)	Arena Simulation Model using URM Sequence
4x8	Binary Variables	28	588	560	-
	Makespan	95.1	95.6	95.6	95.6
	CPU Time (s) ^a	0.484	11.25	0.091	-
	Job Sequence p	4-2-8-5-1-7-3-6	4-2-5-8-1-7-3-6	4-2-8-5-1-7-3-6	
4x10	Binary Variables	45	945	900	-
	Makespan	115.5	115.6	116	116
	CPU Time (s) ^a	6.785	488.7	0.122	-
	Job Sequence p	9-2-5-8-10-4-1-7-3-6	9-6-5-4-10-2-8-1-7-3	9-2-5-8-10-4-1-7-3-6	
4x14	Binary Variables	91	1911	1820	-
	Makespan	154.7	158.8	156.2	156.2
	CPU Time (s) ^a	3600 ^b	3600 ^b	0.235	-
	Job Sequence p	9-12-5-8-7-11-14-10-2-4-1-13-3-6	9-2-8-12-4-14-10-11-5-1-3-7-13-6	9-12-5-8-7-11-14-10-2-4-1-13-3-6	
8x10	Binary Variables	45	3285	3240	-
	Makespan	149.4	154.4	156.7	166.4
	CPU Time (s) ^a	55.07	3600 ^b	3.42	4.8
	Job Sequence p	6-9-2-1-3-4-7-5-10-8	4-9-3-1-2-6-7-5-10-8	6-9-2-1-3-4-7-5-10-8	
12x10	Binary Variables	66	10362	10296	-
	Makespan	192.2	206.3	197.2	227.8
	CPU Time (s) ^a	145.5	3600 ^b	152.97	7.2
	Job Sequence p	6-8-3-2-9-5-10-7-4-1	6-1-2-10-5-3-9-7-8-4	6-8-3-2-9-5-10-7-4-1	
12x12	Binary Variables	105	16485	16380	-
	Makespan	241.9	NFS ^c	NFS ^c	334.2
	CPU Time (s) ^a	6.785	3600 ^b	3600 ^b	12.0
	Job Sequence p	6-8-3-11-2-13-9-5-14-10-12-1-15-4-7	-	6-8-3-11-2-13-9-5-14-10-12-1-15-4-7	
12x15	Binary Variables	300	47100	46800	-
	Makespan	357	NFS ^c	NFS ^c	516.8
	CPU Time (s) ^a	3600 ^b	3600 ^b	3600 ^b	48.0
	Job Sequence p	6-16-8-11-20-4-21-18-17-19-5-10-15-22-14-22-14-2-12-3-25-13-24-9-23-7-1	-	6-16-8-11-20-4-21-18-17-19-5-10-15-22-14-2-12-3-25-13-24-9-23-7-1	

(a)MILP models were solved by using CPLEX 12 in a PC Core 2 Quad parallel processing in 4 threads. (b) Termination criterion (3600 CPU s). (c) No feasible solution found after 3600 sec.

By analyzing the model statistics, we can notice that the solutions generated by the Simulation Model, by using the URM sequence, are very close to the ones found by the ORM and RCURM models, which points out the high performance of the alternative proposed methodology for many small size cases. But, when the model size increases the solution obtained by this approach becomes poor in comparison with ones reported by RCURM and ORM models.

The most important difference between ORM/RCURM-MILP approaches and our Simulation Model lies on the computational time consumed, what is more evident in medium size and large size cases, as 4×14 , 8×10 , 12×10 and 12×12 , 12×15 , 12×25 configurations respectively.

Moreover, for many larger problems only the Simulation Model may find feasible solutions of the entire problem with very low computational cost.

In consequence, the application of the proposed solution strategy to manage the activities of the robot will compare favourably against a MILP mathematical formulation and a MILP-based decomposition method for many large-size problems in the AWS station. Also, the solution generated by this approach can be considered as an initial solution of the whole problem, which may be later enhanced by alternative meta-heuristic or optimization-based methodologies.

CONCLUSIONS AND FUTURE WORK

A novel discrete event simulation model has been developed to simultaneously address the integrated scheduling problem of manufacturing and material-handling devices in the AWS in the semiconductor industry. The proposed model can be easily used to dynamically validate, generate and improve different schedules. We have demonstrated that the proposed solution algorithm for the robot is able to generate very effective results with modest computational effort. For large-sized cases, only our simulation approach found feasible solutions to the problem in a reasonable computational time.

In addition, alternative heuristic rules can be easily embedded into the simulation framework for making convenient timing and sequencing decisions. At the same time, alternative system configurations involving several robots for wafer-handling in the AWS station can be easily considered. As a future work, a hybrid approach lying on the concepts of optimization and simulation tools will be developed in order to improve the generation of the solution for the whole scheduling problem.

ACKNOWLEDGMENTS

Financial support received from Fundação para a Ciência e Tecnologia and Ministério de Ciencia, Tecnología e Innovación Productiva, under the Scientific Bilateral Cooperation Agreement between Argentina and Portugal (2010-2011), from AECID under Grant PCI-D/030927/10, from CONICET under

Grant PIP-2221 and from UNL under Grant PI-66-337 is fully appreciated.

REFERENCES

- Aguirre, A. M., Méndez, C. A., Castro, P. M., 2011. *A Novel Optimization Method to Automated Wet-Etch Station Scheduling in Semiconductor Manufacturing Systems*. Comput. Chem. Eng., doi:10.1016/j.compchemeng.2011.02.14.
- Banks, J., J. S. Carson, B. L., Nelson, D. M. Nicol. 2004. *Discrete-Event System Simulation*. 4th ed. Upper Saddle River, New Jersey: Prentice-Hall, Inc.
- Bhushan, S., and Karimi, I. A., 2003. *An MILP approach to automated wet-etch scheduling*. Industrial and Engineering Chemistry Research, 42(7), 1391-1399.
- Bhushan, S., and Karimi, I. A., 2004. *Heuristic algorithms for scheduling an automated wet-etch station*. Computers and Chemical Engineering, 28, 363-379.
- Geiger, C., Kempf K. G., and Uzsoy, R., 1997. *A tabu search approach to scheduling an automated wet etch station*. Journal of Manufacturing System, 16, 102-116.
- Kelton, W. D., R. P. Sadowski, D. T. Sturrock., 2007. *Simulation with Arena*. 4th ed. New York: McGraw-Hill Inc.
- Law, A. M., 2007. *Simulation Modeling and Analysis*. 4th ed. New York: McGraw-Hill, Inc.
- Zeballos, L. J., Castro, P. M., Méndez, C. A., 2011. *Integrated Constraint Programming Scheduling Approach for Automated Wet-Etch Stations in Semiconductor Manufacturing*. Ind. Eng. Chem., 50, 1705.

Dr. CARLOS A. MENDEZ is a Titular Professor of Industrial Engineering at Universidad Nacional del Litoral (UNL) in Argentina as well as an Adjoint Researcher of the National Scientific and Technical Research Council (CONICET) in the area of Process Systems Engineering. He has published over 100 refereed journal articles, book chapters, and conference papers. His research and teaching interests include modeling, simulation and optimization tools for production planning and scheduling, vehicle routing and logistics.

Dr. PEDRO M. CASTRO is an Assistant Researcher at Laboratório Nacional de Energia e Geologia in Portugal in the area of Process Systems Engineering. His research interests include scheduling of batch and continuous processes, mixed integer and nonlinear optimization, and process integration. He has published over 30 papers in international journals with refereeing and has been invited to give seminars in a few Universities and Research Centers worldwide.

DATA STREAM MANAGEMENT IN INCOME TAX MICROSIMULATION MODELS

Molnár, István^(a), Lipovszki, György^(a, b)

^(a) Department of Computer and Information Systems
School of Business
Bloomsburg University of Pennsylvania
Bloomsburg, Pennsylvania, 17815, U.S.A.

^(b) Department of Mechatronics, Optics and Engineering Informatics
Faculty of Engineering
Budapest University of Technology and Economics
H-1111 Budapest, Muegyetem rkp. 3-9, Hungary

^(a)[Email: imolnar@bloomu.edu](mailto:imolnar@bloomu.edu) ^(b)[Email: lipovszki@rit.bme.hu](mailto:lipovszki@rit.bme.hu)

ABSTRACT

The paper discusses the repetitive direct use of large data sets in microsimulation models. First, the income tax microsimulation models and the related data requirements are discussed. Next, the processing of the applied data set is introduced and the idea of using direct data instead of samples developed along with applied data stream management principles. The discussion of data stream management issues turns subsequently into an analysis of model control aspects of data driven simulation. Finally, some practical experiences and applied technologies are discussed.

Keywords: Data stream management, Income tax microsimulation, Microsimulation model development environment

1. INTRODUCTION

During the past 50 years, microsimulation research and practice focused on solving methodological problems closely related to mathematical modeling, mathematical statistics and economics. Less attention was devoted to model implementation, computational and application efficiency.

Recent developments show a change in focus of the efforts and recognition of the progress in the broad field of IT/IS, including use of system development environments, network oriented applications, collaborative system development and high performance computing. There is a significant stream of efforts in the field, which tries to integrate legacy systems with modern IT/IS, while using well recognized modeling and simulation (M&S) methodologies and tools. The change of obsolete world view and technology however is not fast enough and we still need to wait for a complete breakthrough.

This paper aims to contribute to these efforts and open new perspectives for microsimulation using M&S methodologies and tools widely used in engineering and

sciences. The authors believe firmly that the ideas presented are powerful enough to be successfully applied in M&S of business and economic systems, including all application fields of microsimulation.

2. MICROSIMULATION MODELS' DATA REQUIREMENTS

2.1. Data sources and data analysis

One of the major components of microsimulation models (and therefore of the whole microsimulation modeling environment) are the data related to the micro units: *initial simulation model data*, *intermediate* and/or *final simulation data*. Initial data of microsimulation models are collected from *cross-sectional surveys*, and/or *longitudinal surveys*. Cross-sectional surveys collect data about a sample population for a single period of time (e.g., a survey of tourists' expenditures), while longitudinal surveys collect data about the same sample population (also called panel) for several periods of time (e.g., a household statistical survey). Microsimulation models could also use a *time-series of cross-sectional surveys*, which collect data at periodic intervals (e.g., micro-census). Special techniques (e.g., imputing, merging, synthetic data) have been developed to improve data quality and use additional data sources available. In addition, the microsimulation model consists of a series of *economic indicators* and diverse data, which are also stored for simulation as *variables and/or model parameters* and used for further analysis and computations.

The model behavior in microsimulation models is defined by *algorithms*, which, among others, reflect the micro units' behavior rules, related probabilities, furthermore, describe economic processes and represent their impact. By using this methodology, special care is taken to do the *data analysis* and the *estimation of microsimulation model parameters*.

The microsimulation model is working in an *experimental framework* in order to study the effects of

policy changes on the microsimulation model behavior. Given the model responses of micro-units at unit level, the microsimulation models can estimate aggregate effects and aggregate changes by grouping, creating distributions, tabulating or summing up unit-level individual model results in order to make statements about the various characteristics of the population as a whole, helping to determine “winners” and “losers”.

The modeling process and numerical computation of simulation results are loaded with different kind of errors; e.g., the model is never a perfect representation of the system, parameter-estimations are rarely perfect, while the numerical computation might also contain round-off and/or method-related errors. *Validation* checks the modeling process and its final result, the conceptual model, and whether the model was able to represent the studied system in a satisfactory way. *Verification* checks the computer model, especially whether the computer model is executed properly and the results calculated correctly. It is the responsibility of the modeler to make a final decision about the model and the acceptance/rejection of the model results. Therefore, model verification and validation also use various sophisticated statistical methods and techniques, the results of which can be desirable to store, retrieve and analyze.

Taking the development and the application environment of microsimulation models into account, one can state that *data are collected, stored, retrieved and processed in a distributed way in different databases at different locations, moreover, they are maintained by different, mainly government authorities.* Most of the data are available in the form of different time series, in such a way that data content is hard to define, it might change with time and data integrity and accuracy are difficult to maintain. One of the biggest problems could be the management of the same data content under different names and different data content under the same name. Some of the simulation modeling problems are traditionally solved by using synthetic data sets (e.g., merging, imputing), which means that artificial data sources (and methods) are applied instead of traditional system data sampling. In addition to the “real system’s” data, different types of simulation data are regularly stored and retrieved. These characteristics of microsimulation models’ “messy” data sets are reflections of the characteristics of very large scale systems of the social sciences which are also very “messy”.

2.2. Mathematical models used

The stored data of microsimulation models can be: initial model data, intermediate and final simulation data. These data are stored for further analysis. The description of model behavior in microsimulation models is based on algorithms, which describe the behavior of the micro units and their environment.

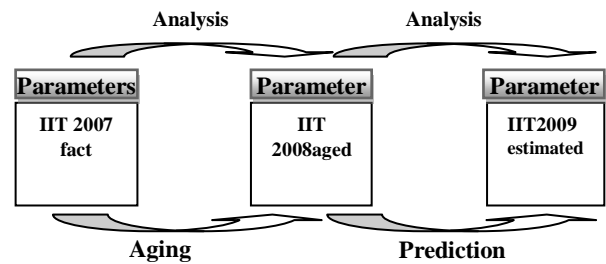


Figure 1: The Individual Income Tax simulation (IIT) process

The estimation of simulation model parameters, the general data analysis in order to determine and/or estimate behavioral rules are important steps of the modeling process. The microsimulation model is embedded in an experimental framework, which provides an environment to study the effects of policy changes on the microsimulation model behavior.

All available data are carefully analyzed using special techniques, which have been developed to improve microsimulation data quality (e.g., imputing, merging, synthetic data). For solving special model verification and validation problems, different methods and techniques were developed, which are not discussed further.

2.3. Computer and Information technology applied

The historically developed major model classes must maintain a significant amount of data and use methods for processing and analyzing these data. Data sources were historically integrated into microsimulation models in one of the following three different ways:

- File processing approach
- Database-oriented approach
- Agent-oriented approach

These approaches are not architecture neutral or network-oriented. Two tier databases are used at most. About a decade ago computer science started to develop new network-oriented technologies usable also for model-based applications, which support the use of heterogeneous hardware and/or software platforms. These network-based technologies became more widespread recently, when technical development allowed for networked multiplatform applications and beyond the networked data access also for distributed computing.

Despite the data-related problems, to the best of the authors’ knowledge as of now there are no significant efforts to use standardized database content, data structure and access and data retrieval for microsimulation models. It is clear that even most recently published studies focus on platform dependent, non-portable microsimulation applications (e.g., PC-oriented or supercomputer-based) and provide related data management solutions. Little attention has been paid to the management of large data sets for distributed microsimulation databases and distributed applications,

moreover, portable and architecture neutral network-oriented technologies have been largely ignored (see Flory and Stöwhase 2011; Sutherland 2011).

3. NEW APPROACH TO MICROSIMULATION DATA MANAGEMENT

3.1. Requirements towards data stream management environments

Based on widely acknowledged simulation and information technology principles, a series of requirements were developed and implemented for the general microsimulation software development environment. These requirements help to establish a broad framework for different microsimulation applications:

- Platform-independent hardware and software solutions based on open standards.
- Data and network security.
- User friendliness, standardized user interfaces.
- Network-oriented data and model access using a database management system.
- Distributed model development, model execution and data analysis.
- Efficiency of software development during the whole software life cycle.

The implementation of the major principles listed, helped to manage the current difficulties of the microsimulation model development process. The implementation uses meta-database and Service-Oriented Technology (SOA), both of which are considered as key technologies for microsimulation. The technological solutions helped to expand the scope of existing research and to introduce two significant additional changes, which are going to be implemented:

- Use only data sources, which were selected based on the principle “straight from the reliable source”.
- The overall efficiency of the microsimulation application can be best supported using role-based and workflow supported application.

3.2. Use of Object-oriented Data Modeling, Meta Databases and Data Warehouses

Large and complex data sets can be best managed using a database management system, which consists of crucial information about all individual microsimulation data available in the data system; i.e., a meta-database, which consists of data about microsimulation data. Based on the applied microsimulation models, the microsimulation data meta-database must reflect the supported microsimulation model classes and therefore is rather model-oriented (see Molnár and Sinka 2006).

The major advantage of using a meta-database for microsimulation data consists of, among others, increased data quality and overall cost-efficiency. Other criteria listed, like network-oriented data access and data analysis, as well as the efficiency of software development require different technologies, which are sometimes integrated with the core database management system (e.g., Oracle’s Business

Intelligence solutions support data warehousing, data analysis, data mining and report generation).

When using a meta-database for microsimulation data, all used data can be best referred to by using the appropriate data references of the meta-database. Because all methods, which can be executed on the microsimulation database data, can also be considered as data, it is evident that the most efficient solution to store the methods themselves is to store them in database(s), i.e., by creating appropriate method-database(s). Having method-databases requires the same data management practice as using a meta-database; i.e., a meta-database for microsimulation methods must be created.

The following information technology solutions are used:

- *Meta database*: a database of data about microsimulation data (how microsimulation data are collected/generated, accessed, processed, etc.).
- *Data warehouse*: data are stored in a special structure based on data-related dimensions (e.g., time, collected by, data content). Data are pre-processed before being stored (e.g., filter, extract, transform, classify, aggregate, summarize).
- *Object-oriented data modeling*: a data modeling paradigm, which applies object-oriented approach and data modeling.
- *Object oriented programming*: a programming paradigm.
- *Service Oriented Architecture*: applications implemented as Web-based components, which offer certain functionality to clients via the Internet.

Meta database and the related object-oriented database technologies help to manage the complex data sets and clean up the “messy” data.

3.3. Use of Service Oriented Architecture and Web services

SOA provides methods for systems development and integration in a standardized environment, where different software system functionalities are offered as independent and interoperable services. Web service architecture is built on open standards and vendor-neutral specifications and it is a way to implement a SOA.

This approach is a shift from a single application usage to a multi-component distributed application running on different platforms using open standard based network communication. Parallel to the authors’ research and the implementation of the microsimulation software development environment, this technology became the foyer of cloud computing.

From the point of view of microsimulation software development, the major emphasis is on model development, use and re-use. Microsimulation algorithms can be developed using high level programming languages (e.g., Java); data analysis and

parameter estimation can be prepared by special mathematical and statistical software tools (e.g., SAS, SPSS). Both components are supported by Web services. Networked DBMS functionality can also be accessed using the high level programming languages provided by the Web service framework.

The Web service components have well-defined interfaces. Once deployed, they can be discovered, used/and reused by consumers (clients, other services or applications) as building blocks. Web service architecture is built on open standards and vendor-neutral specifications.

A significant advantage of this approach is that most elements of the technology are platform independent, and widely available (in part as free or open source software). Further advantages include but are not restricted to the following:

- Legacy systems can be integrated, existing codes re-used,
- Software development and maintenance costs, furthermore operational costs can be reduced,
- New business models can be established and new revenue can be generated, while interfaces with customers and integration with business partners can be improved.

The IT/IS industry is moving rapidly towards SOA and cloud computing in general, therefore the current concerns related to the use of Web services (e.g., there is no network performance guarantee, some standards are still missing) will soon disappear.

SOA is able to satisfy several application requirements listed, such as distributed data processing and model computations, security, platform independency and overall software development efficiency through the whole software life cycle. At the same time, SOA enables a move to cloud computing and the application of data-driven simulation techniques.

3.4. Enabling the use of data-driven simulation

Dynamic Data-Driven Application Simulation (D3AS) is a new paradigm, which enables the application (or simulation) and measurements to become a control system, in which data dynamically control almost all aspects of the long term simulation. (See Figure 2 based on Molnár and Sinka 2010, and NSF 2011.)

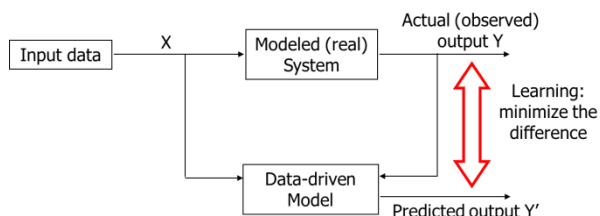


Figure 2: The Structure of Dynamic Data Driven Application

Traditional simulation models run many simulations using static data as initial condition. D3AS runs a very small number of simulations with additional

data “injected” as they become available. Dynamic data are used to determine e.g., whether a “warm start” is needed, whether a rollback in time is required or whether the error of the simulation run is still acceptable. The D3AS approach to modeling focuses on the use of Machine Learning (ML) methods in building models to complement or replace “knowledge-driven” models, which describe the behavior of the system.

D3AS enables to create more accurate models of complex systems, which are adaptive to changing conditions and infer new knowledge (not determined by the initial state and parameters). Possible new application areas are business, engineering, and sciences (e.g., manufacturing process control, resource management, weather and climate prediction, traffic management, social and behavioral modeling).

4. HUNGARIAN INCOME TAX SIMULATOR

Elements of the presented methodology were implemented and the software applied for both a prototype model and the 2009 Hungarian Income Tax Simulator (HITS-2009). The HITS-2009 used real data to determine the possible fiscal impacts of the new linear income tax policy, planned to be introduced (see Molnár et al. 2009).

Based on the data source principles, anonymous 2008 income tax declaration data submitted to the Hungarian Tax Office (HTO) were used for the whole population and not the Household Statistical Panel (HSP) (i.e., no samples were applied, but about 4.5 million individual income tax declarations directly processed). The HITS-2009 is a classical tax policy simulator and as such, it is static, but able to predict the impact of different income tax policy changes, among others the fiscal impact of the generated income tax revenue component of the 2009 state budget.

In the same time, HITS possesses the capability to use updated data resources as a feedback control in a later point of simulation time and establish a major step towards conversion of the static microsimulation model into a dynamic one.

The HITS-2009 model could only have been assessed in June 2010, by which time the individual income tax declarations had already been processed. The authors of Molnár et al. 2009 found that model behavior was consistent with widely accepted results of the related economic theories. However, taking into account that the income tax model was created as a static microsimulation model (e.g., with no behavioral changes of the tax payers), model results must be interpreted and related to the reality with extreme caution. The standardized model output (e.g., additional statistics like the marginal tax rate) provides new opportunities for analysts to get more insights about income tax behavior.

The model results demonstrated clearly that both the methodology and technology are able to live up to the high expectations and result in increased model reliability and accuracy. To demonstrate the

standardized outputs of HITS-2009, Figure 3 and 4 are presented. The figures show the impact of different income tax rates to the budget and to the tax payers, respectively.

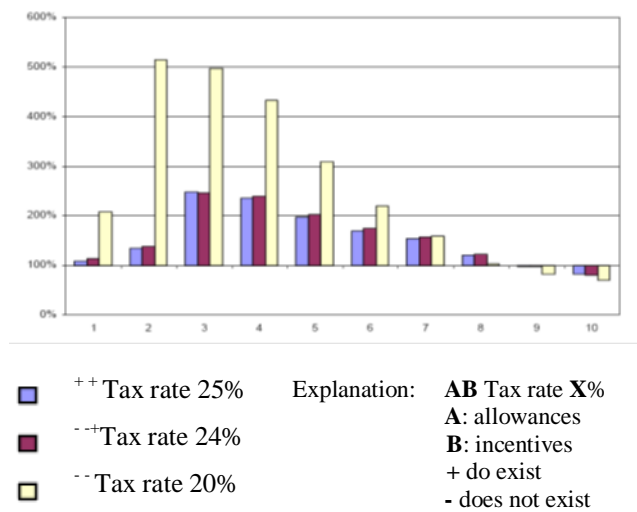


Figure 3: Average annual tax increase/decrease in different deciles in comparison to 2007

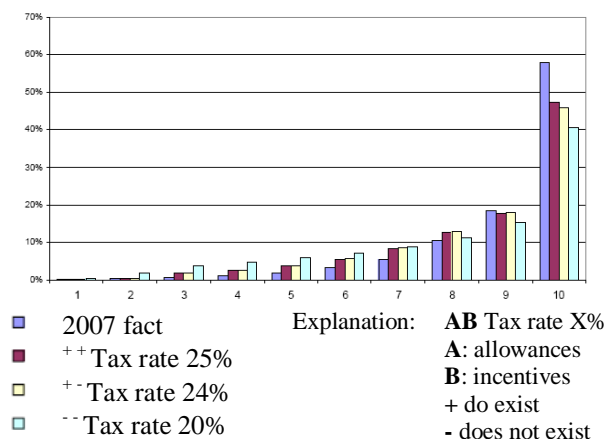


Figure 4: Contribution of income tax deciles to the budget (in case of different tax rates)

5. CONCLUSION

The paper presents a generally applicable methodology and a microsimulation model development environment to create and use data- and/or model-driven microsimulation models within a platform-independent, distributed and networked software environment. The basic idea of the solution is that the data and model complexity is best managed by a multidimensional distributed database, which makes extensive use of meta-database data, which are also implemented in the very same distributed environment using Web-services. In order to increase the flexibility of data storage and data processing, the relationships stored in databases are described using object-oriented data modeling and programming concepts; the computation engine uses the stored data and methods from the same database. The resulting microsimulation model development

environment is able to support the implementation of data-driven and model-driven static and dynamic microsimulation models. The resulting architecture does not require but enables and supports the use of specialized agent-based modeling software.

Using the presented technologies, all relevant user requirements, listed as objectives, can be satisfied. The experiences with simple models showed results upon which one can build real word microsimulation applications. The HITS-2009 model demonstrated the strength of the applied methodology and technology under “real-world” circumstances.

The conversion of a traditional microsimulation model into web-based model applications enables the integration of legacy systems into the new, integrated information systems. The new technology based on SOA provides real solutions for distributed and secure data access, which is a major concern in most government applications, and also ensures distributed model development and analysis. SOA also paves the way toward collaborative model building and use.

The pilot project application of methodology and technology demonstrate the potential of regular use and can be included into microsimulation model bases, forming the first elements of a microsimulation-based Decision Support System. These systems could be extremely useful in scientific research (e.g., study of artificial societies) and state administration (e.g., tax microsimulation).

In order to extend the range of applications and provide wide availability of the microsimulation models, administrative regulation of data access must also be implemented. The use of distributed databases and meta databases in microsimulation models sends a strong signal to the authorities responsible for national statistical data and raises questions related to the efficient and regulated use of these data. The answers to these questions will determine the use of microsimulation models for national policy decision-making and with that the future of distributed state administration applications, which will certainly have a significant impact on any EU-wide regulations.

REFERENCES

- Flory, J. and Stöwhase, S. 2011. MIKMOD-ES - A Static Microsimulation for Model the Evaluation of Personal Income Taxation in Germany. *3rd General Conference of the International Microsimulation Association: Microsimulation and Policy Design*, Stockholm, Sweden. 1-19.
- Molnár, I. and Sinka, I. 2010. State-of-the-art information technology for microsimulation. *International Journal of Technology, Modeling and Management (IJTMM)*. Serials Publications. 1 (1), 41-62.
- Molnár, I. and Sinka, I. 2006. Model-oriented, data-driven architecture for microsimulation. *Simulation News, Europe, Journal on Developments and Trends in Modeling and Simulation*. ISSN: 0929-2268, 16 (3), 37-40.

- Molnár, I., Bátor, D., BelyóP., Sinka, I., Szathmáry, B. and Tóth, Cs. G. 2009. Adó mikroszimuláció (Egyes adónemek mikroszimulációja), ECOSTAT Gazdaságelemző és Informatikai Intézet, Időszaki Közlemények 37.szám, Budapest, ISBN: 978 963 88329 6 2, ISSN: 1418 7892., 1-59.
- NSF (National Science Foundation). 2011. *DDDAS: Dynamic Data Driven Applications Systems*. Available from: <http://www.nsf.gov/cise/cns/dddas/> [Accessed 06.06.2011]
- Sutherland, H. 2011. Address: "EUROMOD: the state of play and a vision for the future" 3rd General Conference of the International Microsimulation Association: Microsimulation and Policy Design, Stockholm, Sweden.

AUTHORS BIOGRAPHY

István Molnár

Educated at the Corvinus University Budapest, Hungary, where he received his M.Sc. and Dr. Oec.(Ph.D.) degrees. He completed his postdoctoral studies in Darmstadt, Germany and took part in different research projects in Germany and Western Europe as Guest Scientist in the 1980s and 1990s. In 1996, he received his C.Sc. (Ph.D.) degree from the Hungarian Academy of Sciences. His main fields of interest are simulation, simulation optimization, software technology and education. He has been a member of different international scientific organizations and editorial boards of scientific journals, publishing houses. Currently, he is the Editor in Chief of *IJTMM* and a member of the Editorial Board of *IJMLO*.

György Lipovszki

Born in Miskolc, Hungary and finished his study at Budapest University of Technology and Economics, where he was graduated in 1975 in electronics sciences. Currently, he is an Associate Professor at the Department of Mechatronics, Optics and Engineering Informatics. His research field is the development of modeling and simulation framework systems in different programming environments. He is a member of the Editorial Board of *IJSTL* and *IJTMM*.

EXPERIMENTAL MANUFACTURING SYSTEM FOR RESEARCH AND TRAINING ON HUMAN-CENTRED SIMULATION

Diego Crespo Pereira^(a), David del Rio Vilas^(b), Rosa Rios Prado^(c), Nadia Rego Monteil^(d), Adolfo Lamas Rodriguez^(e)

^{(a)(b)(c)(d)(e)} Integrated Group for Engineering Research (GII), University of A Coruña (Spain)

^(a)dcrespo@udc.es, ^(b)daviddelrio@udc.es, ^(c)rrios@udc.es, ^(d)nadia.rego@udc.es, ^(e)alamas@udc.es

ABSTRACT

Human performance modelling and simulation requires a multidisciplinary approach for a full understanding of its effects in manufacturing. These areas broadly span production theory, behavioural science and ergonomics. In despite of the advances in each one of them separately, few papers have adopted a global-scope approach. Previous works have mainly focused on the integration of models coming from the different disciplines involved. This paper presents the design and construction of an experimental manufacturing system which allows for conducting research and training on human operations analysis within a controlled environment. Task procedures, supervisory mechanisms and data acquisition systems are arranged so that non desirable variability is restrained to an acceptable level. System architecture was inspired by virtual simulators enabling results analysis in a structured way. The system provides with capability for experimentation in interaction of behavioural and ergonomic effects, model validation research and teaching in simulation and other process improvement tools.

Keywords: human-centred simulation, modelling and simulation of human behaviour, ergonomics.

1. AIM AND PREVIOUS RESEARCH

Flexibility provided by labours is one of the major reasons usually argued for not automating manufacturing operations, especially in expensive labour markets in which cost based decisions may not support this argument. Different production environments can be found in which human work characteristics such as adaptability, responsiveness or learning cannot be efficiently substituted by machines.

In addition, a variety of production circumstances - high product variability, small batches production, customizable design or short product life cycles, among others- require a production line to be easily reconfigured in order to reduce setup costs.

Low inventory systems are another type of systems highly sensitive to variation (Schultz et al. 2003). Flexibility provided by labours is one way of counteracting this drawback and enabling gains from lower inventory costs to be effectively realized. Indeed, self-regulated labours work-pace is the main cause

found by Shultz et al. (1999) to explain the empirical observation presented in previous works (Schonberger 1982): low inventory systems with manual operations do not present as large throughput losses due to blocking and starvation as expected from an analysis in which human performance is modelled in a mechanistic way.

Assembly and disassembly are other production areas that largely rely on human involvement due to high investment costs in automation (Bley et al. 2004). These authors also refer to flexibility and reconfigurability as mandatory needs for ensuring competitiveness within an environment of growing demand on product variety and shortening lot sizes.

Traditionally, analysis of production systems has been mainly focused on technical aspects such as machines, buffers or transportation elements (Baines and Kay 2002). Human resources are introduced in the same way as machines and variation sources related to ergonomics or behaviour are ignored (Neuman and Medvo 2009). However, evidence supports that human performance variability differs from that of machines in several ways (Powel and Shultz 2004). Humans behave as state-dependant resources with the capability to readjust their work-pace depending on the circumstances. Also dynamic changes in the working rate are related to factors such as experience, aging, time of day and other external factors (Baines and Kay 2002). Although processing rates of machines might be satisfactorily modelled by their cycle time and failures distributions, a detailed model of human performance should include both dynamic and state dependant effects. Baines et al. (2004) show how simulation results change once certain dynamic effects are taken into account. Powel and Shultz (2004) demonstrate how a flow line performance depends on the presence of self-regulated behaviour.

Knowledge from ergonomics and behavioural science should be incorporated into traditional operations research models for a proper modelling of human factors. In spite of intensive research has been conducted in each one of these areas separately, their interface with operations research has received less attention in the literature; several authors call for further research to be conducted (Neuman et al. 2006, Schultz et al. 2010).

The majority of the papers published so far deal with the integration of models from either ergonomics or behavioural science. For instance, Neuman et al. (2009) have incorporated factors such as operator's autonomy for resting, individual differences and operators capacity in a discrete events simulation experiment. Their results show significant effects on throughput based on variability levels. Elkosantini and Gien (2009) and Riedel et al. (2009) have incorporated cognitive models for labours decision making in simulation models. These authors conduct feasibility studies and provide guidelines on how to implement them. However, they both point out the need to study the effective actual application of their approaches to real cases as well as the necessity of a proper model validation.

Another approach found in the literature is the execution of experiments in laboratory manufacturing settings. Laboratory experimentation is a common research tool in behavioural science, although we have been able so far to find only three papers that adopt this approach when studying the interaction between technical and behavioural elements in manufacturing. Schultz et al. (1998, 1999, 2003) executed experiments on human performance effects in low inventory systems and work-sharing. They arranged a laboratory flow line that consisted of three serial operations. The tasks consisted of introducing codes in a software application representing customer orders. Experiment subjects were high school students. Another experiment is presented in the paper of Bendoly and Prietula (2008). In this case the process consisted of a single operation in which subjects had to solve TSP instances by means of a software application. Subjects were recruited among students in a business school and thus they had a different profile compared to those in the Schultz's experiment. In both cases the tasks have only a mental workload. Physical workload is negligible, what makes an outstanding difference with many manufacturing environments.

In this paper we present a system designed for conducting research on the effects of human variability in manufacturing and for validation of human performance models. The approach consists of arranging an experimental manufacturing setting in which product and process related variability is kept under control. Thus, human resources variability can be isolated and studied in deep. Comparison between the experimental system output and a virtual simulation model allows for analysing the effects on system behaviour and the errors incurred by the modelling approach.

The system can also be applied for training purposes. Realistic simulation case studies can be proposed to students who can take part as either operators or simulation practitioners. Process improvement tools can be tested and put into practice. This method for teaching in simulation has the benefits of learning by doing. Students may develop greater skills for applying simulation tools. They also gain

insight into how to properly model a system and how to validate results within a controlled environment in which all the relevant factors can be taken into account.

Section 2 describes the systems conceptual design and elements. Possible variability sources are introduced along with an explanation on how to manage them. Section 3 presents an initial experiment conducted based on a roofing slates manufacturing process. Industrial Engineering students have taken part as operators and simulation practitioners. In section 4 some preliminary results obtained from the experiment are shown and discussed. The paper finishes with some concluding remarks.

2. SYSTEM DESIGN

2.1. The process

The designed process has been inspired by a manufacturing plant that produces roofing slates elements (del Rio Vilas et al. 2009). It is a labour intensive process characterized by high levels of product, process and resources variability. Previous research has shown important individual differences in performance and how productivity gains can be achieved when improving ergonomic conditions (Rego et al. 2010).

The experimental manufacturing process consists of five tasks arranged in a closed loop. Four of them constitute the analysed process and the fifth one is disposed in order to close the loop preventing from recirculating starvation or blocking events. The fifth task is converted into an events horizon by means of a security stock of input parts which would be consumed in case the production output was temporary incapable of providing enough input.

Process input and output products are the same, i.e., lots of a fixed amount of slates. The size of these lots will be noted as N_E . Slates are grouped into three types according to two attributes. A fraction p_R of the slates are printed with a red mark on them and the rest of them (fraction $p_G = 1 - p_R$) with a green one. Green slates are divided into two sizes, large size elements with dimensions 32x22mm and small size elements with dimensions 30x20mm and 27x18mm. These formats correspond to the main commercial formats traded by the company. The fraction of large size slates within green type will be noted as p_L and the fraction of small size type p_S . Green slates also display an alphanumeric code printed on their surface made up by two letters and one number. Input lots contain a sequence of slate types randomly generated according to the proportions defined before. Consecutive realizations of the selected slate type are independent between them.

The first task is the classification of slates according to their colour. It is performed in the so-called workstation 1 (WS1). Classified items are batched into lots of size N_R for red slates and N_G for green slates. Every time that a lot is passed to the next station the operator registers it in a software application

called WS1_Register by pressing either the red or green lot corresponding key.

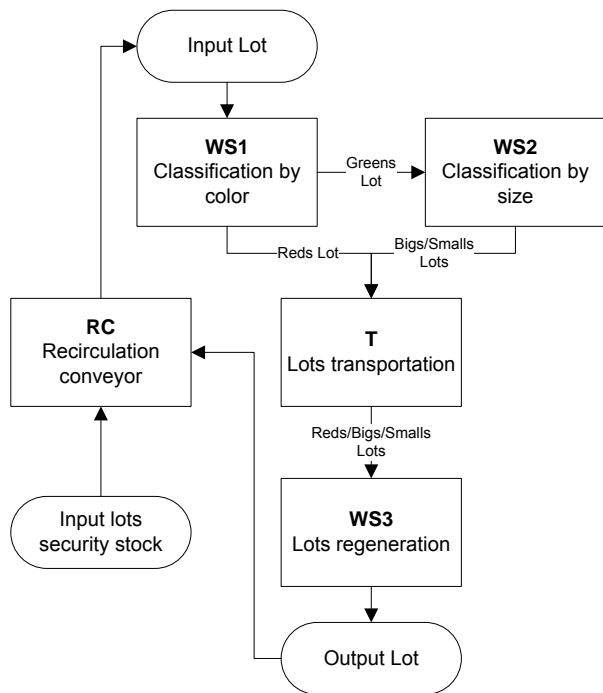


Figure 1: Process diagram.

The second task is performed in workstation 2 (WS2). It consists of the measure and classification of the green slates according to their size. Slates are taken one by one and measured either by means of a reference mark printed on the workplace or at a glance once the operator has acquired experience. Then the slate code is typed on a computer and registered by the application WS2_Register. The slate is finally piled in the corresponding lot upon size. Errors in either typing or classification are penalized so that demand on worker's attention is the highest within the process.

The third task in the process is a transportation one. Classified lots from workstations 1 and 2 are carried up to the workstation 3. A default parking location has been established at an intermediate point between WS1 and WS2 and marked on the floor.

The fourth task has the function of regenerating the input lots for the process. A random sequence of N_E slate types is generated and printed in a monitor by the WS3_Register application. Once a lot is completed it is pushed to a recirculation conveyor which acts as both the source and the sink for the rest of the process. Each time that a lot is pushed, it is registered in the application by pressing a key.

The fifth task is disposed in order to make the WS1 arrival process independent from the WS3 state. Thus the closed loop setting results would not differ from those of an open process. The workplace is functionally equivalent to a conveyor belt in which input lots are moved from WS3 back into the source slot. An auxiliary reserve of input lots is placed beside this station for use in cases of lack of output lots from WS3. It is a supervisory and control oriented stage which

plays an important role in standardizing process conditions and restricting undesired forms of variability. Lot arrivals to WS1 are registered in a control application called Source_Register which also provides functions for managing experimental runs such as time control or workers assignments to workplaces. It will not be analysed as part of the process along with the rest of the tasks.

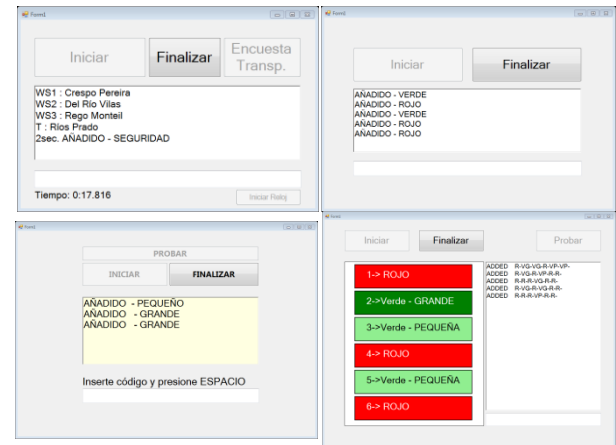


Figure 2: Source_Register, WS1_Register, WS2_Register and WS3_Register screenshots.

A process variant was designed by enabling work-sharing between transporter and WS2. When this collaborative mode is enabled, the transporter assists WS2 labour by typing registries on the computer. Then WS2 operator focuses only on classifying and moving slates whilst he dictates the alphanumeric codes to his/her teammate, so that cycle time is severely shortened. Meanwhile work-sharing is taking place, transporter cannot attend transportation orders from WS1 to WS3 and thus a trade-off between these two operation modes is created.

Tasks design was intended to result in different types according to the degree of physical and mental workload. Table 1 shows a characterization performed by the research team members.

Table 1: Tasks characterization

Task	Physical workload	Mental workload
WS1	Moderate	Moderate
WS2	Moderate	High
T	High	Low
WS3	Moderate	Low

2.2. System layout

The production line was built in the Industrial Engineering laboratory of the Escola Politecnica Superior of Ferrol. Four tables were arranged in line and a fifth one was placed nearby for serving as a security buffer of input lots. Slots were printed on the tables in order to establish fixed locations for working and buffering. Each workstation counts with a computer running the corresponding application. The computers

are connected to a LAN so that they can connect to a MySQL server for storing the registered data. Figure 3 shows a floor plan of the setting. Table 2 shows the function of each slot. Number of parts in them is constrained in order to simulate capacitated buffers.

Table 2: Slots in layout

Slot code	Parts Capacity	Function
S1	1 Input Lot	Pick up point for input lots.
S2	1 Input Lot	Input lots pick up point for operating under bad ergonomic conditions.
WS1	1 Input Lot	Working slot for WS1.
RB	1 Reds Lot	Batching of red slates.
GB	1 Greens Lot	Batching of green slates.
GTB	1 Greens Lot	Connection buffer of greens lots.
WS2	1 or unrestricted Greens Lot	Working slot for WS2.
LGB	1 Large Greens Lot	Batching of large green slates.
SGB	1 Small Greens Lot	Batching of small green slates.
RTB	1 or 2 Reds Lot	Reds lots input buffer to transporter.
LGTB	1 or 2 Large Greens Lot	Large greens lots input buffer to transporter.
SGTB	1 or 2 Small Greens Lot	Small greens lots input buffer to transporter.
RR	Unrestricted Red Slates	Buffer of red slates waiting to be recirculated.
RLG	Unrestricted Large Green Slates	Buffer of large green slates waiting to be recirculated.
RSG	Unrestricted Small Green Slates	Buffer of small green slates waiting to be recirculated.
WS3	1 Input Lot	Working slot for WS3.
RC	5 Input Lots	Recirculation conveyor.

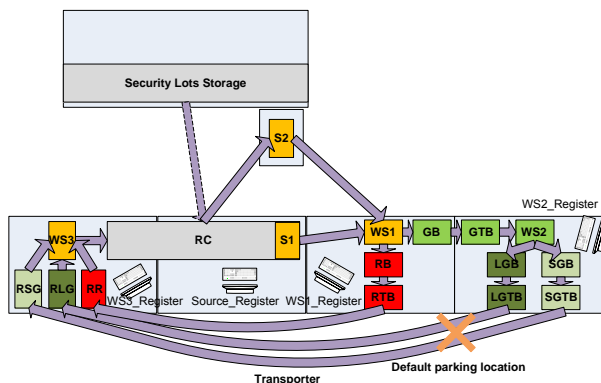


Figure 3: Experimental setting layout.



Figure 4: Experimental Setting in the Industrial Engineering Laboratory.

2.3. Sources of variability

Variability was analysed under a PPR (product, process and resource) approach (del Rio et al. 2009). During the experiment design phase, possible sources of variation were discussed and actions taken in order to avoid non-desirable ones, to control those ones subject of analysis and to trace those considered as non-controllable.

Process variability was limited by defining standardised task procedures which covered the sequence of steps to be performed, the permitted actions and the priorities. Penalties in a reward function together with supervisory mechanisms were put in place. Therefore, operators could not be benefited by deviating from them. The only three exceptions made to this rule were:

1. The transporter was given freedom to choose what lots to prioritize. This was allowed aiming at sampling the different prioritization rules intuitively developed by the subjects.
2. WS2 operators were given freedom on what subtask to perform first: classifying a slate or typing its code on the computer. Although this degree of freedom may increase the effect of individual differences on performance, it is representative of the variability encountered in most of real settings. It also allows the experiment subjects for working in a more comfortable way to them.
3. Under work-sharing enabled, transporter was given freedom to choose when to offer support to WS2 operator and when to stop the cooperation. WS2 operator could refuse the assistance.

However, these sources of variation are human-driven so they will be treated together with the other human resources forms of variability.

Product variability has been intentionally introduced by means of the random composition of input lots. Depending upon their colour, slates flow directly from WS1 into the transporter or they do so through WS2. This sort of variability is present in many

real systems that combine the production of products with different processing steps.

This kind of product variability affects line balancing. WS2 task is slower than WS1 in terms of processing time per slate. Hence, when p_R is high, WS1 is the most congested workstation and the low arrival rate of parts to WS2 causes it to have idle times. On the other hand, when p_R is low there are plenty of green slates to be processed in WS2 and since this is a slower task, it causes WS1 to be blocked. Thus, the system bottleneck location will depend on p_R and it can be altered by simply modifying its value. Furthermore, random temporary variations of average p_R will cause the bottleneck to dynamically change from WS1 to WS2 and vice versa. Although this behaviour increases the variability levels of throughput rates – which may cause human-driven variability to be harder to detect –, it is a very desirable feature when analysing state-dependant effects on human performance. Their impact in system performance is expected to be amplified by the higher variability in elements states.

Finally, human resources variability will be the main subject of study within this paper. Several sources of variability were considered taking into account previous published results.

1. Individual differences. Differences in motivation, skills, ergonomic fitness of the workstation among others, cause workers to perform differently when doing the same task. Some basic personal data was gathered in order to search for individual characteristics that might be correlated with performance. Subjects were asked about their age, height, weight, sex and physical activity.
2. Group differences. Interaction among individuals involves complex dynamics which might either favour or disfavour overall performance (Bendoly et al. 2010). As a way of limiting the group effect on performance, subjects were randomly assigned to workstations. This avoided that assignment to position choices distorted the results. However, other forms of group dependant variation could not be accounted for. For instance, the group response to a change in situational pressure might completely differ depending on whether the group is driven by the Abilene Paradox or Group Think (Bendoly et al. 2010).
3. Learning curves. No subject had ever performed this kind of work before. Hence, experience was simply recorded as the number of runs already done by the individual.
4. Time and day of week. Experiment sessions were executed either during the mornings or afternoons. Day of week varied as well. A randomized assignment of treatments to experimental units was expected to counteract its possible effect in the results.
5. Tiredness. The effect of tiredness in performance is a complex issue. It is also a variable hard to measure. Runs duration was set to 12 minutes in order to dispose a level of tiredness affordable by all the subjects. However, the high physical workload caused them to clearly experience it. It was measured by means of surveying at the end of each run.
6. Motivation. Motivational levels are another hard to measure variable. Different levels of effort by individuals were noticed by us during the experiment execution. Some questions in the enquiries were disposed at this purpose.
7. State-dependant behaviour. Analysing state-dependant behaviour requires collecting data of single realizations of task cycle times together with information of the system state. In order to do so, the data acquisition system was designed to report statistics in a virtual simulator-like fashion. By constructing a list of events occurred in the system, its state at each time could be tracked and therefore its relation to task cycle time studied.

2.4. Parameters setup

An initial production run was performed by the research team members as a means of characterising tasks time distributions. This data fed a simulation model implemented in Delmia Quest 5 R20 (Figure 4). The model was then used for adjusting parameters p_R , N_E , N_R , N_G .

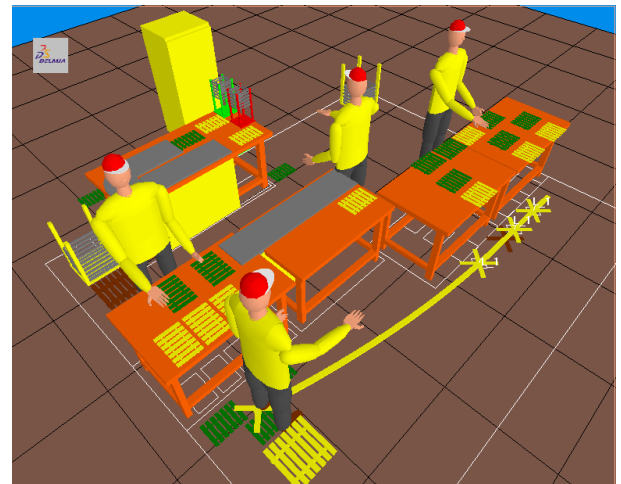


Figure 4: View of the simulation model implemented in Quest.

A solution in which all the workstations had high utilization rates and WS2 was the bottleneck was adopted. WS1 was intended to have a utilization close to WS2 in order to be able to turn it into the bottleneck by a small configuration change. Table 3 and 4 show the adopted configuration and the expected workstations utilization estimated by the model.

Table 3: Process parameters in the adopted solution

Parameter	Value
p_R	60%
N_E	6
N_R	3
N_G	3

Table 4: Process balance estimated by the simulation model (230 runs)

Operator	Utilization
Operator 1	91.74%
Operator 2	96.82%
Operator 3	77.34%
Operator 4	64.81%

3. THE EXPERIMENT

3.1. Subjects

One of the main aspects hampering the possibility of adequately conducting human factors experimentation is in fact the assured and convinced availability of human subjects. Doing so in real manufacturing environments might imply a set of negative consequences in terms of motivation and availability as well as economic and production implications

In this case, the experiment subjects were recruited among Industrial Engineering students of the third year in the Quantitative Methods for Industrial Engineering subject at the University of A Coruna. The contents of this subject span some common operations research methods such as non-linear optimization, meta-heuristics, queuing theory, discrete events simulation and decision theory. It is the first contact of students with the Operations Research field. Students were offered an alternative evaluation plan to the one traditionally followed consisting of a single final exam. A realistic case study in simulation was proposed for the simulation and optimization of the experimental setting. Students who took part in the activity could get half of their total mark upon the quality of their simulation and optimization analysis and their performance in a final experiment execution rated by means of a reward function. A total of eight teams were formed.

3.2. Design

The experiment was conducted in three phases. First phase was aimed at introducing the subjects to the process and the task procedures. It consisted of a single session of four production runs, each one five minutes long. No information regarding the process was given to them previously. Operators were randomly assigned to workplaces and rotated at each run. Thus all of them had a try on every task and reference cycle times could be computed.

The second phase comprised two sessions of three runs each. Runs were twelve minutes long. The manufacturing experiment was conducted during this phase. Eight teams by two sessions of three runs

provided with a total of forty-eight experimental units. Details on the experiment design are given below.

In the fourth session students were evaluated by means of a reward function dependant on throughput rates, work in process levels and errors committed. In this session students could modify selected system parameters: reds proportion (p_R), size of greens lots (N_G), assignment of operators to workstations and capacity of RT, LGT and SGT. Teams were ranked upon score and an additional mark incentive was given accordingly. Duration was set to fifteen minutes.

The experiment factors were selected according to a twofold objective. First goal was to analyse the effect of ergonomics in performance. A single bi-level factor was introduced to this end. Second goal was to study the effect of different manufacturing approaches on behaviour. Four factors were introduced concerning connection buffers size, system state perception, work-sharing and incentives approach. Table 5 shows these factors and their levels.

Table 5: Factors levels

Factor	Reference level (0)	Alternate level (1)
Ergonomics	Good ergonomic conditions in WS1 (Source S1)	Poor ergonomic conditions in WS1 (Source S2)
Inventory	High – Capacity of WS2: ∞ , RT: 2, LGT: 2, SGT: 2.	Low – Capacity of WS2: 1, RT: 1, LGT: 1, SGT: 1.
Perception	Full – Operators have visibility of the whole setting	Restricted – A opaque panel is disposed between WS1 and WS2
Work-sharing	Disabled	Enabled
Approach	Throughput – Reward function dependent on throughput rate	Quality – Reward function dependent on errors committed

A full factorial design was dismissed because of the limited number of units available and not all the possible interactions among factors were regarded of interest. Eight treatments were selected aiming at testing the interactions related to the posed hypothesis. They were divided in three areas of interest: ergonomics, technical elements and incentives. Table 6 shows the treatments. Treatment 1 was established as reference treatment for comparison. Treatment 2 was introduced for testing the ergonomic factor effect and treatment 3 for testing the effect of incentives approach. Treatments 4 to 8 concern technical aspects. Due to system balance, combined work-sharing and low inventory settings were dismissed. The setup cost of starting and ending

cooperation made work-sharing unfavorable under this circumstances. A complete factorial design was employed for the remaining factors.

Table 6: Treatments

Factor	Treatment							
	1	2	3	4	5	6	7	8
Ergonomics	0	1	0	0	0	0	0	0
Inventory	1	1	1	1	0	1	1	0
Perception	1	1	1	0	1	1	0	0
Work-sharing	1	1	1	0	0	0	1	0
Approach	0	0	1	0	0	0	0	0

Factors were randomly assigned to the experimental units under the constraint of not assigning a treatment more than once to each group.

4. PRELIMINARY RESULTS

Experiments were conducted between March and May of 2011. No major incidents happened but for some eventual mistakes committed by the students when following the working procedures. These random errors were recorded as an error rate for each experiment. No significant effect from this error rate in throughput was found.

Data was collected from the software applications, videos and enquiries provided to the students. A preliminary analysis of the data recorded by the applications is now presented.

Applications records provided with lists of events occurred in the system. They spanned entries of lots in the system, exits from WS1, processed items in WS2 and exits in WS3. Thus it was possible to build a basic list of events happened in WS1, WS2 and the system as a whole. Then it was used to plot a graph of buffer contents and to calculate average residence times in the same fashion as the results that can be obtained from simulation software. A demonstration of the conducted results analysis is provided below.

Figure 5 displays the plot of slates in WS1 as a function of time. It includes the contents of buffers WS1, GB and RB plus the slates that are been processed by operator 1. Figure 6 displays a plot of the total residence time in the system as a function of the number of lot exited from WS3. The two observed leaps in residence times correspond to lots that suffered of a delay in WS3 due to starvation.

Tables 7 to 9 summarize the results of throughput rates achieved under the different treatments in sessions two and three. A regression model was fitted by least squares for the output rates of WS1, WS2 and WS3 containing terms for experience effect (numbering the production runs from 1 to 6) and several technical

aspects. It can be noticed that only the experience factor was significant on the overall throughput rate (Table 9). This result contrasts with the expected benefit from increased buffer sizes or work-sharing that was expected. Actually, it was WS2 the one expected to be the one most favoured by the higher inventory conditions. WS2 was the process bottleneck, so by disposing an infinite buffer before it, random starvation caused by WS1 was removed and thus throughput should have been improved. However, evidence does not support this argument. Table 8 shows a high p-value for the low inventory effect, indicating that no significant effect was found. Furthermore, the terms sign is positive, which contrasts with the expected effect. Table 10 presents the effect expected by means of the simulation model. It can be seen that lowering buffers capacity significantly reduces throughput when human resources are modelled in a mechanistic way. Further analysis of results might provide more information on the roots of the observed deviance.

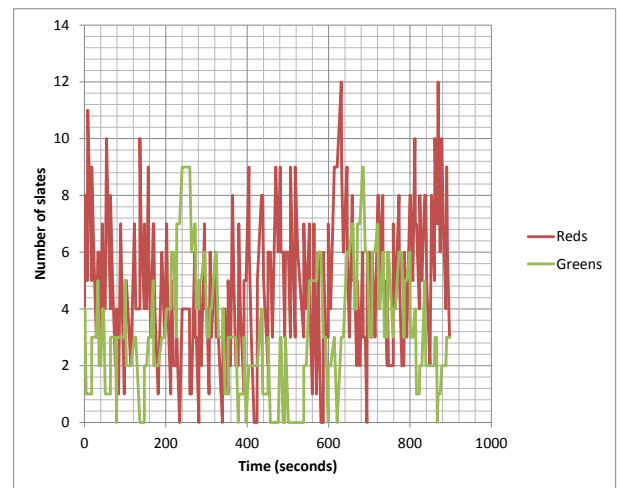


Figure 5: Slates contents in WS1.

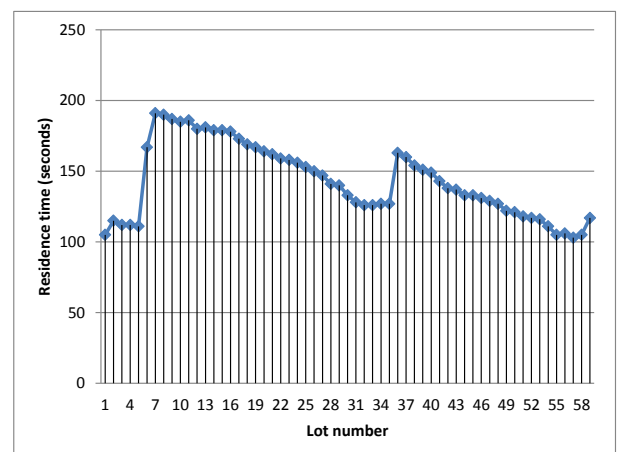


Figure 6: Total residence time of output lots in WS3.

Table 7: Regression model for WS1 output

Coefficient	Estimate	p-value
(Intercept)	0.135	0.000
Experience	0.005	0.004
Low Inventory	-0.024	0.004
Reduced Perception	-0.003	0.753
Work-sharing On	-0.017	0.056
Low Inventory & Reduced Perception	0.006	0.607
Reduced Perception & Work-sharing On	0.001	0.955
R Squared		0.3682
F Statistic		3.982
F test p-value		0.003137

Table 8: Regression model for WS2 output

Coefficient	Estimate	p-value
(Intercept)	0.124	0.000
Experience	0.004	0.003
Low Inventory	0.006	0.395
Reduced Perception	0.003	0.637
Work-sharing On	-0.001	0.937
Low Inventory & Reduced Perception	0.001	0.940
Reduced Perception & Work-sharing On	0.005	0.626
R Squared		0.4312
F Statistic		5.179
F test p-value		4.823E-4

Table 9: Regression model for WS3 output

Coefficient	Estimate	p-value
(Intercept)	0.042	0.000
Experience	0.002	0.002
Low Inventory	-0.003	0.423
Reduced Perception	-0.001	0.855
Work-sharing On	-0.001	0.799
Low Inventory & Reduced Perception	0.003	0.550
Reduced Perception & Work-sharing On	-0.001	0.834
R Squared		0.3039
F Statistic		2.983
F test p-value		0.01642

Table 10: Simulation results for inventory effect on WS2 and Z-test for differences in means.

WS2 Simulation (230 runs)		
Setting	Mean	Std. Deviation
Low inventory	31.44	1.18
High inventory	31.00	1.29
Z statistic		0.3039
p-value		2.983

5. CONCLUDING REMARKS

An experimental manufacturing system has been designed and built in the Industrial Engineering

laboratory of the Escola Politecnica Superior de Ferrol. Sources of process variability other than those originated from human resources have been set up, monitored and kept under control. Thus, controlled variability is then characterized by means of a discrete events simulation model. A redundant data acquisition system has been implemented so that system events are traced and stored in a relational database in a computer simulator-like fashion. Real system production runs statistics are then compared with virtual model ones. Deviations in results are due to effects of human variation. Human performance models can be validated by introducing them in the simulation model and testing whether they actually improve the model prediction capability.

The system can also be applied for training in both simulation and process improvement tools. It provides a controlled environment in which the effect of factors of interest can be studied in depth and isolated from other factors. Data can be collected in large samples hard to obtain in many businesses processes. Model validation can also be performed in detail, directly comparing statistics from the real process with those from the virtual simulator. A major strength of this system is that subjects can take part as both operators and analysts, thus acquiring the two different points of view.

A twofold objective experiment has been conducted with cooperation from Industrial Engineering students of the University of A Coruna. A process inspired by a roofing slates manufacturing company was designed and installed in a laboratory setting. A joint research and educational activity was carried out aiming at testing the effect of ergonomics and organizational factors in manufacturing as well as a practical teaching in discrete events simulation. The students were organized in eight teams. Three initial sessions were run in which they had to work as process operators. Data was recorded and provided to the students once the initial sessions were ended. Then they had to simulate the process and to optimize certain proposed parameters. A final session was run in which the groups implemented their proposals and their results were compared and rewarded in a competitive fashion.

Some preliminary results have been obtained testing the effect on throughput of the studied factors. These factors span individual differences among groups, experience, buffers capacity, work-sharing, process state perception, ergonomic conditions and approach to either quality or quantity. Significant effects have been showed for the inter-groups variation, experience and ergonomic conditions. This is consistent with most of operations research literature on the effects of learning and either individual or group differences. No significant effects from changes in buffers capacity, work-sharing, process state perception and approach could be proved. This result contrasts with that obtained from a simulation model of the process in which human resources were introduced in a mechanistic way. Buffer capacity increase and work-sharing were expected to provide a significant increase in throughput. Although

several limitations have been identified when extending these results to real manufacturing environments, they are consistent with the findings by Schultz et al. (1999, 2003). Human behaviour effects seem to be counteracting the expected benefits of increasing buffer capacities and enabling work-sharing. Further research is needed for assessing the validity of these findings and to deepen in the explanation of the causes.

REFERENCES

- Baines T.S., Kay J.M., 2002. Human performance modelling as an aid in the process of manufacturing system design: A pilot study. *International Journal of Production Research*, 40(10), 2321-2334.
- Baines, T., Mason, S., Siebers, P., Ladbrook, J., 2004. Humans: the missing link in manufacturing simulation? *Simulation Modelling Practice and Theory*, 12, 515–526.
- Bendoly, E., Prietula, M., 2008. In “the zone”. The role of evolving skill and transitional workload on motivation and realized performance in operational tasks. *International Journal of Operations & Production Management*, 28 (12), 1130-1152.
- Bendoly, E., Croson, R., Goncalves, P., Schultz, K., 2010. Bodies of Knowledge for Research in Behavioral Operations. *Production and Operations Management*, 19 (4), 434–452.
- Bley, H., Reinhart, G., Seliger, G., Bernardi, M., Korne, T., 2004. Appropriate Human Involvement in Assembly and Disassembly. *CIRP Annals - Manufacturing Technology*, 53 (2), 487–509.
- Elkosantini, S. and Gien, D.(2009) 'Integration of human behavioural aspects in a dynamic model for a manufacturing system', *International Journal of Production Research*, 47: 10, 2601 — 2623
- Neumann W.P., Winkel J., Medbo L., Magneberg R., Mathiassen S.E., 2006. Production system design elements influencing productivity and ergonomics: A case study of parallel and serial flow strategies. *International Journal of Operations & Production Management*, 26(8), 904-923.
- Neumann W.P., Medbo P, 2009. Integrating human factors into discrete event simulations of parallel flow strategies. *Production Planning & Control*, 20 (1), 3-16.
- R Development Core Team, 2005. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.
- Rego Monteil, N., del Rio Vilas, D., Crespo Pereira, D., Rios Prado, R., 2010. A Simulation-Based Ergonomic Evaluation for the Operational Improvement of the Slate Splitters Work, *Proceedings of The 22nd European Modeling & Simulation Symposium*, pp. 191-200. Fez, Morocco.
- Riedel, R.; Mueller, E.; von der Weth, R.; Pflugradt, N.; , "Integrating human behaviour into factory simulation- a feasibility study," *Industrial Engineering and Engineering Management*, 2009. IEEM 2009. IEEE International Conference on , vol., no., pp.2089-2093, 8-11 Dec. 2009
- del Rio Vilas, D., Crespo Pereira, D., Crespo Mariño, J.L., Garcia del Valle, A., 2009. Modelling and Simulation of a Natural Roofing Slates Manufacturing Plant. *Proceedings of The International Workshop on Modelling and Applied Simulation*, pp. 232-239. September 23-25, Puerto de la Cruz (Tenerife, Spain).
- Powell, S.G., Schultz, K. L., 2004. Throughput in Serial Lines with State-Dependent Behavior. *Management Science*, 50 (8), 1095–1105.
- Schonberger, R.J., 1982. *Japanese Manufacturing Techniques: Nine Hidden Lessons in Simplicity*. The Free Press, New York.
- Schultz, K.L., Juran, D.C., Boudreau, J.W., McClain, J.O., Thomas, L.J., 1998. Modeling and Worker Motivation in JIT Production Systems. *Management Science*, 44 (12), Part 1 of 2, 1595-1607.
- Schultz, K.L., Juran, D.C., Boudreau, J.W., 1999. The Effects of Low Inventory on the Development of Productivity Norms. *Management Science*, 45 (12), 1664-1678.
- Schultz, K.L., McClain, J.O., Thomas, L.J., 2003. Overcoming the dark side of worker flexibility. *Journal of Operations Management*, 21, 81–92.
- Schultz, K.L., Schoenherr, T., Nembhard, D., 2010. An Example and a Proposal Concerning the Correlation of Worker Processing Times in Parallel Tasks. *Management Science*, 56 (1), 176–191.

AUTHORS BIOGRAPHY

Diego Crespo Pereira holds an MSc in Industrial Engineering and he is currently studying for a PhD. He is Assistant Professor of the Department of Economic Analysis and Company Management of the University of A Coruna (UDC). He also works in the GII of the UDC as a research engineer since 2008. He is mainly involved in the development of R&D projects related to industrial and logistical processes optimization. He has also developed projects in the field of human factors affecting manufacturing processes.

David del Rio Vilas holds an MSc in Industrial Engineering and has been studying for a PhD since 2007. He is Adjunct Professor of the Department of Economic Analysis and Company Management of the UDC and research engineer in the GII of the UDC since 2007. Since 2010 he works as a R&D Coordinator for two different privately held companies in the Civil Engineering sector. He is mainly involved in R&D projects development related to industrial and logistical processes optimization.

Rosa Rios Prado works as a research engineer in the GII of the UDC since 2009. She holds an MSc in Industrial Engineering and now she is studying for a PhD. She has previous professional experience as an Industrial Engineer in an installations engineering company. She is mainly devoted to the development of transportation and logistical models for the assessment of multimodal networks and infrastructures.

Nadia Rego Monteil obtained her MSc in Industrial Engineering in 2010. She works as a research engineer at the Engineering Research Group (GII) of the University of A Coruna (UDC) where she is also studying for a PhD. Her areas of major interest are in the fields of Ergonomics, Process Optimization and Production Planning.

Adolfo Lamas Rodriguez graduated from the University of Vigo in 1998. He holds an MSc and a PhD in Industrial Engineering. He combines his research activities in the GII and his position as a senior engineer in the Spanish leading shipbuilding company *Navantia*. He is also Associate Professor in the University of A Coruna.

MODELLING AND SIMULATION OF A WIRELESS BODY AREA NETWORK PROTOTYPE FOR HEALTH MONITORING

Y. Callero^(a), R. M. Aguilar^(b)

^{(a)(b)} Department of Systems Engineering and Automation, and Computer Architecture.
University of La Laguna. Spain

^(a)ycallero@isaatc.ull.es, ^(b)raguilar@ull.es

ABSTRACT

Recent advances in technology, integrated circuits, and wireless communication have allowed the realization of Wireless Body Area Networks (WBANs). Such networks feature smart sensors that capture physiological parameters from people and can offer an easy way for data collection. WBANs also need suitable interfaces for data processing, presentation, and storage for latter retrieval. They are widely used for ubiquitous healthcare, entertainment, and military applications.

Power limitations are one of the weaknesses of the systems that use WBANs. In that sense, this paper proposes a WBAN prototype that uses the ultra-low energy communication protocol ANT to reduce de WBAN power consumption. The simulation of this architecture for health monitoring is presented to validate the availability of the architecture.

Keywords: wireless body area networks, health monitoring, ANT.

1. INTRODUCTION

The most remarkable life-changer in the past decade has been the introduction and the rapid mass adoption of wireless mobile digital devices. The ways in which we listen to and acquire music, e-mail and communicate via phone, access the Internet, and read books and periodicals electronically have all been radically transformed. In stark contrast, the ways in which diseases are monitored and treated have remained relatively static. But, in spite of this, one of the mayor area where the lack of plasticity of the medical profession and health care system in the face of new technology and information is about to be challenged, is the wireless technologies area (Topol, 2010).

One of the reasons of the growth in the adoption of wireless technologies in the health care is the evolution of the body sensors. Body sensors are small devices that are able to measure and control human body parameters. Some recent projects take advantages of these small sensors in overall body monitoring. In this context, emerged a new network approach known as body area networks (BANs). This new infrastructure is comprised of several sensor nodes placed throughout a human body. Each sensor measures specific

physiological or biological parameters. Evolutions on microelectronics construction enabled endow these sensor nodes with the ability to communicate outside world using a wireless module component.

Table 1. Top ten targets for wireless medicine

Disease	Number affected (millions) in USA	Metrics potentially measured
Alzheimer	5	Vital signs, patients' location, activity, balance
Asthma	23	RR, FEV1, blood oxygen level, air quality, pollen count
Breast cancer	3	Presence of suspicious mass, as detected by ultrasounds self-exam
COPD	10	RR, FEV1, blood oxygen level, air quality
Depression	21	Medication compliance, activity, communication
Diabetes	24	Blood glucose level, calories ingested
Heart failure	5	Cardiac pressures, body weight, BP, fluid status
Hypertension	74	Continuous BP, medication compliance
Obesity	80	Weight, blood glucose levels, calorie intake and output, activity
Sleep disorders	40	Sleep phases, sleep quality, RR, apnea, vital signs, blood oxygen level, heart rhythm

Remote monitoring of human body is now possible, using WBANs to access data collected by the sensors. Sensors are implanted or placed in human body to

monitor some behaviors or pathologies, and help patients maintain their health through biofeedback phenomena such as temperature analysis, blood pressure detection, Electrocardiography (ECG), Electromyography (EMG), among others (Pereira, Caldeira, & Rodrigues, 2010). (Topol, 2010) exposed top ten targets of chronic diseases that can be monitored via wireless sensors. These targets are showed in Table 1.

One of the biggest problems regarding sensor networks is power consumption, which is greatly affected by the communication between nodes (Sarpeshkar, 2006)(Adloo, Deghat, & Karimghae, 2009). One solution to this issue is the introduction of aggregation points to the network. The aggregation points reduce the messages exchanged between nodes and saves energy. Usually, they are regular nodes that receive data from neighboring nodes, perform some processing, and then forward the data to the next node. Another way for energy saving is setting the nodes into sleep mode if they are not needed and wake them up when required. Energy saving is one of the most challenge that engineers face when they create wireless body area networks (WBANs).

In this sense, the University of La Laguna Simulation group (SIMULL) proposes the use of an ultra-low energy communication protocol to reduce the WBAN power consumption. This paper presents a study of the availability the low power consumption WBAN trying to model and simulate its behavior. First of all, the paper presents a smart revision of the state of the art of remote health monitoring. After that, a wireless communication protocols review is presented. In next sections, the WBAN prototype modeling and simulation is presented. Finally, conclusions and future work are shown.

2. STATE OF ART

With the growing needs in ubiquitous communications and recent advances in very-low-power wireless technologies, there has been considerable interest in the development and application of wireless networks around humans. A wireless body area network (WBAN) is a radio frequency (RF)-based wireless networking technology that interconnects tiny nodes with sensor or actuator capabilities in, on, or around a human body. Typically, the transmissions of these nodes cover a short range of about 2 m. Complementing wireless personal area networks (WPANs), in which radio coverage is usually about 10 m, WBANs target diverse applications including healthcare, athletic training, workplace safety, consumer electronics, secure authentication, and safeguarding of uniformed personnel. A WBAN can also be connected to local and wide area networks by various wired and wireless communication technologies (Cao, Leung, Chow, & Chan, 2009).

WBAN applications can be categorized based on the type of sensors/actuators, radio systems, network

topologies, and use cases. We enumerate here several pioneer healthcare WBAN research projects, as well as platforms for human-computer interaction (HCI) applications.

2.1. WBANs for healthcare

WBANs extend conventional bedside monitoring to ambulatory monitoring, providing a point of care to patients, the elderly, and infants in both hospital-based and home-based scenarios. Monitoring, autonomous diagnostic, alarm, and emergency services, as well as management of electronic patient record databases can all be integrated into one system to better serve people.

The CodeBlue project at Harvard University (Shnayder, Chen, Lorincz, Fulford-Jones, & Welsh, 2005) considers a hospital environment where multiple router nodes can be deployed on the wall. All nodes use the same ZigBee radio. Patients/caregivers can publish/subscribe to the mesh network by multicasting; there is no centralized or distributed server or database for control and storage. Localization functionality is provided by MoteTrack with an accuracy of 1m, based on the same radio. As a result of mobility and multihop transmissions, the system experiences considerable packet loss and is limited to 40 kb/s aggregate bandwidth per receiver.

Based on the CodeBlue architecture, the Advanced Health and Disaster Aid Network (AID-N) is being developed at Johns Hopkins University (Gao et al., 2007) for mass casualty incidents where electronic triage tags can be deployed on victims. Additional wireless capabilities (e.g., Wi-Fi and cellular networks) are introduced to facilitate communications between personal servers and the central server where data are stored. Furthermore, a web portal is provided to multiple types of users, including emergency department personnel, incident commanders, and medical specialists. A Global Positioning System (GPS) module is employed for outdoor localization, while a MoteTrack system is designed for tracking indoors. However, patients have mobility constraints due to the lack of routers in the network, and a very limited number of sensor nodes can be put on each patient because of the limited bandwidth.

The Wearable Health Monitoring Systems (WHMS) is being developed at the University of Alabama (Milenkovic, Otto, & Jovanov, 2006) and targets a larger-scale telemedicine system for ambulatory health status monitoring. Unlike CodeBlue and AID-N, WHMS has a star-topology network for each patient, which is connected via Wi-Fi or a cellular network to a healthcare provider. The personal server, implemented on a personal digital assistant (PDA), cell phone, or personal computer (PC), coordinates the data collection from sensor nodes using a time-division multiple access (TDMA) mechanism, provides an interface to users, and transfers data to a remote central

server. Physicians can access data via the Internet, and alerts can be created by an agent running on the server. However, the power consumption and cost associated with long-term data uploading can hamper system realization.

2.2. WBANs for HCI

Traditional computer interfaces, like keyboards, mice, joysticks, and touch screens, are all replaceable by potential WBAN devices capable of automatically recognizing human motions, gestures, and activities. Disabled people can benefit from novel WBAN platforms based on a series of miniature sensors. The intra-body communications (IBC) applications proposed in (Ruiz & Shimamoto, 2006) can be used to assist handicapped people. For example, an IBC enabled sensor embedded inside the shoes of a blind person can be used to send voice information such as the current location to him/her by an IBC enabled facility, such as a doorway or crosswalk. IBC enabled eyeglasses that can display texts, working with IBC enabled speakers, can help deaf people comprehend audio broadcast announcements.

Early research efforts at MIT Media Lab have produced MITHril (Pentland, 2004), a wearable computing platform that includes electrocardiography (ECG), skin temperature, and galvanic skin response sensors for wearable sensing and context-aware interaction. MITHril is not a real WBAN in that multiple sensors are wired to a single processor. A later version of this platform, MITHril 2003, extends MITHril to a multi-user wireless distributed wearable computing platform by utilizing Wi-Fi function available on PDAs (i.e., a PDA acts as a personal server and relays data of each person to a central station).

The Microsystems Platform for Mobile Services and Applications (MIMOSA) (Jantunen, 2008) is a research project involving 15 partners from eight different European countries to create ambient intelligence. MIMOSA's approach is similar to WHMS while it exclusively employs a mobile phone as the user-carried interface device. Wibree, later renamed Bluetooth Low Energy technology, and radio frequency identification (RFID) tags are used for connecting local sensor nodes. NanoIP and Simple Sensor Interface (SSI) protocols are integrated into MIMOSA to provide an application programming interface (API) for local connectivity and facilitate sensor readings.

The Wireless Sensor Node for a Motion Capture System with Accelerometers (WiMoCA) (Farella, Pieracci, Benini, Rocchi, & Acquaviva, 2008) project at several Italian universities is concerned with the design and implementation of a distributed gesture recognition system. The system has a star topology with all sensing nodes sending data to a non-sensing coordinator node using a TDMA-like approach, and the coordinator in turn relays the data to an external processing unit using

Bluetooth. The sensing modules, each made up of a tri-axial accelerometer, can be put on multiple parts of the body for motion detection. The radio modules of all nodes work in the 868 MHz European license-exempt band, with up to 100 kb/s data rate. A Java-based graphical user interface (GUI) at the processing unit side interprets the data stream for posture recognition.

3. WIRELESS CONNECTIVITY TECHNOLOGIES IN HEALTH CARE

Standards are an important part of health care systems. One of today's healthcare industries goals is to drive patient related information to be exchanged freely between the various systems in the continuum care.

One area of these systems is the health monitoring information, in which various types of sensors gather and send patient data. In today's health monitoring market, devices and software platforms are manufactured and developed by many different entities. In order for a health monitoring system to be able to support a wide range of vital signs, it is necessary to use a wide variety of measurement devices such as blood pressure monitors, glucose meters, heart rate monitors, weighing scales, ECG sensors and fall detection accelerometers. For each type of device there are a number of companies manufacturing them, but no company makes all of these devices. Therefore, it is necessary to work with many different suppliers in order to provide a complete range of measurement devices to its customers. Unfortunately, unlike other market areas such as the financial industry, healthcare industry is still in its primordial phases when it comes to defining interoperability standards. Particularly when it comes to wireless health monitoring sensor networks, there is yet to appear a definitive standard that is adopted by the majority of the industry. Different device suppliers use different communication technologies (Wi-Fi, Bluetooth, ZigBee, etc.) and even if they all used the same technology for communication they would still use different ways of structuring the messages transmitted over the transport technology. Without standards that define how to structure messages in order to share information and what technologies to use for communication, it will never be possible to truly achieve system interoperability. Fortunately, there are various attempts at defining what technologies and communication standards are to be used between the various products in the health care industry: industry players gather to develop and agree upon guidelines or profiles that define which standards to use on a given situation and how to combine them in order to achieve interoperability. Collaboration between these entities and the use of global standards is essential for the future of health monitoring systems. Nowadays, the communications standardization in health care is guided by two important coalitions: the Continua Health Alliance and the ANT+ Alliance, and their respective sensor wireless technologies of choice: ZigBee, Bluetooth and ANT.

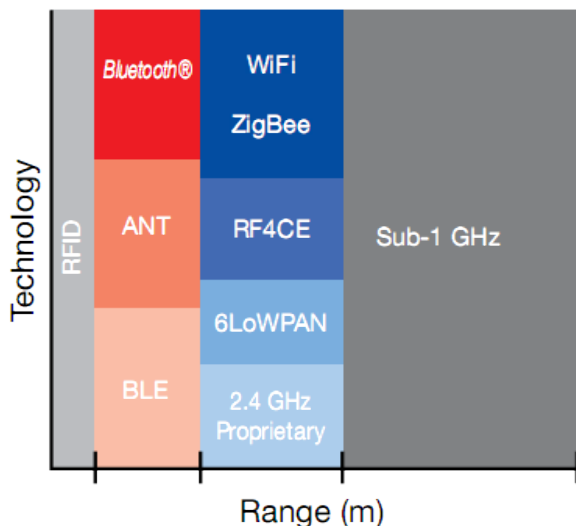


Figure 1 Wireless communication protocols classified by transmission range (Texas Instrument, 2011).

ZigBee is a standards-based technology for remote monitoring, control and sensor network applications. The ZigBee standard was created to address the need for a cost-effective, standards-based wireless networking solution that supports low data-rates, low-power consumption, security, and reliability. ZigBee can be used in any monitoring and control application that requires a wireless link:

- Home, building and industrial automation
- Energy harvesting
- Home control/security
- Medical/patient monitoring
- Logistics and asset tracking
- Sensor networks and active RFID
- Advanced metering/smart energy
- Commercial building automation

Bluetooth wireless technology is one of the most

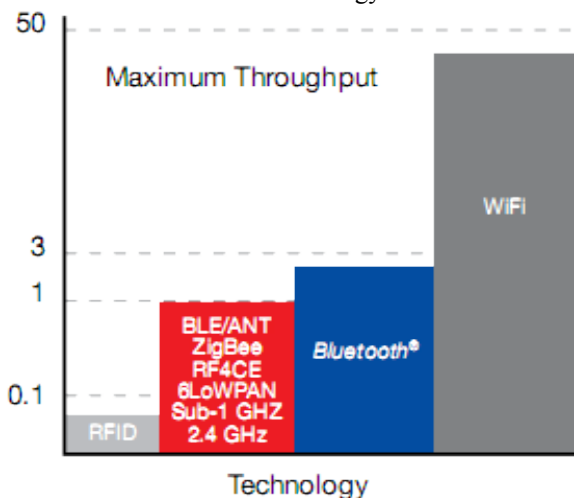


Figure 2 Wireless communication protocols classified by maximum throughput (Texas Instrument, 2011).

prominent short-range communications technologies with an installed base of more than three billion units. Bluetooth is intended to replace the cables connecting portable and/or fixed devices while maintaining high levels of security, low power and low cost. Bluetooth low energy technology (BLE) offers ultra-low power, state-of-the-art communication capabilities for consumer medical, mobile accessories, sports and wellness applications. Compared to classic Bluetooth capabilities, Bluetooth low energy is a connectionless protocol, which significantly reduces the amount of time the radio must be on. Requiring only a fraction of the power consumption of traditional Bluetooth technology, Bluetooth low energy can enable target applications to operate on a coin cell for more than a year. Application areas can be:

- Mobile accessories
- Consumer health/medical
- Sports/Fitness
- Remote controls
- Wireless sensor systems

ANT provides a simple, low-cost and ultra-low power solution for short-range wireless communication in point-to-point and more complex network topologies. Suitable for various applications, ANT is today a proven and established technology for collection, automatic transfer and tracking of sensor data within sports, wellness management and home health monitoring applications. The functionality of ANT enables mobile handheld device manufacturers to deliver ANT+ interoperable sports, fitness and consumer health monitoring products. Application areas can be:

- Sports/Fitness
- Consumer health/medical
- Mobile accessories

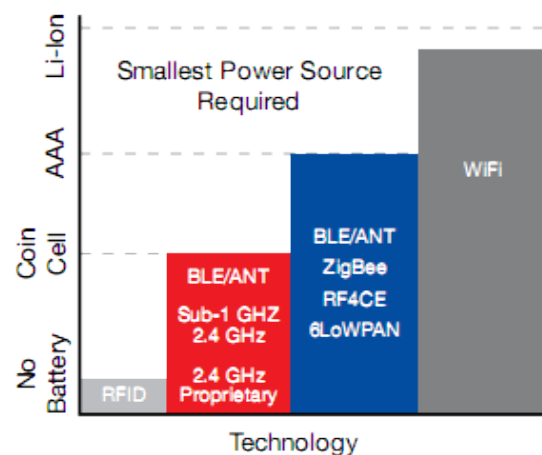


Figure 3 Wireless communication protocols classified by smallest power source requirements (Texas Instrument, 2011).

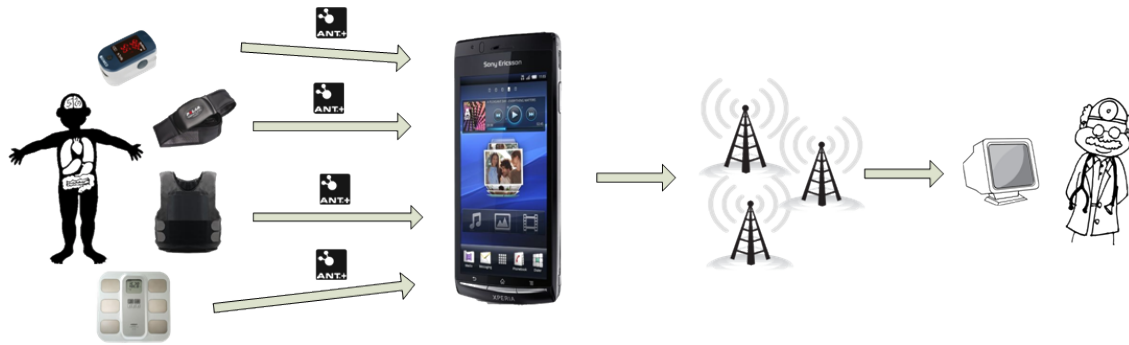


Figure 4. Monitoring system schema

- Wireless sensor systems

Figure 1, Figure 2 and Figure 3 compare these communication protocols in different technical properties such as transmission range or power consumption. As seen in this figures, ANT protocol has one of the smallest power requirements which is optimal for non-invasive sensors. About transmission range, ANT covers the needs of a WBAN and its maximum throughput covers the functionality of a health monitoring system which sensors emits small pieces of information frequently during long periods of time. In conclusion, ANT is an applicable communication protocol for a WBAN prototype centered in health monitoring.

4. MONITORING EXPERIMENT MODELING

4.1. Experiment modeling

The experiment proposed in this paper is centered in the simulation of a system which monitors several physiological parameters of a group of a group of patients in real time simultaneously.

Our case study is centered in the simulation of an elderly care center where patients have several sensors which monitor some physiological parameters and send the data to a mobile phone which store it. An example

of the validity of this case study can be observed in (Love, 2011) where a similar system was developed.

Sensor used in the experiment is a heart rate sensor. This sensor can monitor the heart rate of patients in real time. Using this information it is possible to define an alert system, in the mobile phone, which alerts to the center nurses if the heart rate of the patient reaches dangerous values.

4.1. ANT development kit

ANT development kit Figure 5 offers a comprehensive set of hardware and software to help users to evaluate, design and prototype using ANT technology. ANTAP2DK1 and ANTDKT3 are the two ANT development kits containing the same hardware, with the exception that ANTAP2DK1 features the latest AP2 modules in replacement for AP1 modules in ANTDKT3.

This kit allows us to interface directly to an ANT module, giving you the ability to test and analyze all parameters of the ANT protocol. This includes setting channel parameters (channel type, RF frequency, message rate, pairing bit etc) to set up and monitor different types of ANT channels.

The table below shows the hardware content of the two development kits.

Table 2. ANT Development kit components

Component	ANTAP2DK1	ANTDKT3
ANTAP281M5IB module	2	0
ANT11TS33M5IB module	2	2
ANTAP1M5IB module	0	2
ANT battery board	2	2
ANT I/O interface board	2	2
USB interface board	2	2
CR2032 battery	2	2



Figure 5. ANT development kit (ANT, 2011)

4.2. Mobile framework

A software framework, in computer programming, is an abstraction of a collection of classes, applications and libraries helping the different components to work



Figure 7. The proposed framework together. Frameworks can be seen as software libraries where they are reusable abstractions of code wrapped in a well-defined API. They contain the following three key distinguishing features that separate them from standard libraries: inversion of control; extensibility; and non-modifiable code. Inversion of control dictates that the overall program flows of control is dictated by the framework and not by the caller. The extensibility characteristic allows user to extend it usually by selective overriding or specialized by user code providing specific functionality. The third characteristic concerns with the code itself. In other words, the framework code is not allowed to be modified, although, users can extend the framework and implement new characteristics.

There are numerous mobile devices with different characteristics and operating systems. The proposed mobile framework is centered in the Android OS. ANT+ published and ANT+ API library for Android devices at the beginning of 2011. This library allows us to communicate with the ANT protocol chipset incorporated in mobile devices and threat data given by the ANT+ sensors.

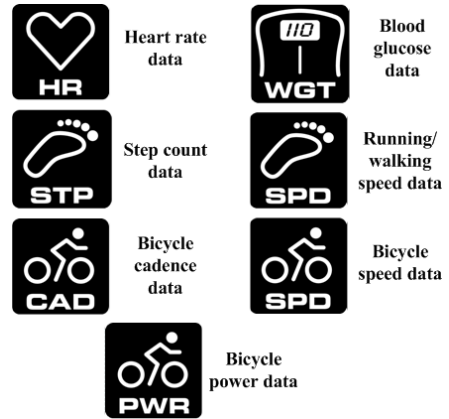


Figure 6. ANT+ simulator available sensors

The experiment mobile device target is a smartphone. Nowadays the only mobile phone vendor which incorporates the capabilities needed to recognize ANT+ communication protocol is Sony Ericsson. The Sony Ericsson Xperia mobile phones family is the first one which can manage ANT+ sensors and its commercialization starts in the second half of 2010.

The mobile phone framework proposed is showed in the Figure 7. Firstly, the sensors data will send to the ANT+ API library. Secondly, an Android application will store this data into mobile phone for its use. Finally, an Android display will manage the stored data and show it in the mobile phone display. In the development of the Android application and display and open source ANT+ demo is used as a base.

5. MONITORING SYSTEM SIMULATION

One of the key tools in ANT+ application development is the ANT+ Simulator (Figure 8). This software tool allows developers to create applications compatible with ANT+ sensors without the need of a physical sensor to generate ANT+ data during development.

The ANT+ Simulator consists of two applications: the ANT+ Sensor Simulator and the ANT+ Display Simulator. The ANT+ Sensor Simulator is used to simulate a variety of ANT+ sensors, generating and broadcasting data according to their respective ANT+ device profiles. The profiles whose can be simulated by the actual version of the simulator (v1.600) are showed in the Figure 6. The ANT+ Display Simulator receives and decodes messages transmitted from a variety of broadcast ANT+ sensors.

Using the ANT+ Sensor Simulator it is possible to emulate a heart rate sensor behavior and its communication with a mobile phone. For the experiment, the Sensor Simulator was configured as a heart rate sensor which transmits data in a broadcast mode. A Sony Ericsson Xperia Arc was used to monitor the results. As seen in Figure 9, the mobile phone is capable to receive and decode the ANT+ sensor transmission so the ANT+ communication via mobile with one sensor is successfully tested. But a WBAN is

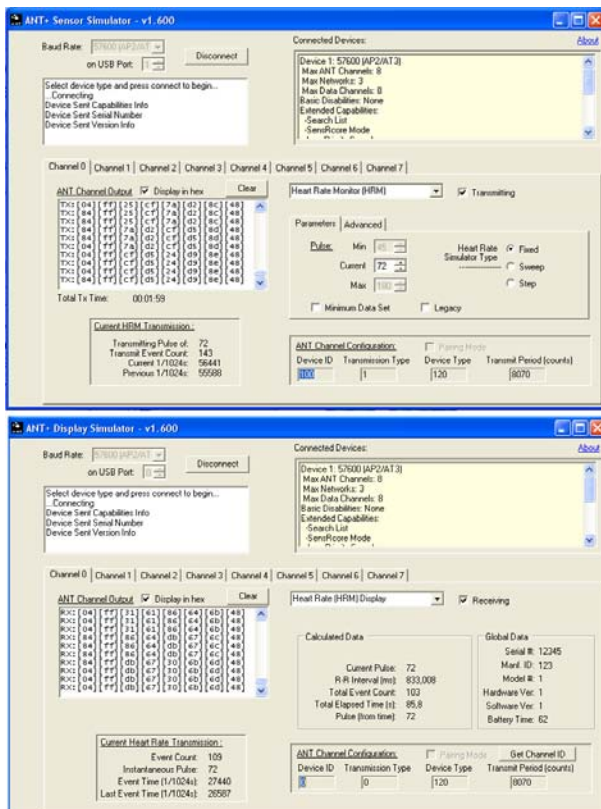


Figure 8 ANT sensor and display simulator capture

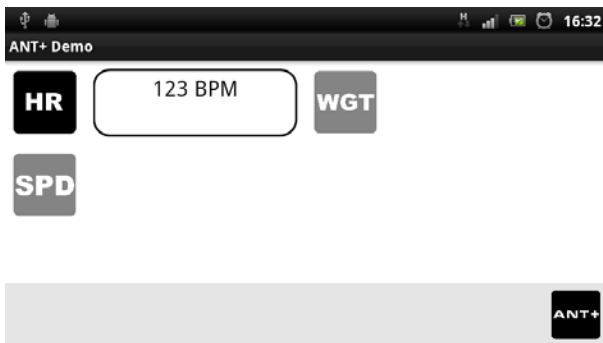


Figure 9 Single sensor simulation result

composed by more than one sensor usually so it is mandatory to probe the correct system behavior using more than one sensor. For this new experiment, the Sensor Simulator was configured with a heart rate sensor and a speed and distance sensor transmitting simultaneously using different communication channels. As seen in Figure 10, Sony Ericsson's device is capable to manage several sensor signals transmitting simultaneously so the multi-sensor ANT+ communication via mobile phone is successfully tested.

Once this basic experiment probes the mobile phones capabilities to work with a WBAN using ANT+ as a communication protocols, it is needed a deeper development work to reach an operational prototype of a health monitoring system. For example, it is necessary to design a more detailed user interface, and define and develop a functional alarm system. This task is planned and will be available in future works but, in this case, is beyond the scope of this paper.

6. CONCLUSIONS

The growth in the adoption of wireless technologies in the health care is a fact nowadays. There are several examples of BANs and WBANs, collecting physiological patients' data, in the scientific literature.

The next step in this continuous advance is the use of ultra-low power communication protocols in the health care monitoring. This kind of protocols allows the reduction in the sensors' size and the rise of the patients' mobility.

In addition, the mobile phone's market starts to incorporate this kind of protocols in its newer products. This fact makes available the possibility of a wireless patients' monitoring all day, not matter if they stay at home, at work...

In this work, a test of a development with ANT+ in this field is presented. An experiment was designed and the basic capabilities of the system are tested to consolidate the technical assumptions of the development. The tested capabilities are the possibility of communicate one or several ANT+ sensors with a commercial mobile phone. This test will be the start point of a health care system monitoring development.

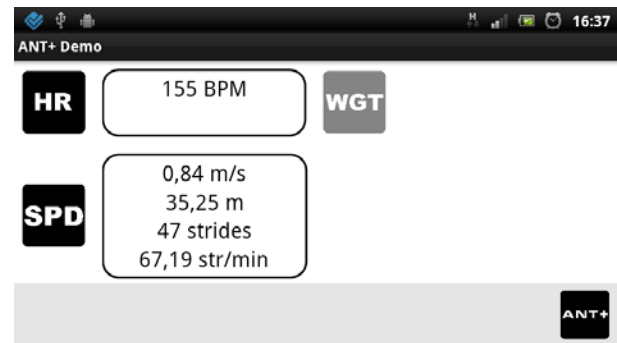


Figure 10 Multi-sensor simulation result

ACKNOWLEDGMENTS

Yeray Callero is being supported by a postgraduate grant from CajaCanarias Canary Island bank.

REFERENCES

- Adloo, H., Deghat, M., & Karimaghaee, P. (2009). Iterative state feedback control and its application to robot control. *2009 IEEE International Conference on Mechatronics, 00*(April), 1-6. IEEE. doi: 10.1109/ICMECH.2009.4957228.
- ANT. (2011). Homepage. Retrieved 2011, from www.thisisant.com.
- Cao, H., Leung, V., Chow, C., & Chan, H. (2009). Enabling technologies for wireless body area networks: a survey and outlook. *Communications Magazine, IEEE, 47*(12), 84-93. IEEE. doi: 10.1109/MCOM.2009.5350373.
- Farella, E., Pieracci, A., Benini, L., Rocchi, L., & Acquaviva, A. (2008). Interfacing human and computer with wireless body area sensor networks: the WiMoCA solution. *Multimedia Tools and Applications, 38*(3), 337-363. doi: 10.1007/s11042-007-0189-5.
- Gao, T., Massey, T., Selavo, L., Crawford, D., Chen, B.-rong, Lorincz, K., et al. (2007). The advanced health and disaster aid network: A light-weight wireless medical system for triage. *Biomedical Circuits and Systems, IEEE Transactions on, 1*(3), 203-216. IEEE. doi: 10.1109/TBCAS.2007.910901.
- Jantunen, I. (2008). Smart sensor architecture for mobile-terminal-centric ambient intelligence. *Sensors and Actuators A: Physical, 142*(1), 352-360. doi: 10.1016/j.sna.2007.04.014.
- Love, J. (2011). *Ultra-Low Power Wireless Quarter*.
- Milenkovic, a, Otto, C., & Jovanov, E. (2006). Wireless sensor networks for personal health monitoring: Issues and an implementation. *Computer*

Communications, 29(13-14), 2521-2533. doi: 10.1016/j.comcom.2006.02.011.

Pentland, A. S. (2004). Healthwear: medical technology becomes wearable. *Computer*, 37(5), 42–49. Published by the IEEE Computer Society. doi: 10.1109/MC.2004.1297238.

Pereira, O., Caldeira, J. M. L. P., & Rodrigues, J. J. P. C. (2010). Body Sensor Network Mobile Solutions for Biofeedback Monitoring. *Mobile Networks and Applications*. doi: 10.1007/s11036-010-0278-y.

Ruiz, J. a, & Shimamoto, S. (2006). Novel communication services based on human body and environment interaction: applications inside trains and applications for handicapped people. *IEEE Wireless Communications and Networking Conference, 2006. WCNC 2006*. (Vol. 4, p. 2240–2245). Ieee. doi: 10.1109/WCNC.2006.1696644.

Sarpeshkar, R. (2006). Invited Talk: Ultra Low Power Electronics for Medicine. *International Workshop on Wearable and Implantable Body Sensor Networks (BSN06)*, 37-37. IEEE. doi: 10.1109/BSN.2006.38.

Shnayder, V., Chen, B.-rong, Lorincz, K., Fulford-Jones, T. R. F., & Welsh, M. (2005). Sensor networks for medical care. *SenSys' 05: Proceedings of the 3rd international conference on Embedded networked sensor systems* (p. 314–314). New York, New York, USA: Citeseer. doi: 10.1145/1098918.1098979.

Texas Instrument. (2011). *Wireless Connectivity Guide. Technology*.

Topol, E. (2010). Transforming Medicine via Digital Innovation. *Science Translational Medicine*.

AUTHORS BIOGRAPHY

YERAY CALLERO was born in Haría, Lanzarote and attended the University of La Laguna, where he studied Engineering Computer Science and obtained his degree in 2008. He is currently working on his PhD with the Department of Systems Engineering and Automation at the same university. His research interests include simulation and health monitoring.

ROSA M. AGUILAR received her MS degree in Computer Science in 1993 from the University of Las Palmas de Gran Canaria and her PhD degree in Computer Science in 1998 from the University of La Laguna. She is an associate professor in the Department

of Systems Engineering and Automation at the University of La Laguna. Her current research interests are decision making based on discrete event simulation systems and knowledge-based systems, intelligent agents, and intelligent tutorial systems.

An Extreme Learning Machine Algorithm for Higher Order Neural Network Models

Dr Shuxiang Xu

School of Computing, University of Tasmania, Locked Bag 1359

Launceston, Tasmania 7250, Australia

Email: Shuxiang.Xu@utas.edu.au

ABSTRACT

Artificial Neural Networks (ANN) have been widely used as powerful information processing models and adopted in applications such as bankruptcy prediction, predicting costs, forecasting revenue, forecasting share prices and exchange rates, processing documents and many more. This paper uses Extreme Learning Machine (ELM) algorithm for Higher Order Neural Network (HONN) models and applies it in several significant business cases. HONNs are neural networks in which the net input to a computational neuron is a weighted sum of products of its inputs. ELM algorithms randomly choose hidden layer neurons and then only adjust the output weights which connect the hidden layer and the output layer. The experimental results demonstrate that HONN models with ELM algorithm offer significant advantages over standard HONN models as well as traditional ANN models, such as reduced network size, faster training, as well improved simulation and forecasting errors.

KEYWORDS: Higher Order Neural Network, Feedforward Neural Network, Extreme Learning Machine, Financial Forecasting.

1. INTRODUCTION

Business is a diversified field with several general areas of specialisation such as accounting or financial analysis. Artificial Neural networks (ANNs) provide significant benefits in business applications. They have been actively used for applications such as bankruptcy prediction, predicting costs, forecast revenue, processing documents and more (Kurbel et al, 1998 ; Atiya et al, 2001 ; Baesens et al, 2003). Almost any neural network model would fit into at least one business area or financial analysis. Traditional statistical methods have been used for business applications with many limitations (Azema-Barac et al, 1997 ; Blum et al, 1991 ; Park et al, 1993).

While conventional ANN models have been bringing huge profits to many financial institutions, they suffer from several drawbacks. First, conventional ANNs can not handle discontinuities in the input training data set (Zhang et al, 2002). Next, they do not perform well on complicated business data with high frequency components and high order nonlinearity, and finally, they are considered as 'black boxes' which can not explain their behaviour (Blum et al, 1991; Zhang et al, 2002 ; Burns, 1986).

To overcome these limitations some researchers have proposed the use of Higher Order Neural Networks

(HONNs) (Redding et al, 1993 ; Zhang et al, 1999 ; Zhang et al, 2000). HONNs are able to provide some explanation for the simulation they produce and thus can be considered as 'open box' rather than 'black box'. HONNs can simulate high frequency and high order nonlinear business data, and can handle discontinuities in the input training data set (Zhang et al, 2002).

HONNs (Higher Order Neural Networks) (Lee et al, 1986) are networks in which the net input to a computational neuron is a weighted sum of products of its inputs. Such neuron is called a Higher-order Processing Unit (HPU) (Lippman, 1989). It was known that HONN's can implement invariant pattern recognition (Psaltis et al, 1988 ; Reid et al, 1989 ; Wood et al, 1996). Giles in (Giles et al, 1987) showed that HONN's have impressive computational, storage and learning capabilities. In (Redding et al, 1993), HONN's were proved to be at least as powerful as any other FNN (feedforward Neural Network) architecture when the orders of the networks are the same. Kosmatopoulos et al (1995) studied the approximation and learning properties of one class of recurrent HONNs and applied these architectures to the identification of dynamical systems. Thimm et al (1997) proposed a suitable initialization method for HONN's and compared this method to weight initialization techniques for FNNs. A large number of experiments were performed which led to the proposal of a suitable initialization approach for HONNs.

Unlike traditional ANN learning algorithms, Extreme Learning Machine (ELM) randomly chooses hidden neurons and analytically determines the output weights (Huang et al 2005, 2006, 2008). Many types of hidden nodes including additive nodes, RBF (radial basis function) nodes, multiplicative nodes, and other non neural alike nodes can be used as long as they are piecewise nonlinear. ELM algorithm tends to generalize better at very fast learning speed: it can learn thousands of times faster than conventional popular learning algorithms (Huang et al 2006).

This paper is organized as follows. Section 2 proposes an ELM learning algorithm for HONN models. Section 3 presents several experiments with results to compare the performance of 4 ANN models: HONN with ELM, standard HONN, traditional MLP (multilayer perceptron) neural network, and RBF (radial basis function) neural network. Finally, Section 4 summarizes this chapter.

2. HONN MODELS WITH ELM ALGORITHM

HONNs (Lee et al, 1986) are networks in which the net input to a computational neuron is a weighted sum of products of its inputs. Such neuron is called a Higher-order Processing Unit (HPU) (Lippman, 1989). The network structure of an HONN is the same as that of a multi-layer FNN. That is, it consists of an input layer with some input units, an output layer with some output units, and at least one hidden layer consisting of intermediate processing units. Usually there is no activation function for neurons in the input layer and the output neurons are summing units, the activation function for hidden layer neurons can be any nonlinear piecewise continuous ones.

Based on a one-dimensional HONN defined in (Zhang et al, 2002), this paper proposes the following ELM algorithm for HONNs. The ELM algorithm was originally proposed by Huang (2006), for Single-Layer Feedforward Neural Networks (SLFN). The main idea of ELM lies in the random selection of hidden neurons with random initialization of the SLFN weights and biases. Then, the input weights and biases do not need to be adjusted during training, only the output weights are learned. The training of the SLFN can be achieved with a few steps and very low computational costs.

Consider a set of S distinct training samples (X_i, Y_i) with $X_i \in R^n$ and $Y_i \in R^m$, where n and m are positive integers. Then an SLFN with N hidden neurons can be mathematically represented by

$$\sum_{i=1}^N O_i \times f(w_i \times X_j + b_i), \quad 1 \leq j \leq S \quad (2.1)$$

with f being the randomly selected neuron activation function, w_i the input weights, b_i the biases, and O_i the output weights.

In case of two-dimensional HONN with a single hidden layer, equation (2.1) becomes

$$\sum_{i=1}^{NP} O_i \times f(w_i \times \begin{bmatrix} X_j \\ H(X_j) \end{bmatrix} + b_i), \quad 1 \leq j \leq S \quad (2.2)$$

where

$$NP = N + C_N^2 \quad (2.3)$$

$$H(X_j) = X_j \otimes X_j^T \quad (2.4)$$

Assume that the single layer HONN approximates the training samples perfectly, then the errors between the estimated outputs are the actual outputs are zero, which means

$$\sum_{i=1}^{NP} O_i \times f(w_i \times \begin{bmatrix} X_j \\ H(X_j) \end{bmatrix} + b_i) = \begin{bmatrix} Y_j \\ H(Y_j) \end{bmatrix}, \quad 1 \leq j \leq S \quad (2.5)$$

where

$$H(Y_j) = Y_j \otimes Y_j^T \quad (2.6)$$

Then the ELM algorithm, when applies to a HONN, states that with randomly initialized input weights and biases, and with the condition that the randomly selected neuron activation function is infinitely differentiable, then the output weights can be determined so that the single layer HONN provides an approximation of the sample values to any degree of accuracy.

3. HONN MODEL APPLICATIONS IN BUSINESS

In this section, the HONN model with ELM algorithm as defined in Section 2 has been used in several financial applications. The algorithm has been implemented in Java, based on the ANN implementation in Matlab version R2009b. The results are given and discussed.

3.1 Simulating and Forecasting Total Taxation Revenues of Australia

The HONN model has been used to simulate and forecast the Total Taxation Revenues of Australia as shown in Figure 3.1. The financial data were downloaded from the Australian Taxation Office (ATO) web site. For this experiment monthly data between Jan 1994 and Dec 1999 were used. The detailed comparison between the following 4 models has been performed: HONN with ELM, standard HONN, traditional MLP (multilayer perceptron) neural networks, and RBF neural networks. The results are illustrated in Table 3.1.

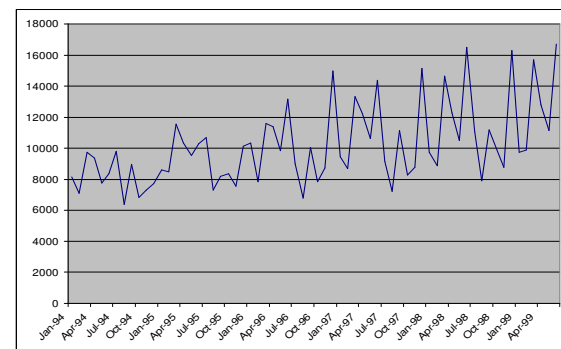


Figure 3.1. Total Taxation Revenues of Australia (\$ million) (Jan 1994 To Dec 1999)

Neural Network	No. HL	HL Nodes	Epoch	RMS Error
HONN with ELM	1	4	3,000	0.085468
Standard HONN	1	4	7,000	0.165874
MLP ANN	1	10	12,000	0.856654
RBF ANN	1	10	12,000	0.987996

Table 3.1. HONN with ELM, Standard HONN, MLP, and RBF to Simulate Taxation Revenues (HL: Hidden Layer. RMS: Root-Mean-Square)

After the 4 ANN models have been well trained over the training data pairs, they were used to forecast the taxation revenues for each month of the year 2000. Then the forecasted revenues were compared against the real revenues for the period and the overall RMS errors reached (in order) were 4.55%, 7.65%, 11.23%, and 12.01%, respectively. It was worth noting that HONN with ELM was the fastest model in terms of convergence.

3.2 Simulating and Forecasting Reserve Bank Of Australia Assets

The HONN with ELM has also been used to simulate and forecast the Reserve Bank Of Australia Assets as shown in Figure 3.2. The financial data were obtained from the Reserve Bank Of Australia. For this experiment monthly data between Jan 1980 and Dec 2000 were used. The detailed comparison between the 4 ANN models is illustrated in Table 3.2.

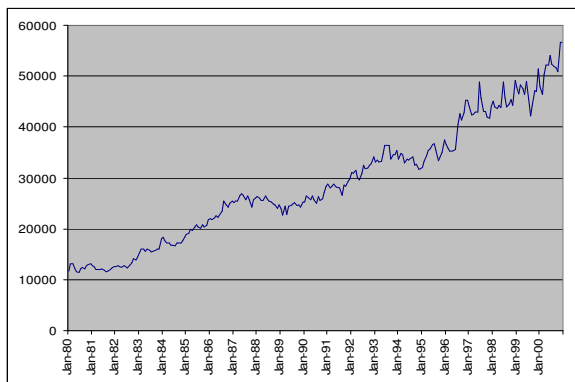


Figure 3.2. Reserve Bank Of Australia Assets (\$ million) (Jan 1980 To Dec 2000)

Neural Network	No. HL	HL Nodes	Epoch	RMS Error
HONN with ELM	1	5	3,000	0.090231
Standard HONN	1	5	7,000	0.196753
MLP ANN	1	14	12,000	0.956454
RBF ANN	1	14	12,000	0.997216

Table 3.2. HONN with ELM, Standard HONN, MLP, and RBF to Simulate Reserve Bank Of Australia Assets (\$ million) (HL: Hidden Layer. RMS: Root-Mean-Square)

After these 4 ANN models have been well trained over the training data pairs, they were used to forecast the Reserve Bank of Australia Assets for each month of the year 2001. Then the forecasted assets were compared against the real assets for the period and the overall RMS errors reached (in order) were 5.05%, 8.62%,

13.23%, and 11.81%, respectively. HONN with ELM was also the fastest model to converge.

3.3 Simulating and Forecasting Fuel Economy

In the next experiment a dataset containing information of different cars built in the US, Europe, and Japan was trained using the 4 ANN models to determine car fuel economy (MPG - Miles Per Gallon) for each vehicle. There were a total of 392 samples in this data set with 9 input variables and 1 output. The dataset was from UCI Machine Learning Repository (2007). The output was the fuel economy in MPG, and the input variables were:

- number of cylinders
- displacement
- horsepower
- weight
- acceleration
- model year
- Made in US? (0,1)
- Made in Europe? (0,1)
- Made in Japan? (0,1)

To compare the performance of the 4 ANN models the dataset was divided into a subset containing 300 samples for training, a subset containing 53 samples for testing, and the last subset of 39 samples for forecasting (or generalization). Each of the 4 ANN models was trained using the training subset, tested using the testing subset (for adjusting certain parameters), and finally applied on the forecasting subset. The experimental results for the forecasting subset are illustrated in the following Table 3.3.

Neural Network	No. HL	HL Nodes	Epoch	RMS Error
HONN with ELM	1	3	2,000	0.060768
Standard HONN	1	3	5,000	0.152359
MLP ANN	1	9	9,000	0.709845
RBF ANN	1	9	9,000	0.870946

Table 3.3. HONN with ELM, Standard HONN, MLP, and RBF to forecast fuel economy (HL: Hidden Layer. RMS: Root-Mean-Square)

4. CONCLUSIONS

In this paper an HONN with ELM model was introduced and applied in business applications such as simulating and forecasting government taxation revenues. Experiments demonstrated that such model offers significant advantages over standard HONN models and traditional ANN models such as reduced network size, faster training, as well as improved simulation and forecasting errors. It appears that HONN with ELM model works faster than the other ANN models due to the nature of the ELM algorithm. As part of the future research, some current cross-validation approaches may be improved and applied so that the forecasting errors could be reduced further down to a more satisfactory level.

Another direction for future research would be the use of an ensemble of HONN models for modelling and simulation.

REFERENCES

- Arai, M., Kohon, R., Imai, H. (1991). Adaptive control of a neural network with a variable function of a unit and its application, *Transactions on Inst. Electronic Information Communication Engineering*, J74-A, 551-559.
- Atiya, A.F., (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results, *IEEE Transactions on Neural Networks*, Volume: 12, Issue: 4, page(s): 929-935.
- Azema-Barac, M., Refenes, A., (1997). Neural Networks for financial applications, in E. Fiesler & R. Beale (eds) *Handbook of Neural Computation*, Oxford University Press (G6:3:1-7).
- Baesens, B., Setiono, R., Mues, C., Vanthienen, J., (2003). Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation, *Management Science*, Volume: 49, Issue: 3.
- Barron, A. R., (1994). Approximation and Estimation Bounds for Artificial Neural Networks. *Machine Learning*, (14):115-133.
- Blum, E., Li, K., (1991). "Approximation theory and feed-forward networks", *Neural Networks*, Vol.4, pp.511-515.
- Burns, T., (1986). The interpretation and use of economic predictions, *Proc. Royal Society A*, pp.103-125.
- Campolucci, P., Capparelli, F., Guarnieri, S., Piazza, F., & Uncini, A. (1996). Neural networks with adaptive spline activation function. *Proceedings of IEEE MELECON 96*, Bari, Italy, 1442-1445.
- Chen, C.T., Chang, W.D. (1996). A feedforward neural network with function shape autotuning, *Neural Networks*, 9(4), 627-641
- Dayhoff, J. E., (1990). *Neural network architectures : an introduction*. New York, N.Y. : Van Nostrand Reinhold.
- Gallant, A. R., White, H. (1988) There exists a neural network that does not make avoidable mistakes. *IEEE Second International Conference on Neural Networks*, San Diego, SOS Printing, I: 657-664.
- Giles, C.L., Maxwell, T. (1987) Learning, invariance, and generalization in higher order neural networks, *Applied Optics*, 26(23), 4972-4978.
- Grossberg, S. (1986). Some nonlinear networks capable of learning a spatial pattern of arbitrary complexity. *Proc. National Academy of Sciences*. 59, 368-372.
- Hammadi, NC, Ito, H. (1998). On the activation function and fault tolerance in feedforward neural networks. *IEICE Transactions on Information & Systems*, E81D(1): 66 – 72.
- Hansen, J.V., Nelson, R.D. (1997). Neural networks and traditional time series methods: a synergistic combination in state economic forecasts, *IEEE Transactions on Neural Networks*, 8(4), 863-873.
- Hinton, G. E. (1989). Connectionist learning procedure, *Artificial Intelligence*, 40, 251 – 257.
- Holden, S.B., and Rayer, P.J.W. (1995). Generalisation and PAC learning: some new results for the class of generalised single-layer networks. *IEEE Transactions on Neural Networks*, 6(2), 368 – 380.
- Haykin, S. S, (1994). *Neural networks : a comprehensive foundation*. New York : Macmillan.
- Hu, Z., Shao, H. (1992). The study of neural network adaptive control systems, *Control and Decision*, 7, 361-366.
- Huang, GB, Siew, CK, "Extreme Learning Machine with Randomly Assigned RBF Kernels," *International Journal of Information Technology*, vol. 11, no. 1, pp. 16—24, 2005.
- Huang, GB, Zhu, QY, Siew, CK, "Extreme Learning Machine: Theory and Applications", *Neurocomputing*, vol. 70, pp. 489-501, 2006.
- Huang, GB, Li, MB, Chen, L, and Siew, CK, "Incremental Extreme Learning Machine With Fully Complex Hidden Nodes," *Neurocomputing*, vol. 71, pp. 576-583, 2008.
- Kawato, M., Uno, Y., Isobe, M., Suzuki, R. (1987) A hierarchical model for voluntary movement and its application to robotics, *Proc. IEEE Int. Conf. Network*, IV, 573-582.
- Kay, A., (2006). Artificial Neural Networks, Computerworld website, Retrieved on 27 November 2006 from <http://www.computerworld.com/softwaretopics/softw/are/appdev/story/0,10801,57545,00.html>
- Kosmatopoulos, E.B., Polycarpou, M.M., Christodoulou, M.A., Ioannou, P.A. (1995). High-order neural network structures for identification of dynamical systems, *IEEE Transactions on Neural Networks*, 6(2), 422-431.
- Kurbel, K., Singh, K., Teuteberg, F. (1998). Search And Classification Of "Interesting" Business Applications In The World Wide Web Using A Neural Network Approach. *Proceedings of the 1998 IACIS Conference*, Cancun, Mexico.
- Lee, Y.C., Doolen, G., Chen, H., Sun, G., Maxwell, T., Lee, H., Giles, C.L. (1986) Machine learning using a higher order correlation network, *Physica D: Nonlinear Phenomena*, 22, 276-306.
- Lippman, R.P. (1989) Pattern classification using neural networks, *IEEE Commun. Mag.*, 27, 47-64.
- Park, J., Sandberg, I.W., (1993). "Approximation and radial-basis-function networks", *Neural Computation*, Vol.5, pp.305-316.
- Picton, P, (2000). *Neural Networks*. Basingstoke : Palgrave.

- Psaltis, D., Park, C.H., Hong, J. (1988) Higher order associative memories and their optical implementations, *Neural Networks*, 1, 149-163.
- Redding, N., Kowalczyk A. and Downs, T., (1993). "Constructive high-order network algorithm that is polynomial time", *Neural Networks*, Vol.6, pp.997-1010.
- Redding, N.J., Kowalczyk, A., Downs, T. (1993). Constructive higher-order network algorithm that is polynomial time, *Neural Networks*, 6, 997-1010.
- Reid, M.B., Spirkovska, L., Ochoa, E. (1989). Simultaneous position, scale, rotation invariant pattern classification using third-order neural networks, *Int. J. Neural Networks*, 1, 154-159.
- Rumelhart, D.E., McClelland, J.L. (1986). *Parallel distributed computing: exploration in the microstructure of cognition*, Cambridge, MA: MIT Press.
- Thimm, G., Fiesler, E. (1997). High-order and multilayer perceptron initialization, *IEEE Transactions on Neural Networks*, 8(2), 349-359.
- UCI Machine Learning Repository, (2007). <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/auto-mpg/auto-mpg.data>, accessed in April 2007.
- Vecci, L., Piazza, F., Uncini, A. (1998). Learning and approximation capabilities of adaptive spline activation function neural networks, *Neural Networks*, 11, 259-270.
- Wood, J., Shawe-Taylor, J. (1996). A unifying framework for invariant pattern recognition, *Pattern Recognition Letters*, 17, 1415-1422.
- Xu, S (2009), A Novel Higher Order Artificial Neural Networks, Proceedings of the Second International Symposium on Computational Mechanics (ISCM II), 30 Nov – 3 Dec 2009, Hong Kong – Macau, pp. 1507-1511. ISBN: 978-0-7354-0778-7.
- Xu S, Chen L (2009), Adaptive Higher Order Neural Networks for Effective Data Mining, Proceedings of The Sixth International Symposium on Neural Networks (ISNN 2009), Wuhan, China, May 26-29, 2009, pp 165 – 173.
- Yamada, T., Yabuta, T. (1992). Remarks on a neural network controller which uses an auto-tuning method for nonlinear functions, *IJCNN*, 2, 775-780.
- Zhang, M., Xu, S., and Lu B., (1999). Neuron-adaptive higher order neural network group models, *Proc. Intl. Joint Conf. Neural Networks - IJCNN'99*, Washington, DC, USA, (Paper # 71).
- Zhang, M., Xu, S., Fulcher, J., (2002). Neuron-Adaptive Higher Order Neural-Network Models for Automated Financial Data Modeling, *IEEE Transactions on Neural Networks*, Vol 13, No. 1.
- Zhang, M., Zhang, J. and Fulcher, J., (2000). "Higher order neural network group models for financial simulation". *Intl. J. Neural Systems*, vol 12, No. 2, pp 123 –142.

INCREASING AVAILABILITY OF PRODUCTION FLOW LINES THROUGH OPTIMAL BUFFER SIZING: A SIMULATIVE STUDY

Vittorio Cesarotti^(a), Alessio Giuiusa^(b), Vito Introna^(c)

Department of Mechanical Engineering, "Tor Vergata" University of Rome, Via del Politecnico 1, Rome - Italy

^(a)cesarotti@uniroma2.it, ^(b)alessio.giuiusa@uniroma2.it, ^(c)vito.introna@uniroma2.it

ABSTRACT

In flow shop highly automated production lines the absence or undersize of inter-operational buffer between consecutive stations is an occurrence as frequent as detrimental for the productivity of the entire production line. A correct sizing of buffers mitigates or even eliminates the propagation, on the entire production line, of small inefficiencies due to stops and / or slowdowns of the single station. This paper describes a simulation approach to investigate the effect buffer between two successive stations and measure its effects in terms of change in the overall efficiency of the line. A wide range of typical production parameter is considered. This allows to extend the paper results to many different production system and to evidence some interesting analogies in production effectiveness behavior depending on buffer size. The introduction of an analytic experimental relation allows to describe the evidenced behavior and to size the buffer without need for further simulations.

1. INTRODUCTION

The increasing competition and attention of the market to the cost of the product have prompted the producers of goods to use more automated forms of production in recent years. This is pursued through more complex workstations that can perform many operations, with the aim of increasing productivity while ensuring the requested level of flexibility in production.

This issue is very important in flow shop dominant sectors (e.g. pharmaceutical, cigarettes, electronic, etc.) and has led to the development of production lines that complement many workstations (even over 20) in succession.

During several years of experience with some leading multinational companies, the authors have noticed that the design of these systems is often exclusively focused on the balance of workstations in ideal operating conditions. This approach neglects the effects of efficiency losses propagation between the workstations, that is dramatically important in this type of systems.

In particular the production rate of each workstation is often characterized by short but frequent interruptions

and delays caused by minor stoppages (e.g. pieces stuck in the machines, block of mechanical parts, temporary reduction of workstations speed, congestion, minor stoppages). The effects of inefficiencies in the single work station can spread along the entire production line slowing down the other machines that otherwise would be able to operate properly. Therefore the lack of a minimum level of independence between workstations in series (belonging to the same production line) can get to stop the production lines (Spinellis 1999) also due to the temporary blockage of a single machine.

If the number of workstations is high, this effect may result in a reduction of the overall performance of the production line even over 30 %, also if the failure of a single work station is limited.

The loss of productivity of the production lines has a wide impact at a strategic organizational level, due to:

- Higher production costs;
- Delays in delivery (customer satisfaction)
- High inventory in stock (interest payable)

2. LITERATURE REVIEW

To ensure the desired performance it is necessary to determine an appropriate level of independence between successive work stations of the production line through the insertion of buffer of opportune size and suitably located. This issue has been widely debated in literature through two different paradigms:

- Buffer Allocation Problem (BAP)
- Buffer Size Problem (BSP)

The problem is studied by the scientific community in order to identify the location and/or the size of the inter-operational buffers, in order to minimize both cost and space, and maximize production line throughput.

2.1. Existing Problem approaches

Historically, there are different approaches to this problem, among which:

- Heuristic, typical of operations research (Hillier 1993, Lutz 1998 and Papadopoulos 2001);
- Survey followed by procedures for sizing (Tempelmeier 2003);
- Mathematics: (Hillier F. S. 1977 and Gutowski 2005);
- Simulation, (Malakooti 1994, Chiadamrong, 2003 and Yamada, 2003);

As stated from (D. Battini 2009) "*The optimal buffer size problem based on the machine availability is a very critical issue (50% of studies analyzed consider machine reliability parameters), but has not been yet sufficiently investigated.*". Furthermore there is a lack in the available literature of benchmarking analysis investigating the change in production line availability depending on the change of buffer dimension.

In particular, most of the literature available today does not provide an approach that fulfills the needs of industrial producers, i.e. an approach that is at the same time practical, operative and easily repeatable by the companies themselves. In fact:

(1) The mathematical approach is often too complex and too hard to repeat by the industrial producers;

(2) "*The dynamical simulation approach is often appreciated and applied by researchers to face the BAP problem under specific working conditions: otherwise, as (Chiadamrong 2003) underlines, no standard formulae or algebraic relations between line throughput and buffer sizes has yet been obtained to help practitioners in the fast and easy design and optimization of buffers, when time constraints avoid the use of simulation (which is often complex and time consuming);*" (D. Battini 2009).

(3) Furthermore, many existing approaches are not related to industrial standard parameters such as: Overall Equipment Effectiveness (OEE), Availability and Performance Efficiency (Samuel H. Huangt 2003). All these aspects cause the inability to easily quantify the productivity lost in the production lines as a result of short failures due to an ineffective buffer sizing.

As noticed by the authors in their professional experience (e.g. pharmaceutical packaging lines, electrical components assembly line, etc.), this gap is strongly felt by the industrial sector, and it leads to buffers generally absent and/or under sized and/or misplaced. Thus, production lines present reduced OEE, which eliminate some of the benefits arising from greater speed automated lines.

Therefore authors believe that in literature there is the space to approach a second specific paradigm, already

introduced by (D. Battini 2009) useful to study the function of buffers in production lines, which is the Buffer Design for Availability (BDFa).

Within BDFa, all the works available in literature up to now aim to assess the maximum buffer size depending on the production parameters. This approach could be comprehensive in flow shop industry if down-time production cost are more relevant than inventory cost, as stated by Gerwish and Goldin in their "*Efficient Algorithms for transfer line design*" (Gershwin 1995) (e.g. food, beverage, etc.).

Rather than maximum buffer size Authors, wants to investigate the trend of the OEE depending on the buffer size. In fact but in other flow shop production lines (such as electronic or pharmaceutical) where inventory cost can be higher than others to find the maximum size of the buffer could not be a sensible solution. Furthermore, not only costs affect the choice of the buffer size, but also others factors that may prompt to a smaller sizing, must be taken into account. For example:

- Operative production constraints.

The value of Work in progress, or production line lead time may need to be under a certain value (Slack 1993) Therefore maximum buffer of size could not allows to respect these constraints;

- OEE trend.

The values of OEE, depending on the Buffer Size, assume an asymptotic trend on its maximum value after a certain value of the buffer. To chose the maximum buffer size, without analyze this trend could bring to increase the size of the buffer of more than 40% in order to obtain an improvement of OEE of only 1%.

By changing the value of the buffer size from zero to the maximum buffer size, Authors will show graphically and analytically the relation that connect the size of the buffer to the OEE of the entire production line.

Therefore this work aims to deliver a tool that give the possibility to estimate the OEE trend depending on the buffer size and to chose right size of buffer according to both inventory cost and all the others industrial factors

The resulting parametric curves obtained by Authors and the analytic model that will be propose in this study could certainly represent a significant step towards the needs of industry, since they allow both the buffer sizing in a simple and immediate way and provide to managers a greater sensitivity on the effects of buffer (in terms of OEE and productivity line), otherwise absent

2.2. The Buffer Design for Availability

The studies regarding the BDFa already available on literature aim to deliver the maximum buffer size that allow to achieve the maximum OEE, taking into account different performance parameters. This is a

very important result. In fact the bigger is the buffer, the highest is the OEE, but once achieved the maximum buffer size no improvement in OEE will be obtained by a further increase in buffer size. Therefore, the value of the max Buffer size depending on the process parameters (such as Availability, stoppage time, speed losses, etc) is a very important information that can allow to do not invest more money in buffer than necessary.

By the way is important also to consider the productivity trend of the production line depending on the buffer size.

In fact, considering synchronous flow shop lines with n series station (figure 1), if the buffers between the stations are null the global efficiency of the system will be the factorization of the single station productivity ($P_1 * P_2 * \dots * P_i * \dots * P_n$). The performance of each station depend on the performance on the other n-1 stations (Complete dependence).



Figure 1: Stations in complete dependence

Instead, if the buffers between the stations are opportune sized with their maximum value (figure 2) the Productivity of the line will be the one of the bottleneck ($\min(P_1, P_2, P_i, \dots, P_n)$).

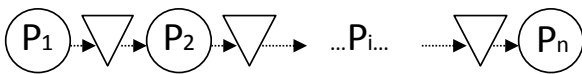


Figure 2: Stations in complete independence

The trend of the OEE between these two bound (Complete dependence and Complete independence) depend directly by the buffer size.

Therefore an enterprise can be interested to:

- size the buffers of the production line in order to maximize its overall availability and then its throughput;
- determine the minimal buffer size that allows to reach a desired level of real throughput, then minimizing buffer occupation, cost, etc.;
- easily know the lost level of efficiency due to absent/undersized buffers;
- know the expected growth trend of the overall production line productivity depending on the increase in buffer size.

Therefore the goal of this study is to investigate the behavior of the production line productivity depending on the buffer size, taking into account:

- Effect of cycle time variability;
- Effect of minor stoppages;

and to provide a tool that allows an experimental buffer sizing.

3. PROBLEM MODELING

The configuration of reference used in this paper consist of two consecutive work stations separated by a buffer, as shown in figure 1. The results could be extended to a line formed of more elementary units.

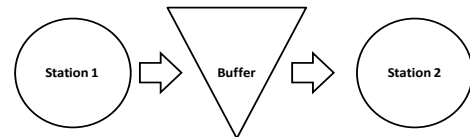


Figure 3: model configuration

The size of the buffer will be provided depending on different typical performance parameters of the line, such as:

- Randomness connected to reduction in Workstation speed;
- Mean Time Between Failures (MTBF);
- Mean Time to Repair (MTTR);
- Standard deviation (as a percentage of MTBF and MTTR).

3.1. Assumption

Authors have identified, considering evidence from literature review (D. Battini 2009), the simulation as the best approach to the problem in terms of robustness and validity of the output solution. The software selected by authors for the simulation model realization is “Rockwell Arena”. The model consist of different modules, already available in the software, that allow to simulate the process. Besides, the right definition of process attributes and variables allows to record the needed data for the further simulation analysis.

Data input of the simulation are typical of the industrial sector and are defined below.

3.1.1. Cycle Time

In automated production line the different station are usually balanced between them, therefore the station are characterized by the same ideal process time, T_i , within the range $[0,01667; 0,25]$. It does not take into account the inefficiencies.

V_i is the ideal throughput achievable in ideal condition (no inefficiencies) in one hour from the production line. It is defined as $60/T_i$. All the time measures are expressed in minutes.

Loss in productivity are considered in the model according to the general notation of OEE (Samuel H. Huangt 2003).

3.1.2. Loss of Quality

In flow shop production line loss of quality due to defect in pieces are usually negligible. Moreover they affect the buffer size only if the detection of wrong units is done between the two stations. Consequently, the loss of quality have been neglected.

3.1.3. Loss of Performance

OEE theory includes in performance losses both the cycle time slowdown and minor stoppages. The latter are discussed within the Availability paragraph because of their feature to be simulated as a stop of the station. Regarding to cycle time slowdown distribution the goal has been to model it in the most general way.

The production stations can be modeled according to the principle of queue theory.(H.T. Papadopoulos 1996). Exploiting the link to this mathematical formulation of the problem is possible to identify the statistical distribution that allows to model cycle time slowdown in order to obtain result with highest general significance.

According to Kingman equation (or VUT equation) (Kingman, 1966) exponential distribution of cycle time (M/M/1) can be used as an upper bound of general distribution in cycle time (G/G/1) for specific range of variability. In accordance to theirs experience in flow shop sector Authors argue that real variability in cycle time is include within this range.

To model the Cycle time variability with an exponential allows to obtain an Upper bound for Buffer size and therefore to gain a more robust size of the buffer.

Performance index P(i) can vary within [0,8;0,00].The performance of the two stations can be different from each other.

3.1.4. Loss of Availability

Availability depend on failures and set up. Set up usually require the stop of the entire production line, therefore it is not considered in this treatment. Also significant failures usually stop the entire, hence only failure till 30 minutes (minor stoppages) are included in this study

The BDFFA aim to eliminate the effect of the unpredictable minor stoppages that may occur during production in one or twice stations , (that can be frequent in a production line). Therefore only this kind of losses will be take into account. When a minor stoppage occurs on the second station a correct size of buffer between the two stations allow to complete all the maintenance with no influence to the first station performance. In presence of minor stoppages in the first station a correct size of buffer allows to feed the second station while maintenance regard the first

Scientific literature(Lawless 1982) provides many models of statistical distribution that can represent different kind of minor stoppage, such as Lognormal (fatigue and material strengths and loading), Weibull (material strength, times-to-failure of electronic and mechanical components, equipment, or systems),

Exponential (behavior of units that have a constant failure rate) or Normal (complex mechanism)

Therefore the variable that require to be modeled for availability are:

- Up-time distribution (Mean Time Between Failures of the station i);
- Down-time distribution (Mean Time to Repair of the station i).

Availability index vary within [0,8;0,99] (lower value of availability in automated lines are infrequent).

Mean Time to repair of stations are supposed within the range [0,01; 30] minute. Because greater downtime are uncommon in automated production lines.

The MTTR of first station is at most equal to MTTR of the second

$$MTTR(1) \leq MTTR(2) \quad (2)$$

Mean Time Between Failures (MTBF) is expressed as function of Availability (table 01).

Both up-time and down-time are assumed as normally distributed.

Standard deviation is expressed as a percentage of respectively Up-time and Down-time average. The range is from 5% till 100%.

All these losses affect the overall productivity of the line reducing the reliable throughput in the time unit. Therefore in the model the OEE have been measured as the Ratio between the Real line throughput and the ideal line throughput.

3.2. Simulation Plan

The simulation is composed of two phases:

1. The first for the definition of Bmax, the minimum buffer size that ensure complete independence between the two stations.
2. The second to study the effect on OEE for a buffer of a size minor then B max.

Both phases have been repeated in two different scenario: The first scenario consider only the losses due to the effect of cycle time variability. It wants to investigate how, in absence of opportune buffer size, the cycle time variability of one station can affect the performance of the other station and so of the entire production line.

Second scenario considers also the effects of failures and theirs variability.

The first phase was carried out by changing the data input (Ti, Ai, Pi) within the range defined above, with an infinite buffer size.. 50 repetitions of the

production period of more than 160 hours for each configuration case was carried out.

All the simulation parameters are briefly reported on table 1.

Table 1: Simulation Parameters

Parameter	Description	Range
Bfix(j)	The chosen buffer size for the specific (j) simulation.	[0;Bmax]
Ti	Ideal Cycle Time to process a piece (min)	[0,01667; 0,25]
Vi	Ideal line throughput in one hour of production (pieces/h)	-
P(i)	Performance represents the speed at which the station (i) runs as a percentage of its designed speed	[0,8;0,99]
A(i)	Availability represent the percentage of scheduled time that the station (i) is available to operate	[0,8;0,99]
MTTR (i)	Mean time to repair of the station i	[0,01;30]
MTBF (i)	Mean time between failures of the station i. It is a function of Availability, where $A(i) = \frac{MTBF}{MTBF + MTTR}$	-
σ (mtbf, mtrr)	As percentage of MTBF or MTTR	[5%;100%]
B Max	The maximum buffer size that allow to decouple globally the two stations	
Throughput	Total number or processed pieces	

In this phase the output data was the maximum buffer size(for all 50 replications) and it is called Bmax.. This value assures the real maximal throughput of the production line (complete independence between the two stations).

The second phase was carried out by using the same data input (Ti, Pi, Ai) configuration, increasing step by step the buffer size Bfix(j) within the range [0; Bmax).

In this phase the output data was the line throughput and the corresponding OEE for each set of parameters Ai, Pi, Ti, Bfix(j). OEE have been measured as the Ratio between the Real line throughput and the ideal line throughput

3.3. Model Validation

Before to start any kind of analysis the model was validated by comparison of the second simulation scenario (more complete) with a work already available in literature. Specifically, the chosen work for validation was the simulative study proposed by (D. BAattini 2009).

The Output Parameter used for the validation of the simulation model was the maximum buffer size obtained under a specific simulations conditions.

The output results of the model in comparison with the available results in literature (once fixed same configuration and same statistical distribution) are briefly showed in table 2.

Table 2: Model Validation

A(a)	A(b)	Mtrr(a)	Mtrr (b)	Model Result	Difference (%)
93%	95%	0,03	14,00	71	1,43%
92%	96%	0,05	12,00	61	0,00%
97%	92%	0,50	20,50	105	0,96%
92%	97%	0,50	20,50	105	0,00%
99%	95%	1,00	26,50	135	0,75%
99%	92%	1,00	25,50	130	0,78%
99%	92%	1,50	21,50	112	2,75%

The statistical significance of the model is high. The obtained R² is of 99,88%.

The Analysis of Variance (ANOVA) on the regression test in order to validate the regression test indicate a P-value of 0,000. Therefore the model is statistically validated.

4. SIMULATION RESULTS

4.1. First Scenario: Effect of cycle time variability

This scenario considers an equal value of the ideal cycle time for the two stations. Ideal cycle times data are distributed as an exponential distribution within this range [0,01667; 0,25]. The performance index is included within the range [0,8;0,99].

investigates how speed losses can affect the OEE of a production line in absence of opportune buffer between stations.

Lack of opportune buffer between the two stations can affect dramatically the availability of the system. In fact also if the ideal cycle times of the two stations are equal, the variability of speed that affect the stations are not necessarily of the same entity because depend on different factor. Furthermore Performance index is an average, therefore it could happens that machine present sometimes reduced speed and sometimes an highest speed. The presence of this effect in two consecutive stations can be mutually compensate or add up.

The simulation of this phenomena In order to evidence the effect of the cycle time variability is useful to express the performance of the production line as the percentage of the maximum achievable value of the OEE. This value is obtained when the losses of one station don't affect the performance of the other. In this situation the OEE of the production line is equal to the minimum OEE of the two stations. obtain general results, the Overall effectiveness of the system could be expressed as a related measure.

For example is possible to express it like the ratio between the achievable performance in terms of OEE

for the selected buffer size (OEE(j)) and the OEE ideally achievable with the max size of the buffer: OEE(B max)

Therefore we can introduce the parameter

$$Rel.OEE(j) = \frac{OEE(j)}{OEE(Bmax)} \quad (3)$$

Where:

- OEE(j) is the value of the OEE corresponding to the buffer size (j).
- OEE(Bmax) is the maximum value of OEE(j) and it is achievable with a buffer size equal to Bmax.

Hence:

- in worst condition Rel.OEE (0)=
 $\frac{OEE(0)}{OEE(Bmax)} =$
 $OEE(1)*OEE(2)/\min(OEE(1);OEE(2));$
- in ideal condition, with Buffer size equal to B max the Rel.OEE (B max)=
 $\frac{OEE(Bmax)}{OEE(Bmax)} = 1$

Figure 4 shows the trend of Rel.OEE(j) depending on the buffer size. The two curves represent the minimum and the maximum simulation results. All the others simulation results are included between these two curves. Maximum curve represents the configuration with the lowest difference in performance index between the two stations, the minimum the configuration with the highest difference.

By analyzing the figure 4 it is clear how an inopportune buffer size affect the performance of the line and how increase in buffer size allows to obtain improve in production line OEE. By the way, once achieved an opportune buffer size no improvement derives from a further increase in buffer.

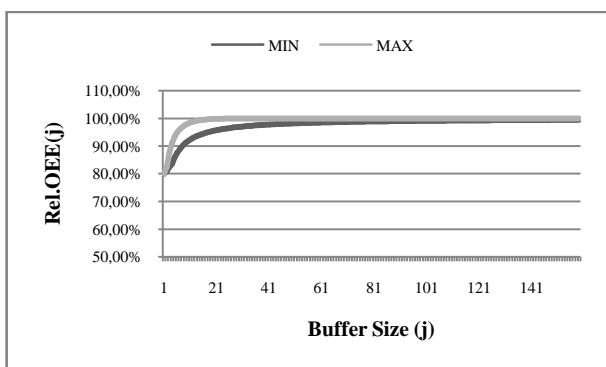


Figure 4: Rel OEE depending on buffer size in system affected by variability due to speed losses

It is important to evidence how the trend of Rel.OEE by change in simulation parameters is really similar between the different cases. Figure 5 shows the deviation between different simulation cases and the value of the curve of the min represented in figure 4.

Figure 5 shows how the gap between the different simulations is negligible. A first assessment of the loss of OEE, depending on buffer size, could be done using the curve of minimum, and the difference with the specific curve of the specific simulation would be negligible in first approximation.

For a example, with a buffer of 17 pieces the Rel.OEE(17) obtained with the simulation with this input parameters (Ti=0,01667; P(a)=0,85; P(b)=0,8) is 97,1%, corresponding to an effective OEE of 77,68%. The value of the Rel.OEE for the same size of the buffer Rel.OEE(17) with the curve of minimum is 94,9%. The deviation between the real curve and the minimum curve of Rel.OEE is 2,2%. To use the curve of minimum as a proxy of the Rel.OEE would bring, in this case, a value of effective OEE of 75,92% with an underestimation of OEE of 1,76%. Then, different curves within all the defined ranges of simulation parameters are available after this study, however the deviation by the specific curve and a first assessment, done by referring to the curve of the minimum is reduced, by increasing the buffers size.

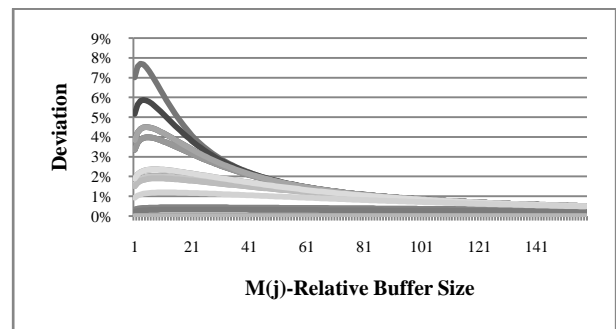


Figure 5: Deviation between the curve of the minimum and other simulation curves

4.2. Second Scenario: Effect of Minor Stoppages

Minor stoppages of the two stations can stop the flow of material processed by the line.

Considering the ratio: $\frac{MTTR(2)}{\frac{Ti}{P(1)}}$ it represent the

maximum amount of material produced by the first station when a failure in second station occurs (P(1) supposed constant and equal to its average).

Likewise the ratio $\frac{MTTR(1)}{\frac{Ti}{P(2)}}$ represents the

maximum amount of material that must be stored on the buffer in order to not affect the second station when a failure in the first occurs.

The maximum of these two ratio could represent a proxy of the buffer size but it does not take into account the effect of variability in MTTR, MTBF, the effect of cycle time speed losses and the moment in which the first failure occurs. Simulation allows to take into account also this effect.

Hence, the buffer size fixed in the specific simulation is expressed as percentage of these ratio

$$M(j) = \frac{B \text{fix}(j)}{\max\left(\frac{MTTR(1)}{P(2)}, \frac{MTTR(2)}{P(1)}\right)} \quad (4)$$

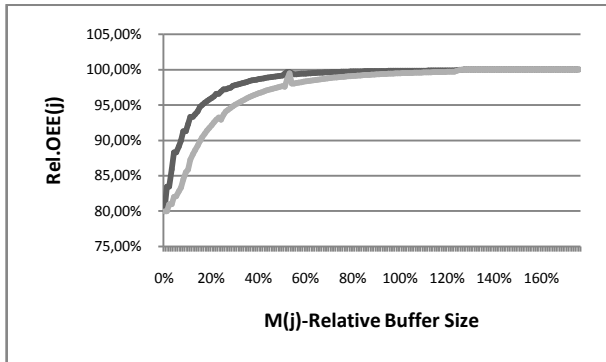


Figure 6: Relative OEE depending on relative buffer size M(j) for a defined value of availability

Authors analyze the behavior of the OEE depending on the buffer size, taking into account all the different configurations of cycle time, different performance yield, and minor stoppages value within the defined range (table 1).

Figure 6 show an example of relative OEE trend depending on the M(j) ratio. The curves represents a configuration where the level of availability is defined, but MTTR and therefore MTBF vary properly within the defined range (table1). For simplicity it reports only two extreme different configurations: Maximum curve represents the configuration with the lowest ideal cycle time (0,01667 min) and minimum curve the configuration with the highest (0,25 min).

Trend like those represented in figure 6 have been developed by authors for each combination of simulation parameters within the defined range (table 1).

The obtained trend are similar for all the Availability level. In fact, Figure 7 shows the deviation between different curves with same level of availability from the curve of the minimum. The proposed figure represent the simulation case with the highest variance between the curves. Hence, considering that simulation results take into account also the effect of variability on MTBF and MTTR (5%), the behavior of the Rel.OEE depending on the relative buffer size is even so regular.

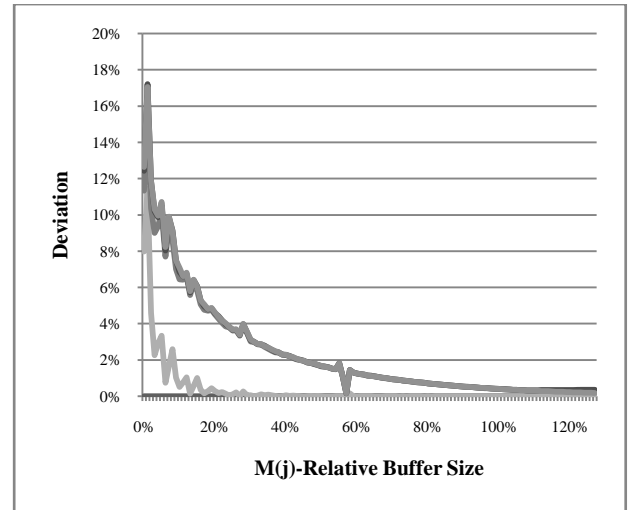


Figure 7: Error in confounding the specific curve

To express the Buffer as a ratio between the chosen buffer size and the amount of material necessary to feed a station during the stoppage of the other evidence how the uncertainty in Performance, and availability affect the buffer size requiring oversized buffer. Nevertheless the frequency of cases in which the Bmax is required is reduced. The differential between different buffer sizes are the cost.

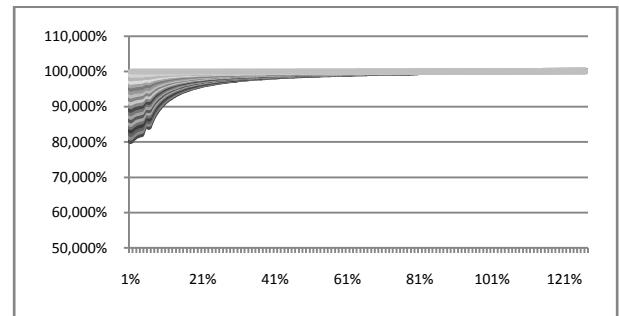


Figure 8: Resume of Rel.OEE depending on M(j) for all the analyzed availability level

The regularity in behavior of the Rel.OEE depending on the M(j) allow to express the evidenced relation between Rel.OEE(j) and M(j) in an analytic way. Trends obtained with simulation analysis (figure 6, 8) are similar to an hyperbole, but more flattened.

Any kind of analytic relation must take into account that in limit configurations, as evidenced in figure 8, the Rel.OEE of the production line is independent from the buffer size. It happens when OEE(station 1) or OEE(station 2) or both tend to 100%.

The analytic relation, obtained by authors to describe properly the considered relation is:

$$\begin{cases} \frac{1-Rel.OEE_j}{1-Rel.OEE_0} = \frac{K_j*(1-M_j)}{M_j} \\ 0,05\% \leq M_j \leq 100\% \end{cases} \quad (5)$$

That can also be expressed as:

$$\begin{cases} Rel.OEE(j) = 1 - \frac{[1-Rel.OEE_0]*[K_j*(1-M_j)]}{M_j*100} & (6) \\ 0,05\% \leq M_j \leq 100\% \end{cases}$$

where:

- $Rel.OEE_j$ is the relative value of O.E.E. obtainable with a (j) size of the buffer (equation 3)
- $Rel.OEE_0$ is the relative value of OEE when buffer size is zero (complete dependence)
- M_j is the percentage dimension of the buffer (equation 4). For values of $M_j > 100\%$, as seen in figure 8, the increase in Rel.OEE is negligible. Therefore this experimental formula is meaningful only for value of $0,05\% \leq M_j \leq 100\%$. If $M_j \leq 0,05\%$ buffer can be considered null, therefore condition of complete dependence occurs, and then $OEE = OEE(\text{station 1}) * OEE(\text{Station 2})$
- K_j is a experimental coefficient that allow to take into account that the behavior of the curve is not a real hyperbole. The value of K_j depends by the value of j and they are reported in Table 4.

Table 4: K_j Value

M_j value	K_j value
0,5%	0,005
1%	0,01
2%	0,019
3%	0,028
4%	0,037
5%	0,039
$\geq 6\%$	0,05

Equations (5) and (6) are the same, but their representation want to evidence two different aspects:

Equation (5) represents the inverse relation between OEE and M_j . When M_j is 100% the buffer is maximum, the second term is null, and the first is null because $REL.OEE = 100\%$, therefore maximum buffer cause maximum REL.OEE.

Equation (6) is easier to use in order to calculate the $Rel.OEE(j)$ for a chosen size of buffer.

For each analyzed configuration the analytic results of the analytic relation (equation 5 and 6) have been compared with simulation result.

The obtained R square index are vey high, and vary, from case to case within [0,945;0,993]. The relative ANOVA output on the significance of regression test

produce a P-value of 0,000. Therefore the formula is statistically validated

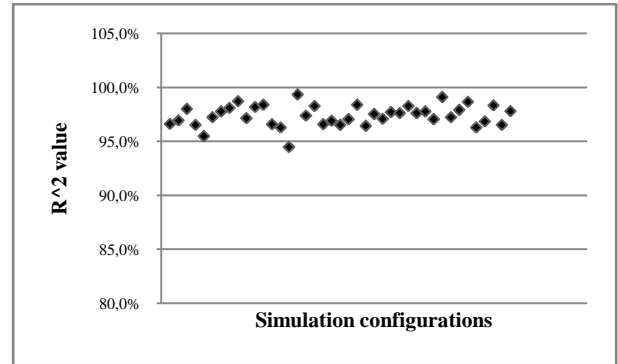


Figure 9: R-square index between Analytic value and simulation results

The formula (5) or (6) allows the buffer size without need for any further simulation.

Two different application of the formula (6) to real industrial cases are summarized in table 4..

Table 4: Formula parameters of two example cases

Cases	MTTR A	MTTR B	T a (Sec)	P A	P B
Case 1	20	30	7	0,9	0,8
Case 2	18	23	15	0,85	0,8

Starting with case 1 we want to show how the Rel.OEE change depending on buffer size, and what are the value assumed by the equation parameter (table 5).

When the buffer size is zero the two station are completely dependent. Therefore the value of the OEE will be the equal to the product of the two performance index ($Pa*Pb=0,8*0,9=0,72$). With a buffer of 3 units M_j is 1,33%, therefore once chosed the right K_j value (0,01) the formula output is $Rel.OEE=92,39\%$, corresponding to a OEE of 73,9%. Hence, a buffer of three unit increase the OEE of the system of 1,9%. For a buffer of 12 units, corresponding to an M_j of 5,19%, once selected the right K_j (0,039) the $Rel.OEE$ is 92,87%, corresponding to an OEE of 74,3%. It means a further increase in OEE of 0,4%. Further increase in buffer size generates increase in OEE.

Table 5: Result description of Formula (6) for Case 1

Buffer size	M_j	K_j	Rel OEE	OEE
0	0%	-	90,00%	72,0%
3	1,30%	0,01	92,39%	73,9%
12	5,19%	0,039	92,87%	74,3%
22	9,51%	0,05	95,24%	76,2%
32	13,83%	0,05	96,88%	77,5%
42	18,15%	0,005	99,77%	79,8%

The same consideration can be done for the case two, and results are briefly shown in Table 6.

Table 6: Result description of Formula (6) for Case 2

Buffer	Mj	Kj	Rel OEE	Oee
0	-	-	85,00%	68,0%
1	1,28%	0,01	88,42%	70,7%
4	5,12%	0,039	89,15%	71,3%
6	7,67%	0,05	90,98%	72,8%
15	19,18%	0,05	96,84%	77,5%
25	31,97%	0,005	99,84%	79,9%

5. CONCLUSION AND FURTHER RESEARCH

The goal of this study have been to deliver an operative tool that allows an effective, but easy sizing of the buffer in flow shop industries also considering all the necessary information regarding the OEE trend.

The wide range of values that have been simulated allows to include in the study a significant amount of different production systems and to evidence some analogies in the behavior between them. Pharmaceutical sector, where authors have already applied this study, is also included. the simulation range.

Big effort in simulation analysis in conjunction with deep knowledge of the physical problem of the buffer design allow the introduction of a analytic relation. The added value of the analytic relation is the possibility to assess immediately, without the need for further simulations, and with strong statistical significance the optimal buffer size for a chosen level of availability. This relation is valid within the wide range of simulated value, and that proximally will be even wider.

In fact studies to obtained a further more general analytic relation has already begun. Many other area of research are possible, such as a deeper analysis on the effect of time variability on buffer size, or the introduction of a analytic method that allows, once define the required OEE; to obtain an estimation of the required buffer size en a easier way, and without recourse to the recursive computation.

Further research could be also carried out by changing the statistical distribution and the method of analysis

Keywords: Availability, Buffer, Buffer Design for Availability, Flow Shop, Simulation.

REFERENCES

- Chiadamrong, N. &. (2003). Using storage buffer to improve unbalanced asynchronous production flow line's performance. *International Journal of Manufacturing Technology and Management* , 149-161.
- D. Battini a, A. P. (2009). Buffer size design linked to reliability performance: A simulative study. *Computers & Industrial Engineering* 56 (2009) , 1633–1641.
- Gershwin, S. G. (1995). *Efficient algorithms for transfer line design*. MIT Laboratory for Manufacturing and Productivity Report LMP-95-005.
- Gutowski, T. (2005). Inventory buffer size. <http://web.mit.edu>.
- H.T. Papadopoulos, C. H. (1996). Queueing theory in manufacturing systems analysis and design: A classification of models for production and transfer lines. *European Journal of Operational Research* , 1-27.
- Hillier, F. S. (1977). On the optimal allocation of work in symmetrically unbalanced production line systems with variable operation times. *Management Science*, 25(8), , 721-728.
- Hillier, F. S. (1993). Some data for applying the bowl phenomenon to large production line systems. *International Journal of Production Research* , 811–822.
- Kingman, J. F. (1966). *On the algebra of queues* . London: Methuen.
- Lawless, J. (1982). *Statistical models and methods for lifetime data*. John Wiley & Sons.
- Lutz, C. M. (1998). Determining buffer location and size in production line using tabu search. *European Journal of Operational Research* , 301–316.
- Malakooti, B. B. (1994). Assembly line balancing with buffers by multiple criteria optimization. . *International Journal of Production Research*, 32(9) , 2159–2178.
- Papadopoulos, H. T. (2001). heuristic algorithm for the buffer allocation in unreliable unbalanced

production line. *Computers and Industrial Engineering* , 261-277.

Samuel H. Huangt, J. P. (2003). Manufacturing productivity improvement using effectiveness metric and simulation analysis. *International Journal Production Research* , 513-527.

Slack, N. (1993). The Flexibility of Manufacturing Systems. *International Journal of Operations & Production Management* , Vol. 7 Iss: 4, pp.35 - 45.

Spinellis, D. D. (1999). Production line buffer allocation: Genetic algorithms versus simulated annealing. *Second international Aegan conference on the analysis and modelling of manufacturing systems*, (pp. 12-101).

Tempelmeier, H. (2003). Pratical considerations in the optimization of flow production systems. *International Journal of Production Research* , 149–170.

Yamada, T. &. (2003). A management design approach to assembly line systems. *International Journal of Production Economics*, 84 , 193–204.

Operations management, Energy Management, Project Management.

Vito Introna has published more than 40 articles on several leading international and national journals, and presented research results in several international conferences.

AUTHORS BIOGRAPHY

Vittorio Cesarotti is Tenured Associate Professor at the University of Rome “Tor Vergata”, holding the Chairs of Operations Management and Quality Management

His Areas of Expertise are Strategy and Organization, Operations Management, Production Planning and Control, Business Excellence, Business Performance Measurement and Improvement, Quality Management and Control, Facility Management, Project Management, Supply Chain Management and Service Science.

Prof. Cesarotti has published one book, seven international case studies, and more than 60 articles on several leading international and national journals, and presented research results in several international conferences.

Alessio Giuiusa is PhD Candidate in Managerial Engineering at the University of Rome “Tor Vergata”.

His Areas of Expertise are Operations Management, Quality Management and Control, Project Management and Service Science.

Vito Introna is Assistant Professor in Industrial plant at University of Rome Tor Vergata.

His Areas of Expertise are Maintenance, Statistical process control, Quality Management and Control,

USING QUERY EXTENSION AND USER FEEDBACK TO IMPROVE PUBMED SEARCH

Viktoria Dorfer ^(a), Sophie A. Blank ^(b), Stephan M. Winkler ^(c),
Thomas Kern ^(d), Gerald Petz ^(e), and Patrizia Faschang ^(f)

^(a,b,c,d) University of Applied Sciences Upper Austria
School of Informatics, Communications and Media, Bioinformatics Research Group
Softwarepark 11, 4232 Hagenberg, Austria

^(e,f) University of Applied Sciences Upper Austria
School of Management, Digital Economy Research Group
Wehrgrabengasse 1-3, 4400 Steyr, Austria

^(a) viktoriam.dorfer@fh-hagenberg.at, ^(b) sophieanna.blank@students.fh-hagenberg.at, ^(c) stephan.winkler@fh-hagenberg.at,
^(d) thomas.kern@fh-hagenberg.at, ^(e) gerald.petz@fh-steyr.at, ^(f) patrizia.faschang@fh-steyr.at

ABSTRACT

PubMed is a search engine that is widely used to search for medical publications. A common challenge in information retrieval, and thus also when using PubMed, is that broad search queries often result in lists of thousands of papers that are presented to the user, too narrow ones often yield small or even empty lists. To address this problem we here present a new PubMed search interface with query extension using keyword clusters generated with evolutionary algorithms to obtain more specific search results. Users can choose to add various words to their query and then rate search results; this scoring is stored in a database to enable learning from user feedback to improve keyword cluster optimization as well as query extensions. We show how users can extend PubMed queries using previously generated keyword clusters, rate query results, and use these ratings for optimizing parameters of the keyword clustering algorithms.

Keywords: Bioinformatical Information Retrieval, PubMed, Query Extension, Keyword Clustering

1. INTRODUCTION

Search queries in PubMed (PubMed 2011) often return a huge number of papers, many of them not relevant for the user. To address this problem of irrelevance, information retrieval provides a suitable solution, namely query extension. User queries are extended with matching terms to result in fewer, but relevant search results. This method is based on clusters of keywords and / or documents; from these clusters words are selected that belong to the same clusters as the search terms defined by the user, and thus queries can be extended automatically. Figure 1 shows a schematic

view of an exemplary keyword clustering solution consisting of three clusters.

In this paper we present a new interface for PubMed search that uses various keyword clusters to extend user queries. These keyword clusters have been generated using various evolutionary algorithms as described in Dorfer, Winkler, Kern, Petz, and Faschang (2010) and Dorfer, Winkler, Kern, Blank, Petz, and Faschang (2011). User ratings of the query results (returned by PubMed for queries that were extended using keyword clusters) will then be used to improve the generation of keyword clusters and query extensions.

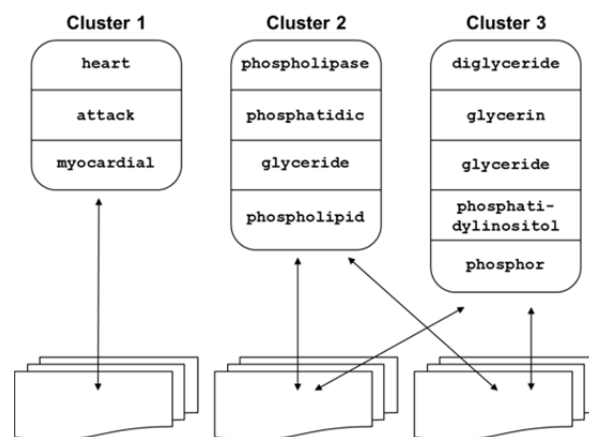


Figure 1: Example of a keyword clustering solution

2. A NEW PUBMED SEARCH INTERFACE

We have developed a PubMed search interface that uses previously generated keyword clusters to suggest words that fit the terms the user entered as query. Several clustering files can be selected to search for suitable extensions.

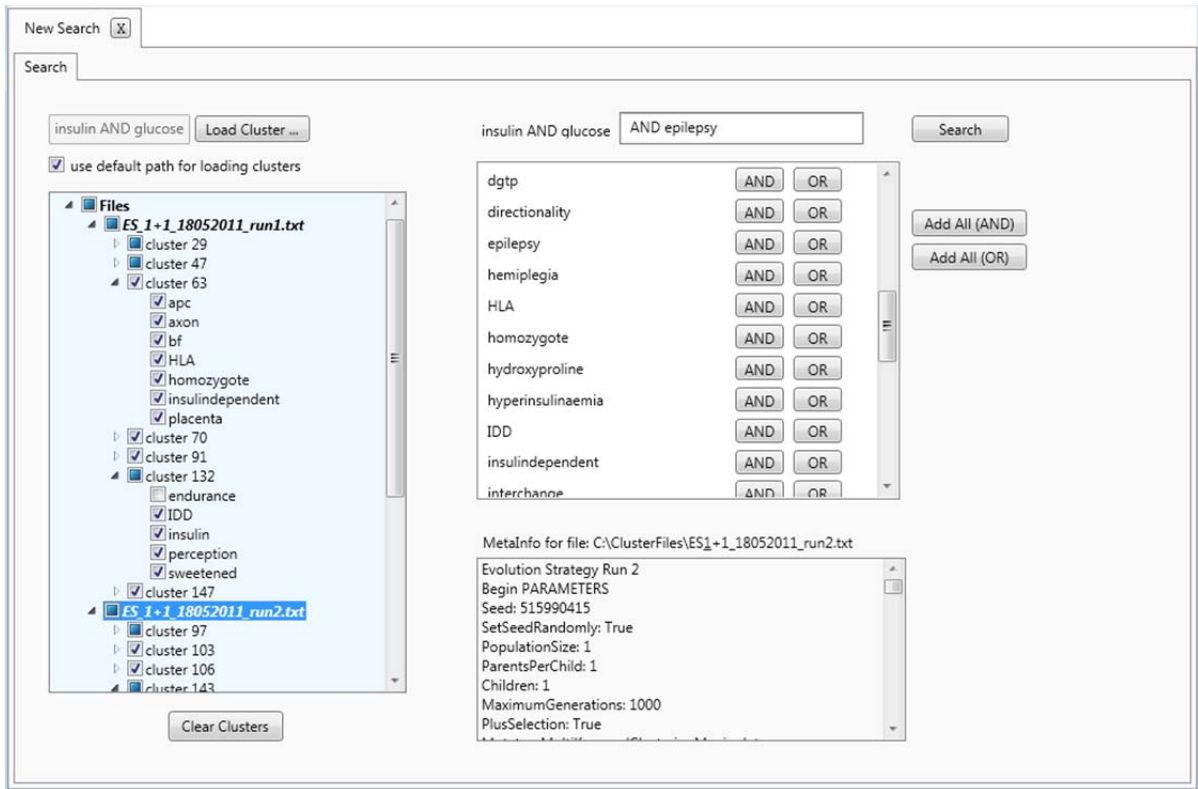


Figure 2: PubMed query extension using previously generated keyword clusters

In Figure 2 a screenshot of the search and extension interface is provided. On the left hand side the keyword clusters containing terms or parts of terms of the user query are shown; these keywords can be selected for query extension.

On the right side the chosen terms can be added to the query using ADD or OR conjunctions. More than one search query can be submitted in parallel to be able to compare results with different selections of keywords extending the original query.

3. KEYWORD CLUSTER GENERATION

As mentioned before, these keyword clusters have been generated using evolutionary algorithms, driven by mutation, crossover and selection (see Dorfer, Winkler, Kern, Petz, and Faschang (2010) for further details on the operators implemented for this problem class) using a fitness function that is used for evaluating solution candidates. This fitness function takes various features into account we consider important for keyword clustering solution candidates and is defined by the following equation (Dorfer, Winkler, Kern, Petz, and Faschang (2010)):

$$F = \alpha \cdot A + \beta \cdot B + \gamma \cdot C + \delta \cdot D + \varepsilon \cdot E + \zeta \cdot G. \quad (1)$$

Parameter A represents the number of clusters a document is assigned to; parameter B the data coverage; C – the cluster confidence – quantifies to which amount the keywords of the cluster are also present in the documents assigned to the cluster, whereas in parameter

D – the document confidence – the amount of keywords in the documents also present in the cluster is represented. Parameter E regards the distribution of the documents in the clusters and G considers the number of generated clusters with respect to the data set size. α , β , γ , δ , ε , and ζ are weighting factors that are necessary to be able to emphasize or neglect specific features, depending on the user's needs.

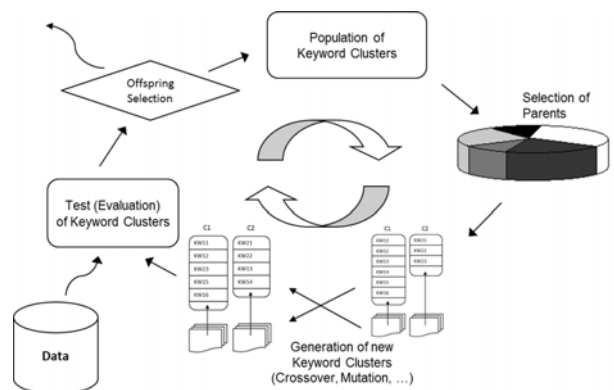


Figure 3: Genetic algorithm with offspring selection optimizing keyword clusters

In Dorfer, Winkler, Kern, Blank, Petz, and Faschang (2011) we have presented a detailed comparison of the performance of various evolutionary algorithms solving this keyword clustering problem (KCP), including the evolution strategy (ES) (Schwefel 1994), the genetic algorithm (GA) (Holland 1975), the

genetic algorithm with offspring selection (OSGA) (Affenzeller, Winkler, Wagner, and Beham 2009), and the elitist non-dominated sorting genetic algorithm (NSGA-II) (Deb, Pratap, Agarwal, and Meyarivan 2002). In Figure 3 we show the most important parts of an evolutionary algorithm solving the KCP (in particular, an OSGA is depicted).

The evolutionary process in ESs is mainly based on mutation and selection, whereas evolution in GAs is directed by the interaction of mutation, crossover and selection; the OSGA is an extension of the standard GA and includes an additional offspring selection step. The NSGA-II is a multi-objective approach, which optimizes various objectives in parallel, as we have identified parameters B, C, and D as potentially contrary.

We have shown that intense parameter settings and weighting factor tuning is necessary to receive good solutions using single-objective algorithms and that ESs and the NSGA-II perform best in terms of achievable solution quality (Dorfer, Winkler, Kern, Blank, Petz, and Faschang (2011)).

The data set used to generate the mentioned keyword clusters has been taken from the 9th Text Retrieval Conference (TREC-9) from the year 2000 (Vorhees and Harman 2000), consisting of 36,890 samples based on PubMed entries including title, abstract, authors, its medical subject headings (MeSH), publication type and source.

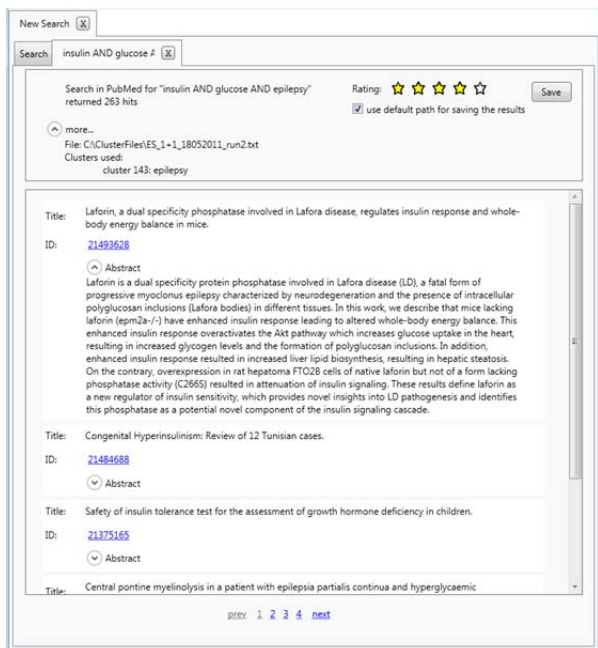


Figure 4: Example of search results with rating possibility

4. USER FEEDBACK

After the search has been performed the user can rate the results (see Figure 4); this rating in combination with the words used for query extension and the algorithm parameters the keyword clusters have been created with (algorithm type, algorithm parameters, documents used for clustering, ...) are stored in a database in order to be able to examine the settings and algorithms which delivered the best results.

Another interesting aspect in the context of biomedical information retrieval is whether there are specific keywords that, when included in the query, never lead to satisfying solutions. By analyzing user feedback these words can be identified and collected in a blacklist to warn the user as soon as such a term is included in the query.

User feedback cannot only be used for the generation of a blacklist, but also to improve the parameter settings of the keyword clustering algorithms. This shall enable an iterative improvement of parameter settings leading to better and more suitable clusters which can then be used in query extension.

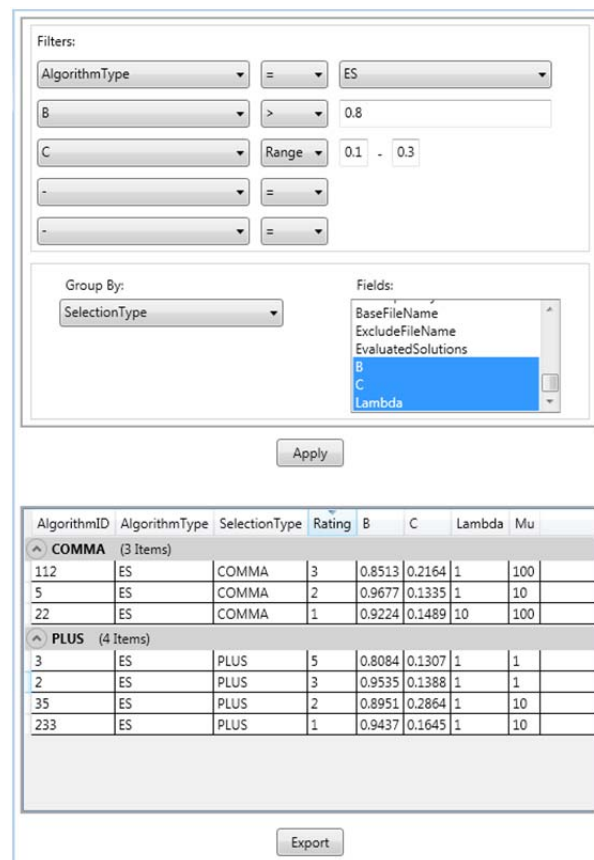


Figure 5: Analysis of algorithm parameters that were applied for forming keyword clusters using various evolutionary algorithms. The keyword clusters have been used for query extension and the search results have been rated; algorithm settings are ranked according to the user's feedback

In Figure 5 we show an example of the examination of rating results and algorithm specific parameter settings to find out whether these settings yield appropriate solutions. All fields of the database can be selected for filtering, including general algorithm parameters (such as algorithm type, population size, and also fitness function specific parameters as – for example – B and D) and of course the user ratings. The previously explained keyword clusters and the evolutionary algorithms used to generate these clusters provide the basis for this analysis. Based on the results of these examinations we want to further improve the algorithm parameters to optimize the generation of keyword clusters; this shall extend our optimization algorithm with a user-driven feedback component.

5. APPLICATION EXAMPLES

Using the best keyword clustering files identified in Dorfer, Winkler, Kern, Blank, Petz, and Faschang (2011), generated by ES, GA, OSGA and the NSGA-II, a detailed analysis on the differences of the generated clusters and on the performance of the different algorithms with the different parameter settings with respect to user feedback can now be performed.

In Figure 6 we provide an application example, in particular, we search for “urethra”. We can clearly see the differences in the clusters generated by ES and NSGA-II. As evolution in evolution strategies solving the KCP is dependent on the fitness function given in Equation 1, quite different keywords are grouped together in contrast to the clusters generated by the NSGA-II, which in this example optimizes only parameters B, C, and D. On the right bottom of the window detailed information on the used algorithm and its parameter settings to generate the cluster is given. All shown keywords can be chosen for query extension, optionally in combination; this selection does not depend on the algorithm the clusters have been generated by.

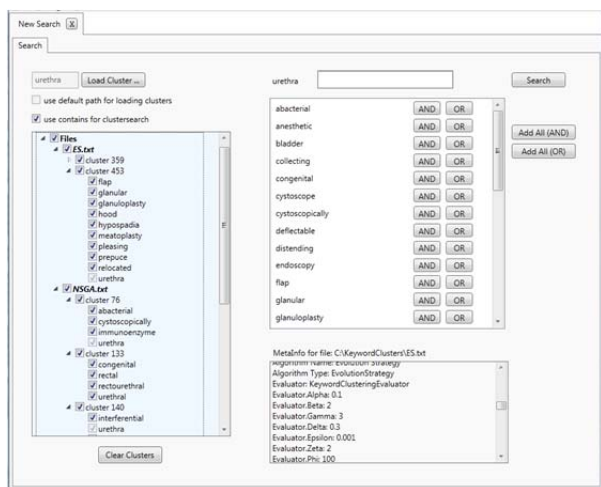


Figure 6: Application example using keyword clusters generated by two different evolutionary algorithms

After users have performed various queries and have rated the obtained results accordingly, we can now have a further look on the performance of the various evolutionary algorithms and their corresponding settings in terms of parameters and also weighting factors. In Figure 6 we can see that here keyword clusters generated by evolution strategy with specific weighting factor settings and also generated by the NSGA-II provided clusters the users ranked best. However, this does not have to be the case in all examples; one can assume that specific user needs and specific combinations of terms are more likely to be produced by a genetic algorithm, for example. This and other questions can be analyzed by applying various filters on the feedback data, beside the possibility to gain a quick insight in which algorithm performs best regardless of the specific user needs.

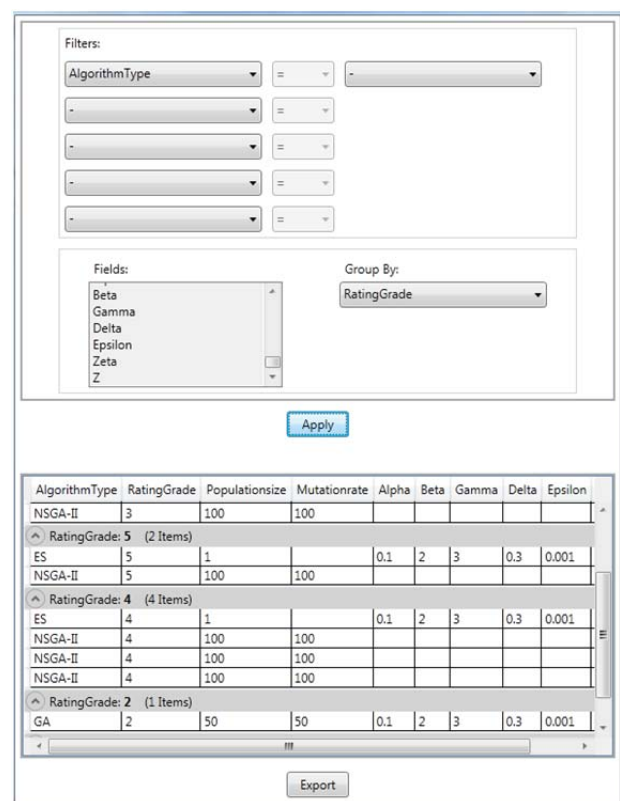


Figure 7: Application example on the analysis of user feedback

As this new search tool shall be used to improve PubMed search results we give another application example here: Assuming – for the sake of simplicity – one wants get some information about the effect of smoking on the lung, a search of these two terms (“smoke AND lung”) would yield 7310 hits. In Figure 8 the words that often occur in combination with at least one of these two terms are depicted.

To be sure not to have missed any paper on this topic, the query can be extended with an OR conjunction on the term “pulmonary” to retrieve also all the papers where pulmonary has been consequently used as a synonym for lung, leading to a publication list of 8384 entries.

Alternatively, a reduction of the result space can be obtained by adding more specific terms inspired by the suggestion of the keyword clusters. Adding for example the keyword “asthma”, PubMed returns only 699 hits, adding the keyword “haemoptysis” leads to only 6 results.

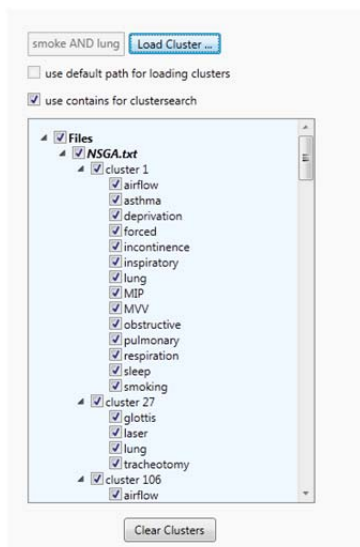


Figure 8: Keyword clusters containing the terms “lung” or “smoke”

6. CONCLUSION

In this paper we have presented a new query interface that used keyword clusters to improve PubMed queries. The used clusters have been generated using various evolutionary algorithms and different parameter settings. In addition to the extension functionality, obtained results can be rated and various analyses can be performed on this kind of user feedback.

We have shown that the various algorithms generate different keyword clusters and that it is therefore reasonable to use several keyword clustering files to choose the words for extension. The provided example described the applicability for extending and narrowing the search space by the use of either synonyms or specifications, depending on the users’ needs in the context of PubMed search.

7. FUTURE WORK

We are currently working on an online version for the proposed tool for improved PubMed querying. This version will enable us to address a broader user clientele and then to do further analyses on their feedback. Having now identified the best performing algorithms we plan to slightly adapt parameters and weighting factors and to perform further keyword clustering to hopefully produce even better keyword clusters. This will be an ongoing process as new keyword clusters will lead to new ratings and so on.

As mentioned before we also plan to design a black list of terms or of combinations of terms and algorithms that are less likely to lead to useful search results. We are also working on the automated

generation of so-called white lists; these are lists of terms that should always be used in combination to obtain meaningful results. Users can then be warned or encouraged to include specific terms or keyword clusters of specific algorithms in their search.

ACKNOWLEDGMENTS

The work presented in this paper was done within the TSCHECHOW project, sponsored by the basic funding program of the University of Applied Sciences Upper Austria.

REFERENCES

- Affenzeller, M., Winkler, S., Wagner, S., Beham, A., 2009. *Genetic Algorithms and Genetic Programming - Modern Concepts and Practical Applications*. Chapman & Hall/CRC. ISBN 978-1584886297. 2009.
- Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6:182-197.
- Dorfer, V., Winkler, S.M., Kern, T., Blank, S.A., Petz, G., and Faschang, P., 2011. On the Performance of Evolutionary Algorithms in Biomedical Keyword Clustering, *Proceedings of the Genetic and Evolutionary Computation Conference*, 10, July 12-16, 2011, Dublin, Ireland.
- Dorfer, V., Winkler, S.M., Kern, T., Petz, G., and Faschang, P., 2010. Optimization of Keyword Grouping in Biomedical Information Retrieval Using Evolutionary Algorithms, *Proceedings of the 22nd European Modeling and Simulation Symposium*, 25-30, October 13-15, 2010, Fes, Morocco.
- Holland, J.H., 1975. *Adaption in Natural and Artificial Systems*. University of Michigan Press
- PubMed, 2011, Available from: <http://www.ncbi.nlm.nih.gov/pubmed>
- Schwefel, H.-P., 1994. *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie*. Basel: Birkhäuser Verlag
- Vorhees, E.M., Harman, D.K., 2000. *NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9)*. Gaithersburg, Maryland: Department of Commerce, National Institute of Standard and Technology.

AUTHORS BIOGRAPHIES



VIKTORIA DORFER is a senior researcher in the field of Bioinformatics at the Research Center Hagenberg, School of Informatics, Communications and Media. After finishing the diploma degree of bioinformatics in 2007 she was a team member of various projects in the field of bioinformatics and software engineering. She is currently working on biomedical information retrieval within the TSCHECHOW project.



SOPHIE A. BLANK received her bachelor's degree in bioinformatics from the Upper Austria University of Applied Sciences in 2011; she is currently a junior researcher at the Research Center Hagenberg, School of Informatics, Communications and Media working on biomedical information retrieval within the TSCHECHOW project.



STEPHAN M. WINKLER received his MSc in computer science in 2004 and his PhD in engineering sciences in 2008, both from Johannes Kepler University (JKU) Linz, Austria. His research interests include genetic programming, bioinformatics, nonlinear model identification, and machine learning. Since 2009, Dr. Winkler is professor at the Department for Medical and Bioinformatics at the Upper Austria University of Applied Sciences, Campus Hagenberg, and since 2011 he is head of the Bioinformatics Research Group.



THOMAS KERN is head of the Research Center Hagenberg, School of Informatics, Communications and Media, Upper Austria University of Applied Sciences (UAS). He finished his studies in Software Engineering in 1998. After some work experience in industry he started his academic career at UAS in autumn 2000 as research associate and lecturer for algorithms and data structures, technologies for knowledge based systems, semantic systems and information retrieval. Since 2003 he has conducted several application oriented R&D projects in the fields of bioinformatics and software engineering.



GERALD PETZ is director of the degree course Marketing and Electronic Business at the University of Applied Sciences in Upper Austria, School of Management in Steyr. His main research areas are Web 2.0, electronic business and electronic marketing; he has also conducted several R&D projects in these research fields. Before starting his academic career he was project manager and CEO of an internet company.



PATRIZIA FASCHANG received her bachelor's degree in electronic business and is a junior researcher at the Research Center Steyr, School of Management, in the area of digital economy. She has worked on several small projects in the field of marketing and electronic business and is currently working on Opinion Mining and Web 2.0 methods within the TSCHECHOW project.

SIMULATION OF THE VESSEL TRAFFIC SCHEDULE IN THE STRAIT OF ISTANBUL

Şirin Özlem^(a), İlhan Or^(a), Birnur Özbaş^(b)

^(a)Boğaziçi University, Istanbul

^(b)Rutgers, The State University of New Jersey

^(a) sirinozlem@yahoo.com, or@boun.edu.tr, ^(b) birnur@ozbas.com.tr

ABSTRACT

In this study, a simulation model is developed via the Arena 11.0 software to mimic the actual Istanbul Strait vessel flow under the established traffic regulations and meteorological conditions. The established practice of uni-directional daytime and two-directional nighttime traffic schedules are reflected and pilot and tugboats services scheduled in the traffic flow direction, visibility, current and storm information are also integrated into the model. The effects of factors such as pursuit distance, vessel profile, pilot availability, arrival rate and visibility over selected performance measures are investigated through scenario analysis and the most important factors are determined as arrival rate of vessels and visibility.

Keywords: Strait of Istanbul, Maritime traffic, Simulation

1. INTRODUCTION

The Istanbul Strait, 31 kilometers in length is one of the narrowest waterways in the world with only 660 meters at its narrowest point (Almaz 2006). Vessels navigating through the Strait have to make many sharp turns (between 45 and even 80 degrees) which carry high risks for the vessels in such a narrow channel (Ulusçu et Al. 2009). The Strait which is situated in the middle of a huge metropolitan area of 15 million residents, features a very heavy maritime traffic (more than 51,000 vessels annually), with more than 15,000 such vessels carrying dangerous cargo; there is also heavy local traffic including more than 2,000 passenger ferry trips daily between the two shores (Gönültaş 2007).

One noteworthy property of the Strait is the prevailing currents which may rise up to 8 knots speed. Other adverse meteorological conditions like fog, wind, rain and storm also increase the difficulty of navigation in the Strait. In dense fog conditions, vessel traffic may be partially or wholly suspended until meteorological conditions improve, which causes dangerous and unwanted pile-ups at the Strait entrances and puts further strains on the maritime traffic management, since it increases navigation problems (Özbaş 2005).

The Vessel Traffic System (VTS) was established in 2004 in order to regulate and guide maritime traffic in the Strait, in accordance with international and

national conventions and regulations, while improving safe navigation, protecting life and environment. Within the framework of this system, vessels desiring to transit the Strait have to submit two reports to the VTS, Sailing Plan 1 (SP-1) and Sailing Plan 2 (SP-2). SP-1 includes all the information about the vessel and must be submitted at least 24 hours before the arrival. SP-2 is of vital importance for planning of vessel passages from the Strait and must be submitted at least 2 hours or 20 nautical miles (whichever comes first) prior to entry into the Strait. The VTS analyze the data in these reports and prepare a safe daily sailing traffic plan (VTS Users' Guide).

2. SIMULATION MODEL

The first step to better understand the risks generated by the maritime traffic in the Strait is to understand and model the maritime actively in the Strait. This study aims to design and develop a simulation model to represent the actual traffic flow in the Strait with regard to the VTS rules and regulations (R&R) and policies that meteorological and geographical conditions, support services (like pilot and tugboats) and frequency, type and cargo characteristics of vessel arrivals (to make a passage through the Strait) with the aim of identifying the impact of such factors on traffic conditions, potential problems and bottlenecks for a less risky transit and overtaking allowance during the passage of vessels on Strait lanes.

2.1. Vessel Classification

The VTS has a specific vessel classification system based on vessel types, cargo characteristics and vessel lengths. In this study a somewhat simplified version of this classification (which is displayed in Figure 1) is used.

The main reason why tankers and dangerous cargo vessels up to 100 meters and LPG-LNG up to 150 meters, tankers and dangerous cargo vessels between 100 and 150 meters and dry cargo carrying vessels between 150 and 300 meters are placed in the same class is that according to the VTS regulations, they have to satisfy the same conditions in entering and navigating the Strait. This way of classification simplifies the understanding of vessel entrance and sailing conditions.

Length (m.)	Draft (m.)	Type				
		Tanker	LNG-LPG	Carrying Dangerous Cargo	Dry Cargo	Passenger Vessels
< 50	< 15	Class E	Class C	Class E	Class D	Class P
50 - 100	< 15					
100 - 150	< 15					
150 - 200	< 15	Class B			Class C	
200 - 250	< 15	Class A				
250 - 300	> 15					
> 300	> 15	Class T6				

Figure 1: Vessel Classification

2.2. The Arrival Process

The Arena Input Analyzer which is a very efficient tool for distribution fitting to data is deployed in fitting interarrival time distributions. Via the Input Analyzer's Fit menu, all probable distributions fitted to the actual data are revealed and "fit all" property estimates the distribution with the minimum square error. After fitting a distribution, a histogram and the probability density function (pdf) superimposed on the histogram summarize the characteristics of the fit (Law and Kelton 2007). To illustrate, the best fitted interarrival time distribution of northbound Class E vessels is found as the Gamma distribution with shape parameter α being 648 and scale parameter β being 0.974. In the summary report of Arena Input Analyzer (as displayed in Figure 2), the shape of the probability density function overlaps with the histogram and just looking at this figure, one gets the feeling that the selected function represents the actual interarrival time data quite well.

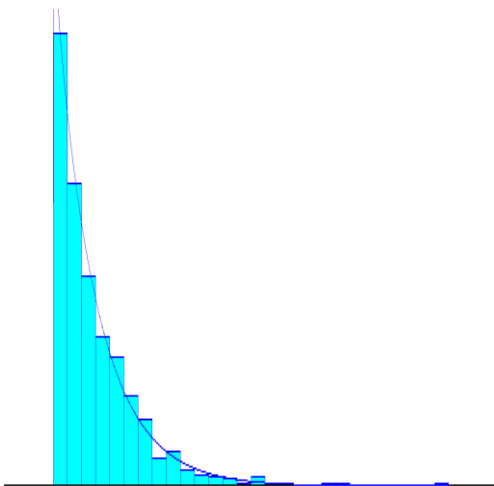


Figure 1: Histogram of northbound Class E interarrivals

2.3. The Istanbul Strait Traffic Rules and Regulations

Vessels enter the Strait either from the north, (traveling south and thus are called as southbound vessels) or from the south (traveling north and thus are called northbound vessels) entrances.

Some R&R related to vessel transit management that are also reflected in the simulation model are as follows:

- There should be at least a 10-minute interval between two consecutive ready to enter vessels from one direction.
- Class A and T6 vessels pass through the Strait only during daytime.
- No vessels are allowed to meet with Class A vessels.
- Class B, C and E vessels should not meet each other during bi-directional nighttime flow.
- There should be at least 75 minutes between two consecutive southbound Class A vessels and at least 90 minutes between two consecutive northbound Class A vessels.
- Passenger vessels are allowed to the Strait regardless of their direction of flow when pursuit distance, meteorological and pilot and tugboat request conditions are satisfied.
- Southbound stopover vessels have priority over northbound stopover vessels, which have priority over any non-stopover vessels.

2.4. Vessel Sequencing

Observations of the 2009 transit data and discussions with the VTS authorities have indicated that the implementation of the regulations regarding pursuit distances between two consecutive vessels of various classes can be parameterized into a set of easily followed rules.

Let θ be the minimum pursuit distance between two consecutive vessels of class D, E, P traveling northbound and let μ be the minimum pursuit distance between two consecutive vessels of class D, E, P traveling southbound. According to the R&R, the minimum pursuit distance between a northbound (southbound) class D, E or P vessel and a class A, B or C vessel sailing in the same direction is also θ (μ). The minimum pursuit distance between two consecutive class C vessels traveling northbound (southbound) is $2*\theta$ ($2*\mu$) and the minimum pursuit distance between a northbound (southbound) class C and a class A or B vessel sailing in the same direction is also $2*\theta$ ($2*\mu$). The minimum pursuit distance between two consecutive A and B vessels traveling northbound (southbound) is respectively $6*\theta$ ($6*\mu$) and $4*\theta$ ($4*\mu$).

2.5. Daytime Vessel Scheduling

As mentioned before, traffic flows from one direction at a time during daytime. The maximum duration of daytime and start time of the daytime traffic differ according to seasons. The first direction of vessel flow

into the Strait at daytime is determined based on the total number of vessels in queues and their waiting time regarding vessel priorities (two hours before the starting time). The formula used for in the determination of starting direction is as follows:

$$S^d = a * \frac{C_a * NQ(A)^{(d)} + C_c * NQ(C)^{(d)} + C_d * NQ(D)^{(d)} + C_e * NQ(E)^{(d)}}{NQ(A)^{(d+d')} + NQ(C)^{(d+d')} + NQ(D)^{(d+d')} + NQ(E)^{(d+d')}} + b * \frac{C_a * WT(A)^{(d)} + C_c * WT(C)^{(d)} + C_d * WT(D)^{(d)} + C_e * WT(E)^{(d)}}{WT(A)^{(d+d')} + WT(C)^{(d+d')} + WT(D)^{(d+d')} + WT(E)^{(d+d')}} \quad (1)$$

where:

S^d : score value of the active direction d

$S^{d'}$: score value in the opposite (passive) direction d'

a : multiplicative constant for number of vessels in queues

b : multiplicative constant for waiting time of vessels in queues

C_a : coefficient for A type vessels

C_c : coefficient for C type vessels

C_d : coefficient for D type vessels

C_e : coefficient for E type vessels

$NQ(i)_{t_s}^{(d)}$: number of i type vessels in queue in active direction d at time $t=t_s$

$NQ(i)_{t_s}^{(d')}$: number of i type vessels in queue in passive direction d' at time $t=t_s$

$WT(j)_{t_s}^{(d)}$: total waiting time of j type vessels in active direction d at time $t=t_s$

$WT(j)_{t_s}^{(d')}$: total waiting time of j type vessels in passive direction d' at time $t=t_s$

This formula is applied for both directions and the direction with higher score is declared as the starting direction of the daytime traffic schedule. Two significant factors influencing the determination of the first direction of daytime flow are the number of vessels in queues and vessel waiting times and they are in different level of significance. (The associated weights a and b are nominated as 0.25 and 0.75 respectively).

Class A and T6 vessels are the most critical vessels in terms of the risks they generate. Therefore, in order to set out the framework for daytime schedule, (after attaining the first direction of daytime traffic), number of Class A vessels transiting from both directions are estimated. In this respect, maximum daytime duration is divided into two, proportion to the number of Class A vessels in northbound and southbound queues.

Starting direction traffic time window length is calculated as:

$$W_d = \frac{NQ(A)_{t_s}^d}{NQ(A)_{t_s}^d + NQ(A)_{t_s}^{d'}} \quad (2)$$

Opposite direction traffic time window length is calculated as:

$$W_{d'} = \frac{NQ(A)_{t_s}^{d'}}{NQ(A)_{t_s}^d + NQ(A)_{t_s}^{d'}} \quad (3)$$

The number of Class A vessels planned to enter the Strait during the starting direction vessel traffic flow is:

$$N_p(d) = \frac{W_d}{6 * \delta(d)} \quad (4)$$

where:

$$\delta(d) = \begin{cases} 0 & \text{if } d \text{ is northbound} \\ \mu & \text{if } d \text{ is southbound} \end{cases} \quad (5)$$

The parameters in the denominator changes with regard to starting direction decision.

The number of Class A vessels planned to enter the Strait during the opposite direction vessel traffic flow is:

$$N_p(d') = \frac{W_{d'}}{6 * \delta(d')} \quad (6)$$

Both $N_p(d)$ and $N_p(d')$ are rounded down to nearest integer numbers.

Waiting time of vessels is adjusted depending on whether they are stopover vessels or not. The adjusted waiting time of vessel j is defined by:

$$W_j^a = c * WT_{t_s}^d(j) \quad (7)$$

where:

$$c = \begin{cases} 1.5 & \text{if } j \text{ is a stopover southbound vessel} \\ 1.25 & \text{if } j \text{ is a stopover northbound vessel} \\ 1 & \text{otherwise} \end{cases} \quad (8)$$

Since passenger vessels have the highest priority in vessel sequencing, the model first searches the Class P queue in the determined direction. If there exist any P vessels in the determined direction and if the visibility conditions and pilot and tugboat demand are satisfied, the one having the maximum elapsed waiting time is allowed to the Strait and the time is incremented as θ (μ) minutes. Meanwhile, if there exist any P vessels on the other side, the one with the maximum elapsed waiting time is allowed to the Strait as well (even though a uni-directional time window is in action). If there is no P vessel in the determined direction, the model searches the Class A queue. If there is any A type vessel in the determined direction, then the pursuit distance requirements, meteorological situations and pilot and tugboat availabilities are checked. When all conditions are fulfilled, the class A vessel having the

maximum elapsed waiting time enters the Strait, otherwise model examines the Class C, E and D vessel queues respectively and allows the one having maximum elapsed waiting time regarding their minimum pursuit distances among class types. As soon as a vessel enters the Strait, again time is incremented as the minimum pursuit distance interval (as θ or μ minutes) and the other distance rules among vessel types are also checked until the last planned A vessel in the active direction enters the Strait.

Since the original daily schedule is made in the morning (two hours before traffic start time), the uni-directional time windows of that schedule are designated to service just the available vessels (especially A vessels) at that time. So, close to the end of the time window of the starting direction, say at time $t = \bar{t}$, the model reviews the number of Class A vessels in queues and revises the original schedule to extend the uni-directional time windows as long as the maximum daytime duration permits. This extended time interval is named as the slack time.

For slack time traffic plan, the number of Class A vessels planned to enter the Strait during the starting and opposite direction uni-directional traffic flow time windows is computed by dividing this apportioned times by the minimum pursuit distance between two consecutive Class A vessel transiting from starting and opposite directions time windows.

The length of slack time is:

$$ST = \text{MAX}(0, DT - (\bar{t} - t_s + W_{d'})) \quad (8)$$

where t_s is the start time of the first direction vessel traffic flow.

The steps for slack time schedule at time $t = \bar{t}$ are as follows:

(i) Number of Class A vessels in the opposite direction at time $t = \bar{t}$ is checked. One important detail at this point is ignoring the number of previously planned vessels in the opposite direction ($N_p(d')$), since they are already scheduled to pass in the original time window determined at plan time. Namely, the new arrivals (since plan time) of class A vessels in opposite direction are:

$$NQ_{\bar{t}}(A)_{SLACK}^{d'} = \text{MAX}(0, (NQ(A)_{\bar{t}}^{d'} - N_p(d'))) \quad (9)$$

(ii) The additional waiting time of new arrival (since plan time) class A vessels in direction d' at time \bar{t} is computed. This can be done by removing the realized waiting time of planned A vessels from total waiting time of Class A in direction d' , that is:

$$WT(A)_{SLACK}^{d'} = WT(A)^{(d')} - WT(A)_{\bar{t}_s}^{d'} \quad (10)$$

(iii) The ratio for number of unscheduled class A vessels in both directions is estimated as:

$$X = \frac{NQ(A)^{(d')}}{NQ(A)_{SLACK}^{d'}} \quad (11)$$

Since \bar{t} represents a time point at which all scheduled vessels in the active direction have already moved into the Strait, the numerator must only contain the new arrival class A vessels since plan time.

i) The ratio for waiting time of unscheduled vessels in direction d and d' at time \bar{t} is calculated as:

$$Y = \frac{WT(A)^{(d')}}{WT(A)_{SLACK}^{d'}} \quad (12)$$

ii) If the amount of slack time is larger than or equal to time length that allows a southbound A vessel transit ($6 * \mu$), the slack time algorithm tries to make use of this time by scheduling one more northbound or southbound class A vessel.

iii) The indicator Z is determined as follows:

$$Z = X * \alpha + Y * b \quad (13)$$

iv) The exact procedure of allocating the slack time to additional northbound and / or southbound class A vessels is as follows:

a. If Z is greater than or equal to 1, it is deduced that the additional class A vessel (planned to pass in the slack time) should be a d -directional vessel and then the equations (11) and (12) are updated. Number of d -directional planned A vessels in slack time ($N(d)_p^{SLACK}$) is incremented by one.

$$N(d)_p^{SLACK} = N(d)_p^{SLACK} + 1 \quad (14)$$

and the slack time length is updated as:

$$ST = ST - 6 * \delta(d) \quad (15)$$

b. If Z is less than 1, it is deduced that the additional class A vessel (planned to pass in the slack time) should be a d' -directional vessel and then the equations (13) and (14) are updated. Number of d' -directional planned A vessels in slack time ($N(d')_p^{SLACK}$) is incremented by one and the slack time length is updated same as equation (15).

$$N(d')_p^{SLACK} = N(d')_p^{SLACK} + 1 \quad (16)$$

(viii) Returning to step (iii), the algorithm proceeds until the end of ST.

By means of this reschedule procedure, more vessels from both directions are scheduled and admitted to transit until the end of the slack time.

At the end of the (extended) starting direction time window (i.e. with the entrance to the Strait of the last scheduled class A vessel from that direction), the traffic is closed from both directions until the last vessel leaves the Strait. Since it takes approximately 30 minutes for a class A at Filburnu (in northbound traffic flow case) or at Boğaziçi Bridge (in southbound traffic case) to completely exit the Strait, the time gap between the last northbound or southbound Class A vessel and the following vessel from the opposite direction should be $6 * \theta + 30$ or $6 * \mu + 30$ minutes, respectively.

At the end of the starting direction time window (i.e. with the entrance to the Strait of the last scheduled class A vessel from that direction), the traffic is closed from both directions until the last vessel leaves the Strait. The start and execution of the vessel traffic flow in the opposite direction traffic is the same as the first direction flow. Vessels are allowed into the Strait until reaching the number of planned A vessels in this direction. If slack time admits any more A vessels in this direction, they also enter the Strait until the start of the nighttime vessel traffic. A typical example for daytime vessel schedule is displayed in Figure 3.

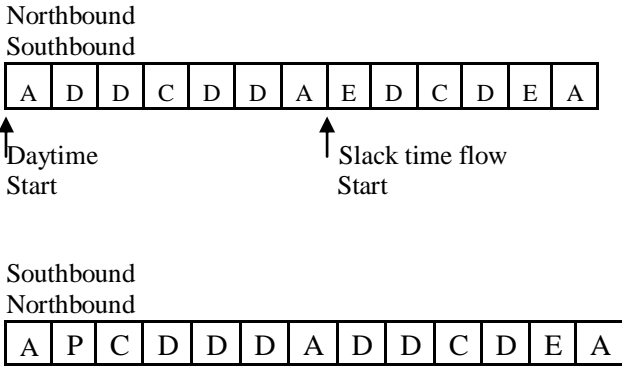


Figure 3: Daytime Schedule

2.6. Nighttime Vessel Scheduling

When daytime traffic ends, the active traffic flow direction remains as the first (active) direction of nighttime traffic. Additionally, unlike daytime uni-directional traffic, at nighttime, there exist two restricted vessel flows (according to the R&R, Class D vessels may enter from the opposite direction when there are such vessels available and meteorological conditions allow since no Class A vessels are allowed from either direction during nighttime).

Number of Class B vessels and the number of all Class C vessels (the ones which will be used for

deciding windows length after sequencing class B vessels) at nighttime plan ($t = t_n$) are updated in starting and opposite directions respectively as follows:

$$NQ_{t_n}^d(B_{up}) = NQ_{t_n}^d(B) + \text{MAX}(0, (NQ_{t_n}^d(C) - (NQ_{t_n}^d(B) - 1)) / 2) \quad (17)$$

$$NQ_{t_n}^{d'}(B_{up}) = NQ_{t_n}^{d'}(B) + \text{MAX}(0, (NQ_{t_n}^{d'}(C) - (NQ_{t_n}^{d'}(B) - 1)) / 2)$$

Then, the tentative time window length in the nighttime active direction is calculated as follows:

$$NW_p(d) = NT * \frac{NQ_{t_n}^d(B_{up})}{NQ_{t_n}^d(B_{up}) + NQ_{t_n}^{d'}(B_{up})} \quad (18)$$

The tentative time window length in the nighttime passive direction is calculated as follows:

$$NW_p(d') = NT * \frac{NQ_{t_n}^{d'}(B_{up})}{NQ_{t_n}^d(B_{up}) + NQ_{t_n}^{d'}(B_{up})} \quad (19)$$

where NT is the total nighttime duration, which is the time gap between the following day's daytime traffic plan start time and the end of the present day's daytime windows.

Accordingly, the number of Class B vessels planned to enter the Strait in the active direction flow is:

$$N_p^B(d) = \min(NQ_{t_n}^d(B), \frac{NW_p(d)}{4 * \delta(d)}) \quad (20)$$

The number of Class B vessels planned to enter the Strait in the passive direction flow is:

$$N_p^B(d') = \min(NQ_{t_n}^{d'}(B), \frac{NW_p(d')}{4 * \delta(d')}) \quad (21)$$

Presuming Class B the most critical group in the nighttime schedule, the length of the northbound and the southbound time windows are outlined by Class B vessels (similar to the role of Class A vessels in daytime scheduling). However, the relatively high population of the abundance of Class C vessels (around 9000 Class C vessels in a year) necessitates the consideration of this class while designing the nighttime traffic plan. Considering that minimum pursuit distance between two Class B vessels is $4 * \theta$ ($4 * \mu$) whereas minimum pursuit distance between a Class B vessel and a Class C vessel is $2 * \theta$ ($2 * \mu$), the duration of nighttime restricted traffic flow time is determined by the number of planned Class B vessels (multiplied by $2 * \theta$ or $2 * \mu$, according to the active direction), and the number of remaining class C vessels (multiplied by θ or μ , according to the active direction).

The total number of Class C vessels planned to enter the Strait after sequencing class B vessels in the active direction flow is:

$$N_p^{(C)}(d) = \max\left(0, \frac{NQ_n^d(C) - (N_p^B(d) - 1)}{2}\right) \quad (22)$$

The total number of Class C vessels planned to enter the Strait after sequencing class B vessels in the passive direction flow is:

$$N_p^{(C)}(d') = \max\left(0, \frac{NQ_n^{d'}(C) - (N_p^B(d') - 1)}{2}\right) \quad (23)$$

Both equations (22) and (23) are rounded down to nearest integer numbers.

The resulting total nighttime vessel traffic duration in the active direction is:

$$NW(d) = \min(NW_p(d), N_p^B(d) * 4 * \delta(d) + N_p^C(d) * 4 * \delta(d)) \quad (24)$$

The resulting total nighttime southbound vessel traffic duration in the passive direction is:

$$NW(d') = \min(NW_p(d'), N_p^B(d') * 4 * \delta(d') + N_p^C(d') * 4 * \delta(d')) \quad (25)$$

Once the scheduled transit of Class B and C vessels is completed, if there is remaining nighttime, Class D and E vessels continue entering the Strait from both directions (with Class E still having higher priority) according to the minimum pursuit distances (θ or μ) rules.

2.7. The Traffic Lanes and Overtaking

In the model, vessels follow two main lanes, (the northbound or the southbound lanes) and the overtaking lane, if permitted, while transiting the Strait. The whole Strait is divided into 22 slices with stations. Slices are at eight cables (0.8 nautical miles \approx 1.482 km.) intervals and in order to sustain a predetermined pursuit distance between vessels each slice is also composed of 2 cables long substations. Since stopping in the Strait for any reason is not allowed, vessels continuously move from one station to another during their stay in the Strait. Overtaking is allowed in the Strait except at the narrowest part, according to these conditions:

- When a vessel is in the overtaking lane, there should be no other vessel in this lane in the opposite direction at least up to the next station.
- There should be at least the pursuit distance between two closest vessel in the overtaking lane traveling the same direction.
- After overtaking is completed, vessels move back to the main lanes.

2.8. Pilot and Tugboat Services

According to the R&R, having a pilot captain on board during the Strait passage is compulsory for vessels longer than 250 meters and optional (though strongly recommended) for other vessels. All vessels express their pilot captain and tugboat needs in their SP-1 and SP-2 reports. There are 20 pilots and 6 tugboats available(as in the real situation).

In the simulation model pilots and tugboats are treated as resources which are seized by vessels at the embarking area in the Strait and released while leaving. In order to meet pilot and tugboat needs, every hour the model searches the number of available pilots (including transferring pilots) in the active direction and requests pilots from the opposite side when it is less than 6. The model also searches the number of available tugboats in the active direction and requests tugboats from the opposite side when it is less than 3. During the nighttime time windows, number of pilots at both sides is equalized to 6 and tugboats to 3 to meet the pilot and tugboat demand. Once a piloted vessel's passage in a certain direction is completed, the pilot is released from its current duty and included in the set of available resources for the opposite direction.

2.9. Visibility Conditions

According to the R&R, when visibility is less than one nautical mile in the Strait (called FogType1, only one-way traffic is permitted and when visibility in the Strait is less than 0.5 mile (called as FogType2), vessel traffic is suspended in both directions. The visibility module in the simulation model reads the fog information from the visibility data of (Almaz 2006) externally. Before a vessel is allowed to enter the Strait from the active direction during daytime, visibility condition is checked; if there is a FogType2 event, the vessel waits until it disappears. FogType1 does not affect daytime flows very much (since almost all vessel activity with the exception of class P vessels is uni-directional anyway); only the class P vessels coming from the opposite (passive) direction are stopped. When a FogType1 occurs at nighttime, however, two-way traffic is suspended.

2.10. Current Conditions

The most dominant current type on the Strait is the southbound surface current caused by level difference between the Black Sea and the Mediterranean Sea. The current module of the simulation model is integrated into the model from the previous study (Almaz 2006). In the study, the most effective southbound current is taken into account and a moving average function is built to estimate a daily base current value. Then, the current level at different regions of the Strait are assigned as predetermined percentages of the base value, based on historical current data. In order to comply with the R&R, when current speed exceeds 4 knots, class A, B, C and E vessels having a speed less than 10 knots are not allowed into the Strait. Moreover, all vessels in these classes have to wait in their queues

(until current conditions stabilize) when current speed exceeds 6 knots.

3. OUTPUTS OF THE MODEL

This model is run for the 13 months time period (between 1 December 2008 and 1 January 2010). The first month is considered as the warm up period. Some performance measures determined for the analysis are:

- R1: The average waiting time of vessels (aggregate and vessel type based);
- R2: Total number of vessels passed;
- R3: Average number of vessels in queues;
- R4: The entire Strait vessel density;
- R5: Pilot utilization;

4. VERIFICATION AND VALIDATION

Due to the fact that the simulation model in this study consists of many submodels integrated into the main traffic model running concurrently, it is difficult to monitor the system. However; with the trace module of Arena, arrival of each vessel, attributes assigned to it, its movement to the anchorage area or to the appropriate queues and its admittance to the Strait are followed clearly, while simultaneously watching entities related to meteorological events affecting the system. Moreover, animation reveals all events in the whole system; therefore, logic errors can be captured easily. Variable indicator of the Arena is also a frequently utilized tool in this study. The change in values of performance measures can directly be traced by variable indicators.

Extreme condition verification is first performed by increasing vessels arrival rates by 20% in a three month simulation run. When compared to the base scenario, average vessel waiting time shows more than fifteen-fold increase (from 541 minutes to 9272 minutes), average number of vessels in queues increase from 52.6 to 1154.4, number of vessels passed increases to 14756 from 12845 and pilot utilization increases from 0.23 to 0.25.

Another extreme conditions effect is reducing the total number of pilots in the model to 12 instead of 20. The model is run for one year with 25 replications and as expected, the pilot utilization, average, maximum waiting time of vessels and number of vessels in queues increased and total number of vessels passed the Strait decreased.

The most conclusive of the validation tests in this study are the output comparisons with the real 2009 data. The results of selected performance measures are sufficiently close to the data 2009 to support the claim that the model mimics the actual system reasonably well. As an example, average waiting times of all vessels in model and in actual data are compared. The results are quite similar to each other, as displayed in Table 2.

Table 2: Comparison of average waiting time of vessels

Waiting Times (in minutes)				
	2009 Data	The Simulation Model		Relative Error (%)
		Average	Half Width	
All Vessels	842	814.4	123.13	-3.28

5. SCENARIO ANALYSIS AND RESULTS

Four factors are selected for the scenario analysis of the simulation model:

- A: minimum pursuit distance (in time units) between vessels
- B: vessel profile
- C: pilot policy
- D: arrival rate

The levels of identified factors for scenario analysis are displayed in Table 3.

Table 3: Main factors and their levels in scenario design

Factor	Name	Low	Average	High
A	pursuit distance	13N-11.5S	13.5N-12S	14N-12.5S
B	vessel profile	base	base	>=150 m
C	Pilot availability	16	20	24
D	arrival rate	base	5% more	10% more

The first factor A with three levels is the minimum pursuit interval between two consecutive vessels (13N for the low setting means 13 minutes interval for northbound vessels and 11.5S means 11.5 minutes interval for southbound vessels). Regarding the vessel profile factor B, the low setting corresponds to the base scenario in which vessels demand pilots according to the pilot request frequency distribution of vessel classes generated based on the 2009 data. In the high setting, in addition to this random pilot demand, all vessels longer than 150 meters are routinely assigned a pilot while passing the Strait. In pilot availability factor C, the number of available pilots is set at 16 for the low level and 20 for the average level (as is the case in the current system) and 24 for the highest level. According to the last factor, regarding the arrival rate of vessels D, the low setting (which is the setting assumed in the base scenario) is taken as the rates estimated in the interarrival distribution for each subclass based on the 2009 data. In the average level, arrival rate of vessels is increased by 5 per cent (compared to the rates estimated based on the 2009 data) and in the high level, vessel arrival rates are increased by 10 per cent.

Accordingly, a total of 54 different scenarios (including the base scenario), are projected and run with 25 replications for a full factorial design. The outputs of these scenarios are gathered from Arena reports, the significant factors and their interactions are investigated through the ANOVA tables in the Design Expert 8.0 software. The percent contribution of each factor on performance measures are displayed in Table 4.

Table 4: Percent contributions of main factors

	A	B	C	D	AD
Average waiting time	38			59	2.4
Total vessels passed	0.3	0.1	0.1	98	0.1
Average transit time				89	
Pilot utilization	0.3	3.1	93	3.5	
Vessel density	0.3	0.1	0.1	99	0.1

In order to track the effects of factors easily, single factor level change in scenarios is investigated through the comparison of scenarios 19, 3, 7 and 16 with the base scenario 1 as can be seen in Table 5.

Table 5. Scenarios with various factor level changes

Scenarios	R1	R2	R3	R4	R5
1	814	51,178	79.9	9.45	0.24
19	608	51,206	59.2	9.46	0.24
3	722	51,204	70.3	9.46	0.31
7	754	51,200	73.4	9.45	0.25
4	2289	56,628	251	10.5	0.26
10	2275	56,624	250	10.5	0.27
12	2170	56,677	237	10.5	0.22
25	663	51,193	64.6	9.45	0.25
49	614	53,880	62.8	9.95	0.22
52	622	53,882	63.7	9.95	0.25

Decreasing pursuit distance to 13.5 minutes for south entrances and to 12 minutes for north entrances (scenario 19), primarily decrease the waiting time (by 25 per cent), decrease the number of vessels in queues by 26.25 per cent, while keeping the total number of vessels passed and vessel density almost the same. Decreasing the number of available pilots from 20 to 16 (scenario 3) increases pilot utilization by 29.2 per cent and decreases waiting time by 11.30 per cent (the reason why the average waiting time decreases is due to decrease in waiting time of Class D vessels, which enter the Strait more frequently while other vessel types remain waiting because of pilot unavailability). Assigning pilots for all vessels longer than 150 meters (scenario 7) increases pilot utilization by 4.1 per cent. Increasing vessel arrival rate by ten per cent (scenario 4) increases total number of vessels passed by 10.64 per cent, average waiting time by 181 per cent, number of vessels in queues by 212 per cent, pilot utilization by 29.2 per cent and vessel density by 10.8 per cent. The effect of two, three and four factor level changes over responses is also investigated. For instance, decreasing pursuit distance to 13.5 minutes for northbound and to 12 minutes for southbound vessels while assigning pilot for all vessels longer than 150 meters (scenario 25) decrease the waiting time by 18.6 per cent when compared to the base scenario; however, waiting time is increased by 9.9 per cent when compared to the single factor level change case involving 13.5 minutes pursuit distance for south entrances and 12 minutes for north entrances (scenario 19). Table 5 also displays that increasing vessel arrival rate by ten per cent and 20 available pilots to 24 in the system while assigning pilot

for all vessels longer than 150 meters (comparison of scenarios 10 and 12) decrease the waiting time by 4.62 per cent and number of vessels in queues by 5.22 per cent. Furthermore, increasing 20 available pilots to 24 in the system while assigning pilot for all vessels longer than 150 meters under five per cent higher arrival rate (comparison of scenarios 49 and 52) have almost same performance measure results.

In another scenario analysis, 4 factors influencing the response variables under high arrival rate conditions (number of arrived vessels increased by 10 per cent) are analyzed. The levels of factor A are 13.5 minutes for northbound and 12 minutes for southbound in low setting and 14 minutes for northbound and 12.5 minutes for southbound in high setting. Regarding the visibility factor (D), the low setting describes the base scenario in which vessels encounter fog events according to the visibility submodel, whereas in the high setting, the fog pattern of the worst case (i.e. the autumn fog realizations which have the longest fog durations) is chosen as the visibility data for the whole year.

In order to track the effects of factors easily as displayed in Table 6, level change in scenarios is investigated compared to the base scenario 1. Decreasing pursuit distance to 13.5 minutes for north entrances and to 12 minutes for south entrances (scenario 7) primarily decrease the waiting time by 41.5 per cent, decrease the number of vessels in queues by 41.6 per cent, while keeping the total number of vessels passed almost the same. Setting low visibility conditions (scenario 13) increases average waiting time by 88.8 per cent yet does not significantly change the total number of vessels passed. The effect of two and three factor level changes over responses may also be investigated in this table. For example, although reducing pursuit distances to 13.5 minutes for northbound passages and 12 minutes for southbound passages and deploying 24 pilots instead of 20, the average waiting time increases by 82 per cent under low visibility conditions (comparison of scenarios 7 and 20) and number of vessels in queues increase by 91 per cent.

Table 6. Scenarios with various factor level changes under high arrival rate conditions

Scenarios	R1	R2	R3	R4	R5
1	2289	56629	250	10.5	0.3
7	1367	56850	146	10.5	0.3
13	4320	56324	474	10.4	0.3
15	4087	56336	448	10.4	0.4
20	2482	56531	279	10.5	0.2
23	2354	56615	262	10.5	0.2

In the full factorial analysis of the related scenarios, the 24 different scenarios are experimented through 25 replications (i.e. the scenario analysis is composed of 600 distinct observations). In the scenario analysis, the

most effective factor on performance measures is observed as visibility conditions. As fog in the Strait becomes stronger, average waiting time of vessels and transit time increase. Moreover, low visibility conditions decrease total number of vessels passed from directions, pilot utilization and vessel density in the Strait.

6. CONCLUSION AND FURTHER RESEARCH

In this study, a simulation model is developed for representing the vessel traffic behavior in the Strait. In this simulation model, maritime rules and regulations about vessel admittance, pursuit distances among vessels, priority levels of distinct vessel types and pilot requirements are all considered. Moreover, submodels representing meteorological conditions such as fog, current and storm are integrated to the model. For validation purposes, the simulation outputs are compared with the actual 2009 data and quite satisfactory results are obtained.

In order to analyze the effects of various factors such as vessel arrival rate, vessel profile, pilot availability and minimum pursuit distances between vessels, on performance measures, 54 scenarios are performed with the full factorial design. The most significant factor for all selected variables is observed as the vessel arrival rate. The minimum pursuit distance between vessels is also significant for most performance measures. The interaction of arrival rate and pursuit distance is effective on the most responses, as well. Pilot availability is principally important for pilot utilization.

Another scenario analysis is conducted when vessel arrival rate is increased by 10 per cent and the visibility factor is added. Results associated with the considered 24 scenarios show that visibility is the most critical factor for performance measures and its interaction with minimum pursuit distance at different levels is also significant for performance measures such as average waiting time of vessels, number of vessels passed and pilot utilization.

This study is planned to be used for risk analysis of the Strait. Incorporating probable vessel accidents and the consequences to the model can have a very beneficial effect for revising the policies and minimizing risk.

REFERENCES

- Almaz, A. Ö., 2006, *Investigation of the Maritime Transit Traffic in the Istanbul Channel through Simulation Modeling and Scenario Analysis*, M.S. Thesis, Department of Industrial Engineering, Boğaziçi University, Istanbul.
- Gönültaş, E., *Analysis of the Extreme Weather Conditions, Vessel Arrival Processes and Prioritization in the Strait of Istanbul through Simulation Modeling*, M.S. Thesis, Department of Industrial Engineering, Boğaziçi University, Istanbul, 2007.
- Law, A. M. and W. D. Kelton. 2007 *Simulation Modeling and Analysis*, McGraw-Hill Press, Singapore.
- Özbaş, B. 2005, *Simulation of Maritime Transit Traffic in the Istanbul Channel*, M.S. Thesis, Department of Industrial Engineering, Boğaziçi University, Istanbul.
- Ulusçu, S. Ö., B. Özbaş, T. Altıok and İ. Or. 2009. "Risk Analysis of the Vessel Traffic in the Strait of Istanbul", *Risk Analysis*, Vol. 20, No. 10, pp. 1454-1472.
- VTS Users Guide, *Turkish Straits Vessel Traffic Service*. 2004. General Management of Coastal Safety and Salvage Administrations, 3rd edition, Istanbul.

AUTHOR BIOGRAPHIES

İLHAN OR was born in Istanbul, 1951. He received his BS (1973), MS (1974) and Ph.D. degrees (1976) from Northwestern University, Evanston, Illinois, USA. He has been a faculty member at Department of Industrial Engineering of Bogazici University, Istanbul, Turkey since 1976. He was a visiting faculty member at Syracuse University (1982-1983) and University of Maryland (1983). He served on the "Naval Research Logistics Quarterly" Journal's Editorial Board between 1993 and 2003. Research areas are environmental and risk management, energy policy and planning, production and maintenance planning. His e-mail address is: or@boun.edu.tr and personal web page is: <http://www.ie.boun.edu.tr/~or/>

ŞİRİN ÖZLEM was born in Bursa, 1985. She received her BS (2008) from Uludag University, Bursa, Turkey and MS (2011) from Bogazici University, Istanbul, Turkey. She is currently a doctoral student at the Department of Industrial Engineering of Boğaziçi University, Turkey. Her e-mail address is: sirin.ozlem@boun.edu.tr

BİRNUR ÖZBAŞ was born in Istanbul, 1980. She received her BS (2003) in Systems Engineering from Yeditepe University, Istanbul, Turkey MS (2005) and PhD (2010) in Industrial Engineering from Bogazici University, Istanbul, Turkey. She is currently a post doctoral associate at Center for Advance Infrastructure and Transportation (CAIT) / Laboratory for Port Security (LPS), Rutgers, The State University of New Jersey, U.S.A. Her e-mail address is: birnur@ozbas.com.tr

NEW GENETIC PROGRAMMING HYPOTHESIS SEARCH STRATEGIES FOR IMPROVING THE INTERPRETABILITY IN MEDICAL DATA MINING APPLICATIONS

Michael Affenzeller, Christian Fischer, Gabriel Kronberger, Stephan M. Winkler, Stefan Wagner

Upper Austria University of Applied Sciences
School for Informatics, Communications, and Media
Heuristic and Evolutionary Algorithms Laboratory
Softwarepark 11, 4232 Hagenberg, Austria

michael.affenzeller@fh-hagenberg.at, gabriel.kronberger@heuristiclab.com, stephan.winkler@fh-hagenberg.at,
christian.fischer@students.fh-hagenberg.at, stefan.wagner@heuristiclab.com

ABSTRACT

In this paper we describe a new variant of offspring selection applied to medical diagnosis modeling which is designed to guide the hypothesis search of genetic programming towards more compact and more easy to interpret prediction models. This new modeling approach aims to combat the bloat phenomenon of genetic programming and is evaluated on the basis of medical benchmark datasets. The classification accuracies of the achieved results are compared to those of published results known from the literature. Regarding compactness the models are compared to genetic programming prediction models achieved without the new offspring selection variant.

Keywords: Medical data mining, Genetic programming, Offspring selection.

1. INTRODUCTION

Genetic Programming (GP) plays an outstanding role among the various data-mining techniques from the field of machine learning and computational intelligence. Due to its model representation, GP is able to produce human interpretable models without taking any assumptions about the nature of the relationship. Also GP-based data analysis has quite good generalization properties. Furthermore, GP is able to simultaneously evolve the structure and the parameters of a model with implicit feature selection. The combination of these aspects makes GP a very powerful and also robust method for various data analysis tasks.

Nevertheless, there are still some aspects in the practical application of GP-based data analysis which leave room for improvement:

GP-based data analysis suffers from the fact that – even if the models are interpretable – the results are often quite complex and far from being unique. Often the models are still quite complex because of the tendency of GP to bloat and also because of introns which is counterproductive in terms of interpretability as well.

One of the reasons for genetic bloat is identified in the tendency of GP to favor more complex hypothesis structures for explaining equivalent correlations (Luke and Panait, 2006). The new proposed offspring selection variant aims to counteract this phenomenon by including additional offspring selection criteria: Instead of only considering the error measure, the enhanced offspring selection (OS) criteria also consider the complexity as well as the number of variables of the candidate hypothesis in order to decide, whether or not a new candidate hypothesis is accepted for the next generation. By this means the hypothesis search should lead not only to models with more predictive power, but also to more compact and more unique models which are easier to interpret. Especially the latter aspects are considered as important in the field of medical data mining where the domain expert should be able to analyze not only the statistical properties of the prediction models but also their usefulness in the medical context.

The effects of the new introduced extended offspring selection formulation for data based modeling are discussed for medical benchmark datasets from the UCI machine learning repository¹.

The rest of the paper is organized as follows: Section 2 describes standard offspring selection with its parameters, its main characteristics, and how it can be integrated into genetic programming. Section 3 discusses specific extensions of offspring selections designed for data based modeling which aim to guide hypothesis search to simpler and easier to interpret models. In section 4 the characteristics of the extended offspring selection variant are discussed exemplarily for medical benchmark data sets. Finally, section 5 summarizes the achieved results and points out future perspectives for future research.

¹ <http://archive.ics.uci.edu/ml/>

2. OFFSPRING SELECTION

The basic principles of offspring selection have been described in (Affenzeller and Wagner 2005). In the meanwhile, offspring selection has been discussed for several benchmark problems from the field of combinatorial optimization, function optimization and data based modeling. The following description of standard offspring selection is taken from (Wagner et al 2010) where the aspect of mutation in offspring selection has been discussed in further detail. In general, offspring selection consists of the following steps:

At first parents are selected for reproduction either randomly or in any other well-known way of genetic algorithms (e.g., fitness proportional selection, linear rank selection, tournament selection). After crossover and optionally mutation have been applied to create a new child solution, another selection step is introduced which considers the success of the applied reproduction procedure. The goal of this second selection step is to continue the search process with offspring which surpass their parents' quality. Therefore, a new parameter called success ratio (*SuccRatio*) is introduced. The success ratio defines the relative amount of members in the next population that have to be generated by successful mating (crossover, mutation).

Additionally, it has to be defined when a solution is considered to be successful: Is a child solution better than its parents, if it surpasses the fitness of the weaker, the better, or some kind of mean value of both? For this purpose a parameter called comparison factor (*cf*) is used to define the success criterion for each created solution as a weighted average of the quality of the worse and the better parent (i.e., if the comparison factor is 0, successful solutions at least have to be better than the worse parent, and if it is 1 they have to outperform the better parent).

For steering the comparison factor, the authors decided to introduce a cooling strategy which is similar to simulated annealing. Following the basic principle of simulated annealing, an offspring only has to surpass the fitness value of the worse parent in order to be successful at the beginning of the search process (*cf* is initialized with 0 or a rather small value). While evolution proceeds solutions have to be better than a fitness value continuously increasing between the fitness of the weaker and the better parent (*cf* is increased in each generation until it reaches 1 or a rather high value). As in the case of simulated annealing, this strategy leads to a broader search at the beginning, whereas at the end the search process becomes more and more directed.

After the amount of successful solutions in the next generation has reached the success ratio, the remaining solutions for the next generation (i.e., $(1-SuccRatio) \cdot |POP|$) are taken from the pool of solutions which were also created by crossover and mutation but did not necessarily reach the success criterion. The actual selection pressure *ActSelPress* at

the end of a single generation is defined by the quotient of individuals that had to be created until the success ratio was reached and the number of individuals in the population:

$$ActSelPress = \frac{|POP| \cdot SuccRatio + |POOL|}{|POP|} \quad (1)$$

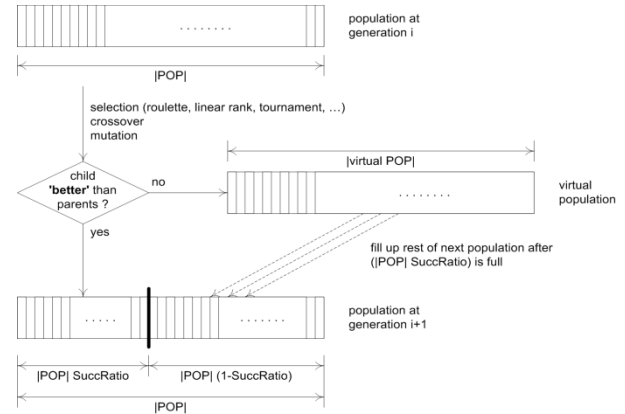


Figure 1: Flowchart of Offspring Selection

Figure 1 shows these basic steps of offspring selection and how they are embedded into a classical genetic algorithm.

Furthermore, an upper limit for the selection pressure (*MaxSelPress*) can be defined as another parameter which states the maximum number of children (as a multiple of the population size) that might be created in order to fulfill the success ratio. With this additional parameter offspring selection also provides a precise detector for premature convergence: If the algorithm cannot create a sufficient number of successful solutions ($SuccRatio \cdot |POP|$) even after $MaxSelPress \cdot |POP|$ solutions have been created, convergence has occurred and the algorithm can be stopped.

If OS is applied with the parameters $cf = 1$ and $SuccRatio = 1$, it is commonly referred to as strict OS. Strict OS has the property that children with worse quality compared to its better parent are automatically discarded and therefore the overall quality of the population steadily increases.

3. NEW OFFSPRING SELECTION FOR DATA ANALYSIS

The standard variant of offspring selection as discussed in Section 2 implements the offspring selection criterion purely on the basis of solution quality. For data based modeling the offspring selection criterion is usually based on the mean squared error (MSE) for classification problems and on the coefficient of correlation R^2 or MSE for regression problems. This means that an offspring solution candidate is considered successful if the MSE or R^2 fitness measure of the candidate offspring is better than the respective fitness measure of the parent solutions. This means that only the quality of the models is considered and not the

simplicity of interpretability of the involved solution candidates.

The main idea of the here discussed offspring selection extension is that not only the quality of the candidate models should be considered but also its compactness in order to combat the bloat. From theoretical bloat analyses (Luke and Panait, 2006) it is known that genetic programming based hypothesis search tends to find rather more complex models in order to achieve the same model quality. Therefore, it seems reasonable to include also model complexity measures into the offspring selection criterion. In that sense, an offspring candidate model is considered successful if it surpasses not only the model quality of its own parents but is also not more complex than its parent models. As model complexity measures we have introduced the number of nodes as well as the number of used input features (variables) of the involved structure trees. In that sense an offspring solution is considered successful not only if it surpasses the model quality of its own parents; additionally, the offspring model must not be more complex than its parent model. Similar to the standard case of offspring selection we have to decide if the criterion compares the resulting offspring to the better, the weaker, or to some intermediate value. In order to handle this aspect we introduce new model complexity comparison factors: Let q_b and q_w be the model qualities of the better and the weaker model, l_b and l_w the length of the shorter (better) and the more complex (worse) model. As a model complexity measure we here use the number of nodes of the two parent structure trees. For the number of variables of the two parent models let v_b be the model using less variables (better) and v_w the model using more variables (worse). Similar to the standard case of offspring selection comparison factors cf_q , cf_l , and $cf_v \in [0, 1]$ define the certain thresholds which distinguish a successful offspring from an unsuccessful offspring based on the characteristic features of the parents. But in contrast to original offspring selection a candidate offspring has to fulfill three criteria instead of one in order to be accepted; it does not only have to be better but also less complex and use less variables. Similar to the standard case a comparison factor of 0 means that it is sufficient to surpass the certain characteristics of the worse parent whereas a comparison factor of 1 means that the candidate offspring has to be better than the better of the two parents.

Obviously it becomes harder to evolve successful offspring solution candidates which results in higher selection pressures on the one hand; on the other hand due to the preference to simpler and more compact models genetic diversity can hardly emerge. Therefore, the additional offspring selection criteria concerning the model complexities and the number of variables should better not be activated from the start. First studies have shown that the new OS variant works a lot better if the additional criteria are activated not until genetic diversity can emerge which usually happens after about

one or two dozen of iterations. Algorithmically we have considered this aspect by introducing further parameters which specify two time windows tw_l and tw_v which specify when the additional length and number of variables criterion should be active.

Summarizing the above mentioned aspects, the here discussed first version of a new offspring selection criterion dedicated to the reduction of bloat can be stated as follows (in the minimization variant for MSE as quality):

$$\begin{aligned} isSuccessful(co, p1, p2, gen) \Leftrightarrow \\ [q(co) < q_w + cf_q(q_w - q_b)] \text{ and} \\ [(l(co) < l_w + cf_l(l_w - l_b)) \text{ or } (gen \notin tw_l)] \text{ and} \\ [(v(co) < v_w + cf_v(v_w - v_b)) \text{ or } (gen \notin tw_v)] \end{aligned}$$

This means that in order to be considered as successful, a candidate offspring (co) has to be better than some intermediate fitness value of its own parents (defined by cf_q) in any case. Additionally, in some predefined time window tw_l an offspring does not only have to be better but also at least as compact than some intermediate compactness value of its own parents and in the same sense there is a time window tw_l where the candidate offspring have to use not more variables than some intermediate value of variables used by its parent models. Therefore, also the actual generation gen has to be considered in order to decide if one of the two time windows is active at the moment.

The empirical discussion of the next section compares the achieved results on the basis of standard classification benchmark datasets for generating prediction models for breast cancer, thyroid, and melanoma.

4. RESULTS

The configurations used for the test runs in table 2, 4 and 6 with Melanoma, Thyroid and Wisconsin datasets are shown in table 1. If not otherwise stated the time windows include all generations. The maximum solution length is 100, maximum solution height is 12. Up to 1000 Generations are created with a maximum permitted selection pressure of 555.

#	Configuration
1	cfq=0,cfl=0
2	cfq=1,cfl=0
3	cfq=0..1,twq=1..100,cfl=0
4	cfq=0,cfl=-100..1,twl=20..100
5	cfq=1,cfl=-100..1,twl=20..100
6	cfq=0..1,cfl=-100..1,twl=20..100
7	cfq=0,cfl=1,twl=20..100
8	cfq=1,cfl=1,twl=20..100
9	cfq=0..1,cfl=1,twl=20..100
10	cfq=0,cfh=0,twh=20..100
11	cfq=1,cfh=0,twh=20..100
12	cfq=0..1,twq=1..100,cfh=0,twh=20..100
13	cfq=0,cfh=-100..1,twh=20..100
14	cfq=1,cfh=-100..1,twh=20..100
15	cfq=0..1,cfh=-100..1,twh=20..100

16	cfq=0,cfh=1,twh=20..100
17	cfq=1,cfh=1,twh=20..100
18	cfq=0..1,cfh=1,twh=20..100
19	cfq=0,cfv=0,twv=20..100
20	cfq=1,cfv=0,twv=20..100
21	cfq=0..1,twq=1..100,cfv=0,twv=20..100
22	cfq=0,cfv=-100..1,twv=20..100
23	cfq=1,cfv=-100..1,twv=20..100
24	cfq=0..1,cfv=-100..1,twv=20..100
25	cfq=0,cfv=1,twv=20..100
26	cfq=1,cfv=1,twv=20..100
27	cfq=0..1,cfv=1,twv=20..100
28	cfq=0,cfl=cfh=cfv=-100..1,twl=twh=twv=20..100
29	cfq=1,cfl=cfh=cfv=-100..1,twl=twh=twv=20..100
30	cfq=0..1,cfl=cfh=cfv=-100..1,twl=twh=twv=20..100

Table 1: Configurations for all datasets

For comparison purposes the same datasets were used in regular Offspring Selection Genetic Algorithms (OSGA) as seen in table 3, 5 and 7. A strict configuration with a comparison factor of 1 is used. Maximum allowed length is 50, 100 and 200; maximum allowed height is 7, 12 and 17 for configurations a, b, and c respectively.

4.1. Results Melanoma Dataset

The results of the performed test runs with the Melanoma dataset are shown in table 2. The regular OSGA results for Melanoma are shown in table 3.

#	Acc.(Tr.)	Acc.(Te.)	Height	Length	Nr.OfVar	MSE(Tr.)	MSE(Te.)	Exec.Time
1	0.924	0.925	9.8	31.6	5.6	0.091	0.074	03:04
2	0.934	0.934	8.2	37.8	6.6	0.069	0.166	03:24
3	0.923	0.907	9.4	36.6	5.8	0.070	0.075	02:30
4	0.935	0.929	9.8	49.8	7.2	0.062	0.068	39:03
5	0.916	0.913	9.0	32.2	5.4	0.078	0.083	02:48
6	0.931	0.926	11.8	68.0	8.8	0.068	0.077	08:37
7	0.927	0.924	8.6	30.4	5.8	0.080	0.075	01:22
8	0.929	0.929	11.0	40.8	7.2	0.063	0.069	03:53
9	0.927	0.914	9.6	32.8	6.6	0.075	0.096	01:12
10	0.924	0.920	7.8	18.2	4.0	0.073	0.077	01:04
11	0.920	0.921	10.6	56.8	8.6	0.072	0.249	02:46
12	0.917	0.921	9.4	32.0	5.4	0.081	0.081	01:55
13	0.937	0.927	12.0	66.4	8.0	0.060	0.066	09:25
14	0.931	0.911	11.4	42.0	7.2	0.065	0.071	09:15
15	0.925	0.914	10.4	50.8	6.4	0.065	0.072	05:33
16	0.921	0.914	11.2	46.2	7.4	0.075	0.077	01:41
17	0.919	0.921	11.2	36.8	5.8	0.078	0.153	11:14
18	0.918	0.915	8.4	34.6	6.2	0.123	0.087	01:10
19	0.927	0.921	11.8	49.8	6.8	0.071	0.073	02:19
20	0.926	0.907	10.6	41.0	6.6	0.071	0.078	02:47
21	0.930	0.918	11.4	45.4	7.2	0.075	0.075	03:12
22	0.944	0.927	12.8	85.0	10.8	0.054	0.074	15:39
23	0.921	0.915	9.4	29.6	4.8	0.072	0.071	03:00
24	0.925	0.918	11.8	67.8	10.0	0.091	0.070	12:02
25	0.920	0.917	11.0	45.8	8.8	0.077	0.074	01:28
26	0.924	0.902	8.6	33.6	6.0	0.111	0.095	03:32
27	0.911	0.911	9.8	41.2	6.8	0.086	0.082	02:12
28	0.930	0.917	10.8	43.2	5.8	0.064	0.070	10:41
29	0.923	0.915	10.2	31.2	6.0	0.104	0.125	03:07
30	0.932	0.922	11.6	65.8	8.8	0.077	0.064	07:07

Table 2: Results with Melanoma dataset

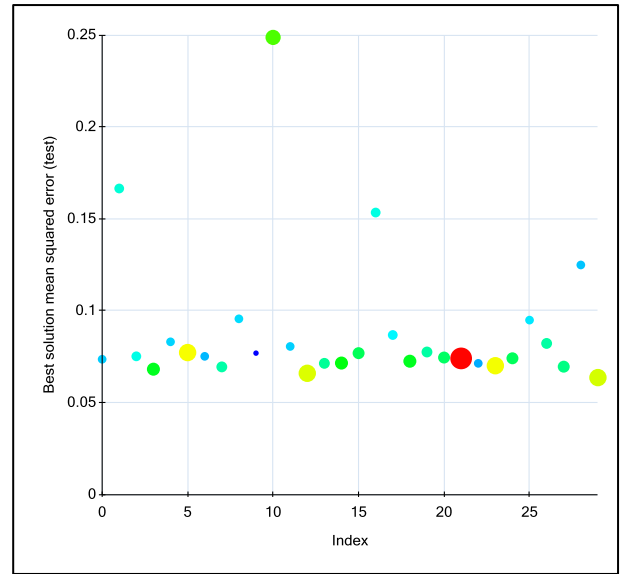


Figure 2: Melanoma Results: Configuration vs. Quality; little-high complexity (blue-red and bubble size)

#	Acc.(Tr.)	Acc.(Te.)	Height	Length	Nr.OfVar	MSE(Tr.)	MSE(Te.)	Exec.Time
a	0.919	0.917	6.8	15.6	3.2	0.086	0.100	02:49
b	0.931	0.919	9.0	35.6	7.0	0.075	0.072	03:46
c	0.927	0.915	13.2	71.0	11.6	0.071	0.078	04:30

Table 3: Regular OSGA results Melanoma

4.2. Results Thyroid Dataset

The results of the performed test runs with the Thyroid dataset are shown in table 4. The regular OSGA results for Thyroid are shown in table 5.

#	Acc.(Tr.)	Acc.(Te.)	Height	Length	Nr.OfVar	MSE(Tr.)	MSE(Te.)	Exec.Time
1	0.970	0.962	6.6	14.2	2.2	90.87	220.17	04:08
2	0.983	0.976	10.6	49.8	6.8	57.37	78.24	08:55
3	0.989	0.987	7.8	22.8	3.2	61.17	74.60	06:48
4	0.991	0.988	12.6	85.0	6.6	32.81	39.22	58:05
5	0.983	0.980	12.4	67.2	7.6	69.52	87.90	13:18
6	0.990	0.986	12.6	76.4	6.0	36.74	52.99	24:02
7	0.941	0.935	9.0	30.2	3.4	220.92	207.82	02:53
8	0.984	0.983	11.4	61.8	6.2	63.17	63.53	05:00
9	0.950	0.945	8.8	23.2	3.0	183.75	194.81	02:17
10	0.943	0.943	7.4	29.6	3.8	197.61	192.17	02:21
11	0.983	0.975	10.4	61.2	6.8	65.27	94.61	05:46
12	0.952	0.947	9.0	32.8	4.8	150.90	170.13	03:19
13	0.992	0.990	12.2	88.0	6.2	36.00	59.93	17:11
14	0.988	0.987	11.4	82.2	7.4	46.20	60.26	33:37
15	0.993	0.991	11.8	76.4	7.4	34.88	45.13	22:04
16	0.938	0.938	10.2	37.4	4.0	199.67	186.36	01:50
17	0.982	0.980	11.4	60.0	6.4	63.38	76.04	03:14
18	0.935	0.936	10.6	48.4	6.2	216.02	236.95	01:34
19	0.942	0.939	9.2	34.2	1.4	161.51	160.31	01:06
20	0.987	0.984	11.8	67.6	6.6	53.26	65.30	13:02
21	0.953	0.952	10.4	44.8	2.2	136.96	171.29	02:10
22	0.993	0.987	12.8	91.6	6.2	32.45	50.64	18:20
23	0.989	0.987	12.8	73.0	7.4	48.39	54.27	45:34
24	0.993	0.989	12.6	92.0	7.0	41.28	46.95	21:06
25	0.943	0.940	8.6	24.0	3.0	230.51	235.51	01:10
26	0.979	0.979	11.2	60.2	7.0	73.97	96.89	03:13

27	0.938	0.935	10.0	40.4	2.0	189.46	189.00	01:43
28	0.993	0.991	11.8	86.0	5.8	39.09	223.87	37:29
29	0.986	0.982	12.8	88.0	7.6	51.26	66.97	09:51
30	0.985	0.983	12.2	75.6	7.6	56.67	90.64	10:13

Table 4: Results with Thyroid dataset

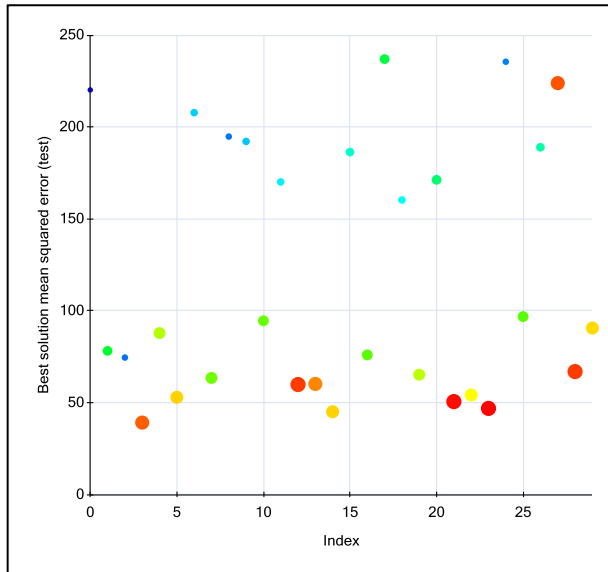


Figure 3: Thyroid Results: Configuration vs. Quality; little–high complexity (blue–red and bubble size)

#	Acc.(Tr.)	Acc.(Te.)	Height	Length	Nr.OfVar	MSE(Tr.)	MSE(Te.)	Exec.Time
a	0.982	0.977	7.8	36.2	4.8	50.67	70.50	08:33
b	0.977	0.972	12.0	68.0	7.6	311.36	98.74	09:36
c	0.990	0.986	17.4	115.4	9.0	53.14	67.22	09:45

Table 5: Regular OSGA results Thyroid

4.3. Results Wisconsin Dataset

The results of the performed test runs with the Wisconsin dataset are shown in table 6. The regular OSGA results for Wisconsin are shown in table 7.

#	Acc.(Tr.)	Acc.(Te.)	Height	Length	Nr.OfVar	MSE(Tr.)	MSE(Te.)	Exec.Time
1	0.956	0.955	9.2	26.4	4.2	0.188	0.204	02:25
2	0.959	0.947	12.4	59.0	6.0	0.168	0.236	03:03
3	0.960	0.955	11.4	39.8	4.4	0.190	0.174	03:28
4	0.960	0.943	12.6	71.4	4.4	0.149	0.211	14:09
5	0.967	0.959	12.0	63.0	6.8	0.151	0.185	13:39
6	0.961	0.934	12.2	67.6	5.6	0.162	0.197	11:21
7	0.944	0.925	8.4	23.6	3.4	0.225	0.238	01:10
8	0.968	0.959	11.2	50.0	6.4	0.159	0.166	04:17
9	0.950	0.952	9.4	23.2	3.2	0.239	0.209	02:20
10	0.947	0.946	7.6	20.4	3.2	0.205	0.232	01:33
11	0.967	0.955	11.8	62.0	5.8	0.153	0.180	03:17
12	0.942	0.934	9.0	31.4	3.8	0.208	0.241	02:12
13	0.963	0.953	12.0	71.6	6.6	0.155	0.189	14:08
14	0.966	0.949	12.0	60.0	6.2	0.146	0.172	04:41
15	0.960	0.937	11.8	58.4	5.6	0.158	0.214	08:53
16	0.947	0.937	8.8	34.8	4.4	0.220	0.239	01:26
17	0.959	0.946	12.4	59.8	6.6	0.183	0.198	02:49
18	0.946	0.925	7.4	15.6	2.8	0.256	0.285	00:59
19	0.944	0.937	10.8	36.4	3.2	0.206	0.202	01:20
20	0.964	0.959	12.0	55.8	6.8	0.161	0.191	02:41

21	0.948	0.938	10.4	30.4	3.2	0.241	0.239	01:28
22	0.965	0.950	12.4	55.4	5.6	0.158	0.217	09:11
23	0.966	0.947	12.2	69.0	6.4	0.144	0.200	03:18
24	0.962	0.950	12.6	79.0	6.0	0.149	0.155	11:06
25	0.940	0.930	9.2	30.8	2.8	0.229	0.255	00:57
26	0.963	0.958	12.2	65.0	6.4	0.175	0.193	03:27
27	0.935	0.921	9.6	30.2	2.8	0.266	0.240	00:53
28	0.966	0.955	11.4	52.8	5.4	0.141	0.213	16:04
29	0.967	0.960	11.8	51.0	6.6	0.148	0.170	02:59
30	0.965	0.960	10.4	53.4	5.0	0.155	0.160	11:01

Table 6: Results with Wisconsin dataset

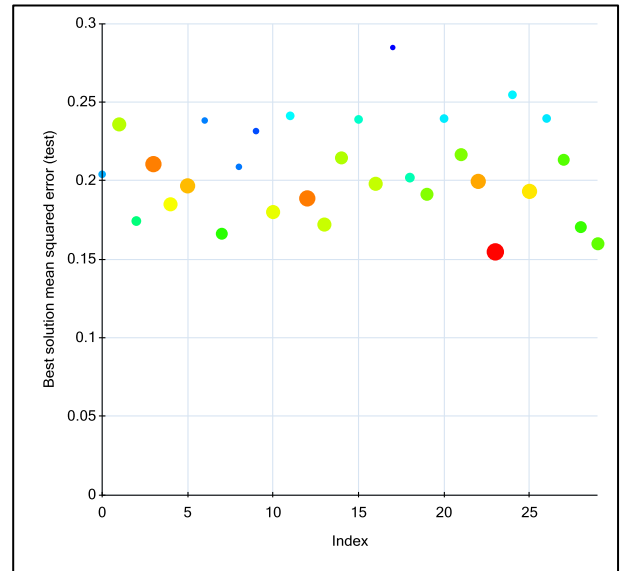


Figure 4: Wisconsin Results: Configuration vs. Quality; little–high complexity (blue–red and bubble size)

#	Acc.(Tr.)	Acc.(Te.)	Height	Length	Nr.OfVar	MSE(Tr.)	MSE(Te.)	Exec.Time
a	0.976	0.966	8.0	36.6	5.6	0.109	0.157	53:56
b	0.979	0.966	13.0	92.8	6.6	0.087	0.139	138:34
c	0.981	0.960	17.8	181.8	8.4	0.085	0.151	64:17

Table 7: Regular OSGA results Wisconsin

5. CONCLUSION

In this paper we have considered the aspects of model interpretability and uniqueness in genetic programming based medical data mining. Due to introns and the bloat phenomenon GP models tend to produce more complex than necessary (Luke and Panait, 2006). In contrast to the so called bloat free GP (Silva, 2011) which allows only those models which do not exceed a certain model complexity we have adapted the concept of offspring selection in a way that the hypothesis search process should be guided towards simple and good prediction models. For this purpose the offspring selection criterion has been extended in a way that it considers not only the model quality in order to decide whether or not a candidate hypothesis should be accepted; in addition also the complexity in terms of number of nodes and the interpretability in terms of number of used variables are considered for the offspring selection criterion. The effects of this approach have been analyzed for some well-known

benchmark problems from the field of medical data mining. The results show that the new offspring selection criterion is quite sensitive in terms of causing premature convergence due to the loss of genetic diversity caused by the complexity limiting aspects in the OS-criterion. Therefore, it remains as a topic for future research to further develop this new way of hypothesis search in a way that a sufficient amount of genetic diversity is maintained in the GP population. One possible way of achieving such kind of behavior might an automated switch on/off of the additional criteria depending on the average model complexity or the diversity in the actual population.

ACKNOWLEDGMENTS

The work described in this paper was done within the Josef Ressel Centre for Heuristic Optimization *Heureka!* (<http://heureka.heuristiclab.com/>) sponsored by the Austrian Research Promotion Agency (FFG).

REFERENCES

- Affenzeller, M. and Wagner, S., 2005. Offspring selection: A new self-adaptive selection scheme for genetic algorithms. *Adaptive and Natural Computing Algorithms*, Springer Computer Science, pp. 218-221.
- Affenzeller, M., Winkler, S., Wagner, S., A. Beham, 2009. *Genetic Algorithms and Genetic Programming - Modern Concepts and Practical Applications*. Chapman & Hall/CRC. ISBN 978-1584886297. 2009.
- Luke S. and Panait L., 2006. *A Comparison of Bloat Control Methods*. Journal of Evolutionary Computation, Vol. 14, No.3 pp. 48-48.
- Silva S., 2011 *Reassembling Operator Equalisation - A Secret Revealed*. Proceedings of GECCO 2011, pp. 1395-1403.
- Wagner, S., Affenzeller M., Beham A., Kronberger G., and Winkler S.M., 2010. Mutation Effects in Genetic Algorithms with Offspring Selection Applied to Combinatorial Optimization Problems. *Proceedings of EMSS 2010*, pp. 48-48.

AUTHORS BIOGRAPHIES



MICHAEL AFFENZELLER has published several papers, journal articles and books dealing with theoretical and practical aspects of evolutionary computation, genetic algorithms, and meta-heuristics in general. In 2001 he received his PhD in engineering sciences and in 2004 he received his habilitation in applied systems engineering, both from the Johannes Kepler University of Linz, Austria. Michael Affenzeller is professor at the Upper Austria University of Applied Sciences, Campus Hagenberg, and head of the Josef Ressel Center *Heureka!* at Hagenberg.



CHRISTIAN FISCHER received his BSc in software engineering in 2009 from the Upper Austria University of Applied Sciences, Campus Hagenberg. He is currently pursuing studies for his master's degree. In the course of his studies he is involved in the project team for the prediction of blood demands in a hospital in cooperation with the Josef Ressel Centre *Heureka!* and the General Hospital Linz.



GABRIEL KRONBERGER authored and co-authored numerous papers in the area of evolutionary algorithms, genetic programming, machine learning and data mining. Currently he is a research associate at the Research Center Hagenberg of the Upper Austria University of Applied Sciences working on data-based modeling algorithms for complex systems within the Josef-Ressel Centre for Heuristic Optimization *Heureka!*.



STEPHAN M. WINKLER received his MSc in computer science in 2004 and his PhD in engineering sciences in 2008, both from Johannes Kepler University (JKU) Linz, Austria. His research interests include genetic programming, nonlinear model identification and machine learning. Since 2009, Dr. Winkler is professor at the Department for Medical and Bioinformatics at the Upper Austria University of Applied Sciences, Campus Hagenberg.



STEFAN WAGNER received his MSc in computer science in 2004 and his PhD in engineering sciences in 2009, both from Johannes Kepler University (JKU) Linz, Austria; he is professor at the Upper Austrian University of Applied Sciences (Campus Hagenberg). Dr. Wagner's research interests include evolutionary computation and heuristic optimization, theory and application of genetic algorithms, machine learning and software development.

ON THE USE OF ESTIMATED TUMOR MARKER CLASSIFICATIONS IN TUMOR DIAGNOSIS PREDICTION - A CASE STUDY FOR BREAST CANCER

Stephan M. Winkler ^(a), Michael Affenzeller ^(b), Gabriel Kronberger ^(c),
Michael Kommenda ^(d), Stefan Wagner ^(e), Witold Jacak ^(f), Herbert Stekel ^(g)

^(a-f) Upper Austria University of Applied Sciences
School for Informatics, Communications, and Media
Heuristic and Evolutionary Algorithms Laboratory
Softwarepark 11, 4232 Hagenberg, Austria

^(g) General Hospital Linz
Central Laboratory
Krankenhausstraße 9, 4021 Linz, Austria

^(a) stephan.winkler@fh-hagenberg.at, ^(b) michael.affenzeller@fh-hagenberg.at, ^(c) gabriel.kronberger@fh-hagenberg.at,
^(d) michael.kommenda@fh-hagenberg.at, ^(e) stefan.wagner@fh-hagenberg.at,
^(f) witold.jacak@fh-hagenberg.at, ^(g) herbert.stekel@akh.linz.at

ABSTRACT

In this paper we describe the use of tumor marker estimation models in the prediction of tumor diagnoses. In previous work we have identified classification models for tumor markers that can be used for estimating tumor marker values on the basis of standard blood parameters. These virtual tumor markers are now used in combination with standard blood parameters for learning classifiers that are used for predicting tumor diagnoses.

Several data based modeling approaches implemented in HeuristicLab have been applied for identifying estimators for selected tumor markers and cancer diagnoses: Linear regression, k-nearest neighbor learning, artificial neural networks, and support vector machines (all optimized using evolutionary algorithms) as well as genetic programming.

In the results section we summarize classification accuracies for breast cancer; we compare classification results achieved by models that use measured marker values as well as models that use virtual tumor markers.

Keywords: Evolutionary Algorithms, Medical Data Analysis, Tumor Marker Modeling, Data Mining,

1. INTRODUCTION, RESEARCH GOALS

In this paper we present research results achieved within the Josef Ressel Centre for Heuristic Optimization *Heureka!*: Data of thousands of patients of the General Hospital (AKH) Linz, Austria, have been analyzed in

order to identify mathematical models for cancer diagnoses. We have used a medical database compiled at the central laboratory of AKH in the years 2005 – 2008: 28 routinely measured blood values of thousands of patients are available as well as several tumor markers (TMs, substances found in humans that can be used as indicators for certain types of cancer). Not all values are measured for all patients, especially tumor marker values are determined and documented only if there are indications for the presence of cancer. The results of empirical research work done on the data based identification of estimation models for cancer diagnoses are presented in this paper: Based on patients' data records including standard blood parameters, tumor markers, and information about the diagnosis of tumors we have trained mathematical models for estimating tumor markers and cancer diagnoses.

In previous work (Winkler et al. 2010; Winkler et al. 2011) we have identified classification models for tumor markers that can be used for estimating tumor marker values on the basis of standard blood parameters. These tumor marker models (also referred to as virtual markers) are now used in combination with standard blood parameters for learning classifiers that can be used for predicting tumor diagnoses. Our goal is to show to which extent virtual tumor markers can replace tumor marker measurements in the prediction of cancer diagnoses.

These research goals and the respective modeling tasks are graphically summarized in Figure 1.

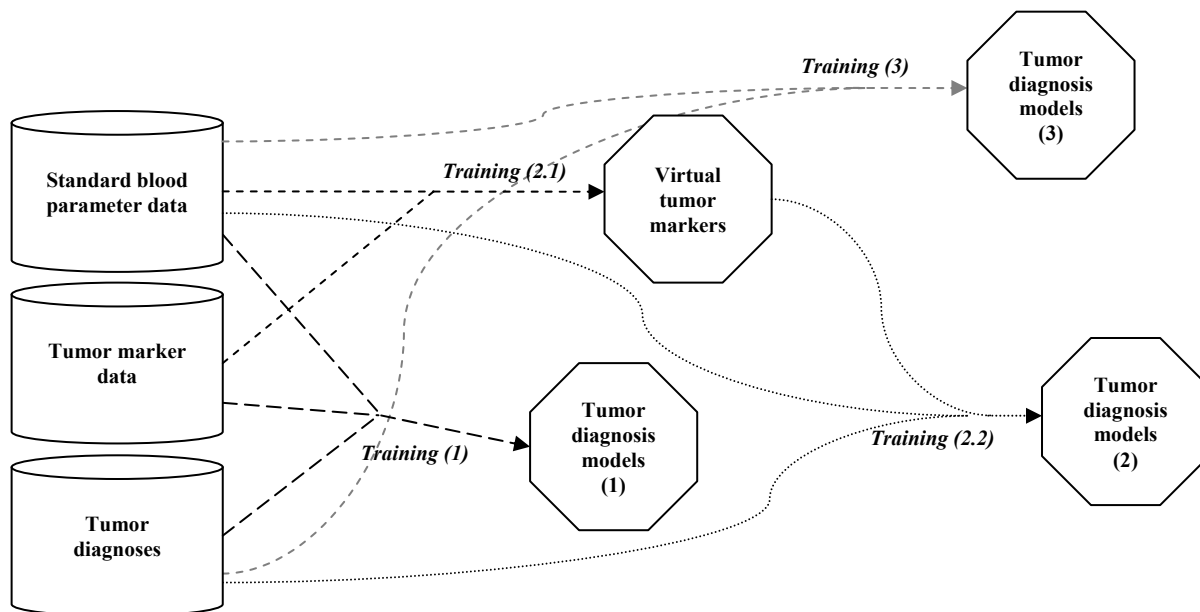


Figure 1: Modeling tasks addressed in this research work: Tumor markers are modeled using standard blood parameters and tumor markers (training scenario 1); tumor diagnosis models are trained using standard blood values and on the one hand tumor marker data and on the other hand using estimated tumor markers (scenario 2); alternatively we also train diagnosis estimation models only using standard blood parameters and diagnosis information (scenario 3).

- First, models are trained for estimating tumor diagnoses using the full set of available data: Standard blood parameters, tumor marker data and diagnoses information are used.
- Second, machine learning is applied for learning estimation models for tumor markers. Concretely, we identify classification models that predict tumor marker values for given standard blood parameters either as “normal” or as “elevated”. Subsequently models are trained for estimating tumor diagnoses using standard blood parameters and diagnoses data, but instead of tumor marker values the estimated tumor marker classifications (calculated using models learned in the first modeling step) are used.
- Third, we also train diagnosis estimation models only using standard blood parameters and diagnosis information.

In the following section (Section 2) we summarize the machine learning methods that were applied in this research project, in Section 3 we give an overview of the data base that was used, and in Section 4 we document the modeling results that could be achieved. This paper is concluded in Section 5.

2. MACHINE LEARNING METHODS APPLIED

In this section we describe the modeling methods applied for identifying estimation models for tumor markers and cancer diagnoses: On the one hand we apply hybrid modeling using machine learning algorithms and evolutionary algorithms for parameter optimization and feature selection (as described in Section 2.1), on the other hand we use genetic

programming (as described in Section 2.2). In (Winkler et al. 2011), for example, these methods have also been described in detail.

2.1. Hybrid Modeling Using Machine Learning Algorithms and Evolutionary Algorithms for Parameter Optimization and Features Selection

Feature selection is often considered an essential step in data based modeling; it is used to reduce the dimensionality of the datasets and often conducts to better analyses. Given a set of n features $F = \{f_1, f_2, \dots, f_n\}$, our goal here is to find a subset of F , F' , that is on the one hand as small as possible and on the other hand allows modeling methods to identify models that estimate given target values as well as possible. Additionally, each data based modeling method (except plain linear regression) has several parameters that have to be set before starting the modeling process.

The fitness of feature selection F' and training parameters with respect to the chosen modeling method is calculated in the following way: We use a machine learning algorithm m (with parameters p) for estimating predicted target values $est(F', m, p)$ and compare those to the original target values $orig$; the coefficient of determination R^2 function is used for calculating the quality of the estimated values. Additionally, we also calculate the ratio of selected features $|F'|/|F|$. Finally, using a weighting factor α , we calculate the fitness of the set of features F' using m and p as

$$fitness(F', m, p) = \alpha \cdot |F'|/|F| + (1 - \alpha) \cdot (1 - R^2(est(F', m, p), orig)). \quad (1)$$

In (Alba et al. 2007), for example, the use of evolutionary algorithms for feature selection optimization is discussed in detail in the context of gene selection in cancer classification. We have now used

evolutionary algorithms for finding optimal feature sets as well as optimal modeling parameters for models for tumor diagnosis; this approach is schematically shown in Figure 2: A solution candidate is here represented as $[s_1 \dots s_m p_1 \dots p_q]$ where s_i is a bit denoting whether feature F_i is selected or not and p_j is the value for parameter j of the chosen modeling method m . This rather simple definition of solution candidates enables the use of standard concepts for genetic operators for crossover and mutation of bit vectors and real valued vectors: We use uniform, single point, and 2-point crossover operators for binary vectors and bit flip mutation that flips each of the given bits with a given probability. Explanations of these operators can for example be found in (Holland 1975) and (Eiben 2003).

We have used strict offspring selection (Affenzeller et al. 2009): Individuals are accepted to become members of the next generation if they are evaluated better than both parents.

In (Winkler et al. 2011) we have documented classification accuracies for tumor diagnoses using this approach for optimizing feature set and modeling parameters.

The following machine learning algorithms have been applied for identifying estimators for selected tumor markers and cancer diagnoses: Linear regression, k-nearest neighbor learning, artificial neural networks, and support vector machines.

2.2. Genetic Programming

As an alternative to the approach described in the previous sections we have also applied a classification algorithm based on genetic programming (GP, Koza (1992)) using a structure identification framework described in Winkler (2008) and Affenzeller et al. (2009).

We have used the following parameter settings for our GP test series: The mutation rate was set to 20%, gender specific parents selection (Wagner 2005) (combining random and roulette selection) was applied as well as strict offspring selection (Affenzeller et al. 2009) (OS, with success ratio as well as comparison factor set to 1.0). The functions set described in (Winkler 2008) (including arithmetic as well as logical ones) was used for building composite function expressions.

In addition to splitting the given data into training and test data, the GP based training algorithm used in our research project has been designed in such a way that a part of the given training data is not used for training models and serves as validation set; in the end, when it comes to returning classifiers, the algorithm returns those models that perform best on validation data.

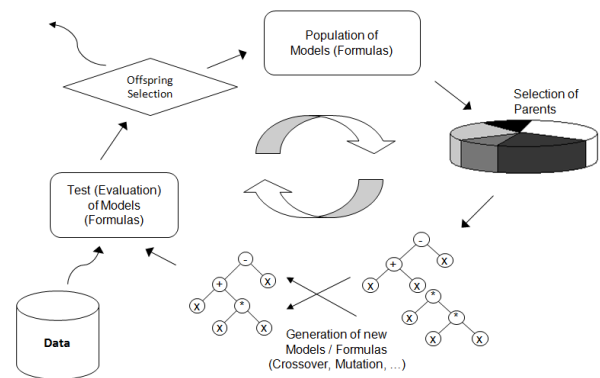


Figure 2: Genetic programming including offspring selection.

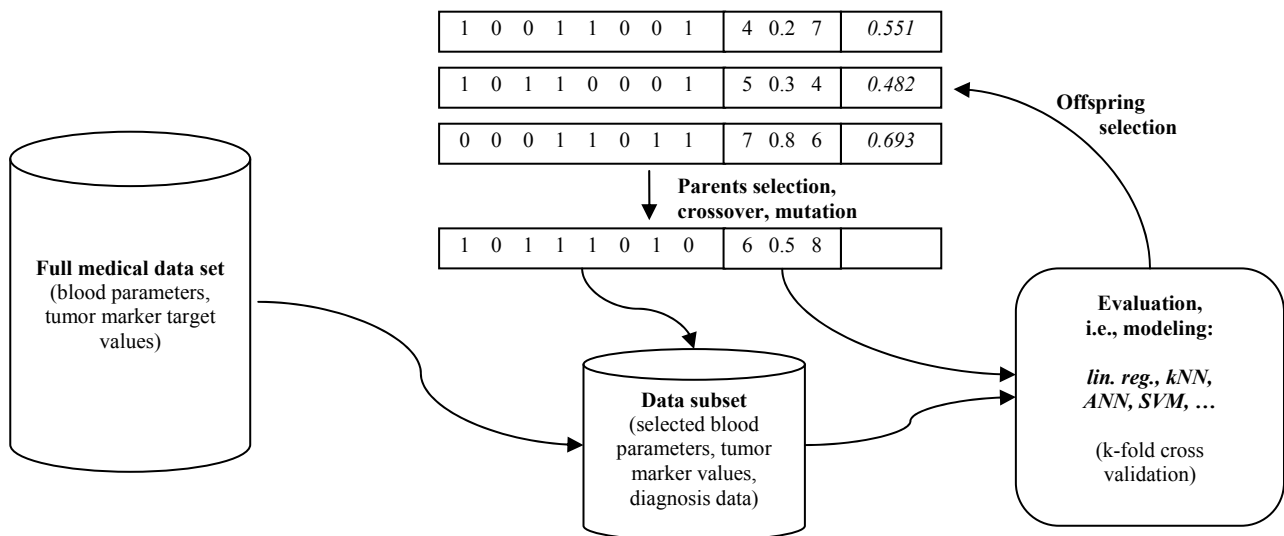


Figure 3: A hybrid evolutionary algorithm for feature selection and parameter optimization in data based modeling. Machine learning algorithms are applied for evaluating feature sets.

3. DATA BASIS

3.1. General Information

The blood data measured at the AKH in the years 2005-2008 have been compiled in a database storing each set of measurements (belonging to one patient): Each sample in this database contains an unique ID number of the respective patient, the date of the measurement series, the ID number of the measurement, standard blood parameters, tumor marker values, and cancer diagnosis information. Patients' personal data were at no time available for the authors except for the head of the laboratory.

In total, information about 20,819 patients is stored in 48,580 samples. Please note that of course not all values are available in all samples; there are many missing values simply because not all blood values are measured during each examination. Further details about the data set and necessary data preprocessing steps can for example be found in Winkler et al. (2010) and Winkler et al. (2011), e.g.

Standard blood parameters include for example the patients' sex and age, information about the amount of cholesterol and iron found in the blood, the amount of hemoglobin, and the amount of red and white blood cells; in total, 29 routinely available patient parameters are available.

Literature discussing tumor markers, their identification, their use, and the application of data mining methods for describing the relationship between markers and the diagnosis of certain cancer types can be found for example in Koepke (1992) (where an overview of clinical laboratory tests is given and different kinds of such test application scenarios as well as the reason of their production are described) and Yonemori (2006).

Information about the following tumor markers is stored in the AKH database: AFP, CA 125, CA 15-3, CA 19-9, CEA, CYFRA, fPSA, NSE, PSA, S-100, SCC, and TPS.

Finally, information about cancer diagnoses is also available in the AKH database: If a patient is diagnosed with any kind of cancer, then this is also stored in the database. Our goal in the research work described in this paper is to identify estimation models for the presence of the breast cancer, cancer class C50 according to the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10).

3.2. Data Preprocessing

Before starting the modeling algorithms for training classifiers we had to compile a data set specific for breast cancer diagnosis:

- First, blood parameter measurements were joined with diagnosis results; only measurements and diagnoses with a time delta less than a month were considered.
- Second, all samples containing measured values for breast cancer are extracted.

- Third, all samples are removed that contain less than 15 valid values.
- Finally, variables with less than 10% valid values are removed from the data base.

This leads to a data set specific for breast cancer; this *BC* data set consists of 706 samples with 324 samples (45.89%) labeled with '0' and 382 (54.11%) labeled with '1'.

The following variables are stored in this so compiled data set *BC*: Age, sex, tumor diagnosis (0/1), ALT, AST, BSG1, BUN, C125 (TM), C153 (TM), , CBAA, CEA (TM), CEOA, CH37, CHOL, CLYA, CMOA, CNEA, CRP, FE, FER, GT37, HB, HDL, HKT, HS, KREA, LD37, MCV, PLT, RBC, TBIL, TF, and WBC. Three tumor markers (C125, C153, and CEA) are available in *BC*.

4. MODELING TEST SERIES

We have trained models for estimating breast cancer diagnoses using genetic programming as described in Section 2.2 with strict OS and population size 200; as rather small and compact models are preferred by the authors' medical cooperation partners, the maximum size of the evolved models was set to 20, 35, and 50 nodes. For modeling and test results achieved using bigger model structures please see Winkler et al. (2011).

For training virtual tumor markers we have used the hybrid modeling approach described in Section 2.1: 5-fold cross validation was applied, linear regression (linReg), k-nearest-neighbor (kNN) learning, artificial neural networks (ANNs), and support vector machines (SVMs) have been used as machine learning algorithms, and their feature selections and modeling parameters have been optimized by an evolutionary algorithm. Details about these machine learning approaches and their implementation can for example be found in Winkler et al. (2010) and Winkler et al. (2011).

We have used the implementations in the open source framework HeuristicLab (Wagner (2009)) (<http://dev.heuristiclab.com>).

4.1. Modeling Strategies

For training estimation models for breast cancer we have applied four different strategies: One using measured tumor markers, one using virtual tumor marker classifiers (combined with OR-conjunctions), one using virtual tumor marker classifiers (combined with majority voting), and finally one not using tumor markers at all.

4.1.1. Strategy I: Using standard blood parameters and measured tumor markers

Measured tumor markers were used as well as standard blood parameters, classification models for breast cancer diagnoses were trained using GP as described above.

This corresponds to scenario 1 as described in Section 1.

4.1.2. Strategy II: Using standard blood parameters and virtual tumor marker classifiers

We have used hybrid modeling as described in Section 2.1 for creating virtual tumor marker classifiers on the basis of the data available in the BC data set. The population size for the optimization algorithm (a GA with strict OS) was set to 20, the maximum selection pressure to 100, and α to 0.1; the test classifications of the so identified best virtual tumor marker classifiers were used as estimated tumor marker values. Classification models for breast cancer diagnoses were trained using GP as described above using virtual tumor markers as well as standard blood parameters.

This corresponds to scenario 2 as described in Section 1.

We applied each machine learning algorithm used here (namely linear regression, support vector machines, neural networks, and kNN classification) twice in each modeling process, leading to eight estimated binary classifications for each tumor marker; these classifications were combined into one binary classification variable for each tumor marker using either an OR conjunction or majority voting:

Strategy II.a: Using OR: If any of the classifiers for a tumor marker return 1 for a sample, then this sample's virtual tumor marker is 1; else it is set to 0.

Strategy II.b: Using majority voting: If more than the half of the classifiers for a tumor marker return 1 for a sample, then this sample's virtual tumor marker is 1; else it is set to 0.

4.1.3. Strategy III: Using only standard blood parameters

In this strategy no tumor markers were used; instead, only standard blood parameters were available for training classification models for breast cancer diagnoses using GP as described above.

This corresponds to scenario 3 as described in Section 1.

4.2. Test Results

As already mentioned, each strategy was executed using five-fold cross validation; we here report on the average classification accuracies on test samples ($\mu \pm \sigma$) which are also shown in Figure 4:

- Strategy I: 0.777 ± 0.104
- Strategy II.a: 0.713 ± 0.107
- Strategy II.b: 0.752 ± 0.042
- Strategy III: 0.699 ± 0.113

5. CONCLUSIONS

In the test results summarized in Section 4 we see that the virtual tumor markers have turned out to be able to improve classification accuracy for the modeling application described in this paper: Whereas classifiers not using tumor markers classify approximately 70% of the samples correctly, the use of virtual tumor markers (combined using majority voting) leads to an increase of the classification accuracy to ~75%. Still, virtual tumor markers have in this example not been able to replace

measured tumor markers perfectly, as the classification accuracy of models using measured TMs reaches ~77.7%.

Future work on this topic shall include the investigation of virtual TMs for other types of diseases; furthermore, we will also focus on the practical application of the here presented research results in the treatment of patients.

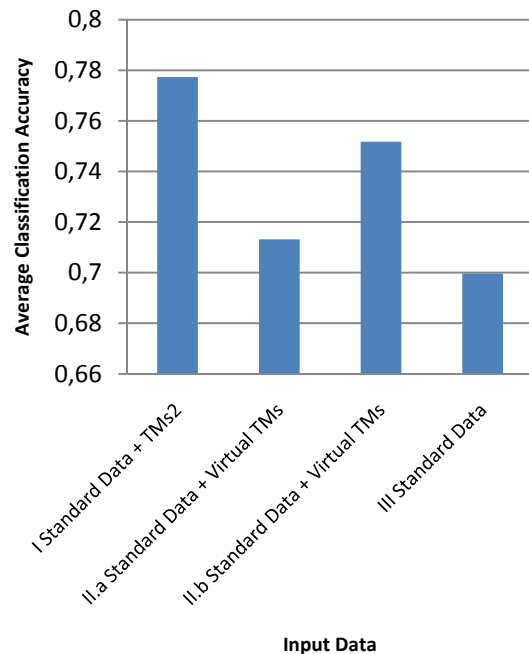


Figure 4: Average test accuracies achieved using the four modeling strategies described in Section 4.1

ACKNOWLEDGMENTS

The work described in this paper was done within the Josef Ressel Centre for Heuristic Optimization *Heureka!* (<http://heureka.heuristiclab.com/>) sponsored by the Austrian Research Promotion Agency (FFG).

REFERENCES

- Affenzeller, M., Winkler, S., Wagner, S., A. Beham, 2009. *Genetic Algorithms and Genetic Programming - Modern Concepts and Practical Applications*. Chapman & Hall/CRC. ISBN 978-1584886297. 2009.
- Alba, E., García-Nieto, J., Jourdan, L., Talbi, E.-G., 2005. Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. *IEEE Congress on Evolutionary Computation 2007*, pp. 284 – 290.
- Eiben, A.E. and Smith, J.E. 2003. Introduction to Evolutionary Computation. *Natural Computing Series*, Springer-Verlag Berlin Heidelberg.
- Holland, J.H., 1975. *Adaption in Natural and Artificial Systems*. University of Michigan Press.
- Koepke, J.A., 1992. Molecular marker test standardization. *Cancer*, 69, pp. 1578–1581.

- Rechenberg, I., 1973. *Evolutionsstrategie*. Friedrich Frommann Verlag.
- Schwefel, H.-P., 1994. *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie*. Basel: Birkhäuser Verlag.
- Wagner, S., 2009. *Heuristic Optimization Software Systems - Modeling of Heuristic Optimization Algorithms in the HeuristicLab Software Environment*. PhD Thesis, Institute for Formal Models and Verification, Johannes Kepler University Linz, Austria.
- Winkler, S., Affenzeller, M., Jacak, W., Stekel, H., 2010. Classification of Tumor Marker Values Using Heuristic Data Mining Methods. *Proceedings of Genetic and Evolutionary Computation Conference 2010, Workshop on Medical Applications of Genetic and Evolutionary Computation*, pp. 1915–1922.
- Winkler, S., Affenzeller, M., Jacak, W., Stekel, H., 2011. Identification of Cancer Diagnosis Estimation Models Using Evolutionary Algorithms – A Case Study for Breast Cancer, Melanoma, and Cancer in the Respiratory System. *Proceedings of Genetic and Evolutionary Computation Conference 2011, Workshop on Medical Applications of Genetic and Evolutionary Computation*.
- Winkler, S., 2009. *Evolutionary System Identification - Modern Concepts and Practical Applications*. Schriften der Johannes Kepler Universität Linz, Reihe C: Technik und Naturwissenschaften. Universitätsverlag Rudolf Trauner. ISBN 978-3-85499-569-2.
- Yonemori, K., Ando, M., Taro, T. S., Katsumata, N., Matsumoto, K., Yamanaka, Y., Kouno, T., Shimizu, C., Fujiwara, Y., 2006. Tumor-marker analysis and verification of prognostic models in patients with cancer of unknown primary, receiving platinum-based combination chemotherapy. *Journal of Cancer Research and Clinical Oncology*, 132(10), pp. 635–642.

AUTHORS BIOGRAPHIES



STEPHAN M. WINKLER received his PhD in engineering sciences in 2008 from Johannes Kepler University (JKU) Linz, Austria. His research interests include genetic programming, nonlinear model identification and machine learning. Since 2009, Dr. Winkler is professor at the Department for Medical and Bioinformatics at the University of Applied Sciences (UAS) Upper Austria at Hagenberg Campus.



MICHAEL AFFENZELLER has published several papers, journal articles and books dealing with theoretical and practical aspects of evolutionary computation, genetic algorithms, and meta-heuristics in general. In 2001 he received his PhD in engineering sciences and in 2004 he received his habilitation in applied systems engineering, both from the Johannes Kepler University of Linz, Austria. Michael Affenzeller is professor at UAS, Campus Hagenberg, and head of the Josef Ressel Center *Heureka!* at Hagenberg.



GABRIEL KRONBERGER received his PhD in engineering sciences in 2010 from JKU Linz, Austria, and is a research associate at the UAS Research Center Hagenberg. His research interests include genetic programming, machine learning, and data mining and knowledge discovery.



MICHAEL KOMMENDA finished his studies in bioinformatics at Upper Austria University of Applied Sciences in 2007. Currently he is a research associate at the UAS Research Center Hagenberg working on data-based modeling algorithms for complex systems within *Heureka!*.



STEFAN WAGNER received his PhD in engineering sciences in 2009 from JKU Linz, Austria; he is professor at the Upper Austrian University of Applied Sciences (Campus Hagenberg). Dr. Wagner's research interests include evolutionary computation and heuristic optimization, theory and application of genetic algorithms, and software development.



WITOLD JACAK received his PhD in electric engineering in 1977 from the Technical University Wroclaw, Poland, where he was appointed Professor for Intelligent Systems in 1990. Since 1994 Prof. Jacak is head of the Department for Software Engineering at the Upper Austrian University of Applied Sciences (Campus Hagenberg) where he currently also serves as Dean of the School of Informatics, Communications and Media.



HERBERT STEKEL received his MD from the University of Vienna in 1985. Since 1997 Dr. Stekel is chief physician at the General Hospital Linz, Austria, where Dr. Stekel serves as head of the central laboratory.

AUTOMATIC SELECTION OF RELEVANT DATA DURING ULTRASONIC INSPECTION

T. Merazi Meksen^(a), M. Boudraa^(a), B. Boudraa^(a)

^(a)University of Science & Technology H. Boumedienne
BP 32, El Alia, Bab Ezzouar, Algiers, Algeria.

t_merazi@yahoo.fr

ABSTRACT

In recent years, research concerning the automatic interpretation of data from non destructive testing (NDT) is being focused with an aim of assessing embedded flaws quickly and accurately in a cost effective fashion. This is because data yielded by NDT techniques or procedures are usually in the form of signals or images which often do not present direct information of the condition of the structure.

Signal processing has provided powerful techniques to extract from ultrasonic signals the desired information on material characterization, sizing and defect detection. The imagery available can add additional and significant dimension in NDT information and for exploiting information.

The task of this work is to minimize the volume of data to process replacing ultrasonic images type TOFD by sparse matrix, as there is no reason to store and operate on a huge number of zeros. A combination of two types of neural networks, a perceptron and a Self Organizing Map of Kohonen is used to distinguish between a noise signal from a defect signal in one hand, and to select the sparse matrix elements which correspond to the locations of the defects in the other hand. This new approach to data storage will provide an advantage for the implementations on embedded systems.

Keywords: workstation design, work measurement, ergonomics, decision support system.

Keywords- Automatic testing; Materials; Ultrasonics; Neural Networks.

1. INTRODUCTION

The use of non destructive testing (NDT) allows the analysis of internal properties of structures without causing damage to the material. Various methods have been developed to detect defects in structure and to evaluate eventually their locations, sizes and characteristics. Some of these methods are based on analysis of the transmission of different signals such as ultrasonics, acoustic emission, thermography, x-radiography, eddy current (Cartz 1995). In the last

decade,

ultrasonic techniques have shown to be very promising for non destructive testing (Blitz 1996) and they are becoming an effective alternative to radio-graphic tests. X-ray widely used to detect and sizing discontinuities, presents the disadvantage to produce ionising radiation and needs to develop a film, which takes some times to provide the results.

Operators are often required to acquire and interpret large volumes of complex inspection data. So, automated signal analysis systems are finding increasing applications in a variety of industries where the diagnostics is difficult. Ultrasonic data can be displayed as images and can add additional and significant dimension in NDT information and thus for exploiting in applications. Many advanced image processing algorithms have provided powerful techniques to extract from ultrasonic images the desired information on sizing and defect detection (Chen 2004; Merazi, 2006; Jasiuniene 2007). But all these methods require considerable amount of computation, making them difficult for real-time operation.

Many mechanized inspection techniques, sensors, and systems for automating defect detection and location have been developed (Cchatzakos 2007, Martin 2007, Berke 2000; Moles 2005; Shuxiang 2004). However, the location and sizing of a defect is an almost entirely manual process: An operator will mark on the scan, using a mouse, where the component echoes lie, and thus where defect lies. The apparatus will then perform the correction and give an indication of the defect size according to what has been indicated by the operator.

Hardwares have been developed and integrated tools of image processing are implemented in order to completely automate the control. Most of these algorithms are computationally intensive, so it is desirable to implement them in high performance reconfigurable systems. Recently, Field Programmable Gate Array (FPGA) technology becomes a viable target for the implementation of algorithms for image processing. (Sato 2009; Johnston 2004; Nelson 2000). E. Ashari developed a method for NDT image binarization by thresholding, implemented in FPGA (Ashari 2004). K.Appiah uses a single chip of FPGA in

order to extract background models presented in an image and to reduce inspection time (Appiah 2005) and D. J Durlington uses a reconfigurable features FPGA performing a variety of operations in hardware, the control program being executed on a microprocessor (Durlington 1997).

On ultrasonic images the zone of interest is often very small in comparison with the image dimensions. This make sense to use a special matrix type called sparse matrix, where only the non zero elements are stored. Not only this reduces the amount of data to store, but also operations of this type of matrix can take advantage of the a-priori knowledge of the positions of non-zero elements to accelerate the calculations (Pissanetzky; Duff 1987).

The aim of this work is to minimize the data to store and to process in order to save memory and computational time. An original approach for data acquisition and representation, which consists on sparse matrix construction instead of an ultrasonic image type TOFD (Time Of Flight Diffraction) is described. It is based on the TOFD technique but avoids the image formation. The sparse matrix is built by combination of a perceptron and a self organizing map algorithm of Kohonen in order to select a defined number of samples from the signals.

Section 2 and 3 in this paper describe respectively ultrasonic non destructive inspection and TOFD technique. In section 4, the two types of Neural Networks used in this work are developed, namely perceptron and Self Organizing Map of Kohonen. Experimental measurements and application of combination of the neural nets are described in section 5. Section 6 concerns the conclusion.

Papers that don't adhere to the guidelines provided in this template will be returned to authors for appropriate revision.

2. ULTRASONIC INSPECTION

The basic components of an ultrasonic inspection system are a pulser/receiver, the cabling, the transducers, and the acoustic/elastic wave propagation and scattering (Figure 1).

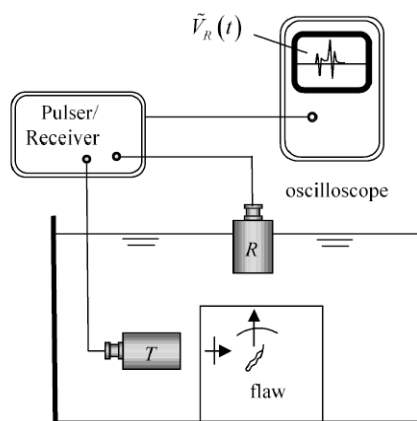


Figure 1: Signal acquisition system for ultrasonic inspection

The pulser section of the pulser/receiver generates short electrical pulses which travel through the cabling to the transmitting transducer. The transducer converts these electrical pulses into acoustic pulse at its acoustic output port, which can be or not be in contact with the material under control. In the latter case, a liquid (couplant) is used to facilitate the transmission of ultrasonic vibrations from the transducer to the test surface. This ultrasonic beam is also transmitted into the solid component being inspected and interacts with any flaw that is present. The flaw generates scattered wave pulses travelling in many directions, and some of these pulses reach the receiving transducer which converts them into electrical pulses. These electrical pulses travel again through cabling to the receiver section of the pulser/receiver, where they are amplified and displayed as a received A-scan voltage $V_r(t)$ as a function of the time. Figure 2 shows an example:

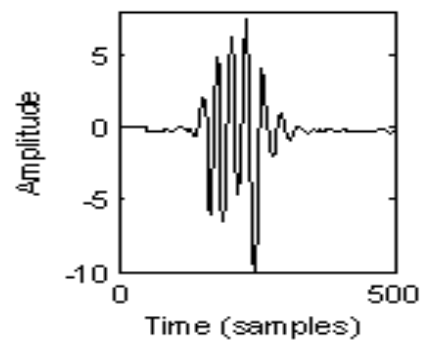


Figure 2: Example of an ultrasonic signal

3. THE TOFD TECHNIQUE

TOFD technique uses the travel time of a diffracted wave at the tip of a discontinuity (Silk1984). The TOFD research concludes that technique is portable, fast, reliable, accurate and inexpensive in the defect detection and sizing. Further, inspection can be semi or fully automated for the defect detection in metal structures. Two transducers, one as a transmitter and the second as a receiver are moved automatically step by step according to a straight line and the diffracted signals are recorded and displayed as images. Those images provide different texture patterns for the detected defects and automatic texture segmentation is investigated using different techniques to improve the detection and classification of defects.

In his thesis, J. Sallard demonstrates that a generator of a hole, can be assimilated to a top of a crack (Sallard 1999). So, a test block containing a hole has been used in this work to test this method (figure 3).

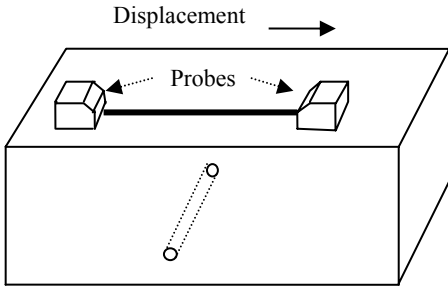


Figure 3: Test block with an artificial defect

Figure below shows the result obtained scanning a test block containing an artificial crack. Every row is constituted of a samples of a reached signal.

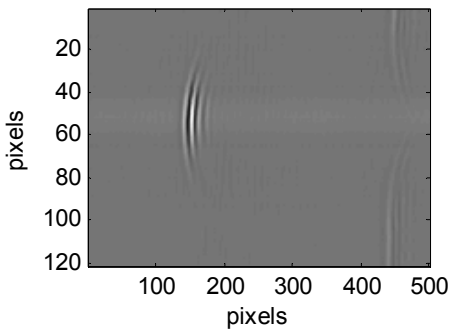


Figure 4: TOFD image showing a defect

IV. ARTIFICIAL NEURAL NETWORKS

The Artificial Neural Networks (ANN) parallel, distributive computational structure is reminiscent of the human neural system. In an ANN structure, many simple nonlinear processing elements, called neurons, are interconnected via weighted synapses to form a network inputs. The functionality is mostly dependent on the values of the weights which can be updated over time, causing the neural network to adapt and possibly “learn”. The learning process is of different types: supervised learning, unsupervised learning, self-organized learning. In a supervised approach, the network is fed with necessary input and the appropriate output for the specified inputs is given the output is achieved together with a global error function. The computed output is compared to the desired output to evaluate the performance of the neural network. The computed error function is then used to update the weights with an aim of achieving output that is close to the desired output.

In contrast to the supervised learning, unsupervised learning or self organizing learning does not require any assistance of desired outputs or an external teacher. Instead, during the training session, the neural network receives a number of different patterns and discovers significant features in these patterns and learns how to classify input data into appropriate categories.

4.1 The Perceptron

The Perceptron is a binary classifier that maps its input x (a real-valued vector) to an output single binary value. If two sets are linearly separable, this classification can be used to decide whether a given vector belongs to one class or another (Rosenblatt 1962).

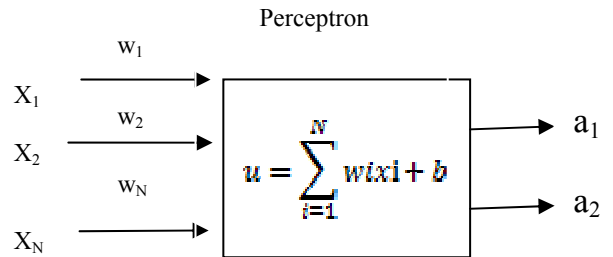


Figure 5: Functional description of a perceptron

The function of each neuron is to compute a weighted sum of all synapse inputs, add the sum to a predefined bias and pass the result through a nonlinear sigmoidal (threshold) function f whose output ranges between 0 and 1:

$$a = f(u) = \frac{1}{1 + \exp(-u)} \quad (1)$$

In this work, inputs (X_1, X_2, \dots, X_N) correspond to the N signal samples. The outputs a_1 and a_2 , are respectively the defect-signal-classe and the noise-signal-classe.

4.2 Kohonen Self-Organizing Maps

Kohonen Self Organizing Maps (SOM) is a widely used ANN model based on the idea of self organized or unsupervised learning (Kohonen 1988). The SOM network is a data visualization technique, which reduces the dimensions of data through a variation of neural computing networks. It is a non parametric approach that makes no assumptions about the underlying population distribution and is independent of prior information. The problem that data visualization attempts to solve is that humans simply cannot visualize high dimensional data. So, SOM goes about reducing dimensions by producing maps of usually one or two dimensions which plot the similarities of the data by grouping similar data items together. Thus, SOM's accomplish two things: they reduce dimensions, and display similarities. In this work, this property allows to select a fixed number of relevant data to store and to process. Figure 6 demonstrates basic structure of self-organizing Kohonen map:

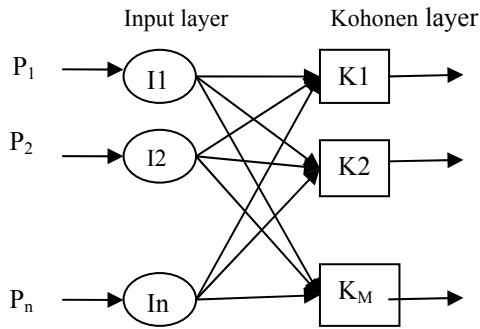


Figure 6: Self Organizing Kohonen Map network

The network has input and output layers of neurons that are fully interconnected among themselves. At each step of training phase an output layer's neuron with weights that best match with input data (usually in a minimum Euclidian distance sense) is proclaimed as the winner. The weights of this neuron and its neighborhood neurons are then adjusted to be closer to the presented input data. The algorithm is described as follows:

1- Initialize W with uniform-random values.

2- For each input vector P , compute the distance d between the vector P and the M weight vectors based on the mean square root:

$$d = (P_k - W_l)^2 \quad l = 1, 2, 3, \dots, M \quad (2)$$

3- Select the vector W_x such that W_x satisfies Equation (3) :

$$(P_k - W_x)^2 = \min(P_k - W_l)^2 \quad l = 1, 2, 3, \dots, M \quad (3)$$

4- Update W_x using Equation 4:

$$W_x(t+1) = W_x(t) + \alpha(X_k - W_x) \quad (0 < \alpha < 1) \quad (4)$$

5- Go to step 2 until $W_l \approx P_l$

In the early learning stage, α is set about 0.8. As the learning progresses, α gradually becomes closer to 0.

V. MEASUREMENTS AND RESULTS

According to the principle of the TOFD technique, two transducers (2 Mhz) are moved step by step by 5 mm each time, straight a line. At each position of the probes, the reached diffracted signal is first analyzed by a perceptron in order to determine if it is a "defect signal" or "a noise signal".

In the next step, every defect signal (classe a_1) is processed and the sample corresponding to the maximum of the amplitude is detected. This sample corresponds to the time of flight of the reached ultrasonic signal (cross on figure bellow).

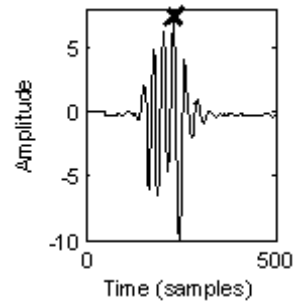


Figure 7: Example of ultrasonic reached signal. The horizontal coordinate of the maximum (indicated by the cross) corresponds to the time of flight of the reached ultrasonic wave

From every signal of defect-signal class, the coordinates p_i and p_j corresponding respectively to the time of flight and the position of the probe, according to the straight line of the displacement, are stored.

At the end of this process, a set of points $P(p_i, p_j)$ are determined and their number equals the number of the signals in class a_1 .

The Self Organization Map of Kohonen is applied with those points as inputs in order to reduce their number to a defined one depending on the desired sparse matrix dimension (30 in this work). Figure below shows the positions of the output elements resulting, obtained using the signals that form the TOFD image on figure 4.

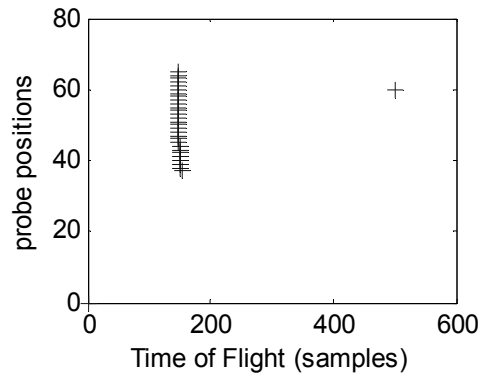


Figure 8: Sparse Matrix obtained instead of the TOFD image of figure 4.

Instead of 120x500 pixels of the TOFD image, this method selects 30 elements which are sufficient to describe the pattern presented in the image. This number is defined by the Self organizing Map algorithm outputs and is independent of the quantity of initial data. The economy of memory is important and the defect location and characterization will be faster when analyzing only the sparse matrix elements.

V. CONCLUSIONS

In this work, a method to detect and locate cracks by analyzing a sparse matrix built from TOFD signals has been described. A first layer of a neural net selects a

point from the reached signal if an echo signal is presented in the zone of interest. The co-ordinates of this point which correspond respectively to the probe position and the time of flight of the signal are stored and used as inputs for a self organizing map network. Outputs will represent a group of points corresponding to the defect presented in the structure. Results of the application of this technique have been promising in terms of speed, and robustness. This would make the proposed system suitable for implementation in situations requiring near real-time processing and interpretation of large volumes of data giving thus an important help in the decision making.

REFERENCES

- Appiah K and Hunter A. 2005. A Single-Ship FPGA Implementation of Real time Adaptive Background Model. *Proceeding of IEEE International Conference on Field Programmable Technology. FPT'05*, pp. 95-102. Singapore, December, 11-14.
- Ashari E. and Hornsey R., 2004 'FPGA Implementation of Real Time Adaptive Image Thresholding', Thesis presented to University of Waterloo, Ontario, Canada.
- Berke M and Kleinert W.D. 2000. Portable Work Station for Ultrasonic Weld Inspection, *15th World Conf. of Non Destructive Testing, WCNDT2000*, Roma, Italy.
- Cartz, L. 1995. *Nondestructive testing, Radiography ultrasonics, Liquid penetrant, Magnetic Particle, Eddy Current*. ASM International.
- Cchatzakos P. and Markopoulos Y. P. 2007. Towards Robotic Non Destructive Inspection of Industrial Pipelines. *4th International Conference on NDT, HSNDTint 2007*, Chania-Crete, Greece Oct. 11-14.
- Chen C.H. 2004. Advanced Image Processing Methods for Ultrasonic NDE Research. *World Congress of Non Destructive Testing, Proc. WCNDT 2004*, Aug. 30-Sep. 3, 2004, pp 39-43. Montreal, Canada.
- Darlington D.J. and Campbell, D.R.. 1997 Reconfigurable FPGAs for data Compression in Ultrasonic Non Destructive Testing, *IEE Colloquium on DSP chips in Real Time Measurement and Control*. 25 September, Leicester, UK.
- Duff I.S and Erisman A.M., 1987, *Direct Method for Sparse Matrix*, Clarendon Press, New York, USA.
- Jasiuniene J. (2007), Ultrasonic Imaging Techniques for Non Destructive Testing of Nuclear Reactors, cooled by liquid Metals: Review. *ULTRAGRAS*, Vol.62, N° 3, pp.39-43.
- C.J Johnston and K.T Gribbon, 2004. 'Implementing Image Processing Algorithms on FPGAs', *Proceeding of 11 Electronic New Zeland Conference ENZCon'04*, pp. 118-123, Palmerston North, New Zeland.
- Kohonen T. 1988, *Self Organization and Associative Memory*, Springer Verlag, Heidelberg..
- Martin C.J and Gonzalez Bueno R.. 2007. Ultrascope TOFD : Un sistema compacto para captura y procesamiento de imagenes TOFD. *IV Conferencia Panamericana de END, PANNDT 2007*, Buenos Aires, Argentina, October 22-26,
- Merazi Meksen T & Boudraa M, (2006), Application of the Randomized Hough Transform on Ultrasonic images in Non Destructive Testing. *WSEAS transactions on signal processing* Vol.2, Issue 8, pp. 1053-1056 Aug. 06.
- Moles M. (2005), Portable Phase Array Application, *3rd, Middle East Nondestructive Testing Conference & Exhibition MENDT*. Manama, Bahrein, 27-30 Nov.
- Nelson A.E., 2000. *Implementation of Image Processing Algorithm on FPGA Hardware*. Thesis in Electrical Engineering, Faculty of the graduate school of Vanderbilt.
- Pissanetzky, *Sparse Matrix Technology*, <http://www.scicontrols.com>
- Rosenblatt F (1962) *Principles of neurodynamics*. New York, Spartan.
- Sallard J. 1999. *Etude d'une Méthode de Déconvolution Adaptée aux Images Ultrasonores*. Thesis presented at the Institut National Polytechnique de Grenoble. France.
- Satoh K and Tad J., 2009. Three dimensional ultrasonic Imaging Operation using FPGA, *IEICE Electronics Express*. Vol. 2, N°2, pp 84-89.
- Shuxiang J. 2004, Development of an Automated Ultrasonic Testing System, *SPIE Proceeding, 3rd International Conference on Experimental Mechanics*, Vol.5852, Singapore.
- Silk M.G., 1984. The Use of Diffraction-Based Time of Flight Measurements to Locate and Size Defects, *British Journal of NDT*, Vol 26, Pages 208-213, May 1984.
- Yella S. and Dougherty M.S., 2006, Artificial Intelligence Techniques for the automatic Interpretation of Data from Non Destructive Testing *Insight*, Vol. 48, N°1, January 2006.

CT BASED MODELS FOR MONITORING BONE CHANGES IN PARAPLEGIC PATIENTS UNDERGOING FUNCTIONAL ELECTRICAL STIMULATION

Páll Jens Reynisson^(a), Benedikt Helgason^(b), Stephen J. Ferguson^(b), Thordur Helgason^{(c),(d)}, Rúnar Unnþórsson^(a), Páll Ingvarsson^(e), Helmut Kern^(f), Winfried Mayr^(g), Ugo Carraro^(h) and Paolo Gargiulo^{(c),(d)}

a. School of Engineering and Natural Sciences, Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland,

b. Institute for Surgical Technology and Biomechanics, University of Bern, Switzerland

c. Department of Development and Consultancy UTS, Landspítali-University Hospital, Reykjavik, Iceland.

d. Department of Biomedical engineering, University of Reykjavik, Iceland.

e. Department of Rehabilitation Medicine, Landspítali-University Hospital.

f. Ludwig Boltzmann Institute of Electrostimulation and Physical Rehabilitation, Department of Physical Medicine, Wilhelminenspital. Vienna, Austria.

g. Medical University of Vienna, Center of Biomedical Engineering and Physics

h. Laboratory of Translational Myology of the University of Padova, Department of Biomedical Sciences, Padova, Italy.

^(a)pall@bestreykjavik.com, ^(b)benedikt.helgason@artorg.unibe.ch, ^(b)Stephen.Ferguson@istb.unibe.ch
^{(c),(d)}thordur@landspitali.is, ^(a)runson@hi.is ^(e)palling@landspitali.is, ^(f)helmut.kern@wienkav.at
^(g)winfried.mayr@meduniwien.ac.at, ^(h)ugo.carraro@unipd.it, ^{(c),(d)}paologar@landspitali.is,

ABSTRACT

Spinal Cord Injured (SCI) paraplegics suffer from pathological changes in the lower extremities such as muscle degeneration, hormonal alterations and bone resorption. The aim of the present study was, with the help of finite element and image analysis, to evaluate the osteogenic response of the patella of paraplegic patients undergoing Functional Electrical Stimulation (FES). For this purpose, a patient that began a daily home based FES treatment 5 years after paralysation was monitored. Bone mechanical parameters were compared at the beginning of FES and after 3 years of treatment.

According to our results, it is possible to conclude that application of long term FES treatment on denervated, degenerated muscles can be beneficial for bone growth in bones attached to the stimulated muscles.

Keywords: Finite Element Modelling, Functional Electrical Stimulation, Flaccid Paraplegia, Bone Strains, Patella.

1. INTRODUCTION

Spinal Cord Injured (SCI) paraplegics suffer from pathological changes in the lower extremities, such as muscle degeneration, hormonal alterations and bone resorption (Maimoun et al 2006). Loss of bone mineral density (BMD), which is one of the symptoms of osteoporosis (Marcus et al 2008), results in bones becoming more fragile, with an increased risk of

fracture in the paraplegic's extremities as a consequence (Maimoun et al 2005; Lazo et al 2001; Jiang et al 2006). In order to decrease this acceleration of tissue deterioration, electrotherapy such as Functional Electrical Stimulation (FES) has been proposed (Gargiulo, 2008; Gargiulo et al 2009, 2011; Kern 2002, 2005; Gallasch & Rafolt et al 2005; Mandl et al 2008). However, recent studies claim that the lack of osteogenic response in paralyzed extremities to electrically evoked exercise during sub-acute and rehabilitation/recovery phases, could not be fully explained, and may warrant further evaluation (Clark et al 2007).

CT data can be used to monitor bone changes in paraplegic patients by quantification of morphological parameters. Additionally CT based finite element models can provide useful information on the structural changes that result from the changes in bone morphology when direct mechanical testing is not possible. The absence of mechanical stimulation in the lower extremities of paraplegic patients makes the patella bone, ideal for monitoring such changes resulting directly from external stimulation.

The aim of the present study was thus to perform a CT based evaluation of the osteogenic response of the patella of paraplegic patients undergoing FES.

2. MATERIAL AND METHODS

A pre-treatment CT dataset of the lower extremities for a paraplegic patient (a 32 year old male suffering

complete flaccid ThXI syndrome with paralysis and areflexia in the legs and medium atrophy) was used to create a FE model of the patient's right patella. A QUASAR phosphate phantom (Modus Medical Devices Inc., London, Ontario, Canada) was used to calibrate the images. The construction of the FE model was carried out in several steps briefly explained as follows:

I) Creation of a 3D triangular surface mesh (STL) through semi-automatic segmentation in MIMICS (Materialize Interactive Medical Image Control System, Leuven, Belgium) using a Hounsfield (HU) threshold of 200 to determine the boundary between bone and soft tissue. II) Creation of a solid model (IGES) in SOLIDworks (Dassault Systèmes SolidWorks Corp., Concord, Massachusetts, USA). III) Creation of a 10-node tetrahedral FE mesh in ANSYS Workbench (ANSYS, Canonsburg, Pennsylvania, USA). IV) FE model imported into ANSYS (ANSYS, Canonsburg, Pennsylvania, USA), where FE equations are solved.

Rigid boundary conditions were applied at the patella ligament and the patella tendon attachments points. Force (F_r) was applied on the model where the distal femur contacts the patella. For determining this force a biomechanical model of the knee joint (Ward 2004; Ward 2005) was used. The joint moments were measured during stimulation at the beginning and three years into the FES treatment, with a non invasive pendulum test (Gallasch & Rafolt et al 2005). Moment arms and force directions in the sagittal plane were derived from the CT data.

Isotropic, linear elastic, heterogeneous material properties were assigned to each node in the model with an in-house MATLAB script (The Mathworks, Natick, MA) based on the NI material mapping method introduced by Helgason et al. (2008b). The relationship (1) between Young's modulus (E) and apparent bone density (ρ) was taken from Morgan et al. (2003):

$$E = 6850\rho^{1.49} \text{ (MPa)} \quad (1)$$

Poisson's ratio was set to 0.3.

A CT dataset acquired after 3 years of daily home based FES treatment was also available for the same patient. Using the procedure described above, this CT dataset was used to create another FE model for comparison to the pre-treatment situation. After FE simulations, the equivalent Von Mises element strains were derived from the FE solutions and compared as shown in Fig. 1. Additionally, Young's modulus distribution, patella bone total volume, weight and average Young's modulus were derived directly from the CT data as shown in Fig. 3 and Tab. 1. A third CT dataset of a healthy, 37 year old male was available for comparison to the patient results but FE simulation was not carried out for this individual.

3. RESULTS

The maximum load on the patella during FES, derived from the biomechanical knee model, was found to be $F_r = 60 \pm 2$ N at the start of FES treatment and $F_r = 123 \pm 3$

N after 3 year of FES treatment. The volumetric strain histograms derived from the FE simulations are illustrated in Fig. 1, with a Frost interval of [200, 2000] micro strains (Frost HM, 1987) and, Rubin and Lanyon threshold of 1000 micro strains, (Rubin & Lanyon, 1987) indicated but these authors reported these thresholds being relevant for maintaining bone mass. 71% of the total bone volume was found to be strained beyond 200 micro strains and 5% beyond 1000 micro strains at the start of FES treatment. The corresponding results after three years of FES were 71% and 19%. The calculated patella weight (mg), bone volume (mm^3) and average Young's modulus (MPa) per element are presented in Tab. 1. Fig. 2 and 3 illustrate a comparison between the Young's modulus distribution for the healthy subject and the patient before and after FES treatment.

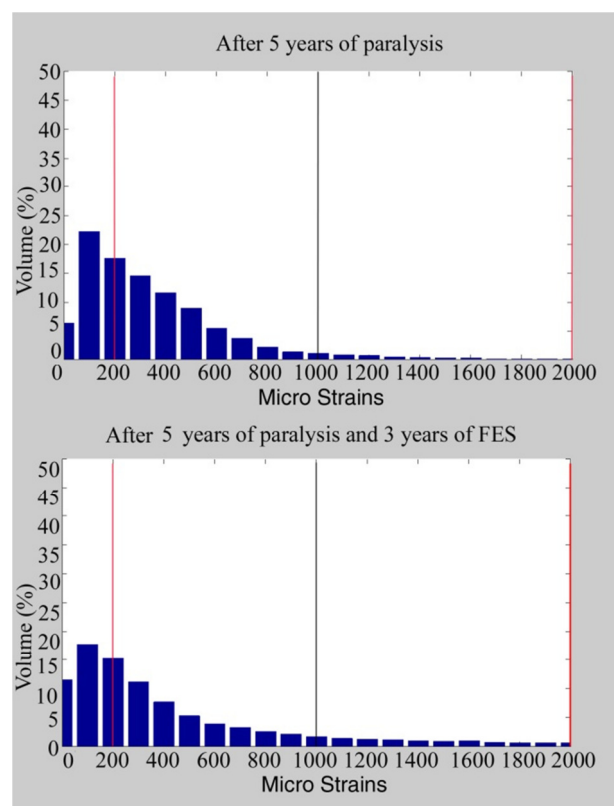


Figure 1. Mechanical strain stimulation according to FE simulations, before FES treatment and after 3 year of FES treatment. The red lines indicate the Frost interval [200, 2000] micro strains (Frost HM 1987) but the grey line indicates the Rubin and Lanyon threshold of 1000 micro strains.

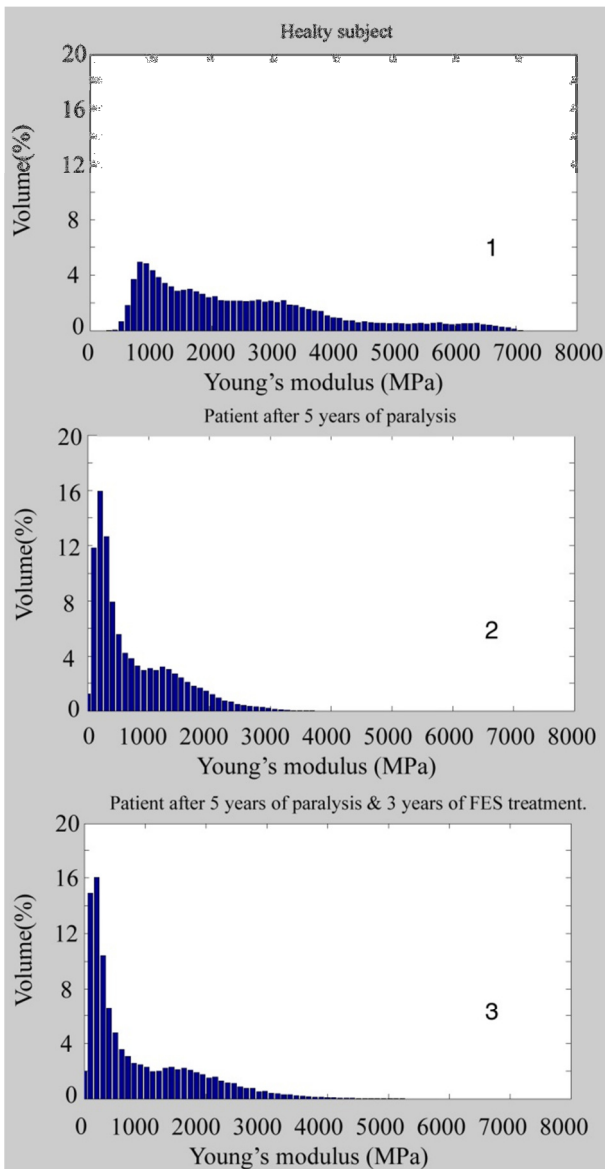


Figure 2. Histograms of Young's modulus distribution. 1. Patella of a healthy subject. 2. Paraplegic patient after 5 years of paralysis. 3. Same Paraplegic patient after 5 years of paralysis and 3 years of FES treatment.

Trabecular bone			
Sets	Weight (mg)	Average Young's Modulus (MPa)	Volume (mm ³)
Healthy Subject	7210	2490	14100
5.y of paralysis	4820	790	13100
5.y of paralysis and 3 y of FES	5340	960	13600

Table 1. Trabecular bone quantities (HU range 200-1000 HU); Total weight (mg), average Young's (MPa) and total volume.

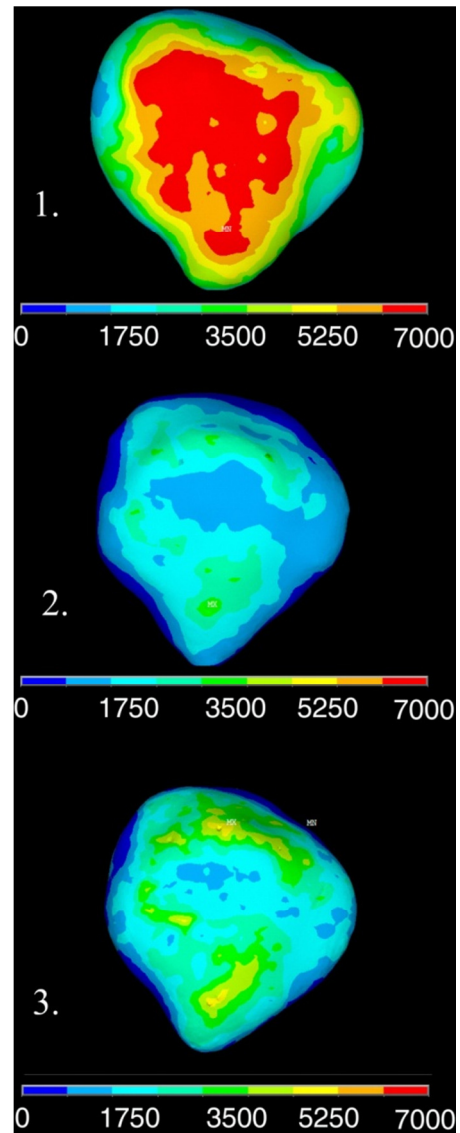


Figure 3. Young's modulus distribution (MPa). Anterior view of the right patella. 1. Patella of a healthy subject. 2. Paraplegic patient after 5 years of paralysis. 3. Same Paraplegic patient after 5 years of paralysis and 3 years of FES treatment.

4. CONCLUSION

The aim of the present study was to evaluate the osteogenic response of the patella of a paraplegic patient undergoing FES using the finite element method and image analysis. The results, as shown in Fig. 2, indicate that bone strain stimulus at the start of treatment was sufficient for bone formation according to published thresholds for bone maintenance (Rubin & Lanyon 1984 and Frost 1987). The results also indicate, that even after 3 years of FES treatment, strain stimulus was larger than was found at the beginning of treatment. This suggests that there is potential for further bone formation.

Comparing the stiffness for the pre- and post-treatment situation as shown in Fig. 3, indicates that during FES the bone is adapting to the loads being applied, especially at the patella ligament and

quadriceps femoris insertion points. The trabecular bone appears to adapt to the load increase by increasing both mass and volume which results in increased average Young's modulus. However, the influence of long term disuse on the cortical shell can clearly be seen in Fig. 3, where Young's modulus distribution in the patella for the healthy and paralysed subjects is compared.

According to our results it is possible to conclude that application of long term FES treatment on denervated degenerated muscles can be beneficial for bone growth in bones attached to the stimulated muscles.

ACKNOWLEDGMENTS

This work has been supported by The University Hospital research fund of Iceland Landspítali, Keilir Institute of Technology and Department of Biomedical engineering, University of Reykjavik.

REFERENCES

- Clark JM, Jelbart M, Rischbieth H, Strayer J, Chatterton B, Schultz C, and Marshall R., 2007. *Physiological effects of lower extremity functional electrical stimulation in early spinal cord injury: lack of efficacy to prevent bone loss*. Spinal Cord, 45: 78-85.
- Frost HM., 1987. *Bone "mass" and the "mechanostat": a proposal*. Anat Rec, 219(1) :1-9,
- Gallasch E & Rafolt D, Kinz G, Fend M, Kern H, and Mayr W., 2005. *Evaluation of FES-Induced Knee Joint Moments in Paraplegics with Denervated Muscles*. Artificial Organs, 29(3) :207-211,
- Gargiulo P, Helgason T, Reynisson PJ, Helgason B, Kern H, Mayr W, Ingvarsson P., Carraro U., 2011. *Monitoring of Muscle Recovery in Spinal Cord Injury Patients Treated With Electrical Stimulation Using Three-Dimensional Imaging and Segmentation Techniques: Methodological Assessment*, Artificial organs, 35(3): 275-281.
- Gargiulo P, Vatnsdal B, Ingvarsson P, Knútsdóttir SS, Guðmundsdóttir V, Yngvason S, Kern H, Carraro U, Thordur Helgason., 2009. *Computational methods to analyze tissue composition and structural changes in denervated muscle undergoing therapeutic electrical stimulation*, Basic Applied Myology, 19 (4):157-161.
- Helgason B, Perilli E, Schileo E, Taddei F, Brynjólfsson S, and Viceconti M., 2008a *Mathematical relationship between bone density and mechanical properties: A literature review*. Clinical Biomechanics, 23(2):124, 135-146.
- Helgason B, Taddei F, Pálsson H, Schileo E, Cristofolini L, Viceconti M, and Brynjólfsson S., 2008b. *A modified method for assigning material properties to FE models of bones*. Medical Engineering & Physics, 30(4):444-453.
- Jiang SD, Dai LY, and Jiang LS, *Osteoporosis after spinal cord injury*., 2006. Osteoporos Int, 17(2):160-192.
- Kern H, Hofer C, Modlin M, Forstner C, Raschka-Högler D, Mayr W, Stöhr H., 2002. *Denervated muscles in humans: limitations and problems of currently used functional electrical stimulation training protocols*. Artificial Organs, 26(3): 216–218.
- Kern H, Salmons S, Mayr W, Rossini K, Carraro U., 2005. *Recovery of long-term denervated human muscles induced by electrical stimulation*. Muscle Nerve, 31(1):98–101.
- Lazo MG, Shirazi P, Sam M, Giobbie-Hurder A, Blacconiere MJ, and Muppidi M., 2001. *Osteoporosis and risk of fracture in men with spinal cord injury*. Spinal Cord, 39(4):208-214.
- Mandl T, Meyerspeer M, Reichel M, Kern H, Hofer C, Mayr W, Moser E., 2008. *Functional Electrical Stimulation of Long-term Denervated, Degenerated Human Skeletal Muscle: Estimating Activation Using T2-Parameter Magnetic Resonance Imaging Method*. Artificial Organs, 32(8): 604–608.
- Maïmoun L, Fattal C, Micallef J-P, Peruchon E, P Rabischong., 2006. *Bone loss in spinal cord-injured patients: from physiopathology to therapy*. Spinal Cord 44:203–210.
- Morgan EF, Barnes GL, Einhorn TA. *The Bone Organ System: Form and Function*., 2008. third edition, OSTEOPOROSIS Chapter 1:3-25.
- Morgan EF, Bayraktar HH, and Keaveny TM. *Trabecular bone modulus-density relationships depend on anatomic site*., 2003. Journal of Biomechanics, 36(7) :897-904.
- Parfitt AM., 2008. *Skeletal Heterogeneity and the Purposes of Bone Remodeling: Implications for the understanding of Osteoporosis*. third edition, OSTEOPOROSIS Chapter 2:71-89.
- Rubin CT, & Lanyon LE. *Kappa Delta Award Paper. Osteoregulatory nature of mechanical stimuli: Function as a determinant for adaptive remodeling in bone*., 1987. Journal of Orthopaedic Research Volume 5(2):300-310.
- Tadashi S, Kaneko, Jason S, Bell, Marina R, Pejčic, Jamshid Tehranzadeh, Joyce H, Keyak., 2004. *Mechanical properties, density and quantitative CT scan data of trabecular bone with and without metastases*. Journal of Biomechanics, 37(4) :523-530.
- Thimas SJ. *Relative electron density calibration of CT scanners for radiotherapy treatment planning*., 1999. The British Journal of Radiology, 72(860) :781-786.
- Ward SR and Powers CM. *The Influence of Patella Alta on Patellofemoral Joint Stress During Normal and Fast Walking*., 2004. Clin. Biomech, 19(10) :1040-1047.
- Ward SR, Terk MR, Powers CM. *Influence of patella alta on knee extensor mechanics*., 2005. Journal of Biomechanics, 38(12) :2415-2422.

AUTHORS BIOGRAPHY

Páll Jens Reynisson is a lecturer and engineer at Keilir Institute of Technology. He earned his bachelor's degree in mechanical and industrial engineering at University of Iceland in 2007 and obtained a MSc. in biomedical engineering from Reykjavik University in 2010.

Benedikt Helgason is a research associate at the Institute for Surgical Technology and Biomechanics, at the University of Bern in Switzerland. He earned his bachelor's degree in civil engineering from the University of Iceland in 1993, a master's degree in civil engineering from the Technical University of Denmark in 1996 and a Ph.D. degree in biomedical engineering at the University of Iceland in 2008.

Stephen J. Ferguson is currently the head of the Biomechanics Division at the Institute for Surgical Technology and Biomechanics, at the University of Bern in Switzerland. He earned his bachelor's degree in mechanical engineering at the University of Toronto in 1991, a master's degree from the same university in 1994 and a Ph.D. degree from Queen's University in 2000. Stephen J. Ferguson has been appointed a full professor at the ETH in Zurich from 1st of July, 2011.

Thordur Helgason is Associate professor at Reykjavik University and works as biomedical engineer and researcher at the University Hospital of Iceland. He studied electrical engineering at University of Iceland, and obtained a Ph.D. at Technical University of Karlsruhe, Germany. His research interests are in the field of electrical stimulation biomedical technologies.

Rúnar Unnþórsson is head of Keilir Institute of Technology. He studied mechanical engineering at University of Iceland, and obtained a CS in 1997, MSc in 2002 and Ph.D in 2008. Rúnar Unnþórsson has been teaching at University of Iceland since 2001. Rúnar is appointed assistant professor at the University of Iceland from 1st of august, 2011.

Páll Ingvarsson is neurologist specialised at Goteborg University, Sweden. He works at Department of Rehabilitation Medicine, Landspítali-University Hospital, Reykjavik, Iceland.

Helmut Kern is head of the "Department of Physical Medicine and Rehabilitation of the Wilhelminenspital" (Vienna, Austria) since 1984 and director of the research institute "Ludwig Boltzmann Institute of Electrical Stimulation and Rehabilitation" since 1988.

Winfried Mayr is Associate Professor of Biomedical Engineering and Rehabilitation Technology. He works at Medical University of Vienna Center for Medical Physics and Biomedical Engineering. His research interests are in the field of Functional electrical stimulation (mobilization of paraplegics, phrenic pacing, EMG-controlled stimulation, pelvic floor applications, denervated muscles, application of FES in microgravity and bed-rest) Implant technology (Microelectrical and electromechanical implants).

Ugo Carraro is Associate Professor of General Pathology at the Faculty of Medicine of the University of Padova, since 1983. Editor-in-Chief of Basic and

Applied Myology/The European Journal of Translational Myology since 1991, he founded and chair since 2005 the University of Padova Interdepartmental Research Center of Myology. **Paolo Gargiulo** is assistant professor at Reykjavik University and works as biomedical engineer and researcher at the University Hospital of Iceland. He studied electrical engineering at University Federico II in Naples, and obtained a PhD at Technical University of Vienna, Austria. His research interests are in the field of electrical stimulation medical modeling and rapid prototyping for clinical applications.

A HIGH RESOLUTION DISTRIBUTED AGENT-BASED SIMULATION ENVIRONMENT FOR LARGE-SCALE EMERGENCY RESPONSE

Glenn I. Hawe, Graham Coates, Duncan T. Wilson, Roger S. Crouch

School of Engineering and Computing Sciences, Durham University, United Kingdom

g.i.hawe@durham.ac.uk, graham.coates@durham.ac.uk, d.t.wilson@durham.ac.uk, r.s.crouch@durham.ac.uk

ABSTRACT

This paper describes the architecture of an agent-based simulation environment for large-scale emergency response. In an effort to increase “model payoff”, it uses two representations for both the environment and the agents. Around each incident, where topographical information is necessary, an *operational level simulator* program models the environment using the OS MasterMap topography and Integrated Transport Network (ITN) layers. At these incident sites, first responder agents are modelled with a rich repertoire of actions. The remaining area of interest, encompassing the locations of relevant resource bases (e.g. ambulance stations, hospitals and fire stations) which are outside of the incident sites, is modelled using only transport network information by a *tactical level simulator* program. This program also simulates the tactical level agents, communicates with each operational level simulator, and provides a viewer. A separate Pre-Simulator program allows new scenarios to be set up with ease.

Keywords: agent-based simulation, emergency response

1. INTRODUCTION

The simulation of emergency scenarios is an important part of the preparedness stage of the emergency management cycle (Haddow 2010). *In silico* simulation in particular finds numerous applications within emergency preparedness and response (Jain and McLean 2003, Longo 2010).

Agent-based models are a popular way of simulating responses to large-scale emergencies (Khalil *et al* 2009). Roberts (2010) describes an agent-based model (ABM) design as:

“...one in which analogs of those real-world entities that are to be modeled are represented as software agents, or objects, *at a level of detail and resolution necessary to address the questions the model is required to answer.*”

The importance of finding the appropriate level of resolution in an ABM is emphasized by Grimm *et al.* (2005):

“Finding the optimal level of resolution in a bottom-up model’s structure is a fundamental problem. If a model is too simple, it neglects essential

mechanisms of the real system, limiting its potential to provide understanding and testable predictions regarding the problem it addresses. If a model is too complex, its analysis will be cumbersome and likely to get bogged down in detail.”

This leads to the concept of the “Medawar zone” (Grimm *et al.* 2005), a region of complexity in which the ABM is not only useful for its intended purpose, but is also structurally realistic. Together, the usefulness and structural realism of an ABM determine the “model payoff”. The Medawar zone may be seen as an application of the Aristotelian notion of “the golden mean”, the most desirable region between two extremes, to ABM design.

Whilst North and Macal (2007) assert that “realistic agent behaviors are the key to agent-based modeling”, they also state that “properly specified agent environments are critical for correct agent operation.” Thus it is important that both the agents and their environment are appropriately represented.

In the remainder of this section, we briefly discuss different representations of agents and their environment in ABMs for large-scale emergency response, and consider their relation to model payoff. We also briefly discuss the high-level software organization of existing ABMs. Then in Section 2 we propose the use of two different representations for both the environment and the agents: one for around incident sites, and another for elsewhere. In Section 3 we describe our software which uses these two representations. Section 4 provides a summary, and describes future planned developments and use of the software.

1.1. Agent representations for large-scale emergency response

The range of complexity which is possible when implementing agents, from simple production systems to cognitive architectures, is discussed by Gilbert (2007). An agent consists of a state (variables) and behavior (methods). The behavior of an agent manifests itself through the actions it decides to make. At the highest level, decision making may be *descriptive* or *normative* (Peterson 2009). Depending on which is used to implement the behavior of the agents, one of two conceptually different ABMs may arise:

1. *Descriptive*: An ABM in which the behaviors of agents are designed to mimic that of their real-life counterparts.
2. *Normative*: An ABM in which agents have behavior which is not based on reality. Instead their behavior is designed to be optimal, according to some criteria.

An example of an ABM which has agents whose behavior is defined by normative decision making is the RoboCup Rescue Simulation (RRS) (Skinner and Ramchurn 2010). The aim in the annual RRS competition is to design *ex novo* behaviors for police, fire brigade and ambulance agents, along with their control centers, to optimize an objective function which combines the health of civilians and damage to property. As Carley *et al* (2006) says, “it is concerned with designing smart algorithms, not with investigating a current human social system as it exists and designing a public policy for it.” Such “smart algorithms” are often quite complex. A simple, yet crude, measure of the complexity of agent representation could be the number of lines of code taken to implement it. For example, thousands of lines of Java code were used to implement agent behaviors in RoboAkut (Akin 2010), the winning entry in the RRS 2010 competition. An interesting approach which yields agents of different complexities is described by Runka (2010). Agents are represented by decision trees, which evolve using genetic programming (Koza 1992). Different fitness functions yield different sized trees, corresponding to agents of different complexity.

This paper is concerned with agents whose behavior is designed to mimic that of their real-life counterparts. A variety of ABMs exist which implement such agents, and complexity can vary greatly. For example, the rescuer agents in the ABS SimGenis (Saoud 2006) are quite simple in their implementation (despite being described as having “perceptive and cognitive intelligence”). A set of heuristic rules determine their behaviors. The casualty agents are even simpler, having only a discrete-valued health state, the evolution of which is modelled using a Markov chain. In PLAN-C (Narzisi 2007), rescue agents are also quite simple. For example, the pseudo-code in (Mysore 2006) is only a few lines long. Although pseudo-code, it is low-level enough to suggest that the actual (Java) code would not be significantly longer. More complex are the rescue agents in the AROUND project (Chu 2009), which learn their behavior from their human counterparts through interactive sessions. Weights in a utility function, which combines multiple objectives, are adjusted so as to select actions in a manner which is most consistent with that of the human experts. State of the art cognitive architectures, such as Soar (Lehman 2006) and ACT-R (Anderson 2007), do not appear to have yet been applied to large-scale emergency response ABMs.

Just from these examples, it is evident that a range of complexities are possible for representing agents, and

different research groups differ in what they deem appropriate. Müller (1999) gives eleven general guidelines for choosing the right agent architecture to apply to a specific problem, one of which is “Do not break a butterfly upon a wheel” (i.e. do not waste effort in developing complex agents when simpler agents suffice). While Sun (2006) points out that most social simulation tools “embody very simplistic agent models, not even comparable to what has been developed in cognitive architectures”, Gilbert (2006) questions when cognitive architectures are needed anyway. Grimm *et al* (2005) point out that many ABMs “try only one model of decision-making and attempt to show that it leads to results compatible with a limited data set”, and point out the flaws in doing so.

One characteristic which existing ABMs for large-scale emergency response do share however is that within each ABM, the representation of any particular agent is *constant*. For example, whether a firefighter agent is leaving the fire station, travelling to an incident scene, or inside the inner cordon, it is modelled using the same code, and thus at the same level of complexity, throughout the simulation.

1.2. Environment representations for large-scale emergency response

Some large-scale emergencies, such as earthquakes, may cause widespread damage to the environment, whilst others, such as terrorist bombs, may cause localized damage. Some may even cause no damage to the environment, e.g. human pandemics. Thus, the most appropriate representation for the environment is dependent on the type of large-scale emergency being simulated, and in particular the damage it causes.

For example, in the ABM EpiSimS (Del Valle 2006), which models the effect of different policies on the spread of human pandemics, only the transport network is modelled. RRS on the other hand, which simulates the response to an earthquake, models the buildings as well, as many will be damaged and need to be considered during the response effort. The representation of the environment in RRS is explored by Sato and Takahashi (2011). They found that the representation of topography influenced simulation results: when modelled at a lower resolution, buildings were found to take a longer time to burn; at a higher level of resolution, gaps between the smaller buildings prevented fire from spreading.

In the context of military simulations, which use triangulated irregular networks (TIN) to model terrain, Campbell *et al.* (1997) point out that “users naturally favour high resolution, high fidelity models because of the realism they offer, but computers that run the simulations may not be able to store and process the amount of data that is associated with these high resolution models.” They propose “by identifying tactically significant (and insignificant) terrain, we can more effectively manage the TIN budget by suggesting areas of terrain that should be modelled at high and low fidelity”, i.e. the use of *different resolutions* within a single model.

1.3. Software organization of ABMs for large-scale emergency response

Many ABMs are single-process programs. However, some do make use of multiple programs, and in particular distributed memory parallelism.

RRS makes use of *functional* parallelism. Different sub-simulators, each of which simulates one aspect of the earthquake response scenario (such as a fire sub-simulator, and a flood sub-simulator), run on separate processors. Using more processors enables more functionality to be modelled; however it does not allow larger areas to be simulated. As well as functional parallelism, the IDSS ABM (Koto 2003), which is also designed for earthquake response simulation, uses data parallelism: large geographical regions are split into smaller ones, which are simulated in parallel. The use of up to 34 machines is reported for modelling an area affected by an earthquake.

Data parallelism is also used in EpiSimS to split the large environment, spanning five US counties, into smaller regions, which are then simulated in parallel following a master-slave model. Del Valle *et al.* (2006) report distributing a single simulation over 106 processors.

EpiSimS also reports the use of separate programs for “enhanced pre- and post-processing” (Del Valle *et al.* 2006). An “*InitializeHealth*” program allows the user to specify different probability distributions on the population being modelled. A “graphic user interface enables it to be used by nonprogrammers”. This program is part of a suite of programs that are combined into a “*Set-up Wizard*”. A set of scripts, which call programs such as gnuplot and Excel, are then used for carrying out post-processing on the output files generated by the simulation.

2. IMPROVING MODEL COMPLEXITY FOR LARGE-SCALE EMERGENCIES

In this section, we propose the use of more than one representation for both agents and their environment, when simulating large-scale emergency response.

2.1. Agent representation

In the vicinity of an incident site, where the casualties are (possibly trapped), first responders carry out a wide range of actions. Using the National Occupational Standards for firefighters in the U.K. (U.K. Firefighter NOS 2005) as an example, four broad groups of activities may be identified (out of nine) as being directly relevant at the time of an emergency. These four groups are highlighted in bold in Table 1.

Using the detailed descriptions of these four activity groups, twelve distinct actions may be identified, as shown in Figure 1. Of these twelve distinct actions, only one is relevant away from the actual incident sites: “driveVehicle” (the action necessary to arrive at the incident site). Thus, away from the incident sites, it is unnecessary to model the full repertoire of twelve actions for firefighters. In this regime, the behavior of firefighters reduces to movement

Table 1: Activities in the National Occupational Standards for FireFighters in the U.K.

Ref	Activity
FF1	Inform and educate your community to improve awareness of safety matters
FF2	Take responsibility for effective performance
FF3	Save and preserve endangered life
FF4	Resolve operational incidents
FF5	Protect the environment from the effects of hazardous materials
FF6	Support the effectiveness of operational response
FF7	Support the development of colleagues in the workplace
FF8	Contribute to safety solutions to minimize risks to your community
FF9	Drive, manoeuvre and redeploy fire service vehicles

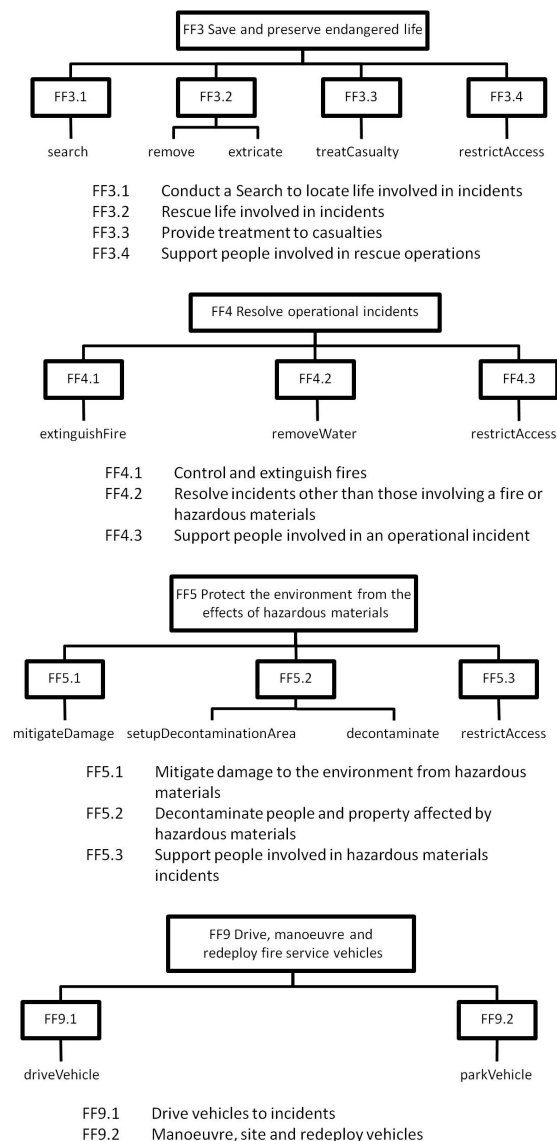


Figure 1: Identification of FireFighter agent actions

along the transport network, and the ABS is more akin to a traffic simulation. The same arguments hold for other first responder agents, such as paramedics and police.

2.2. Environment representation

The preceding discussion leads us to two different representations of the environment also. Around incident scenes, information is required (cues) in order for agents to discern which action to perform. The Ordnance Survey MasterMap (OS MasterMap 2011) topography and Integrated Transport Network (ITN) layers are used to model the topography and transport network of rectangular regions centred around each individual incident scene.

Away from the incident scenes however, as the only action which agents are performing is the action of moving, only the transport network needs to be represented. Thus, only the ITN layer is used to model the larger area which surrounds the incident sites. This area is sufficiently large to capture all the hospitals, fire stations and police stations that may be involved in resolving the incident.

Figure 2 illustrates our approach for the case of the London 2005 bombings. The locations of the bomb explosions are modelled using the topography and ITN layers. Edgware Road (Figure 2 (a)) and Liverpool Street (Figure 2 (b)) are modelled as separate 1 km² regions. As the Tavistock Square and King's Cross/Russell Square incidents were quite close, they are modelled together in a larger 3 km² (1.5 km x 2 km) region (Figure 2 (c)). To capture all the hospitals and fire stations used, the road network in the larger 60 km² area is modelled using the ITN layer as shown.

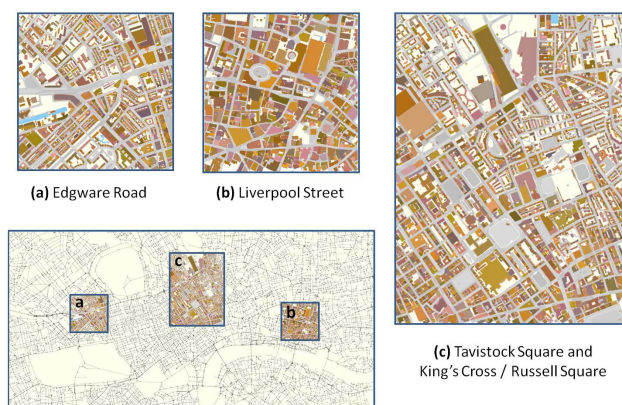


Figure 2: Two representations for the environment

3. SOFTWARE ORGANIZATION

In this section, we describe the high-level organization of the software used to set up and perform simulations, using the different representations mentioned in the previous section.

Three separate programs make up the agent-based simulation environment:

1. A Pre-Simulator program, which is used to set up the emergency to be simulated.
2. A Tactical level simulator program, which:
 - a. simulates tactical level agents,
 - b. simulates first responders when they are travelling along the road network, to and from incident sites,
 - c. provides a viewer for the simulation,
 - d. communicates with the operational level simulator programs.
3. An Operational level simulator program, one instance of which simulates one individual incident site.

Figure 3 shows how these programs are organized. The Pre-Simulator program is used to set up the scenario to be simulated. The details are written to xml files which then serve as input to the Tactical level Simulator program.

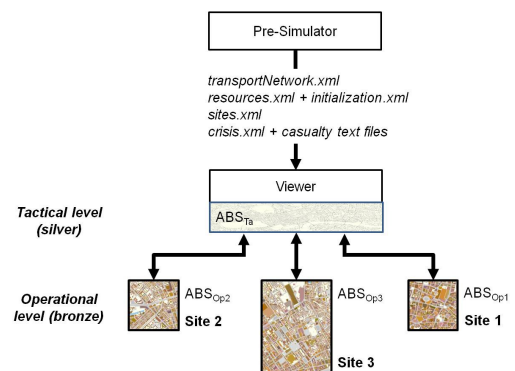


Figure 3: High-level software organization

More precisely, the Pre-Simulator is used to:

1. Specify the OS ITN file defining the transport network, and (optionally) specify a folder containing OS StreetView raster files (OS StreetView 2011) covering the same area. This information is saved to *transportNetwork.xml*.
2. Identify nodes on this transport network which represent locations of resource bases (hospitals, fire stations, ambulance stations and police stations). This information is saved to *resources.xml*.
3. Set the resources available at each resource base (e.g. the number of ambulances at each ambulance station). This information is saved to *initialization.xml*.
4. Initialize the positions and crew of each individual resource. This information is also saved to *initialization.xml*.
5. Specify the individual incident sites, and the OS MasterMap topography file(s) which define the topography in each. This information is saved to *sites.xml*.
6. Set up the incidents (including casualty information held in text files) at each incident site. This information is saved to *crisis.xml*.

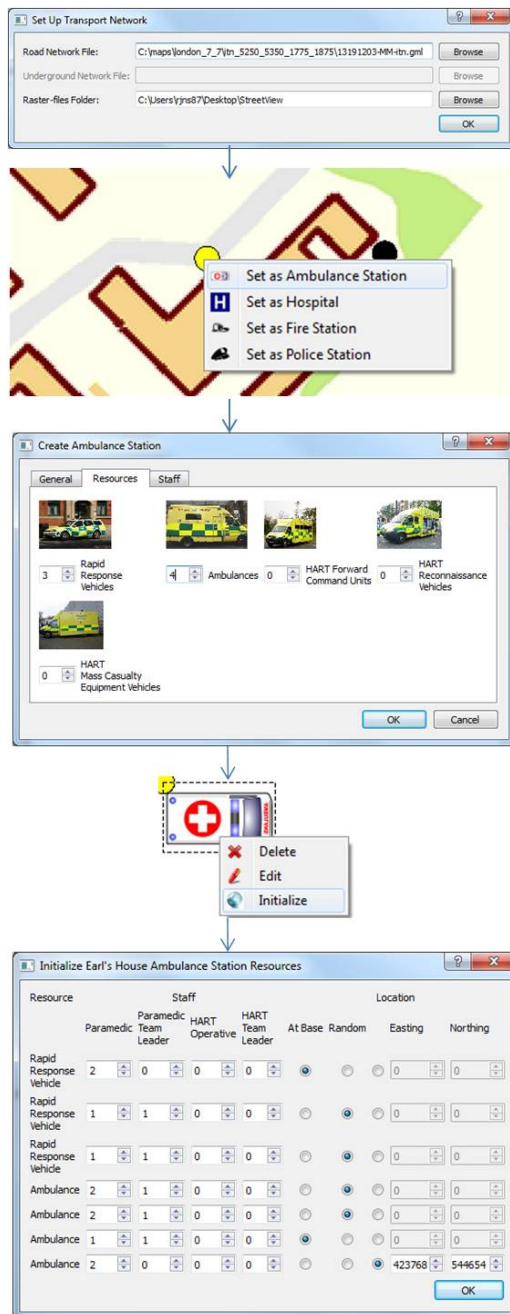


Figure 4: Setting up the transport network, identifying resource bases, and initializing resources in the Pre-Simulator.

Steps 1-4 of this process are illustrated in Figure 4, whilst steps 5-6 are illustrated in Figure 5. Note that, once the road network is loaded and displayed (Step 1), the user must select nodes as resource bases (Step 2). As it is difficult to identify the appropriate nodes using the road network alone, the first dialog in Figure 4 allows the user to (optionally) specify a folder containing OS StreetView raster files. If specified, these are superimposed onto the Pre-Simulator view which shows the transport network, allowing the user to easily identify the nodes of interest.

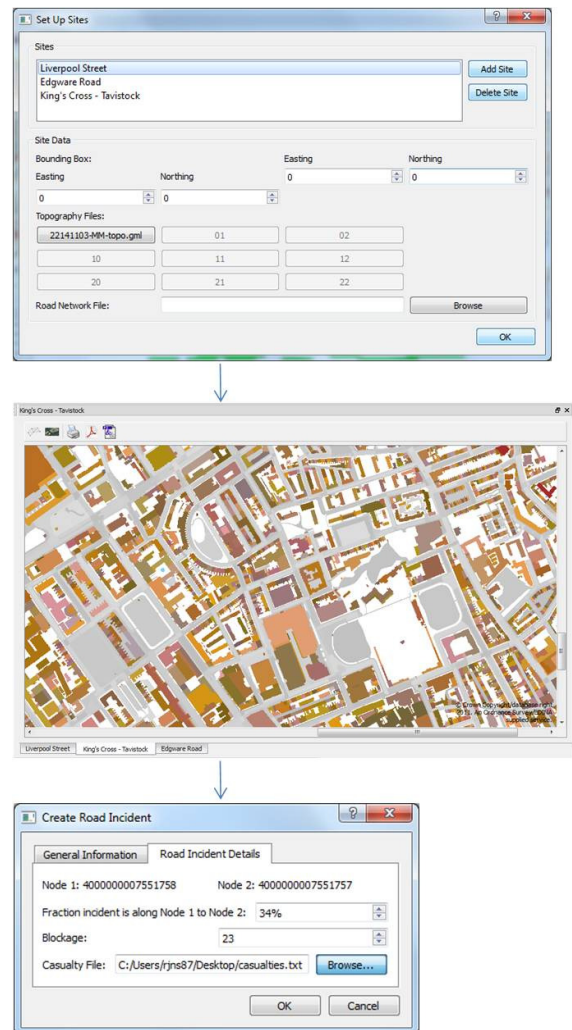


Figure 5: Setting up the incident sites, and creating incidents in the Pre-Simulator.

The tactical level simulator takes the xml input files written by the Pre-Simulator as command line parameters, and uses them to create the virtual environment and populate it with agents. It also provides a viewer of the environment, as shown in Figure 6. Each incident site has its own dockable window showing the topography of the area as defined by its OS MasterMap topography files. These windows are docked in region “a” in Figure 6. The area outside the incident sites is represented by the transport network, but is visualized using the OS StreetView maps. This window is labeled “b” in Figure 6.

The tactical level simulator also simulates the tactical level agents (in the strategic-tactical-operational command structure used for major incidents in the U.K. (LESLP 2007)). These agents are responsible for issuing a plan (usually a predetermined attendance) to the available resources. This is represented in the form of an evolving Gantt chart, which is shown in the window labeled “c” in Figure 6. Finally, the window labeled “d” in Figure 6 shows how the estimated total number of fatalities evolves with time.

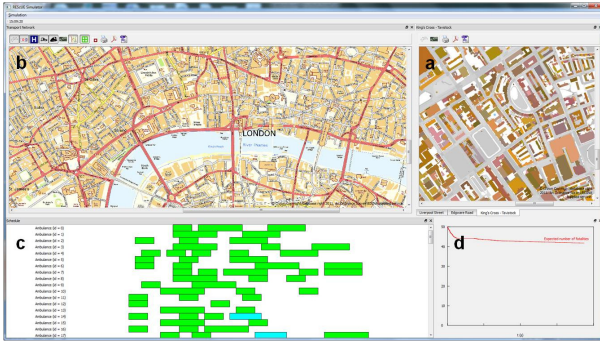


Figure 6: Screenshot of the viewer

The road network is represented as a graph in the tactical level simulator, using the Boost graph library (Siek, Lee and Lumsdaine 2002). This allows the use of Dijkstra’s algorithm, to determine responder agents’ paths to and from the incident sites (once issued with their part of the plan).

When an agent enters an incident site, it stops being simulated in the tactical level simulator, and starts being simulated inside the appropriate operational level simulator. Its representation changes from a basic agent which can merely move along a transport network, to a more sophisticated agent which can perceive its environment, as shown in Figure 7, and select among a wide range of actions to perform. The actions for FireFighters have already been given in Figure 1. In a similar manner, eight actions for Paramedic agents and fourteen actions for Police agents have been identified from their National Occupational Standards and Major Incident Plans.

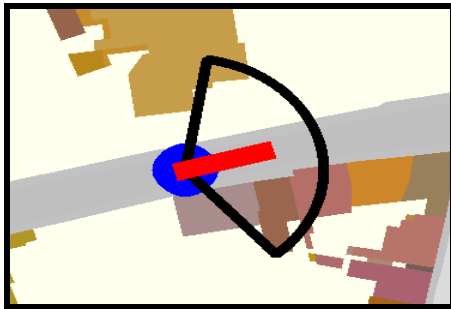
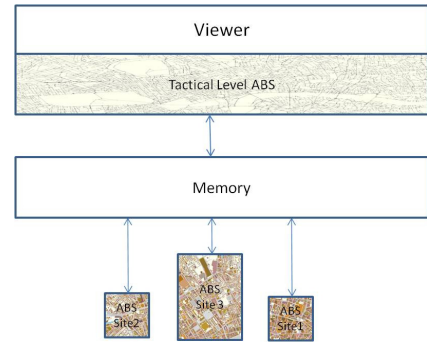
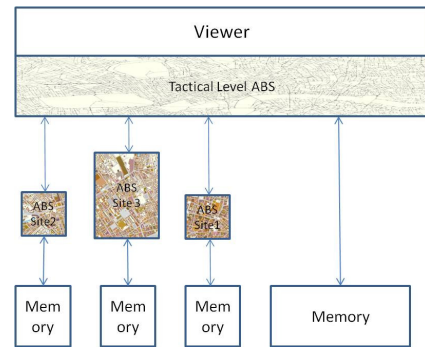


Figure 7: An agent perceiving its environment in the operational-level simulator

Finally, the tactical level simulator and operational level simulator(s) communicate with one another. The way they do this depends on whether the programs are running on the same machine or different machines, as shown in Figure 8. When on the same machine (Figure 8 (a)), they communicate using shared memory, using the QSharedMemory class from Qt (Qt 2011). When on different machines (Figure 8 (b)), they communicate using sockets, using the QTcpServer and QTcpSocket classes from Qt.



(a) Shared memory



(b) Distributed memory

Figure 8: Inter-process communication

4. SUMMARY AND FURTHER WORK

The high-level architecture of an agent-based simulation environment, designed specifically for emergency response has been described. In an effort to target the most appropriate level of model complexity, it uses two different representations for both the agents and their environment. Here we describe four further ongoing developments.

First, although agents have a rich repertoire of actions available in the operational level simulators, their action selection mechanism is still basic. Efforts are underway to model these mechanisms using naturalistic decision making (Klein 2008), in particular using a recognition-primed decision (RPD) model (Klein 2003, Warwick *et al* 2001). This has been shown to correspond well to the decision making of emergency first responders, such as firefighters (Burke and Hendry 1997, Klein *et al* 2010).

Second, parallel to the implementation of an RPD model of decision making for operational level agents, validation and verification will be carried out. Practitioners from local Emergency Planning Units, involved from the initial stages of the project, will provide face validation, whilst past case studies, such as the London 2005 bombings, will be used for retrodiction.

Third, a post-processor program will be developed to enable the analysis and understanding of simulation results.

Finally, the agent-based simulation environment is just one of two software components in the “REScUE” project (Coates *et al* 2011), being carried out at Durham University. The other component is a decision support system (DSS). The DSS has a two way communication with the tactical level simulator. It receives information about the emergency from tactical level agents as it becomes available, and uses this to generate plans for the responder agents. It then communicates these plans back to the tactical level agents, who may or may not decide to issue them to the operational level agents (who may or may not decide to adhere to the plan, depending on their most up-to-date knowledge of the emergency situation). It is a goal of the REScUE project to identify how to formulate near-optimal plans quickly, especially in the case of rapidly evolving, large-scale, unprecedented events where the practice of predetermined attendances and adhering to standard operating procedures may be far from optimal.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the funding provided by the UK’s EPSRC (EP/G057516/1).

Further, the authors thank practitioners from the Emergency Planning Units from Cleveland and Tyne & Wear, Co. Durham & Darlington Civil Contingencies Unit, Government Office for the North East, Fire and Rescue Services from Co. Durham & Darlington and Tyne & Wear, North East Ambulance Service, and Northumbria Police.

REFERENCES

- Akin, H. L., Yilmaz, O. and Sevim, M. M., 2010. RoboAKUT 2010 Rescue Simulation League Agent Team Description. Available from: <http://roborescue.sourceforge.net/2010/tdps/agents/roboakut.pdf>
- Anderson, J., 2007. *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press.
- Burke, E. and Hendry, C., 1997. Decision making on the London incident ground: an exploratory study, *Journal of Managerial Psychology*, 12 pp. 40-47.
- Campbell, L., Lotwin, A., DeRico, M. M. G. and Ray, C., 1997. The Use of Artificial Intelligence in Military Simulations, *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Vol. 3, pp. 2607-2612, October 12-15, Orlando, Florida.
- Carley, K. M., Fridsma, D. B., Casman, E., Yahja, A., Altman, N., Chen, L.-C., Kaminsky, B. and Nave, D., 2006. BioWar: scalable agent-based model of bioattacks. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans* 36 (2) pp. 252-265.
- Coates, G., Hawe, G.I., Wilson, D.T. and Crouch, R. S., 2011. Adaptive Co-ordinated Emergency Response to Rapidly Evolving Large-Scale Unprecedented Events (REScUE). *Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management – ISCRAM 2011*. May 8-11, Lisbon, Portugal.
- Chu, T.-Q., Boucher, A., Drougal, A., Vo, D.-A., Nguyen, H.-P. and Zucker, J.-D., 2008. Interactive Learning of Expert Criteria for Rescue Simulations, *Lecture Notes in Artificial Intelligence*, 5347 pp. 127-138.
- Del Valle, S., Kubicek, D., Mniszewski, S., Riese, J., Romero, P., Smith, J., Stroud, P. and Sydoriak, S., 2006. EpiSimS Los Angeles Case Study. Technical Report LAUR-06-0666, Los Alamos National Laboratory.
- Gilbert, N., 2006. When Does Social Simulation Need Cognitive Models? In: R. Sun, ed. *Cognition and Multi-Agent Interaction*, Cambridge University Press, pp. 428-432.
- Gilbert, N., 2007. Computational Social Science: Agent-based social simulation. In: D. Phan and F. Amblard, eds. *Agent-based modeling and Simulation*. Bardwell Press, Oxford, pp. 115-134.
- Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., Thulke, H.-H., Weiner, J., Wiegand, T. and DeAngelis, D. L., 2005. Pattern-Oriented Modeling of Agent-Based Complex Systems: Lessons from Ecology, *Science*, 310 (5750) pp. 987-991.
- Haddow, G. D., Bullock, J. A. and Coppola, D. P., 2010. *Introduction to Emergency Management*, Butterworth-Heinemann.
- Jain, S. and McLean, C., 2003. A Framework for Modeling and Simulation for Emergency Response, *Proceedings of the 2003 Winter Simulation Conference*, pp. 1068-1076. December 7-10, New Orleans, Louisiana, USA.
- Khalil, K.M., Abdel-Aziz, M. H., Nazmy, M. T. and Salem, A. M., 2009. The Role of Artificial Intelligence Technologies in Crisis Response. *Proceedings of the 14th International Conference on Soft Computing*, pp. 293-298. June 18-20, Brno University of Technology, Czech Republic.
- Klein, G., 1998. *Sources of Power: How People Make Decisions*. MIT Press.
- Klein, G., 2008. Naturalistic Decision Making, *Human Factors* 50 (3), pp. 456-460.
- Klein, G., Calderwood R. and Clinton-Cirocco, A., 2010. Rapid Decision Making on the Fire Ground: The Original Study Plus a Postscript, *Journal of Cognitive Engineering and Decision Making*, 4 pp. 186-209.
- Koza, J. R., 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press.
- Koto, T. and Takeuchi, I., 2003. A distributed disaster simulation system that integrates sub-simulators. *Proceedings of the First International Workshop on Synthetic Simulation and Robotics to Mitigate Earthquake Disaster*. July, Padova, Italy.
- Lehman, J. F., Laird, J. and Rosenbloom, P., 2006. A Gentle Introduction to Soar, an Architecture for Human Cognition: 2006 Update. Available from:

- <http://ai.eecs.umich.edu/soar/sitemaker/docs/misc/GentleIntroduction-2006.pdf>
- LESPL, 2007. *London Emergency Services Liason Panel Major Incident Procedure Manual*, The Stationery Office. Available from: http://www.leslp.gov.uk/docs/Major_incident_procedure_manual_7th_ed.pdf
- Longo, F., 2010. Emergency Simulation: State of the Art and Future Research Guidelines, *SCS M&S Magazine*, Vol. 1.
- Müller, J., 1999. The Right Agent (Architecture) to do the Right Thing. *Lecture Notes in Artificial Intelligence*, 1555, pp. 211-225.
- Mysore, V., Narzisi, G. and Mishra, B., 2006. Agent Modeling of a Sarin Attack in Manhattan. *Proceedings of the First International Workshop on Agent Technology for Disaster Management*, pp. 108-115. May 8-12, Hokkaido, Japan.
- Narzisi, G., Mincer, J., Simth, S. and Mishra, B., 2007. Resilience in the Face of Disaster: Accounting for Varying Disaster Magnitudes, Resource Topologies, and (Sub) Population Distributions in the PLAN C Emergency Planning Tool. *Lecture Notes in Computer Science* 4659, pp. 433-446.
- North, M. J. and Macal, C. M., 2007. *Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation*. Oxford University Press.
- OS MasterMap, 2011. Ordnance Survey MasterMap Available from: <http://www.ordnancesurvey.co.uk/oswebsite/products/os-mastermap/index.html>
- OS StreetView, 2011. Ordnance Survey StreetView Available from: <http://www.ordnancesurvey.co.uk/oswebsite/products/os-streetview/index.html>
- Qt, 2011. A Cross-Platform Application and UI Framework. Available from: <http://qt.nokia.com/>
- Peterson, M., 2009. *An Introduction to Decision Theory*, Cambridge University Press.
- Roberts, D., 2010. Distributed Agent Based Modeling, *Linux Journal*. Available from: <http://www.linuxjournal.com/content/distributed-agent-based-modeling>
- Runka, A., 2010. *Genetic Programming for the RoboCup Rescue Simulation System*. M.S. Thesis, Brock University, Ontario.
- Saoud, N. B-B., Mena, T. B., Dugdale, J., Pavard, B. and Ahmed, M. B., 2006. Assessing large scale emergency rescue plans: An agent-based approach, *The International journal of Intelligent Control Systems*, 11 (4) pp. 260-271.
- Sato, K. and Takahashi, T., 2011. A Study of Map Data Influence on Disaster and Rescue Simulation's Results In: Q. Bai and N. Fukuta, ed. *Advances in Practical Multi-Agent Systems*. Springer Berlin /Heidelberg, pp. 389-402.
- Siek, J. G., Lee, L-Q. and Lumsdaine, A., 2002. *The Boost Graph Library*, Addison Wesley.
- Skinner, C. And Ramchurn, S., 2010. The RoboCup Rescue Simulation Platform, *Proceedings of the 9th International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 1647-1648. May 10-14, Toronto, Canada.
- Sun, R., 2006. Prolegomena to Integrating Cognitive Modeling and Social Simulation, In: R. Sun, ed. *Cognition and Multi-Agent Interaction*, Cambridge University Press, pp. 3-26.
- U.K. Firefighter NOS, 2005. Skills for Justice - National Occupational Standards. Available from: <http://www.skillsforjustice-ipds.com/nos.php>
- Warwick, W., McIlwaine, S., Hutton, R. and McDermott, P., 2001. Developing computational models of recognition-primed decision making, *Proceedings of the 10th conference on computer generated forces*. May 15-17, Norfolk, VA, USA.

AUTHORS BIOGRAPHY

Glenn I. Hawe is a Research Associate in the School of Engineering and Computing Sciences at Durham University, where he is currently developing an agent-based simulation environment for large-scale emergency response. He has a PhD in Electronics and Electrical Engineering from the University of Southampton, and an MPhys in Mathematics and Physics from the University of Warwick.

Graham Coates is a Senior Lecturer in the School of Engineering and Computing Sciences at Durham University. He has a PhD in Computational Engineering Design Coordination from Newcastle University, and a B.Sc. in Mathematics from Northumbria University. Recently, his research interest in coordination has been extended into the area of emergency response, and an EPSRC (UK) grant has enabled a small team to be brought together to carry out a three year study.

Duncan T. Wilson is a PhD student in the School of Engineering and Computing Sciences at Durham University, where he is currently working on a decision support system for large-scale emergency response. He has a B.Sc in Mathematics and an M.Sc in Operational Research from the University of Edinburgh. Prior to Durham he worked for the U.K. Government Operational Research Service.

Roger S. Crouch is Head of the School of Engineering and Computing Sciences at Durham University. In 1994 he took up a lectureship (then senior lectureship) at Sheffield University where he set up the Computational Mechanics Group. In 2005 he moved to Durham. His research work has focused on four areas: (i) structural integrity of nuclear reactor vessels under elevated temperature and over-pressure, (ii) computational plasticity of geomaterials, (iii) wave slamming on breakwaters and (iv) simulating reactive processes in groundwater transport.

EIGENFREQUENCY BASED SENSITIVITY ANALYSIS OF VEHICLE DRIVETRAIN OSCILLATIONS

Oliver Manuel Krieg^(a), Jens Neumann^(b), Bernhard Hoess^(c), Heinz Ulbrich^(d)

^(a) BMW Group Research and Technology, Hanauer Str. 46, 80992 Munich, Germany

^(b) BMW AG, Hufelandstr. 4, 80788 Munich, Germany

^(c) BMW Group Research and Technology, Hanauer Str. 46, 80992 Munich, Germany

^(d) Institute of Applied Mechanics, Technical University of Munich, Boltzmannstraße 15, 85748 Garching, Germany

^(a)Oliver.M.Krieg@bmw.de, ^(b)Jens.JE.Neumann@bmw.de, ^(c)Bernhard.Hoess@bmw.de,
^(d)Ulbrich@amm.mw.tu-muenchen.de

ABSTRACT

For vehicle drivetrain design, there is a serious conflict of objectives between the oscillation phenomena shuffle and cyclic irregularities. The purpose of this paper is to illustrate a methodology to analyse and visualise the sensitivities of drivetrain eigenfrequencies in order to solve this conflict of objectives for the drivetrain design process for selected vehicle drivetrain concepts. Starting with a complex and detailed non-linear model, different simplifications are performed to finally visualise the sensitivities of relevant eigenfrequencies. This provides a profound understanding of the dynamic behaviour of the vehicle and enables engineers to identify parameter combinations that solve or palliate this conflict of objectives. Two exemplary measures are derived and the palliation effect is examined with the complex simulation models.

Keywords: sensitivity, vehicle transient behaviour, eigenfrequency, oscillations

1. INTRODUCTION

Due to the increasing CO₂ requirements of the vehicle fleet consumption, there is a considerable trend of vehicle manufacturers to optimize the vehicle fuel economy. Nevertheless, there is a serious conflict of objectives for the design of vehicle drivetrains due to the fact that vehicle dynamics and comfort aspects also have to be taken into account. Therefore, the vehicle performance needs to meet the requirements and perturbing oscillations must not exceed an admissible threshold.

In order to predict the behaviour of a future vehicle, complex non-linear models are used for the simulation of different components and settings. The pure implementation of a complex model however is not sufficient to gain a parameter setup that meets the requirements. The mere amount of parameters often prevents a straight-forward approach by simply adjusting one parameter after the other to approximate a good setup.

Understanding and visualising the relevant sensitivities can seriously improve the design process since it helps identifying parameters that approximate good vehicle setups, gives profound understanding of the dynamics of the system and facilitates solutions for the described conflicts of objectives. Furthermore, no optimisation algorithms are needed, which usually require distinct optimisation criteria and boundaries that often do not exist explicitly, and will not necessarily lead to an improved understanding of the dynamics of the system. A methodology is presented in this paper performing different simplifications, analysing eigenfrequency sensitivities based on (Dresig and Holzweißig 2010) and deriving measures to palliate the conflict of objectives for shuffle and cyclic irregularity for low frequencies for the vehicle drivetrain design process.

Numerous works are concerned with the discussed oscillation phenomena of a vehicle drivetrain. A complete overview would go beyond the scope of this paper. Therefore only selected works are presented here. In (Bencker 1998), experimental and simulative studies on shuffle are performed to identify palliative measures for the drivetrain. An analysis of a different engine torque excitation for shuffle follows in (Hülsmann 2007). Various works are concerned with active control of shuffle, e.g. (Best 1998), (Lefebvre, Chevrel, and Richard 2003), (Richard, Chevrel, and Maillard 1999). A holistic analysis of driveline oscillations due to cyclic irregularity is presented in (Gosdin 1985). Here, a parameter optimisation is achieved for the vehicle driveline for predefined boundaries. Various works are concerned with mechanical, semi-active or active components reducing cyclic irregularity, e.g. (Reik, Fidlin, and Seebacher 2009). New components for palliation as well as active control algorithm for shuffle control can both profit from the presented methodology. For the former, the discussed conflict of objectives is still present and any palliation helps the effectiveness of an additional

drivetrain component. For the latter, improved shuffle behaviour reduces the control requirements.

Starting point is a detailed model based on the physical behaviour of the different elements of the drivetrain. This model including measurement comparisons is presented in Section 2. The desired vehicle behaviour and existing conflicts of objectives are described in Section 3. In Section 4, different model simplifications are derived and the eigenfrequency sensitivity analysis is performed. The results of the presented methodology are then used to identify palliative measures for the behaviour of a three cylinder drivetrain, which is derived from the six cylinder drivetrain from Section 2. Finally, the results are summarised in Section 5.

2. DRIVETRAIN MODELLING

The examined prototype vehicle is a vehicle fitted with a six cylinder turbocharged engine. First, the engine model is presented, followed by the mechanical drivetrain model. Finally a measurement comparison is illustrated in this section.

2.1. Thermodynamic Engine Implementation

The cylinder volume is the core element of the engine model. Here, the combustion takes place and the mechanical work is transferred to the crank drive. It is basically a homogeneous volume following the first law of thermodynamics applied to open systems (Müller and Müller 2005):

$$\dot{U} = \dot{Q} + \dot{E} + \dot{W} \quad (1)$$

Here, \dot{U} represents the first deriviate of the internal energy with respect to time, \dot{Q} the heat flow and \dot{W} the mechanical power done on the system. \dot{E} represents the inner energy flow of matter entering and leaving the system.

The conservation of mass and the caloric theory corresponding to the cylinder gases must also be taken into account for the model. Additional components as valves, a combustion model, heat transfer elements or the crank drive are also required to model the physical behaviour of the engine. A possible implementation of these elements is described in (Krieg, Förg, and Ulbrich 2011).

Deviant from the filling and emptying approach presented in (Krieg, Förg, and Ulbrich 2011), the fluid oscillations of the intake and exhaust manifold are also considered here. Intake and exhaust manifold are modelled as sequence of pipes, which are discretised homogenous volumes. Incorporating detailed intake and exhaust manifold models can increase the quality of the simulation results for certain operating conditions but leads to more complex models and longer simulation duration. The implementation of these pipes follows the conservation of mass and energy and, in addition to the

filling and emptying approach, also the conservation of linear momentum:

$$\frac{\partial}{\partial t}(\rho A \omega) + \frac{\partial}{\partial x}(\rho A \omega^2) + \frac{\partial}{\partial x}(pA) = -\rho A k_f \quad (2)$$

Here, ρ represents fluid density, A the pipe cross section area, ω the fluid velocity and p the fluid pressure. Furthermore, t represents time, x position and k_f is a coefficient for friction. A solution for the partial differential equation requires a discretisation method, e.g. the finite volume method (Dick 2009). A possible implementation of the pipes is presented in (Miersch 2003). Components implemented as characteristic maps and not physically are the turbine and the compressor of the turbocharger. Here, the mass flow is estimated according to the turbocharger shaft speed and the fluid pressure of the incoming and outgoing pipes of these components, as derived from measured data. The implementation of turbochargers is described e.g. in (Baines and Fredriksson 2007).

2.2. Mechanical Drivetrain Implementation

For the engine crank drive model, an analytical and a multibody approach are possible implementations. A deduction of an analytical implementation according to the projective Newton-Euler equations and an evaluation of both implementation approaches is presented in (Krieg, Förg, and Ulbrich 2011).

All shafts of the drivetrain are modelled as rotational inertias and springs with small damping. The mechanical model of the drivetrain follows from Figure 1. Here, j_i represents the inertias of the drivetrain, c_i the stiffnesses and φ_i the degrees of freedom. The parameters u_3 and u_6 represent the gear ratio and the final drive ratio. T_{eng} represents the engine torque, which is applied to the crankshaft j_2 .

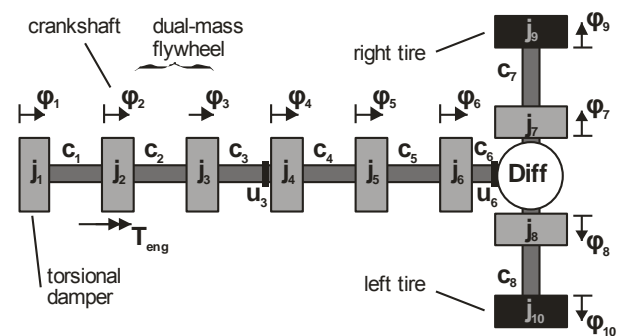


Figure 1: Mechanical Drivetrain System

Tires are implemented as elements that calculate the force between wheel and street according to the differential speed of both. Data for tires are usually measured with a dedicated tire rig and implemented via a curve fitting algorithm. The longitudinal tire force follows according to (Pacejka and Bakker 1992):

$$F_{tire} = s \cdot \beta(F_{load}^{tire}) \quad (3)$$

$$s = \frac{r_{tire} \cdot \omega_{tire} - v_{vehicle}}{r_{tire} \cdot \omega_{tire}} \quad (4)$$

Here, r_{tire} represents the tire radius, ω_{tire} the tire rotational speed and $v_{vehicle}$ the vehicle speed. The tire load F_{load}^{tire} of the rear tire is calculated according to the balance of momentum of the accelerated vehicle, as illustrated in Figure 2. The variable $F_{gravity}$ represents the mass force due to gravity, h represents the height of the centre of mass of the vehicle and l_f and l_r the horizontal distance between front and rear wheel contact point to the centre of mass of the vehicle. The balance of momentum at the front wheel contact point calculates the rear axle load F_{load}^{rear} and thus the tire load according to:

$$F_{load}^{rear} = \frac{F_{tire} h + F_{gravity} l_f}{l_f + l_r} = 2 F_{load}^{tire} \quad (5)$$

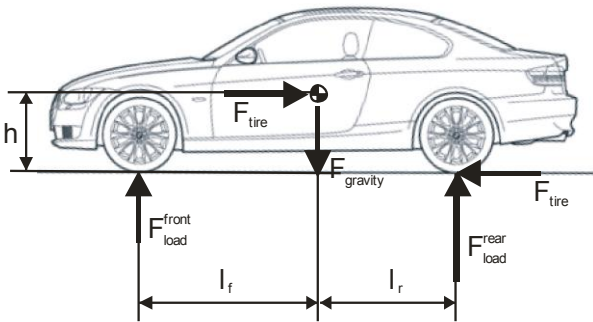


Figure 2: Vehicle Balance of Momentum

Tire force is a quantity proportional to differential speed s , so it can also be seen as non-linear damper with high damping coefficient. The function β represents the curve fitting algorithm, as shown in (Pacejka and Bakker 1992). The tire models illustrated in Figure 1 also contain an inertia that represents the rims.

The equations for the differential follow according to:

$$T_{in} = \frac{T_{out}^{right}}{2} = \frac{T_{out}^{left}}{2} \quad (6)$$

$$\omega_{in} = \frac{\omega_{out}^{right} + \omega_{out}^{left}}{2} \quad (7)$$

Here, T_{in} represents the torque of the input shaft and T_{out}^{right} and T_{out}^{left} of the right and left output shaft. Furthermore, ω_{in} represents the rotational speed of the

input shaft and ω_{out}^{right} and ω_{out}^{left} of the right and left output shaft of the differential. The parameters for all shafts and additional compliant elements are measured on component rigs. All inertias are corrected according to the gear ratio for the equivalent degree of freedom.

2.3. Measurement Comparison

In order to verify the correctness of the drivetrain model, an accurate measurement comparison is required. Therefore, a Tip-In manoeuvre was measured, i.e. the acceleration pedal of a prototype vehicle with constant speed is quickly acted from a defined part load throttle to full throttle. The first gear is engaged here.

The measurement results and the corresponding simulations for the vehicle are illustrated in the following figures. Figure 3 illustrates the intake manifold pressure, which is derived from a dedicated pressure sensor. For the simulated and measured manoeuvre, the acceleration pedal is acted at $t = 0.5$ s. The figure shows that there is good consistence for the intake manifold pressure between measurement and simulation.

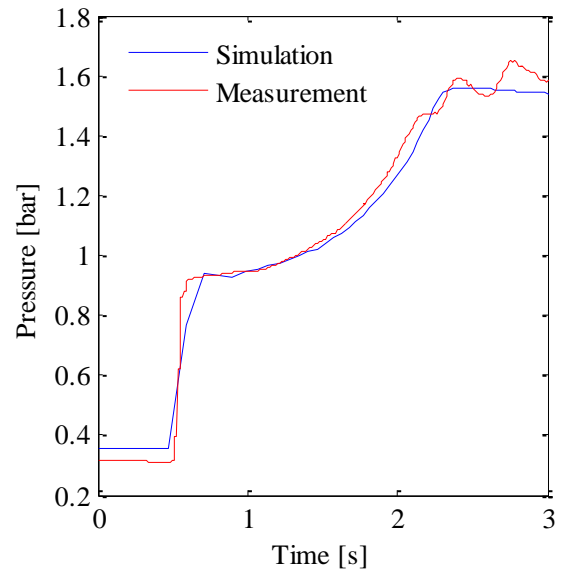


Figure 3: Measurement Comparison of the Intake Manifold Pressure

Figure 4 illustrates the engine torque. The measured engine torque is actually derived from an engine model on the engine control unit (ECU), which estimates the current torque according to diverse measured data, e.g. crankshaft speed or the intake manifold pressure. Apparently, there is also good consistence between measured and simulated engine torque. The first rise of the engine torque is very steep and results in an excitation that is similar to a torque step function. This occurs because of the rapid filling of the intake manifold and the cylinders after increasing the throttle diameter and the subsequent conversion into mechanical work by combusting a larger mass of air

and fuel. The following rise of the engine torque is a consequence of the turbocharger, which has a certain time delay because of its inertia.

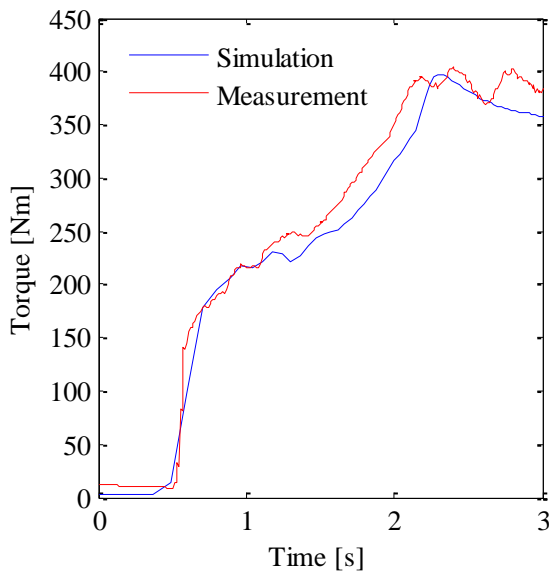


Figure 4: Measurement Comparison of the Engine Torque

A measure comparison for the mechanical drivetrain is also required, as follows in Figure 5. Here, the drivetrain model is acted by the measured engine torque. The longitudinal acceleration of the vehicle is illustrated. The drivetrain model shows good consistence between measurements and simulation. Note that several control algorithms for the damping of oscillations within the ECU are not considered here and were switched off for the measurements to avoid masking the drivetrain behaviour with interference of these algorithm interactions for examined frequencies.

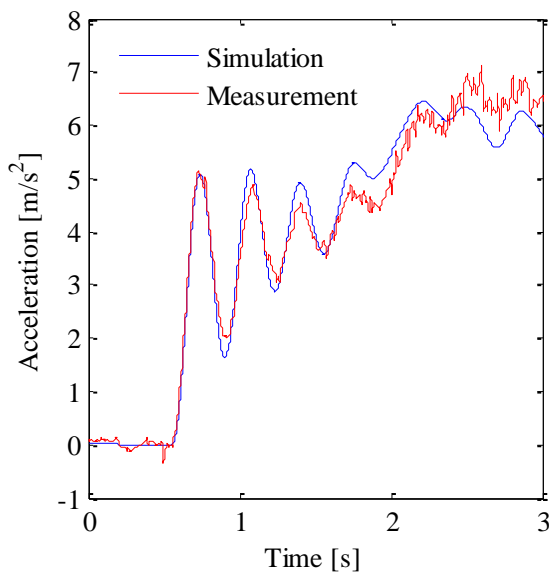


Figure 5: Measurement Comparison of the Vehicle Acceleration

3. DESIRED VEHICLE BEHAVIOUR

Figure 5 illustrates the vehicle longitudinal acceleration for a Tip-In manoeuvre. The excitation of the drivetrain via the engine similar to a step function results in a dominant stimulus of the first eigenfrequency with approximately 2.5 Hz for the vehicle longitudinal acceleration.

In fact, also higher eigenfrequencies of the drivetrain are stimulated by the step excitation. Nevertheless, for higher frequencies less energy is brought into the system for a step function excitation, higher frequencies oscillate with smaller magnitudes and they are quickly damped due to their higher rotational velocity. These effects explain the dominance of the first eigenfrequency for the vehicle longitudinal acceleration.

The desired vehicle behaviour is described from a driver perspective. For the discussed Tip-In manoeuvre, a driver experiences oscillation of the longitudinal vehicle acceleration, also referred to as shuffle, buckling or Bonanza effect, as perturbing. Particularly the height of the oscillation amplitudes is considerable to the driver. The frequency of this oscillation on the other hand is of less importance. Shuffle oscillations are usually between 2-5 Hz and in fact it is difficult for the driver to resolve a difference between these low frequencies. A quick decline of the oscillation is of higher importance. Additionally, the driver appreciates a steep rise in the vehicle acceleration curve.

For a higher frequency of the first drivetrain eigenfrequency, the oscillation amplitude is lower, the damping effect is stronger and the acceleration curve is steeper. As a result, a driver prefers higher frequencies for the first eigenfrequency of the vehicle drivetrain.

From this point of view, a consequence for the vehicle drivetrain design could be to choose e.g. shafts with high stiffness in order to move the first eigenfrequency to higher frequencies. This perspective however is too narrow for a drivetrain design. Driving a vehicle in other use cases also results in exciting higher drivetrain frequencies. In particular for drivetrains with three or even less cylinder engines, the second eigenfrequency is excited by the cyclic irregularity of the engine during stationary operation for low engine speeds $n < 1,500$ rpm in 5th or 6th gear. For higher engine speed or engines with more cylinders, the cyclic irregularity of the engine is also a problem, but the oscillations are transferred to the driver with smaller amplitudes. For these oscillation phenomena, other effects than the eigenfrequency affect the comfort perception of the driver as well. These aspects are not examined here. The focus of this paper is shuffle and cyclic irregularity due to excitation of drivetrain eigenfrequencies.

Even though shuffle and cyclic irregularity are examined with different gears, they illustrate a serious conflict of objectives for the vehicle drivetrain design. A simple optimization of shuffle, e.g. choosing stiffer shafts, would result in a higher frequency of the first and second eigenfrequency of the drivetrain. A higher

second eigenfrequency however is then closer or identical to the cyclic irregularity of the engine for certain vehicle speeds. Vice versa, a simple optimization of the cyclic irregularity behaviour could worsen shuffle. An efficient drivetrain design therefore must take both effects into account and resolve this conflict of objectives.

4. SENSITIVITY ANALYSIS

The eigenfrequency methodology is presented in this section. First, simplifications of the drivetrain model are performed, followed by the application of the sensitivity analysis based on (Dresig and Holzweißig 2010). Subsequently, palliative measures are derived and examined with the detailed model.

4.1. Simplifications

In order to examine the eigenfrequencies, simplifications are performed here. The thermodynamic engine model is of less interest since it essentially defines the excitation, the interaction is negligible. The crank drive is simplified as part of the crankshaft inertia. An additional simplification is concerned with the tire inertia and the vehicle mass. In relation to the drivetrain inertia, the vehicle mass is very high. Therefore the error caused by attaching the vehicle mass to the inertial frame is small. Furthermore, the tires are neglected and the rims are also fixed to the inertial frame. Due to the fact that there is generally only low damping for the drivetrain, damping effects are neglected for the sensitivity analysis. These simplifications actually influence the eigenfrequencies of the drivetrain and therefore a final comparison of the simplified model with the detailed drivetrain model is required.

The simplified drivetrain model is a one dimensional chain of inertias and springs:

$$\mathbf{M} \ddot{x} + \mathbf{C} x = 0 \quad (8)$$

Here, x represents the vector for the rotational positions and \ddot{x} the vector for the rotational accelerations of the inertias. The matrix of stiffness \mathbf{C} follows according to:

$$\mathbf{C} = \begin{bmatrix} c_1 & -c_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -c_1 & c_1 + c_2 & -c_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -c_2 & c_2 + c_3 & -u_3 c_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -u_3 c_3 & u_3^2 c_3 + c_4 & -c_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -c_4 & c_4 + c_5 & -c_5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -c_5 & c_5 + c_6 & -\frac{u_6}{2} c_6 & -\frac{u_6}{2} c_6 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{u_6}{2} c_6 & \frac{u_6^2}{4} c_6 + c_7 & \frac{u_6^2}{4} c_6 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\frac{u_6}{2} c_6 & \frac{u_6^2}{4} c_6 & \frac{u_6^2}{4} c_6 + c_8 & 0 \end{bmatrix} \quad (9)$$

The matrix of inertia \mathbf{M} is diagonal according to:

$$\mathbf{M} = \text{diag}(j_1, j_2, j_3, j_4, j_5, j_6, j_7, j_8) \quad (10)$$

There are only small deviations between the two models of first and fifth gear, e.g. gearing stiffness c_3 or gearing ratio u_3 . The simplified model is illustrated in Figure 6.

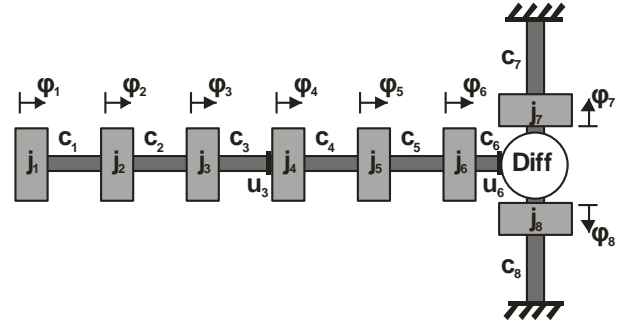


Figure 6: Reduced Mechanical Drivetrain Model

4.2. Sensitivity Algorithm

The presented sensitivity algorithm is based on (Dresig and Holzweißig 2010). A modal decomposition of the differential equation system of Equation (8) returns the modal inertia μ_i and stiffness γ_i as follows (Ulbrich 1996):

$$\gamma_i = v_i^T \cdot \mathbf{C} \cdot v_i \quad (11)$$

$$\mu_i = v_i^T \cdot \mathbf{M} \cdot v_i \quad (12)$$

Here, v_i represents the eigenvector of the eigenfrequency ω_i . The eigenfrequency ω_i is then calculated according to:

$$\omega_i^2 = \frac{\gamma_i}{\mu_i} = \frac{v_i^T \cdot \mathbf{C} \cdot v_i}{v_i^T \cdot \mathbf{M} \cdot v_i} \quad (13)$$

Small variations of the inertias $\Delta \mathbf{M}$ or the stiffnesses $\Delta \mathbf{C}$ lead to small variations of the eigenfrequency $\Delta \omega_i$, and the change of the eigenvector is negligible $\Delta v_i \approx 0$:

$$\tilde{\omega}_i^2 = \omega_i^2 + \Delta \omega_i^2 = \frac{v_i^T \cdot (\mathbf{C} + \Delta \mathbf{C}) \cdot v_i}{v_i^T \cdot (\mathbf{M} + \Delta \mathbf{M}) \cdot v_i} \quad (14)$$

Assuming that variations of the inertias are much smaller than the inertias themselves $\Delta \mathbf{M} \ll \mathbf{M}$, and $\Delta \omega_i$ can be estimated after some transposition according to (Dresig and Holzweißig 2010):

$$\Delta \omega_i^2 \approx \omega_i^2 \left(\frac{v_i^T \cdot \Delta \mathbf{C} \cdot v_i}{v_i^T \cdot \mathbf{C} \cdot v_i} - \frac{v_i^T \cdot \Delta \mathbf{M} \cdot v_i}{v_i^T \cdot \mathbf{M} \cdot v_i} \right) \quad (15)$$

The variation matrices for stiffness $\Delta\mathbf{C}$ and inertias $\Delta\mathbf{M}$ are a sum of the variations of all inertias Δj_l and stiffnesses Δc_k :

$$\Delta\mathbf{M} = \sum_l \mathbf{M}_{l0} \cdot \frac{\Delta j_l}{j_l} \quad (16)$$

$$\Delta\mathbf{C} = \sum_k \mathbf{C}_{k0} \cdot \frac{\Delta c_k}{c_k} \quad (17)$$

The matrices \mathbf{C}_{k0} and \mathbf{M}_{l0} represent the introduced matrices for stiffness and inertias with all elements equal to zero, except element k or l respectively, e.g. for the inertias:

$$\mathbf{M}_{l0} = \text{diag}(0, \dots, 0, j_l, 0, \dots, 0) \quad (18)$$

This finally leads to the following sensitivity coefficients for stiffness φ_{ki} and inertia κ_{li} for eigenfrequency ω_i of element k or l respectively:

$$\Delta\omega_i^2 \approx \omega_i^2 \left(\sum_k \varphi_{ki} \cdot \frac{\Delta c_k}{c_k} - \sum_l \kappa_{li} \cdot \frac{\Delta j_l}{j_l} \right) \quad (19)$$

$$\varphi_{ki} = \frac{\mathbf{v}_i^T \cdot \mathbf{C}_{k0} \cdot \mathbf{v}_i}{\mathbf{v}_i^T \cdot \mathbf{C} \cdot \mathbf{v}_i} \quad (20)$$

$$\kappa_{li} = \frac{\mathbf{v}_i^T \cdot \mathbf{M}_{l0} \cdot \mathbf{v}_i}{\mathbf{v}_i^T \cdot \mathbf{M} \cdot \mathbf{v}_i} \quad (21)$$

These coefficients describe the variation of the eigenfrequency for a small relative parameter variation. Thus they are regarded as sensitivity coefficients for that eigenfrequency.

4.3. Drivetrain Eigenfrequency Analysis

For the presented drivetrain models for the first eigenfrequency of the first gear and the second eigenfrequency of the fifth gear, the sensitivity coefficients are illustrated in the following figures.

In order to solve the discussed conflict of objectives, the task now is to move the first eigenfrequency of the first gear to higher frequencies and vice versa move the second eigenfrequency of the fifth gear to lower frequencies. For a spring and mass system, an eigenfrequency is moved to higher frequencies by decreasing inertias or increasing stiffnesses, compare Equation (19). Figure 7 illustrates the sensitivity coefficients of the inertias and Figure 8 those of the stiffnesses for shuffle. The shuffle eigenfrequency is $\omega_1 = 2.5$ Hz.

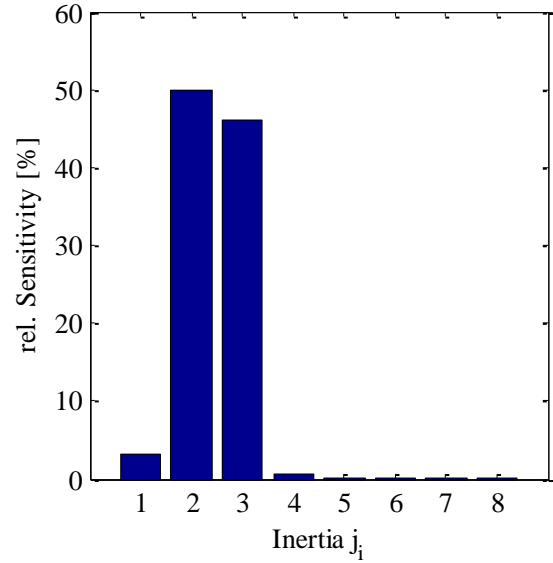


Figure 7: Sensitivity of Inertias for Shuffle

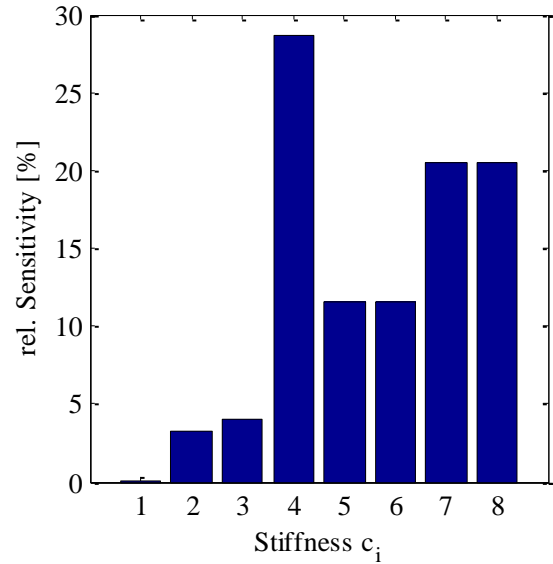


Figure 8: Sensitivity of Stiffnesses for Shuffle

For the shuffle eigenfrequency, there is a high sensitivity for the crankshaft inertia and the inertia of the dual-mass flywheel j_2 and j_3 . Furthermore, there is a high sensitivity for the rubber joint and the sideshaft stiffness c_4 , c_7 and c_8 .

The same sensitivity analysis was performed for the relevant eigenfrequency for cyclic irregularity $\omega_2 = 17.1$ Hz, as illustrated in Figure 9 for the inertias and Figure 10 for the stiffnesses.

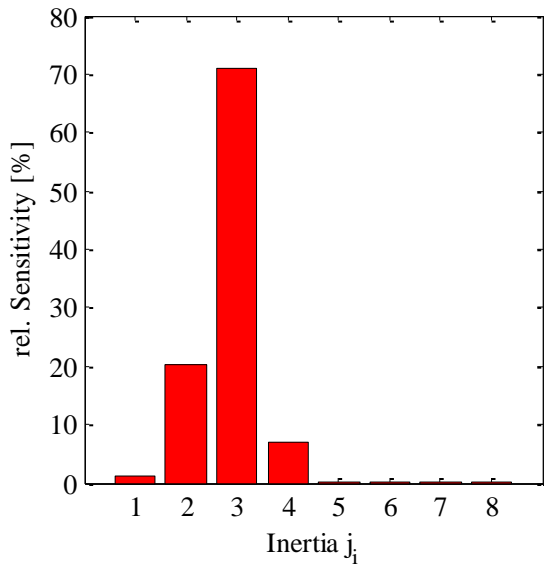


Figure 9: Sensitivity of Inertias for Cyclic Irregularity

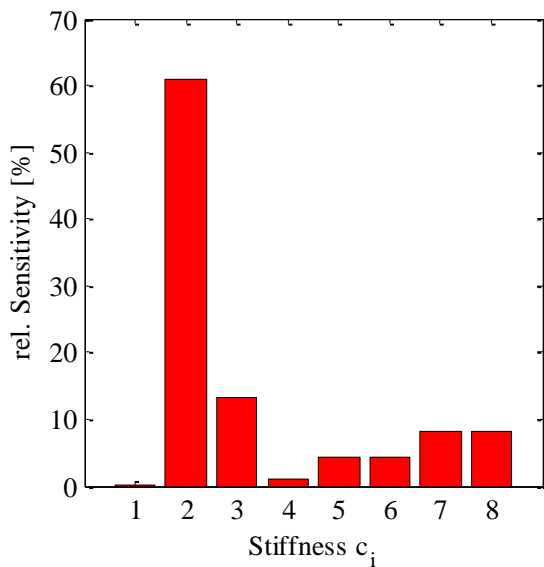


Figure 10: Sensitivity of Stiffnesses for Cyclic Irregularity

Figure 9 illustrates that there is a high sensitivity of the eigenfrequency for cyclic irregularity for the secondary mass of the dual-mass flywheel j_3 . Furthermore, the stiffness of the dual-mass flywheel c_2 also shows high sensitivity for that eigenfrequency.

With the presented diagrams, it is now possible to identify combinations that raise the first eigenfrequency and decrease the second eigenfrequency as well. In order to derive a parameter combination as palliation measure, a first parameter is selected from the illustrations, which has a high sensitivity for shuffle and a low sensitivity for cyclic irregularity. Additionally, a second parameter is chosen with low sensitivity for shuffle and high sensitivity for cyclic irregularity.

4.4. Modifications of the Mechanical Drivetrain

Two different combinations will be discussed in the following section in order to illustrate the methodology

principle. From Figure 8 and Figure 10 it is apparent that the stiffness c_7 , i.e. the stiffness of the right sideshaft, has a major influence on the shuffle eigenfrequency and a minor influence on the eigenfrequency of cyclic irregularity. On the other hand, stiffness c_2 , i.e. the stiffness of the dual-mass flywheel, has a minor influence on the shuffle eigenfrequency and a major influence on the eigenfrequency of cyclic irregularity. A promising combination to solve the conflict of objectives could now be to increase stiffness c_7 and decrease stiffness c_2 . In order to contain the symmetry of the drivetrain, stiffness c_8 , i.e. the stiffness of the left sideshaft, also needs to be modified according to stiffness c_7 . Here, the stiffnesses of the driveshafts are doubled and the stiffness of the dual-mass flywheel decreased to half of the original value. This results in modification 1:

$$c_2^{\text{mod1}} = 0.5c_2 \quad (22a)$$

$$c_7^{\text{mod1}} = 2c_7 \quad (22b)$$

$$c_8^{\text{mod1}} = c_7^{\text{mod1}} \quad (22c)$$

Additionally, another modification is examined to ensure that modification 1 is not a coincidence. The new parameter setup is referred to as modification 2:

$$c_4^{\text{mod2}} = 2c_4 \quad (23a)$$

$$c_7^{\text{mod2}} = 0.8c_7 \quad (23b)$$

$$c_8^{\text{mod2}} = c_7^{\text{mod2}} \quad (23c)$$

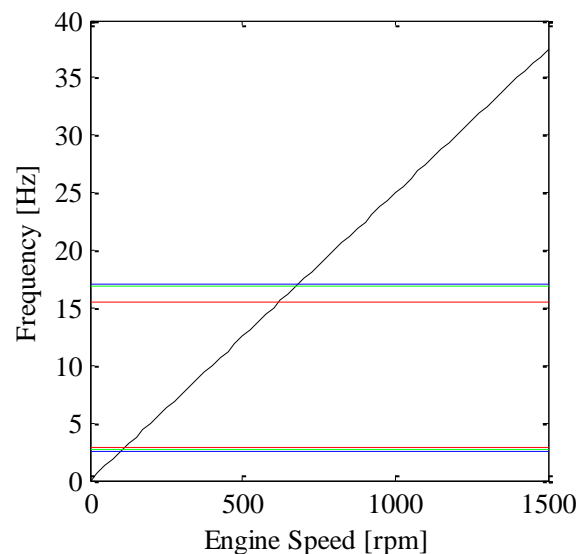
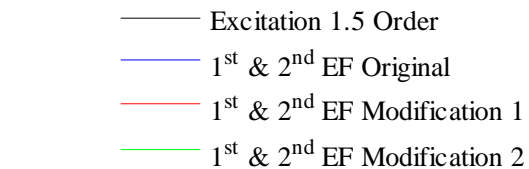


Figure 11: Campbell Diagram

The values for the parameter variation are derived so that the effect of one measure does not overcompensate the effect of the other. The Campbell diagram in Figure 11 shows the result of the drivetrain eigenfrequencies (EF) for both modifications.

The shuffle eigenfrequency of modification 1 is $\omega_1^{\text{mod1}} = 2.8 \text{ Hz}$, the eigenfrequency for cyclic irregularity is $\omega_2^{\text{mod1}} = 15.5 \text{ Hz}$. Furthermore, the shuffle eigenfrequency of modification 2 is $\omega_1^{\text{mod2}} = 2.6 \text{ Hz}$, the eigenfrequency for cyclic irregularity is $\omega_2^{\text{mod2}} = 16.9 \text{ Hz}$. Obviously, the aimed task to move both eigenfrequencies in the desired directions worked for both modifications. A weakness of the described methodology is the difficulty in predicting the precise frequency of the eigenfrequency of the new setup. Hence, the methodology helps to identify, which parameters should be modified in which direction, but not how high the parameter variation should be.

4.5. Simulative Measure Verification

Due to the performed simplifications it is finally required that the results also persist for the detailed model. The eigenfrequency of cyclic irregularity as examined here for low frequencies is a problem for the vehicle drivetrain because it is close to the excitation of the engine. For a three cylinder engine, it is particularly difficult since it excites the drivetrain with the 1.5 order. The excitation of a three cylinder engine is also illustrated in Figure 11. A detailed three cylinder model is derived in this work from the presented six cylinder model of section 2. Therefore, three cylinders are removed and additional modifications for the exhaust and the intake manifold are performed to generate the three cylinder model. In particular, parts of the exhaust and intake manifold system with two parallel paths of the six cylinder model are removed. Those parts with one common path are physically divided in half, e.g. the throttle cross section or the intake manifold volume. This model is used to obtain a realistic engine excitation.

Figure 12 illustrates the engine torque of the model for a Tip-In manoeuvre. This diagram shows that the three cylinder engine has a comparable steepness of the first torque raise which is followed by an engine torque which is approximately half as high as the six cylinder engine.

In Figure 13 the acceleration of the vehicle model is illustrated. Here, the original drivetrain model is excited with the three cylinder torque of Figure 12. Furthermore, the drivetrain model modification 1 is also excited by this engine torque. First remark is that due to the lower torque the absolute values of the acceleration and its oscillations are lower than those of the drivetrain of Section 2. Nevertheless, since the oscillations are observed in relation to the mean acceleration, the oscillations are still a problem.

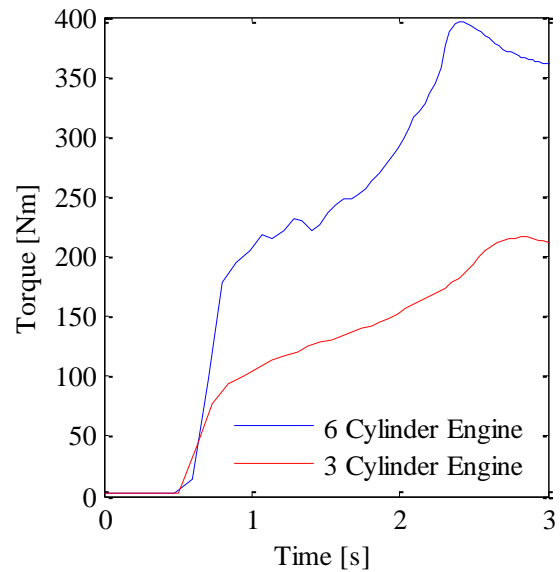


Figure 12: Engine Torque for Original and 3 Cylinder Engine

Additionally, Figure 13 shows that the aimed increase of the shuffle eigenfrequency worked for modification 1. Consequently, also the amplitude of the oscillations is reduced and the suggested measure worked for the simulation of shuffle as expected.

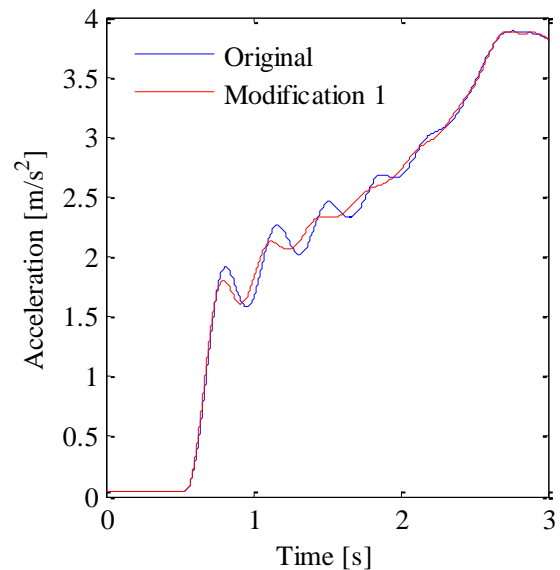


Figure 13: Vehicle Acceleration with Modification 1

Next step is the simulation of cyclic irregularity. Here, the three cylinder engine is again used to generate the excitation. For cyclic irregularity, a constant excitation for a dedicated engine speed is required. According to Figure 11, a low engine speed results in an excitation close to the eigenfrequency. For that reason $n = 1,000 \text{ rpm}$ is used as excitation. Figure 14 illustrates the torque acting on a distinct drivetrain shaft, which is relevant for cyclic irregularity, for the original drivetrain and the drivetrain modification 1 in order to evaluate oscillations due to cyclic irregularity.

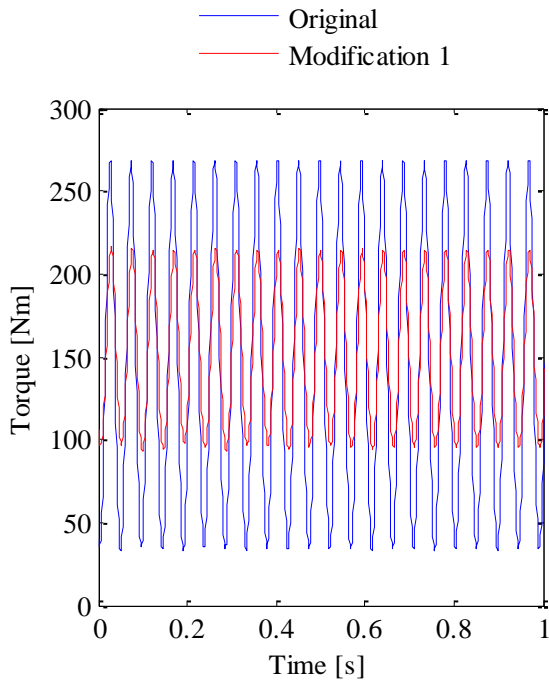


Figure 14: Cyclic Irregularity for $n = 1,000$ rpm, Modification 1

The illustration shows that the amplitude of the torque oscillation of the modified drivetrain decreased from 70 Nm to 30 Nm. Thus, the exemplary system modification worked in the desired way and helped to palliate the conflict of objectives between shuffle and cyclic irregularity.

For modification 2, the drivetrain was excited in the same way as described for modification 1 above. Figure 15 and Figure 16 illustrate the simulation results for drivetrain modification 2.

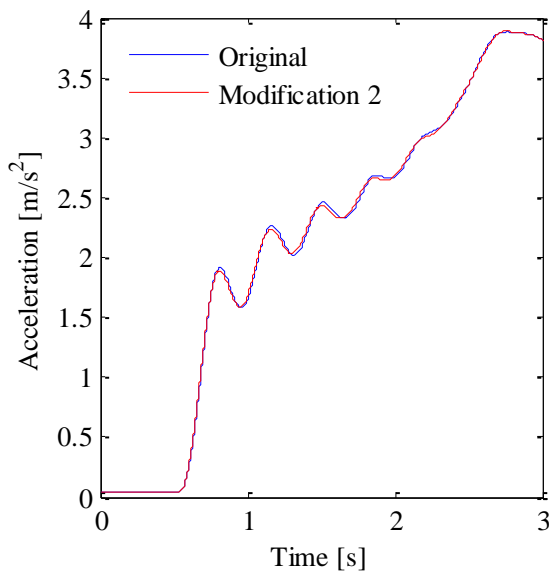


Figure 15: Vehicle Acceleration with Modification 2

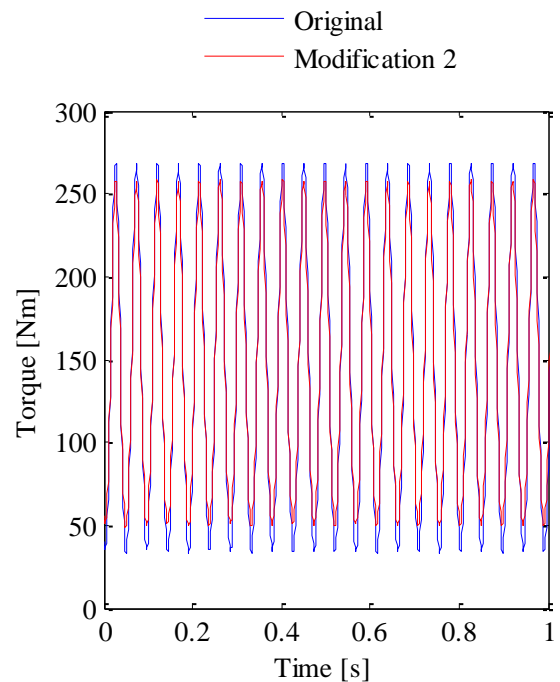


Figure 16: Cyclic Irregularity for $n = 1,000$ rpm, Modification 2

The vehicle acceleration in Figure 15 shows that the effect of modification 2 for shuffle is marginal. The reason for that is the minimal change of $\omega_1^{\text{mod}2}$. The decrease of the first stiffness almost completely compensates the increase of the second stiffness. Improved behaviour is observable for cyclic irregularity in Figure 16. The effect of modification 2 is smaller than that of modification 1.

5. CONCLUSION

For vehicle drivetrain design, oscillations represent a major aspect in order to gain a setup that meets the comfort requirements of the driver. Shuffle and cyclic irregularities due to the engine excitation cause a serious conflict of objectives. A methodology is presented that can help to solve this conflict of objectives for selected drivetrain setups, based on an understanding of the drivetrain eigenfrequency sensitivities.

First, a detailed model of the drivetrain is presented in Section 2, containing a complex engine model for the vehicle excitation and a mechanical drivetrain model including tires. A measurement comparison for the engine torque, intake manifold pressure and the vehicle acceleration is presented to verify the correctness of the models. In Section 3, the desired vehicle behaviour and the conflict of objectives for the palliation of shuffle and cyclic irregularity is described. The methodology is presented in Section 4, starting with a simplification of the drivetrain model and an eigenfrequency sensitivity analysis based on (Dresig and Holzweißig 2010). The following visualisation facilitates the identification of palliation measures for the described conflict of objectives. Two exemplary palliation measures are

derived from these illustrations. In order to show the conflict of objectives, a three cylinder engine is derived from the examined models, which particularly suffers from the described conflict, since the excitation frequency is close to the second drivetrain eigenfrequency for low engine speed. The setup of the mechanical drivetrain is not changed any further to illustrate the methodology principle. Simulations for the detailed models show that the suggested, exemplary measures can palliate the described conflict of objectives for the observed drivetrain.

The example shows improved vehicle behaviour for the performed time-based simulations. Nevertheless, a final measurement comparison is required. It must eventually be clarified for other conditions of use that the suggested measures will not worsen the vehicle behaviour for those use cases. This can also include oscillations of cyclic irregularity for higher frequencies since other effects affect the driver perception here as well.

It shall also be remarked that the described method is not the usual method to derive a drivetrain setup. In particular, the vehicle mass of the three cylinder engine remained identical to the six cylinder engine. Also the gear and final drive ratios remained the same. For a drivetrain design in the industry, the gear and final drive ratio, the vehicle mass and additional parameters are adjusted to achieve a coherent vehicle setup. In this paper, these adoptions are neglected in order to demonstrate the methodology principle. It is the dedicated objective of the presented method to palliate the conflict of objectives between shuffle and cyclic irregularity and not to derive a drivetrain setup for new vehicle concepts.

Despite to an optimisation algorithm, the comparison of the sensitivities for the different use cases provides a profound understanding of the dynamic vehicle drivetrain behaviour. Furthermore, optimisation algorithms require distinct optimisation criteria and boundaries that often do not exist explicitly. In particular, for the heterogeneous design process of a vehicle with different responsibilities spread around research and development departments, a methodology providing a profound understanding of the dynamic behaviour is superior compared to a singular optimum.

REFERENCES

Baines, N., Fredriksson, C., 2007. The Simulation of Turbocharger Performance for Engine Matching. In: Pucher, ed., H., Kahrstedt, J., ed. *Motorprozesssimulation und Aufladung 2: Engine Process Simulation and Supercharging*. Expert, Berlin: pp. 101-111.

Bencker, R., 1998. *Simulationstechnische und experimentelle Untersuchung von Lastwechselphänomenen an Fahrzeugen mit Standardantrieb*. Munich: Hieronymus.

Best, M.C., 1998. Nonlinear optimal control of vehicle driveline vibrations. *UKACC International Conference on Control '98 (Conf. Publ. No. 455)*.

Dick, E., 2009. Introduction to Finite Volume Methods in Computational Fluid Dynamics. In: Wendt, J. F., ed. *Computational Fluid Dynamics*. Springer, Berlin: pp. 87-103.

Dresig, H., Holzweibig, F., 2010. *Dynamics of Machinery: Theory and Application*. Berlin: Springer.

Gosdin, M., 1985. *Analyse und Optimierung des dynamischen Verhaltens eines Pkw-Antriebsstranges*. Düsseldorf: VDI.

Hülsmann, A., 2007. *Methodenentwicklung zur virtuellen Auslegung von Lastwechselphänomenen in PKW*. Munich: Technische Universität München, Dissertation.

Krieg, O.M., Förg, M., Ulbrich, H., 2011. Simulation of Domain-Coupled Multibody and Thermodynamic Problems for Automotive Applications. *Proceedings of Multibody Dynamics 2011*. July 4-7, Brussels, Belgium.

Lefebvre, D., Chevrel, P., Richard, S., 2003. An H-Infinity-Based Control Design Methodology Dedicated to the Active Control of Vehicle Longitudinal Oscillations. *IEEE Transactions On Control Systems Technology*, Volume 11, No. 6: pp. 948-956.

Miersch, J., 2003. *Transiente Simulation zur Bewertung von ottomotorischen Konzepten*. Munich: Hieronymus.

Müller, I., Müller, W.H., 2005. *Fundamentals of Thermodynamics and Applications*. Berlin: Springer.

Pacejka, H.B., Bakker, E., 1992. The Magic Formula Tyre Model. *Vehicle System Dynamics: International Journal of Vehicle Mechanics and Mobility*, Volume 21, Issue S1: pp. 1-18.

Reik, W., Fidin, A., Seebacher, R. 2009. Gute Schwingung – Böse Schwingung: *Proceedings of 6. VDI-Fachtagung Schwingungen in Antrieben*: pp. 3-15.

Richard, S., Chevrel, P., Maillard, B. 1999. Active Control of future vehicles drivelines. *Proceedings of the 38th Conference on Decision & Control*: pp. 3752-3757.

Ulbrich, H., 1996. *Maschinendynamik*. Wiesbaden: Teubner.

IMPROVING JOB SCHEDULING ON A HETEROGENEOUS CLUSTER BY PREDICTING JOB EXECUTION TIMES USING HEURISTICS

Hannes Brandstätter-Müller^(a), Bahram Parsapour^(b), Andreas Hölzlwimmer^(c), Gerald Lirk^(d), Peter Kulczycki^(e)

^(a, b, c, d, e)Upper Austria University of Applied Sciences
School of Informatics, Communication, and Media
Department of Bioinformatics

^(a)[hannes.brandstaetter@...](mailto:hannes.brandstaetter@fh-hagenberg.at) ^(b)[bahram.parsapour@...](mailto:bahram.parsapour@fh-hagenberg.at) ^(c)[andreas.hoelzlwimmer@...](mailto:andreas.hoelzlwimmer@fh-hagenberg.at) ^(d)[gerald.lirk@...](mailto:gerald.lirk@fh-hagenberg.at)
^(e)peter.kulczycki@fh-hagenberg.at

ABSTRACT

In this paper, we propose the scheduling system for the *Bioinformatics Resource Facility Hagenberg (BiRFH)*. This system takes advantage of the fact that the facility offers tailored solutions for the customers, which includes having a limited amount of different programs available. Additionally, the BiRFH system provides access to different hardware platforms (standard CPU, GPGPU on NVIDIA Cuda, and IMB Cell on Sony Playstation machines) with multiple versions of the same algorithm optimized for these platforms. The BiRFH scheduling system takes these into account and uses knowledge about past runs and run times to predict the expected run time of a job. That leads to a better scheduling and resource usage. The prediction and scheduling use heuristic and artificial intelligence methods to achieve acceptable results.

The paper presents the proposed prediction method as well as an overview of the scheduling algorithm.

Keywords: algorithms, bioinformatics, high performance computing, molecular biology

1. INTRODUCTION

Scheduling and resource management are fundamental tasks when running a high performance computing system. Resource management and scheduling systems for different processor technologies and architectures in a single cluster are not very common although they offer great possibilities to the user. Our system software “Bioinformatics Resource Facility Hagenberg” (BiRFH) allows efficient control and management of the so-called “Meta-Heterogeneous Cluster”. BiRFH allows one not only to drive classic heterogeneous clusters (i. e., systems that comprise nodes that vary just in CPU speed and RAM size), it allows one to integrate and operate different processor architectures simultaneously. Our resource facility currently consists of standard Intel CPUs (Intel 2005), NVIDIA GPUs (NVIDIA 2010), and IBM Cell Broadband Engines (IBM 2006). There are many strategies available for scheduling jobs on a cluster. We focus on the special case where jobs are based on a small number

of algorithms with different input data. This allows runtime prediction using heuristics and therefore an improved deadline scheduling.

The “Bioinformatics Resource Facility Hagenberg” is a resource management and scheduling system targeting the needs of microbiology and bioinformatics related high performance computing. The main features are: (1) integration and management of different hardware platforms, (2) scheduling jobs so that the available hardware is used, and (3) making the management and job creation more accessible to non-technical users.

To fully enable the necessary features, and supported by the fact that the BiRFH service is designed to be used with a limited number of algorithms rather than allowing arbitrary code to be uploaded and executed like on many other standard compute clusters, BiRFH requires the algorithms to be adapted and to support some defined methods. The BiRFH framework seeks to mitigate the necessary development effort. The system is based on a framework that is to be included in each algorithm if possible, i. e., if the source code is available for modification. Including the framework directly into the source allows the use of more advanced features and more control over the program. Should the source for a program be unavailable, the framework also supports the creation of wrapper programs that can in turn execute the desired program. Regardless of how the framework is applied, it allows the use of the best available hardware for the selected algorithms with the trade-off of higher development effort to enable the algorithm on as many hardware platforms as possible.

As a further feature, BiRFH uses heuristic scheduling and resource management algorithms in order to optimize the cluster’s throughput.

There are many scheduling systems and resource managers for high-performance cluster computing available (see Table 1). Most of them are designed for uniform hardware. Almost no system allows the coupling of compute platforms having different processor architectures in a way that e. g. allows the migration of a running algorithm from one type of platform to a different one. BiRFH offers this possibility for algorithms with available source

code by implementing a data exchange that can also be used for hibernation, i. e., the freeing of used system resources by writing the current state of the calculations to the hard drive.

There are some other approaches to enable heterogeneity for compute-intensive applications. These include OpenCL (Munshi 2011) and C++ Accelerated Massive Parallelism (AMP) (Sutter 2011; Moth 2011). These focus on enabling single applications flexible access to any available computing device rather than distributing instances of algorithms over the available hardware. The BiRFH approach focuses more on user guidance and supporting the storage of data on the remote compute system. It is notable, however, that Microsoft’s C++ AMP moves the definition of heterogeneity more in the direction of heterogeneous platforms than the previously common heterogeneous systems, i. e., including different processor architectures.

BiRFH focuses on algorithms and computations for biomolecular and bioinformatics applications. The reasons for focusing especially on bioinformatics lie (1) in the near-exponential growth of available data in the currently booming field (Howe et al. 2008), (2) in the demands for making high performance computing available to non-technical users and (3) the availability of bioinformatics knowledge in the project team. Moreover, the BiRFH system supports a scheduling mechanism that is (1) based on the temporal behavior of algorithms and (2) also based on the size and the inner structure of input data to be processed.

Feature	Condor	SGE	SLURM	Maui	MPI2	BiRFH
Workload Manager	✓	✓	✓	✓		✓
Cycle Scavenging	✓					
Heterogeneous Platforms	✓		(✓)		(✓)	✓
Priority Based Scheduling	✓	✓	✓			✓
Hibernation Support	(✓)		✓			✓
Resource Based Scheduling		✓		✓		✓
Advanced Resource Reservation				✓		✓
Topology Awareness			✓			

Table 1: Some features of well-established resource managers and scheduling systems, in addition to the BiRFH system, from (Hoelzlwimmer 2010).

2. DATA MINING AND HEURISTICS

The terms *makespan* and *flowtime* (Pinedo 2008) are commonly used when classifying the success of the output of a scheduler. More advanced scheduling techniques, i. e., most scheduling methods beyond simple load balancing, require a knowledge of the expected time to complete a task. Some solutions (Xhafa and Abraham 2008) require an estimation by the user, which is sometimes too complicated a task for non-IT scientists, and on the other hand does not account for different versions of an algorithm optimized for the available heterogeneous hardware. Therefore, the *BiRFH* approach does not require run time estimations by the user.

Most algorithms exhibit some form of correlation between the input data, other given parameters and the time the program needs to run. The *BiRFH* system gathers performance data on the various algorithm implementations as they are executed. The collected data consists of the runtime measurements, i. e., how long the execution of the task took in real-time. Additionally, the consumed CPU time is also measured. Should the hardware be under-subscribed, algorithms can be executed during this idle time to generate additional measurements, especially for combinations of parameters that are expected to complete “holes” in the available data. This performance data, combined with data about the input and other parameters, is used by machine learning algorithms to predict the temporal behavior of subsequent runs, especially for new parameter combinations or input data.

As the parameter values, especially the input data, usually consist of file names or other non-numeric values, it can not be used directly for the prediction of the expected runtime. Therefore, the *BiRFH* framework calls a method that is expected to be provided for each algorithm. This method should provide a numeric value that represents each non-numeric parameter value. If, for example, one algorithm parameter is an input file name, the method produces a number signifying the “weight”, or expected impact on the runtime, of the contained data. Simple file size is sometimes not sufficient to use when predicting the impact on run time, e. g. when processing a file containing sequences in FASTA format, sometimes the number of sequences, other times the average or maximum length of the sequences is more significant. The amount of computation time to produce this number should be kept within a reasonable time frame, but is left to the algorithm developer to decide. If a simple file size is not sufficient, then partial, sampling or even a total file analysis should be done. This is only appropriate if the computational run time is very long compared to the time needed to load the data from the file, and not just slightly longer than the full analysis of the input file itself would take. The next chapter, 2.1 Heuristics, contains an example dataset and shows how the prediction works.

2.1. Heuristics

The machine learning algorithms that provide the best results on the current training data sets are Artificial Neural

Networks (ANNs). These are very versatile and produce accurate predictions for the available data. Other evaluated machine learning algorithms are the regression algorithms offered by the Weka toolkit (Hall et al. 2009).

To evaluate the available regression algorithms, a sample data set with 2400 measurement instances was created. This data set consists of measurements of a simple algorithm that reads a file containing several FASTA formatted sequences. Then, some string operations are performed and an output file is written. This algorithm is executed with three different input files (one with 500 sequences, one with 1000 and another one with 1500) and also with different parameters influencing the string operations. The collected data can be presented as CSV:

```

wall, InSize, InCount, MaxMut, WinSize, GC-Cont
2288.74, 89725, 500, 1, 5, 0
3279.26, 89725, 500, 1, 5, 0.8
2540.13, 181156, 1000, 1, 5, 0
4627.02, 564117, 1500, 1, 5, 0.7
5089.33, 564117, 1500, 5, 16, 0.6
...

```

Wall represents the wall clock time used to complete this algorithm. InSize is the size of the input file, InCount is the number of FASTA sequences, MaxMut, WinSize and GC-Cont are parameters that influence how the input is processed and may or may not have an influence on the run time. These is the data available after the algorithm runs have finished, and as mentioned in the chapter above, some of these data points are available before the real computation is started. In this case, everything except the wall clock time and the InCount is available before the algorithm is started for the calculation run. Getting the file size is not as compute intensive and can substitute the exact count of input sequences for this case.

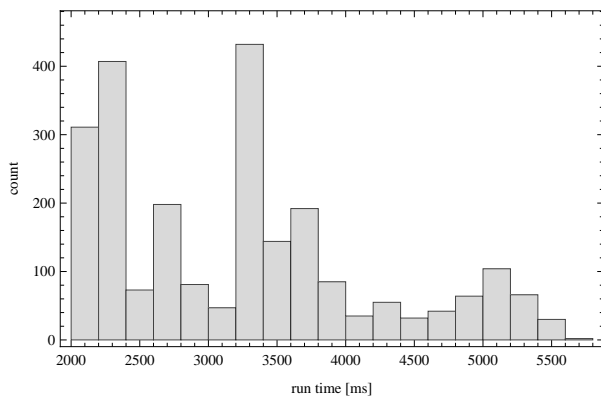


Figure 1: Run times of 2400 calls to the same algorithm with different input data and parameters

Looking at the run times gathered, see Figure 1, it is easy to spot three distinct peaks which indicate the run time effect of the two different input file sizes. There is also a third peak in the upper run time regions that is most likely caused by one of the parameters. The remaining variance due to other parameter variations. This reinforces

the assumption that a regression method can most likely predict the run times accurately from this data set.

Weka yielded the following results with 10-fold cross validation:

Classifier	CC	MAE	RMS	RAE
MultilayerPerc.	0.9256	280.06	358.72	50.3%
LinearRegr.	0.8714	385.02	464.75	69.7%
IsotonicRegr.	0.7620	502.95	613.38	65.8%
SimpleLinRegr.	0.6626	551.43	709.47	72.2%
PACERegr.	0.8713	385.04	464.79	50.4%
LibSVM	0.8452	403.02	507.48	52.7%

Table 2: Results in the Weka Toolkit for various prediction methods: MultilayerPerceptron, LinearRegression, IsotonicRegression, SimpleLinearRegression, PACERegression, LibSVM (CC: Correlation Coefficient, MAE: Mean absolute error, RMS: Root mean squared error, RAE: Relative absolute error)

Taking one of the validation results to better visualize the resulting prediction quality.

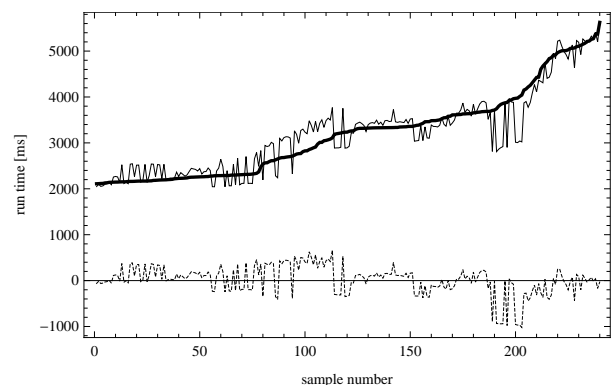


Figure 2: Actual and predicted values of a validation run using MultilayerPerceptron

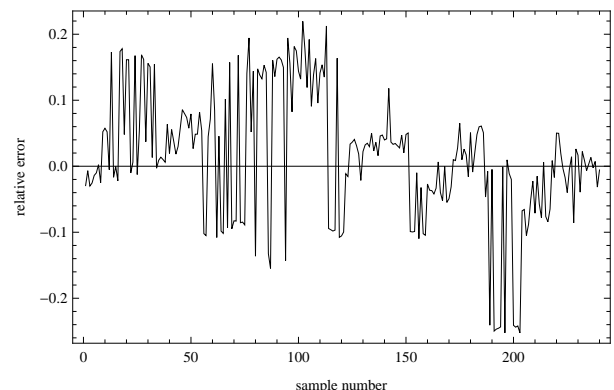


Figure 3: Relative errors of the validation run using MultilayerPerceptron

Using libSVM, the results are similar but with a higher margin of error. Figures 4 and 5 show the best result achieved by tweaking the libSVM parameters “cost”, “gamma” and “epsilon” with *epsilon-SVR* in a similar fashion as above with *MultilayerPerceptron*.

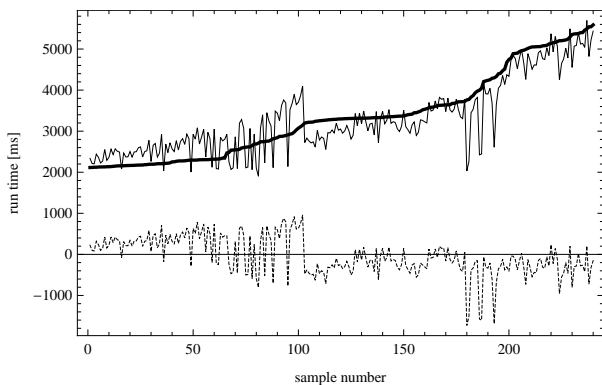


Figure 4: Actual and predicted values of a validation run using libSVM

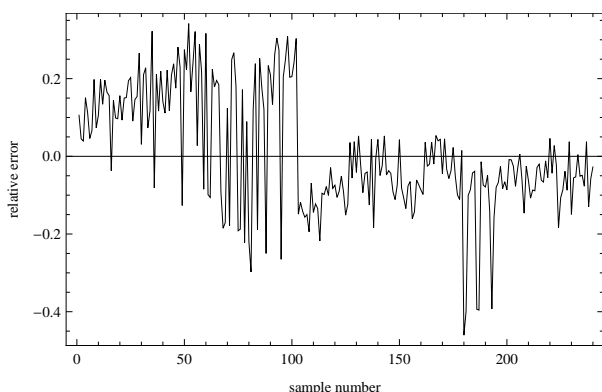


Figure 5: Relative errors of the validation run using libSVM

Standard ANN training algorithms usually require the basic structure to be predefined, i. e., not only the number of inputs and outputs, but also the number of internal layers, the nodes per layer and how the nodes are connected. This would prevent maximal flexibility of the training for the expected diversity of the dataset, as the complexity is an important factor in the success of the neural network training: too little complexity, and the solution quality suffers, but too much complexity would lead to over-fitting during the training. To work around this limitation and make the ANN approach viable for a multitude of different algorithms, training algorithms that start with an empty network and add complexity as needed to reach a neural network for the given training set are used in the BiRFH system. The FANN library (Nissen 2003), which was chosen as ANN implementation, provides such a training algorithm called Cascade2 (Nissen 2007).

3. SCHEDULING

The task of scheduling and resource management working together is to ensure that computation tasks are completed in a timely manner and that the available hardware is used as efficient as possible. There are many strategies available to fulfill these requirements.

As mentioned before in chapter 2, makespan and flowtime are used as indicators for the quality of a scheduling attempt. Having a reliable prediction for the expected run time enables more advanced scheduling features. *Earliest Deadline First (EDF)* is one of the most popular and widely used scheduling strategies, and has some parts in common to the BiRFH approach. The *EDF* strategy requires tasks to have a deadline, usually either the latest possible start time or the latest acceptable finishing time. The scheduler then orders the tasks by their deadlines and executes those with the earliest deadline first. Usually resource managers can also interrupt running tasks should a new task with an earlier deadline become available (Kruk et al. 2007).

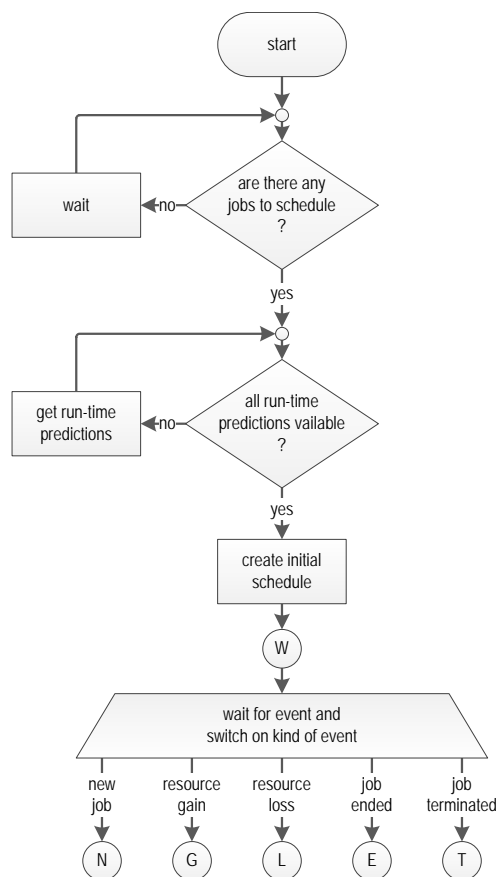


Figure 6: A simplified flowchart of the scheduling system

The basic program flow in the scheduling logic is displayed in figure 6. After an initial check for pending jobs, and calculating the run time estimation for these, the scheduler distributes these jobs among the available resources according to the scheduling policy. Then, the scheduler waits for one of the following events to occur:

(N) new jobs are submitted to be scheduled and executed, (G) an additional machine comes online and is available to execute jobs, (L) a currently available machine goes offline, (E) a currently running job ends or a running job terminates unexpected (T). The scheduler then handles these events accordingly and returns to the waiting state. Some paths lead to a rescheduling, others do not require a reordering. Depending on the circumstances, the reordering can affect the currently running processes as well.

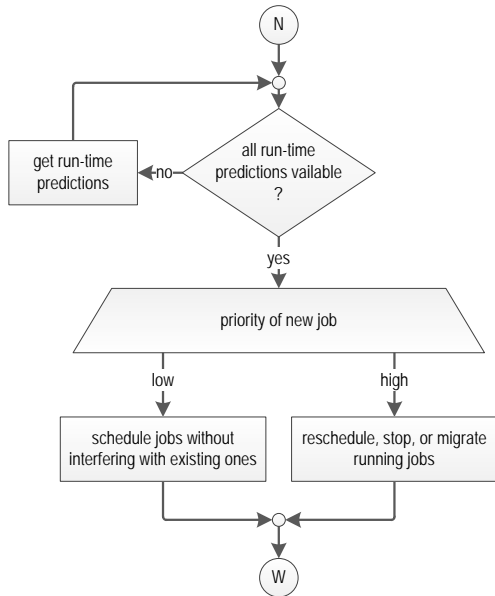


Figure 7: Scheduling for newly submitted jobs

In the path (N) new jobs are submitted, these are again run through the prediction. After the predictions are available for all new tasks, these are scheduled according to their priority. If their priority is default normal, then no special additional steps have to be performed and the tasks are scheduled without interfering with currently running tasks. Should the new tasks have high priority, the currently running tasks are included in the rescheduling, which could lead to low priority tasks to be paused, migrated or even aborted and restarted at a later point.

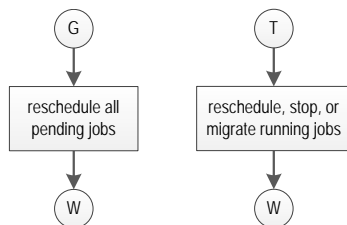


Figure 8: Scheduling for the case of the gain of a resource (left) or an erroneous job termination (right)

Path (G) (Figure 8, left) is caused by new hardware resources coming available for the system. The special case where these resource has been known to the system

already is handled in the third path. This path assumes that the resource is new or has been offline long enough that it can be regarded as new. All pending jobs are rescheduled, as are all currently running jobs if there is a significant reduction in overall time to completion.

The next path, (L), deals with the sudden, unplanned loss of a resource: all currently running and pending jobs of this resource are rescheduled over the remaining resources. Should the resource come back online in time, i. e., before the “replacement job” is completed, and the job on that resource is healthy (i. e., was not broken by the loss of connectivity), the now duplicate backup job is canceled and the pending jobs are rescheduled.

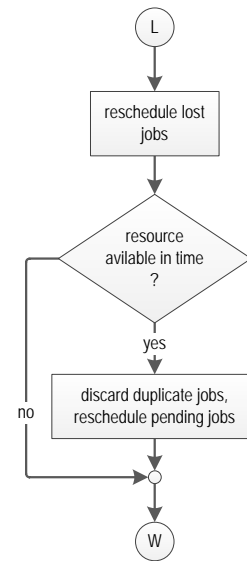


Figure 9: The scheduler handles the unexpected loss of a resource

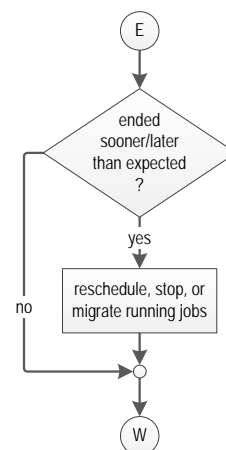


Figure 10: Scheduling when a job ends without error

The remaining two paths are triggered every time a job ends. If the job ends cleanly (E), the run time is compared to the estimate, if the difference is below the

threshold, no further scheduling is needed. If the difference exceeds the threshold, a scheduling run is performed to ensure optimal resource usage. If the job did not terminate cleanly (T) (see Figure 8), subsequent dependent runs are removed from the queue and an error notification is sent to the originating user. The remaining jobs, including the running ones, are then rescheduled again. Under one special condition, in the case where two algorithms are running simultaneously and one algorithm produces output that is immediately used by the second algorithm—called streaming read/write—the consuming job has to be terminated if the producing job experiences an error and terminates.

The scheduling algorithm itself is designed to be fully modular. It reads the information about the pending jobs as well as the current system status from a database, and after reordering the jobs according to the scheduling rules and policies, writes the new orders back to the database, where the resource manager reads them and relays the orders to the compute nodes. That way, different scheduling strategies and optimizations can be evaluated without big changes in the whole BiRFH system.

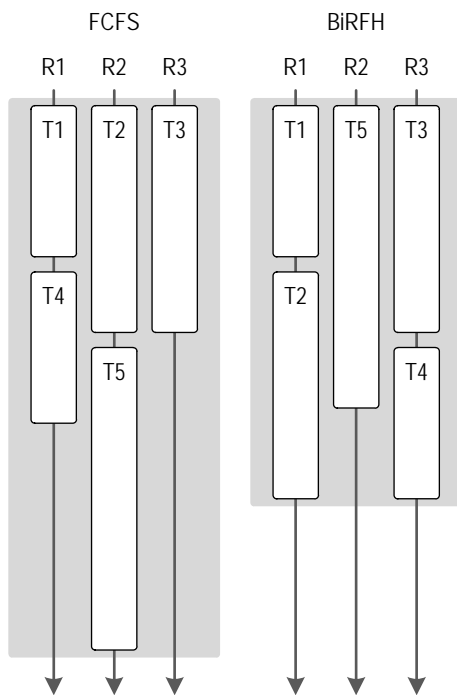


Figure 11: A simple scenario where knowledge of run times yield better scheduling than a basic First Come First Serve

Knowing about the expected runtime can improve the utilization and therefore reduce the overall time needed to complete several tasks. Figure 11 displays a hypothetical comparison of a simple EDF variant, where the tasks are prioritized in the order that they become available, i. e., a First Come First Serve scheduling. In this scenario, 5 tasks with different runtimes are available: Tasks T1 and T4 need 2 time units, tasks T2 and T3 take 3 time units

to finish and task T5 is the longest with 4 time units. The simple FCFS scheduling on the left produces an execution sequence that would take 7 time units. Knowledge of the expected run times can rearrange the tasks in such a way that the total execution time would be 5 time units.

The above example is based on the assumption of three completely identical compute nodes. Our scheduling approach aims to improve this strategy for *Meta-Heterogeneous Clusters*, i. e., clusters consisting of different hardware architectures like CPU, GPUs and others, by considering multiple versions on the different hardware platforms and deferring some algorithms expecting a more powerful platform to become available. This is enabled by the run time predictions of the heuristic analysis.

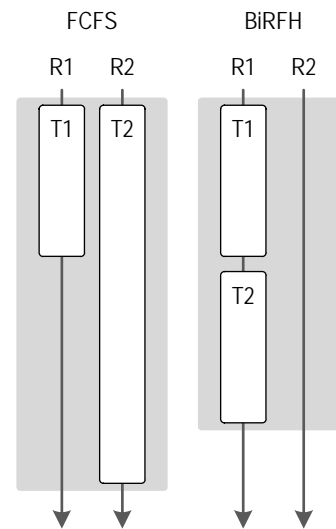


Figure 12: Advantage of knowledge of runtimes with two different hardware platforms

A simple scenario (see Figure 12) where this would be of benefit: Task T2 is a task that finishes fastest on the first system, e. g. a GPU, but the GPU is currently occupied by task T1. Using a standard deadline first algorithm, this task would now run on the second available system, e. g. the CPU. If the R1 node is expected to be ready before the task on the R2 system would be finished, and the sum of the expected calculation for both tasks on the R1 system would put the finishing time earlier, the task would be put on hold instead of using the free CPU.

Even if not all the tasks are known from the beginning, knowledge of the runtimes enables better scheduling in many cases. Figure 13 shows an example: the compute nodes R1 and R2 are again different hardware platforms, which results in different runtimes for the tasks. The FCFS version also displays a possible variant of the basic FCFS strategy, where tasks only are scheduled to hardware platforms where they take the least time to finish. This is shown with T3 (transparent on the lower left). Naively, this should yield better results, but even in this example, using all the available resources still provides better results. The BiRFH result on the right again shows optimized resource usage with reduced overall runtime.

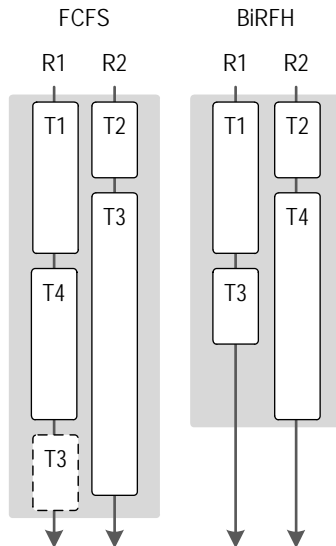


Figure 13: Prioritizing algorithm versions on heterogeneous platforms with minimal makespan in the FCFS strategy does not yield better results

Some scheduling scenarios require a complete rearrangement of the tasks, even running tasks could be blocking an optimal arrangement. In these cases, there has to be an assessment if the running task should be canceled and moved to a different hardware or if it would be better to wait for the task to finish. Knowledge of the expected finishing time can be very helpful in these cases. Figure 14 visualizes a scenario where a new task T3 is added after the tasks T1 and T2 have already been started. This task is only available on the hardware platform of R2, but this system is being used by T2. Therefore, T2 is canceled and restarted on node R1 after T1 has finished. Even though some computational effort is lost, the overall time is significantly shorter than waiting for T2 to finish.

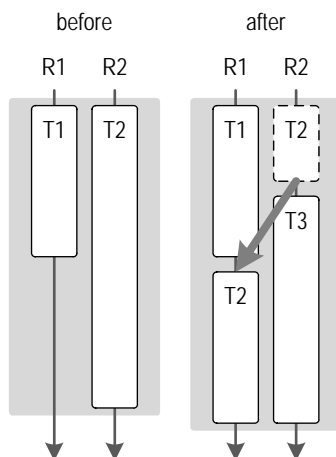


Figure 14: Canceling and rescheduling a task.

Instead of losing the computational progress so far it would be preferable to keep it while migrating to a different hardware platform. This can be enabled when the source code of the algorithm is available. If the algorithm

can be adapted to some kind of stepwise progress, the framework defines the methods to pause and serialize the current memory content to the hard disk for transfer. The counterpart implementation on the target hardware then can read the progress and continue with the calculations.

All the above examples target the least “real world” run time, or minimal makespan, for optimization. Other possible targets, especially when using heterogeneous hardware, could be the conservation of power. If a deadline is set and multiple hardware platforms are able to execute the task, the job could run on a platform that consumes less power during the calculations. The power consumption factor, combined with the run time, can be added to the optimization parameters.

As the overall design of the system allows for the individual models to be easily interchangeable, the prediction model and scheduling methods can gradually be replaced by new ones that yield better results. The underlying database offers a simple interface that enables this modularity. Therefore, future development also encompasses the implementation and evaluation of different prediction models and scheduling strategies.

4. CONCLUSION AND FUTURE WORK

The use of heuristics can improve the scheduling quality given some special circumstances, like, in our case, the limited number of different algorithms and the degree of heterogeneity of the hardware.

Future work includes the implementation and evaluation of other computational intelligence strategies as well as the inclusion of the *pipeline* concept, i. e., the consideration of follow-up calculations or calculations using different algorithms based on the same input data, which are required by follow-up calculations.

The research project “Bioinformatics Resource Facility Hagenberg” is supported by grant #821037 from the Austrian Research Promotion Agency (FFG).

REFERENCES

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. “The WEKA Data Mining Software: An Update.” *SIGKDD Explorations* 11:10–18.

Hoelzlwimmer, Andreas. 2010, September. “A Scheduling Framework Prototype for Heterogeneous Platform High Performance Computing.” Master’s thesis, University of Applied Sciences, Hagenberg.

Howe, Doug, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P. Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, Simon Twigger, Owen White, and Seung Yon Rhee. 2008. “Big data: The future of biocuration.” *Nature* 455:47–50.

IBM. 2006. “Cell Broadband Engine Programming Handbook.” Technical Report, IBM.

Intel. 2005, November. Intel Pentium 4. <http://www.intel.com/support/processors/>

- pentium4/sb/cs-007993.htm. last visit: July 28, 2010.
- Kruk, Lukasz, John Lehoczky, Kavita Ramanan, and Steven Shreve. 2007, December. Heavy Traffic Analysis for EDF Queues with Reneging.
- Moth, Daniel. 2011, June. Blazing-fast code using GPUs and more, with C++ AMP. Session at AMD Fusion Developer Conference 2011, http://ecn.channel9.msdn.com/content/DanielMoth_CppAMP_Intro.pdf.
- Munshi, Aaftab. 2011, January. The OpenCL Specification, Version 1.1.
- Nissen, S. 2003, October. "Implementation of a Fast Artificial Neural Network Library (FANN)." Technical Report, Department of Computer Science, University of Copenhagen.
- Nissen, Steffen. 2007, October. "Large Scale Reinforcement Learning using Q-SARSA(λ) and Cascading Neural Networks." Master's thesis, University of Copenhagen.
- NVIDIA. 2010, June. NVIDIA CUDA Reference Manual Version 3.1. http://developer.download.nvidia.com/compute/cuda/3_1/toolkit/docs/CudaReferenceManual.pdf. last visit: June 28, 2010.
- Pinedo, Michael L. 2008. *Scheduling: Theory, Algorithms, and Systems*. 3rd Edition. Springer.
- Sutter, Herb. 2011, June. Heterogeneous Parallelism at Microsoft. Keynote at AMD Fusion Developer Summit 2011, http://developer.amd.com/documentation/presentations/assets/4-Sutter_Microsoft-FINAL.pdf.
- Xhafa, Fatos, and Ajith Abraham. 2008. Chapter Metaheuristics for Grid Scheduling Problems of *Metaheuristics for Scheduling in Distributed Computing Environments*, edited by Fatos Xhafa and Ajith Abraham, 1–38. Springer.

AGENT-BASED SIMULATION OF ELECTRONIC MARKETPLACES WITH ONTOLOGY-SERVICES

Maria João Viamonte^(a), Nuno Silva^(a), Paulo Maio^(a)

^(a) GECAD - Knowledge Engineering and Decision Support Research Group
Institute of Engineering of Porto
Portugal

^(a) {mjv, nps, pam}@isep.ipp.pt

ABSTRACT

Agent technology has been successfully applied to the Electronic Commerce domain, but the diversity of the involved actors leads to different conceptualizations of the needs and capabilities, giving rise to semantic incompatibilities between them. It is hard to find two agents using precisely the same vocabulary. They usually have a heterogeneous private vocabulary defined in their own private ontology. In order to provide help in the conversation among different agents, we are proposing what we call ontology-services to facilitate agents' interoperability. More specifically, this work presents a multi-agent market simulator with ontology services. The system includes agents that provide services that allow other agents to communicate with each other in order to reach an agreement, ensuring that both parties are able to understand the terms of negotiation.

Keywords: Intelligent Agents, Simulation, Electronic Markets and Ontology Mapping

1. INTRODUCTION

With the increasing importance of Electronic Commerce across the Internet, the need for software agents to support both customers and suppliers in buying and selling good/services is growing rapidly. It is becoming increasingly evident that in a few years the Internet will host a large number of interacting software agents. Most of them will be economically motivated, and will negotiate a variety of good and services. It is therefore important to consider the economic incentives and behaviors of ecommerce software agents, and to use all available means to anticipate their collective interactions. Even more fundamental than these issues, however, is the very nature of the various actors that are involved in Electronic Commerce transactions. The involved actors lead to different conceptualizations of the needs and capabilities, giving rise to semantic incompatibilities between them. It is hard to find two agents using precisely the same vocabulary. They usually have a heterogeneous private vocabulary defined in their own private ontology. This leads to different conceptualizations of the needs and capabilities, giving rise to semantic incompatibilities between them. This problem is referred to as the

ontology problem of electronic negotiations (Viamonte and Silva, 2008).

Consequently, given the increasingly complex requirements of applications, the need for rich, consistent and reusable semantics, the growth of semantically interoperable enterprises into knowledge-based communities; and the evolution; the adoption of semantic web technologies need to be addressed (Silva and Rocha, 2004). In that sense, a suitable approach to address this interoperability problem relies on the ability to reconcile vocabulary used in agents' ontologies. In literature, this reconciliation problem is referred as Ontology Matching (Euzenat and Shvaiko, 2007).

In order to provide help in the conversation among different agents, we are proposing what we call ontology-services to facilitate agents' interoperability. More specifically, this work presents the AEMOS - Agent-Based Electronic Market with Ontology-Service System, a multi-agent market simulator with ontology services. The system includes agents that provide services that allow other agents to communicate with each other in order to reach an agreement, ensuring that both parties are able to understand the terms of negotiation.

The AEMOS system is an innovative project (PTDC/EIA-EIA/104752/2008) supported by the Portuguese Agency for Scientific Research (FCT).

2. AEMOS SYSTEM

AEMOS system is an Agent Based Electronic Market where agents can customize their behaviors adaptively by learning each users preference model and business strategies.

Unlike traditional tools, agent based simulation does not postulate a single decision maker with a single objective for the entire system. Rather, agents representing the different independent entities in electronic markets are allowed to establish their own objectives and decision rules. Moreover, as the simulation progresses, agents can adapt their strategies, based on the success or failure of previous efforts.

AEMOS includes a complex simulation infrastructure; able to cope with the diverse time scales of the supported negotiation mechanisms and with several players competing and cooperating with each

other. In each situation, agents dynamically adapt their strategies, according to the present context and using the dynamically updated detained knowledge (Viamonte, Ramos, Rodrigues and Cardoso, 2006). AEMOS is flexible; the user completely defines the model he or she wants to simulate, including the number of agents, each agent's type, ontology and strategies. Figure 1, figure 2 and figure 3 shows the AEMOS System Interface.

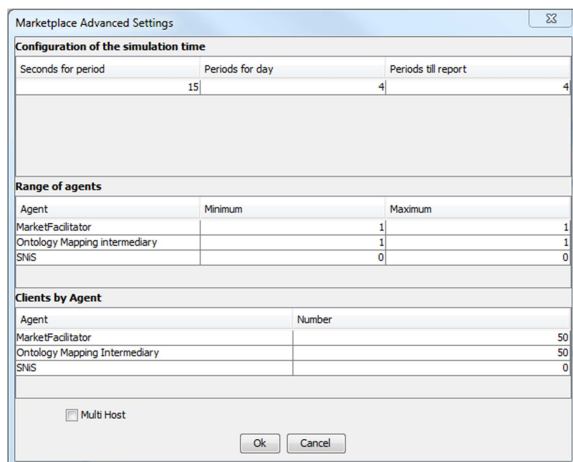


Figure 1: AEMOS system Interface – Internal Market Configuration

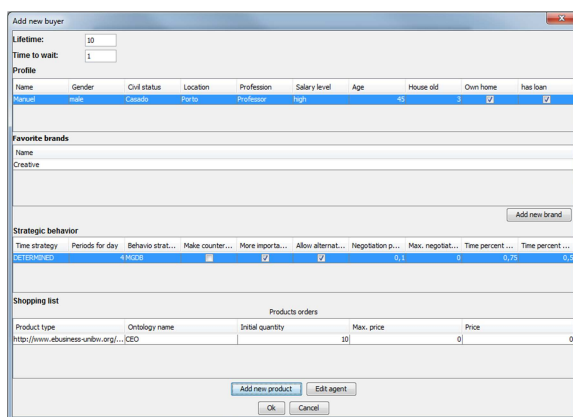


Figure 2: AEMOS system Interface – Buyer Configuration

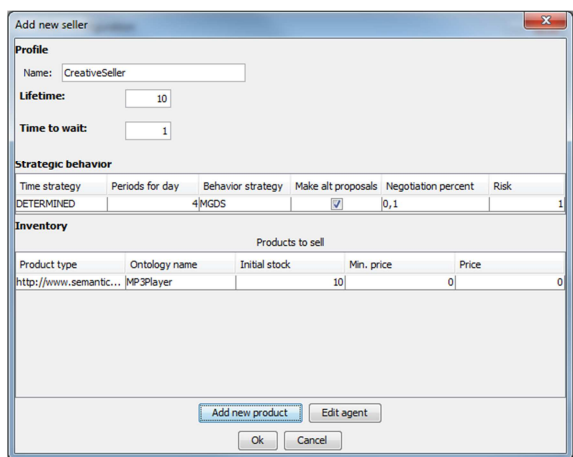


Figure 3: AEMOS system Interface – Seller Configuration

The simulator was developed based on “A Model for Developing a MarketPlace with Software Agents (MoDeMA)” (Viamonte, Ramos, Rodrigues and Cardoso, 2006). The following steps compose MoDeMA:

- Marketplace model definition, that permits doing transactions according to the Consumer Buying Behavior Model;
- Identification of the different participants, and the possible interactions between them;
- Ontology specification, that identifies and represents items on transaction;
- Agents architecture specification, and information flows between each agents module;
- Knowledge Acquisition, defining the process that guarantees the agent the knowledge to act on pursuit of its role;
- Negotiation Model, defining the negotiation mechanisms to be used;
- Negotiation Protocol, specification of each negotiation mechanism rules;
- Negotiation Strategies, specification and development of several negotiation strategies;
- Knowledge Discovery, identification and gathering of market knowledge to support agents' strategic behavior.

2.1. Multi-Agent Model

Multi-agent model includes three main types of actors as illustrated in figure 4.



Figure 4: AEMOS system layers

External Agents represent entities whose behavior is intended to be simulated and studied. There are two different types of external agents:

- Buyers (B) are agents representing entities desiring to acquire products;
- Sellers (S) are agents representing entities desiring to sell products.

Internal Agents provide services that allow external agents to communicate with each other in order to reach an agreement, ensuring that both parties are able to

understand the terms of negotiation. The main internal agents are:

- Market Manager (MM) is responsible for the market management. Manages all internal agents, register external agents and manages agents associations. In order to participate in the market an agent must first register with the MM agent. Usually there is only one MM agent per marketplace;
- Market Data Manager (MDM) registers information about all external agents participating in the market. When an external agent register in the market, the MDM agent collects its information, which is later provided when necessary. This agent is also responsible for writing statistical reports that enable to validate the correct functioning of the market. Normally there is only one MDM agent per marketplace;
- Market Facilitator (MF) is the agent responsible for the information integration process in the messages exchanged between external agents. It is an intermediate agent during the negotiation process that ensures, or tries to ensure that both parties are able to understand each other. Multiple MF agents can exist per marketplace. These agents are initialized by the MM agent when necessary. When an external agent is registered an MF agent is associated, from that moment all messages related to the negotiation process are sent for the associated MF;
- Ontology Matching intermediary (OM-i) is the agent that supports the information integration process. For that, this agent request the services (e.g. perform the information transformation according with the approved alignment) provided by several ontology matching specialized agents. Multiple OM-i agents can exist per marketplace, being initialized by the MM agent when necessary. When a MF agent is initialized an OM-i agent is associated, from that moment all the requests related to the information integration are sent to the associated agent.

2.2. Bilateral Contracts at AEMOS

In bilateral contracting B agents are looking for S agents that can provide them the desired products at the best price. We adopt what is basically an alternating protocol (Faratin, Sierra and Jennings, 1998).

Negotiation starts when a B agent sends a request for proposal. In response, a S agent analyses its own capabilities, current availability, and past experiences and formulates a proposal.

Seller's agents can formulate two kinds of proposals: a proposal for the product requested; or a proposal for a related product, according to the B agent preference model.

$PP_{g_i^{Agts \rightarrow Agtb}}^{DT}$ represents the proposal offered by the S agent $Agts$ to the B agent $Agtb$ at time T, at the negotiation period D for a specific product.

The B agent evaluates the proposals received with an algorithm that calculates the utility for each one, $U_{PP_{g_i}^{Agtb}}^{Agts}$; if the value of $U_{PP_{g_i}^{Agtb}}^{Agts}$ for $PP_{g_i^{Agts \rightarrow Agtb}}^{DT}$ at time T is greater than the value of the counter-proposal that the B agent will formulate for the next time T, in the same negotiation period D, then the B agent accepts the offer and negotiation ends successfully in an agreement;

otherwise a counter-proposal $CP_{g_i^{Agtb \rightarrow Agts}}^{DT}$ is made by the B agent to the next time T.

The S agent will accept a buyer counter-proposal if the value of $U_{CP_{g_i}^{Agts}}^{Agtb}$ is greater than the value of the counter-proposal that the S agent will formulate for the next time T; otherwise the S agent rejects the counter-proposal.

On the basis of the bilateral agreements made among market players and lessons learned from previous bid rounds, both agents (B and S) revise their strategies for the next negotiation rounds and update their individual knowledge module.

3. THE ONTOLOGY-SERVICES MODEL

To provide a transparent semantic interoperability between all Electronic Commerce actors, an ontology-services infrastructure was added to AEMOS. Thus, the AEMOS system architecture recognizes three new types of actors:

- Ontology Matching Service (OM-s) agent is able to specify an alignment between two ontologies based on some ontology matching algorithm. There are several OM-s on the marketplace, each one providing the same service but based on distinct matching approaches (e.g. syntactic, lexical, structural);
- Ontology Matching Information Transformation (OM-t) agent is responsible to transform any information represented according to one ontology (i.e. source ontology) to a target ontology using an already specified alignment between those two ontologies. Multiple OM-t agents can exist per marketplace. When an OM-i agent is initialized an OM-t agent is associated, from that moment all the requests related to the information translation are sent to the associated agent;
- Ontology Matching Repository (OM-r) agent registers the agreed ontology alignments specified between agents' ontologies. These alignments are applied to enable further agents' interactions.

These actors deploy a set of complementary features among themselves whose goal is to automate and improve the quality of the results achieved in the

electronic commerce transactions. The OM-i agent is responsible to manage all these services and consequently to hide the resulting complexity of that task from the marketplace (namely from the MF agent).

Figure 5 depicts the types of interactions between the marketplace internal agents (i.e. MF and OM-i) and the external agents (i.e. B and S).

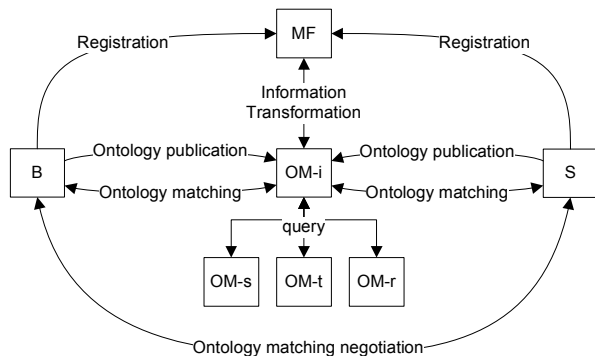


Figure 5: Marketplace actors and interactions

The Registration phase is initiated by the B or S agent and allows these agents to identify themselves to the marketplace and specify their roles and services.

The Ontology Publication phase is the set of transactions allowing B and S to specify their ontologies to the marketplace.

The Ontology Matching phase is the set of transactions driven by OM-i to align the ontologies of agents B and S. This phase is crucial for agents' interoperability and depends on the agents' ontology matching capabilities. By ontology matching (OM) capabilities of an agent we mean the ability it has to generate an alignment between two ontologies reflecting its own preferences and interests (e.g. alignment requirements), being such alignment achieved by its own or in cooperation with a set of ontology matching specialized agents (available or not in the marketplace). Yet, agents having OM capabilities may (or may not) have ontology matching negotiation (OMN) skills. Therefore, according to the agents' OM capabilities and OMN skills, for each pair of agents B and S six distinct scenarios are possible:

1. None of the agents' have OM capabilities;
2. Only one of the agents (say AgOM) have OM capabilities:
 - a. AgOM do not have OMN skills;
 - b. AgOM have OMN skills;
3. Both agents have OM capabilities:
 - a. None of the agents have OMN skills;
 - b. Only one of the agents have OMN skills;
 - c. Both agents have OMN skills.

On the first scenario, the OM-i agent is fully responsible for the ontology matching task. Even though, OM-i agent may take into consideration a set of preferences about the ontology matching process specified by both agents. Therefore, OM-i generates a

single alignment between agents' ontologies which need to be accepted by both agents.

On the other hand, on scenario 3c each agent generates its own alignment according to its internal preferences. Due to agents' different preferences and interests, the resulting alignments may have contradictory and inconsistent perspectives about candidate correspondences. Conflicts about correspondences are addressed by agents through the generic argumentation process described in (Maio, Silva, and Cardoso 2011a). In (Maio, Silva, and Cardoso 2011b) it is described how agents can exploit that argumentation process for ontology matching purposes. At the end, both agents need to inform the OM-i agent about the agreed alignment.

On scenarios 2a, 2b, 3a and 3b the agent lacking OM capabilities needs to delegate such responsibility to the OM-i agent. Yet, because the other agent has OM capabilities two distinct alignments exist. Resulting conflicts about correspondences are addressed either: (i) solely by OM-i agent if none of the agents have OMN skills (scenario 2a and 3a) or (ii) by a negotiation process between the agent with OMN skills and the OM-i agent in representation of the other agent (scenario 2b and 3b).

The Information Transformation phase is the set of information transactions through OM-i that transforms (i.e. converts) information described according to the sender's ontology to be described according to the receiver's ontology. This process is very methodical in accordance to the specified ontology alignment.

4. AN ONTOLOGY MAPPING EXAMPLE

For this example we consider a simple market with only two external agents (one B and one S).

The B agent intends to purchase 10 units of the same product (mp3 player) using for its representation the Ontology Consumer Electronics Ontology (CEO).

The S agent provides the desired product in sufficient quantity however uses for its representation the Ontology MP3Player (MP3_Player).

The interaction between agents is shown in Figure 6.

When MF receives a request for proposals for a product, as there are no S agents who use the same ontology as the B agent, MF makes a request to the OMi to suggest S agents that may be able to satisfy the request. OMi selects S agents that use some of the ontologies that can be mapped with the B agent's ontology. From this selection results a list of S agents (in this case only one) where for each S agent is associated a proposal for the mapping of its ontology with the B agent's.

Then the MF asks B and S agents for approval of the proposed mapping. If both approve, confirms the approval to OMi and ask him to represent the B request data according to the S ontology. The transformed data is replaced on the original request and it's forward to the S agent.

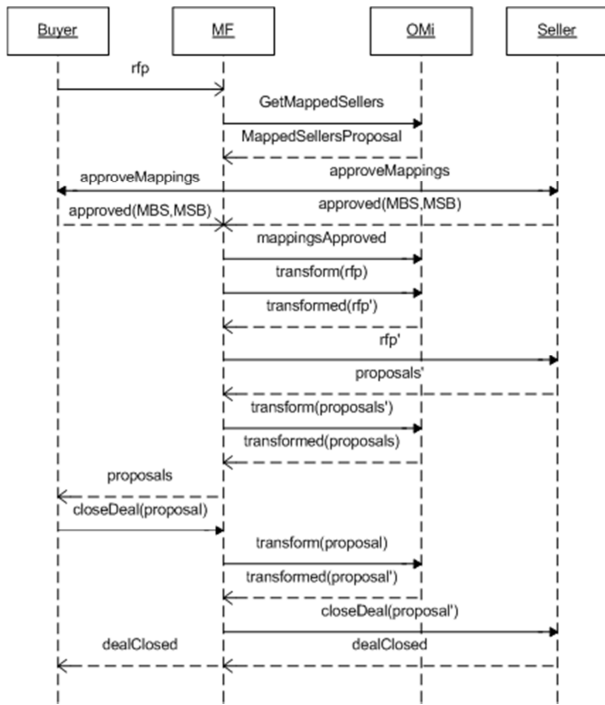


Figure 6: Data Integration Interaction

If the S formulates proposals, the MF makes a new request to the OMi so it represents the proposals data according to the B ontology.

During this process it's registered the approved mapping by the agents for the type of product. When a new communication is made by these agents related with this type of product (e.g. to close deal) the information is transformed using the approved mapping.

In case that the Players (B and S agents) don't approved the ontology mapping proposed by the system an Ontology Matching Negotiation Process (Maio, Viamonte and Silva, 2011) is used in order to obtain an agreement. It is envisaged that in the ontology matching negotiation phase agents adopt the argument-based negotiation process presented in (Maio, Viamonte and Silva, 2011).

In real scenarios with more Players (B and S agents) the process described above is replicated.

5. IMPLEMENTATION

The AEMOS system was developed in Open Agent Architecture (OAA) (<http://www.ai.sri.com/~oaa/>) and in Java.

The OAA platform, figure 7, is a framework for integrating a community of heterogeneous software agents in a distributed environment. It is structured to minimize the effort involved in creating new agents, written in various languages and operating platforms; to encourage the reuse of existing agents; and to allow the creation of dynamic and flexible agent communities. The OAA's Interagent Communication Language is the interface and communication language shared by all

agents, no matter which machine they are running on or which programming language they are programmed in. OAA is not a framework specifically devoted to develop simulations; some extensions were made to make it more suitable, such as the inclusion of a clock to introduce the time evolution mechanism of the simulation.

Each agent is implemented in Java, as a Java thread. The model can be distributed over a network of computers, which is a very important advantage to increase simulation runs for scenarios with a huge amount of agents.

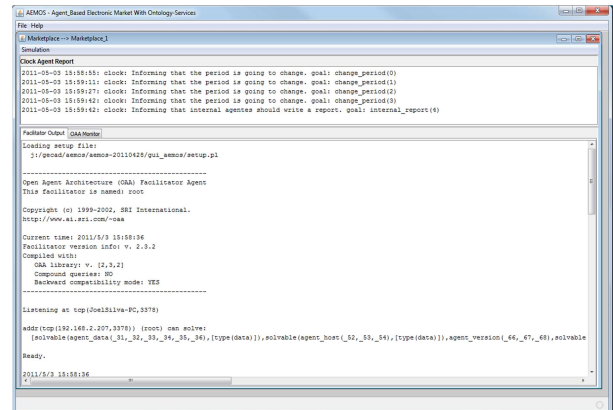


Figure 7: The OAA Facilitator

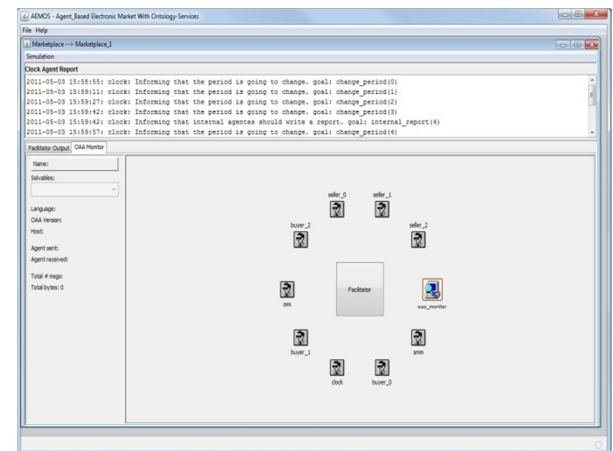


Figure 8: The AEMOS MarketPlace

6. CONCLUSIONS

AEMOS project is an innovative project that proposes a semantic information integration approach for agent-based electronic markets based on ontology-based technology, improved by the application and exploitation of the trust relationships captured by the social networks.

We intent face the problem of the growth of electronic commerce using software agents to support both customers and suppliers in buying and selling products. The diversity of the involved actors leads to different conceptualizations of the needs and capabilities, giving rise to semantic incompatibilities between them.

Ontologies have an important role in Multi-Agent Systems communication and provide a vocabulary to be used in the communication between agents. It is hard to find two agents using precisely the same vocabulary. They usually have a heterogeneous private vocabulary defined in their own private ontology. In order to provide help in the conversation among different agents, we are proposing what we call ontology-services to facilitate agents' interoperability. More specifically, AEMOS project proposes an ontology-based information integration approach, exploiting the ontology mapping paradigm, by aligning consumer needs and the market capacities, in a semi-automatic mode, improved by the application and exploitation of the trust relationships captured by the social networks.

Yet, it is our conviction that the marketplace must encourage agents to play an important role in the required matching process. Even though, that cannot be a mandatory issue and therefore the marketplace must be equipped to deal with agents having different ontology matching capabilities. It is envisaged that by taking part in the matching process agents may become more confident in the underlying communication process and in face of that consider the electronic commerce exchanged data (e.g. RFP and PP) more reliable (safe) and consequently become more proactive in the marketplace.

ACKNOWLEDGMENTS

The authors would like to acknowledge FCT, FEDER, POCTI, POSI, POCI, POSC, POTDC and COMPETE for their support to R&D Projects and GECAD Unit.

REFERENCES

- Viamonte, M. J. and Silva, N., 2008. Semantic Web-Based Information Integration Approach for an Agent Based Electronic Market. *Semantic Web Methodologies for E-Business Applications: Ontologies, Processes and Management Practices*. Chapter VIII (150-169). Publisher: IDEA Book, Editor(s): Roberto García.
- Silva, N. and Rocha, J., 2004. Semantic Web Complex Ontology Mapping. *Web Intelligence and Agent Systems Journal*, vol. 1, no. 3, p. 235—248, 2004.
- Euzenat, J. and Shvaiko, P., 2007. *Ontology Matching*. First ed., vol. 1, 1 vols. Heidelberg, Germany: Springer-Verlag, 2007.
- Viamonte, M.J., Ramos, C., Rodrigues, F. And Cardoso, J, 2006. ISEM: A Multi-Agent Simulator For Testing Agent Market Strategies. *IEEE Transactions on Systems, Man and Cybernetics – Part C: Special Issue on Game-theoretic Analysis and Stochastic Simulation of Negotiation Agents*, vol. 36, no. 1, pp. 107-113.
- Faratin, P., Sierra, C. and Jennings, N., 1998. Negotiation Decision Functions for Autonomous Agents. *Int. J. Robotics and Autonomous System*, 24 , 3, 1998, 159-182
- Maio, Paulo, Nuno Silva, and José Cardoso. 2011a. "EAF-based Negotiation Process." in *The 4th International Workshop on Agent-based Complex Automated Negotiation (ACAN) at AAMAS*. Taipei, Taiwan.
- Maio, Paulo, Nuno Silva, and José Cardoso. 2011b. "Generating Arguments for Ontology Matching." in *10th International Workshop on Web Semantics (WebS) at DEXA*. Toulouse, France.
- (CEO) Ontology Consumer Electronics Ontology. Available from: <http://www.ebusiness-unibw.org/ontologies/consumerelectronics/v1.owl>
- (MP3_Player) MP3_Player. Available from: http://daisy.cti.gr/svn/ontologies/AtracoProject/AtracoUserProfile/Y2Integration-FeelComfortable/MP3_Player.owl

AUTHORS BIOGRAPHY

MARIA JOÃO VIAMONTE is an adjunct professor of informatics at the School of Engineering of the Polytechnic Institute of Porto. Her research interests are in multi-agent simulation, agent mediated electronic commerce and decision support systems. She received her PhD in electrical engineering from the University of Trás-os-Montes and Alto Douro. She is coordinator and technical leader of several research projects Contact her at GECAD - Knowledge Engineering and Decision Support Research Group at the School of Engineering of the Polytechnic Institute of Porto, Rua Dr. António Bernardino de Almeida, 4200-072 Porto, Portugal; mjv@isep.ipp.pt

NUNO SILVA is Coordinator Professor of informatics at the School of Engineering of the Polytechnic Institute of Porto. His research interests are Information Integration, Knowledge Engineering and the Semantic Web. He received his PhD in electrical engineering from the University of Trás-os-Montes and Alto Douro, Portugal. He is coordinator and technical leader of several research projects. Contact him at GECAD - Knowledge Engineering and Decision Support Research Group at the School of Engineering of the Polytechnic Institute of Porto, Rua Dr. António Bernardino de Almeida, 4200-072 Porto, Portugal; nps@isep.ipp.pt

PAULO MAIO is an assistant professor of computer engineering at the Polytechnic Institute of Porto's Institute of Engineering. He is also a PhD student at the University of Trás-os-Montes and Alto Douro. Current research interests are focus in the ontology matching problem applied to multi-agent systems promoting agents' interoperability. Contact him at Instituto Superior de Engenharia do Porto, Departamento de Engenharia Informática, Porto, Portugal; pam@isep.ipp.pt

KEY ISSUES IN CLOUD SIMULATION PLATFORM BASED ON CLOUD COMPUTING

Lei Ren^(a), Lin Zhang^(a), Yabin Zhang^(b), Yongliang Luo^(a), Qian Li^(c)

^(a)School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China;

^(b)Beijing Simulation Center, Beijing 100854, China;

^(c)School of Information Technology, Shandong Institute of Commerce and Technology, Jinan 250103, China

^(a)leo.renlei@gmail.com, ^(b)zhangyabin@139.com, ^(c)be_jesica@126.com

ABSTRACT

Networked modeling & simulation platform can provide important support for collaborative simulation applications by integrating simulation resources over networks. Facing the increasing rich simulation resources over networks, the new challenge is that users need to get simulation services on demand simply and use simulation resources in a more efficient, transparent, and ubiquitous way. This paper presents Cloud Simulation Platform (CSP) that introduces the idea of Cloud Computing into networked modeling & simulation platform. We discuss the key issues in CSP including the operating principle of Cloud Simulation model, CSP architecture and components, and the key technologies in CSP. CSP provides a promising solution to the problems that networked modeling & simulation platform is facing, and this paper presents a map for future research on CSP.

Keywords: cloud simulation, cloud computing, cloud manufacturing, high performance simulation, networked modeling and simulation platform

1. INTRODUCTION

The rapid development of Modeling & Simulation (M&S) technology has profound influence on understanding the complicated world in recent years. M & S platform technology provides effective support for modeling, computing, analyzing, evaluating, verifying, and forecasting in many application domains such as simulation in military training, industry, and economics. To deal with the large amount of simulation resources distributed over networks, networked M&S platform technology has become one of the most powerful tools to support sharing and collaboration of distributed simulation resources (Zhang and Chen 2006), especially in large-scale simulation applications such as virtual prototyping of space shuttle. With the rapid development of distributed computing technology, more and more types of simulation resources can be connected into networks and take part in collaborative simulation process. This has resulted in a sharp increase in complexity of networked simulation systems, as well as increasing difficulty in making use of distributed and

heterogeneous simulation resources (Bohu 2007). Despite the increasingly rich simulation resources over networks, it is even more difficult for users to find what they just need to complete their simulation tasks. Now networked M&S platform technology is facing new challenges from a user-centric viewpoint.

One of the great challenges is that simulation users need to take full advantage of a variety of simulation resources in distributed and heterogeneous environments efficiently and transparently (Lei and Lin 2010). Simulation resources are the basis for modeling and simulation, and they include models, computing devices, storage, data, knowledge, software, and simulator needed by simulation systems. The simulation resources usually distribute in dispersed locations across different organizations over networks. To leverage the required simulation resource, users need to check its location, and then negotiate with the owner before they can access to it. This is a very inefficient operation mode for resources sharing. In addition, the distributed simulation resources often run in heterogeneous environments. Diverse hardware platform, operating system, and programming environment set up obstacles for users to integrate them together to support collaborative simulation applications. So the key problem is how to shield distribution and heterogeneity of simulation resources to provide a transparent access mechanism and build an efficient resources sharing environment.

Moreover, users need to get on-demand simulation services according to their personalized requirements and use them ubiquitously. Currently the networked M&S technology has shifted the focus from computing-centric angle to user-centric. However, current complex simulation systems always bring about a heavy burden caused by the deployment and configuration work. Users have to spend much time installing hardware drivers, operating systems, and software tools to establish a specific simulation application. The system development process is cumbersome and cannot meet user's need of customizing a simulation system simply and rapidly. In addition, there is a growing demand for using mobile network terminals (e.g., pad computer, smart phone, etc.) to participate simulation anywhere

anytime. Users don't want to know too many technical details about how to call a simulation function remotely, as well as how to integrate diverse simulation resources to implement a specific function. Thus users can pay more attention to simulation applications themselves instead of technical details. Therefore, the networked M&S technology had better provide a new paradigm by which simulation resources could be integrated to construct a simulation system according to the needs of users flexibly and dynamically. And the simulation system can be accessed and used through ubiquitous terminals anytime anywhere.

To address these issues, the idea of Cloud Computing (Buyya *et al.* 2009) may provide an opportunity to give impetus to the development of networked M&S technology. Cloud Computing refers to a pattern of IT service delivery and utilization. In Cloud Computing, users may access the scalable IT resources on demand via networks by using a computer, smart phone and other interactive terminals, and they don't need to download and install applications on their own devices because all IT resources (computing and storage) are maintained by cloud servers. In this paper we introduce the idea of Cloud Computing into networked M&S technology and present Cloud Simulation Platform (CSP). The paper firstly presents the operating principle of Cloud Simulation model. Then the CSP architecture and the components are illustrated, and the key technologies in CSP are discussed to indicate the future research areas. The key issues discussed in this paper can contribute to the future research on CSP.

2. RELATED WORK

There are many different definitions for Cloud Computing. For example, it refers to a large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet (Foster *et al.* 2008). Cloud Computing is considered as a new business paradigm describing supplement, consumption and delivery model for IT services by Utility Computing (Mladen 2008) based on the Internet. The typical examples of public Utility Computing include Google AppEngine, Amazon Web Services, Microsoft Azure, IBM Blue Cloud, Salesforce. So far, Cloud Computing is regarded as the sum of IaaS (Infrastructure as a Service), PaaS (Platform as a Service), and SaaS (Software as a Service). Cloud Computing providers can deliver on-demand services covering IT infrastructure, platform, and software via virtualization and service encapsulation technology. Thus, it can meet the needs of ever-rising scale of computing and storage of consumers and lead to decrease in IT investment cost at the same time. Cloud Computing introduces a promising model and technical approach to the problems networked M&S platform facing. For example, virtualization technology (Thomas *et al.* 2005) can support deep encapsulation of a logic

entity of IT infrastructure (e.g., CPU, memory, disk, and I/O) and software into a pool of virtual machines, thus it can achieve high efficient and transparent utilization of resources.

3. CLOUD SIMULATION MODEL

Figure 1 illustrates the operating principle of Cloud Simulation model. A simulation cloud refers to a cluster of virtualized simulation resources and services. The physical simulation resources can be mapped into virtualized resource templates by virtualization technology (Lei and Lin 2010). For example, package ANSYS software, Window XP, 4 CPUs, 1G memory, and etc. into a virtual machine template. And the functions of a simulation resource can be encapsulated into standard services that are loosely coupled and interoperable. The simulation clouds can shield the complexity caused by the distribution and heterogeneity of resources, and accomplish unified management of the standard services.

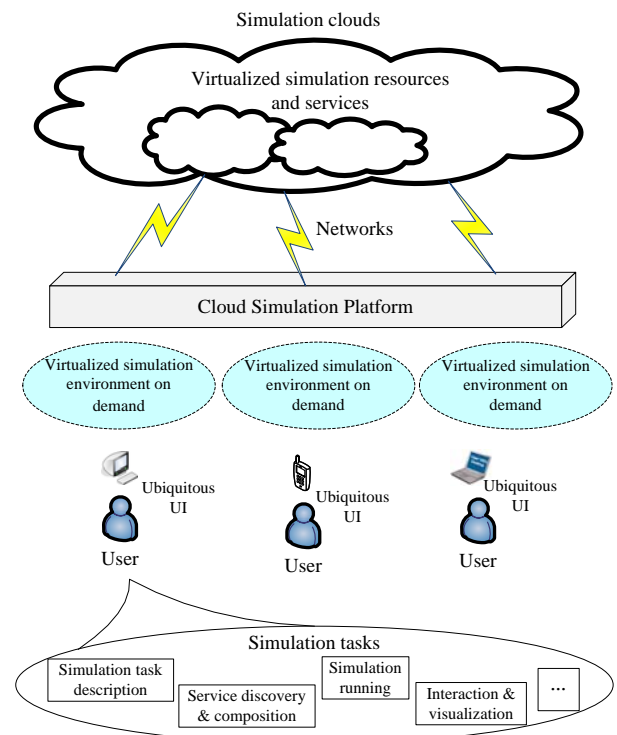


Figure 1: Cloud Simulation Model

Users don't need to deploy their own simulation environments involving hardware and software meeting specific needs. A user can submit a request of simulation tasks to CSP by using ubiquitous interactive terminals (e.g., smart phone, pad computer, etc.), and the request will be parsed by CSP to generate formal requirements of simulation resources. CSP launches a search to match the needed resources in the simulation clouds according to the resource requirements. The selected resources then will be accumulated to establish a virtualized simulation environment, and CSP can complete the deployment of the resources automatically by the way of instantiating the virtualized resource templates. In the virtualized simulation environment,

users feel like facing a real system designed for his specific needs, and the simulation functions can be acquired in a unified form of service, e.g., web service (Papazoglou *et al.* 2007), to support collaborative simulation. In run-time simulation process, CSP can assemble services and schedule resources dynamically over networks, and users can interact with the virtualized system as well as monitor the visual feedback remotely. Once a crash or unrecoverable error occurred in some simulation resource, CSP can carry out live migration (Christopher *et al.* 2005) from the trouble resource to another healthy one. This ability of

fault tolerance can achieve a high reliable collaborative simulation, and the whole process is transparent to users.

4. CLOUD SIMULATION PLATFORM ARCHITECTURE

Figure 2 shows the layered architecture of CSP. It consists of four layers: *Virtualized Resource Layer*, *Middleware Layer*, *Simulation Service Layer*, and *Ubiquitous Portal Layer*. CSP can integrate a broad range of simulation resources as Figure 2 illustrates, and provide support for simulation applications such as military training, product development, and conceptual prototype verification.

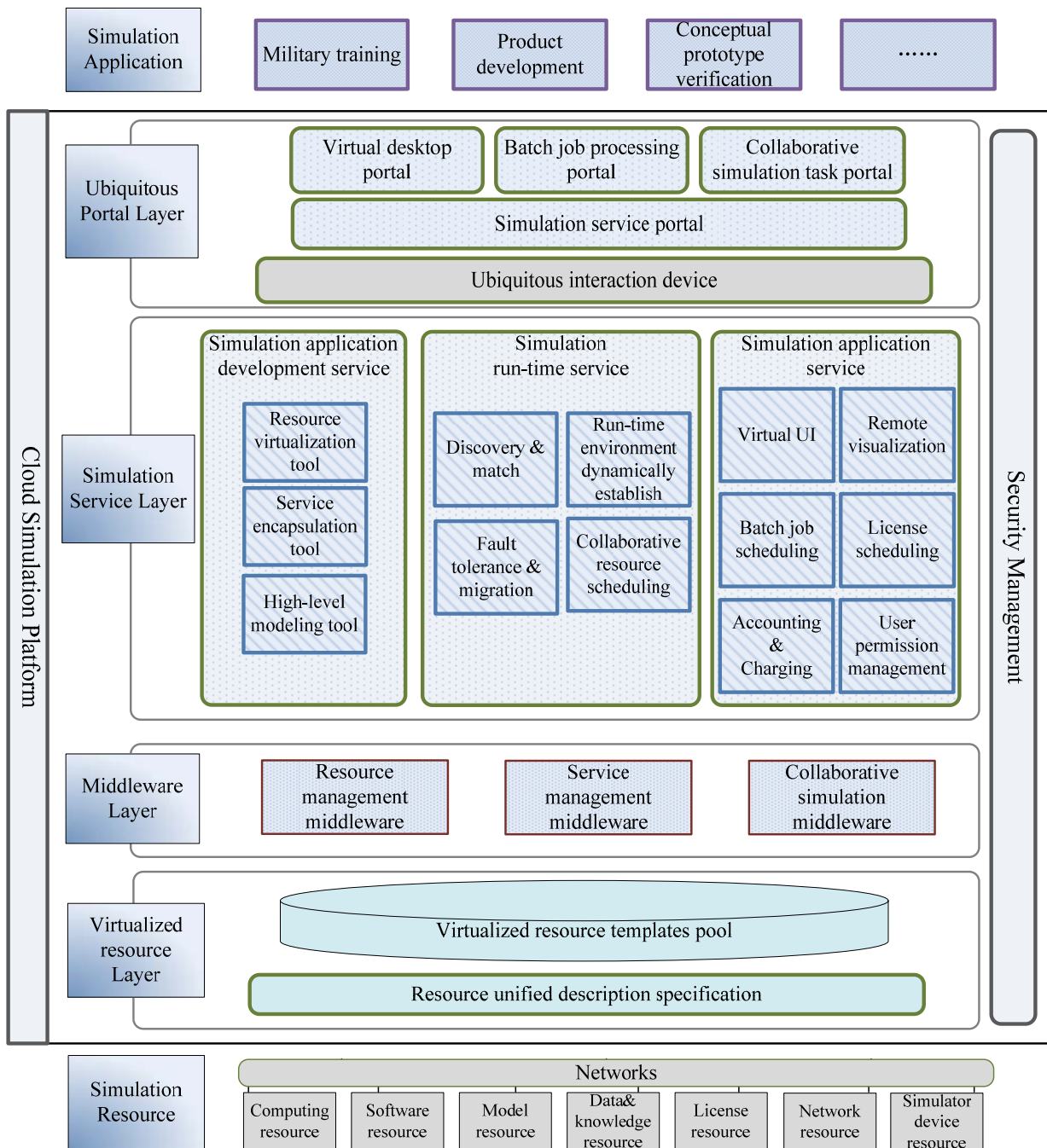


Figure 2: Cloud Simulation Platform architecture

4.1. Virtualized Resource Layer

The *Virtualization Resource Layer* is responsible for connecting a variety of simulation resources over networks and mapping them to virtualized templates according to a unified description specification. The simulation resources consist of computing resource, software resource, model resource, data resource, knowledge resource, license resource, network resource, and simulator device resource.

The physical resources such as computing resource, software resource, and model resource can be mapped to virtual machine templates where the fine-grained resources (e.g., CPU, memory, etc.) can be regrouped to create new virtualized resources. These virtualized resources are logical partition and the abstract composition of the real ones, and the templates should be described according to the unified description specification. The provider of the resources can register the templates in the registration center of CSP to establish a virtualized resource pool. The pool seems like a cloud where the raindrops of simulation resources aggregate and move.

4.2. Middleware Layer

The *Middleware Layer* serves as middleware with the common functions that links the *Virtualization Resource Layer* with the *Simulation Service Layer*. This layer includes three types of middleware to implement resource management, service management, and collaborative simulation, respectively.

The resource management middleware is used to manage the virtualized resource templates pool as well as the physical resources. It takes charge of allocating resources dynamically according to the scheduling commands issued from the upper layer. The distributed and heterogeneous resources are under control of this middleware and the real-time running status can be monitored.

The service management middleware acts as a service container to support unified service description, registration, discovery, match, composition, and remote call. This middleware shields the complexity caused by distribution and heterogeneity, and makes the functions of the resources independent from specific infrastructure.

The collaborative simulation middleware offers support for run-time simulation process management. This middleware serves as a special service-oriented RTI that is responsible for maintaining simulation run-time status, allocating data in federations and controlling the synchronization of multiple tasks.

4.3. Simulation Service Layer

The *Simulation Service Layer* is the core functional component to provide frequently-used services supporting simulation applications. The services are divided into three categories: simulation application development service, simulation run-time service, and simulation application service.

The simulation application development service is used to design and customize simulation applications with three tools. The resource virtualization tool is used

to encapsulate simulation resources into virtualized templates, thus the corresponding physical resources can be managed through the exposed interfaces. The service encapsulation tool is used to package the functional interfaces of the simulation resources into standard services, and this tool provides a development environment for users to customize the domain-specific services. The high-level modeling tools gives support for users to complete high-level modeling work of simulation tasks with interactive visual tools.

The simulation run-time service is responsible for the run-time process management. There are four types of services supporting run-time simulation. The discovery & match module is used to search the suitable services and resources in response to simulation requests. Another module is responsible for establishing the virtualized simulation environment dynamically meeting the needs of simulation tasks, and deploying the resources automatically. The collaborative resource scheduling module is used to parse the description of coordinated tasks, manage the time steps of concurrent tasks, and dispatch the services and resources to complete the simulation tasks.

The simulation application service provides the support of common functions that simulation applications need. The module of virtual UI is used to generate personalized virtual desktop interfaces. Users can access CSP and interact with applications through virtual UI by network browsers. The remote visualization module is responsible for rendering visual graphics that show the live image of simulation task progress. The module of batch job scheduling is used to manage the task queue and schedule the tasks. The license scheduling module is used to manage the licenses of commercial software, including license reservation, dispatch and recovery. The accounting & charging module is used to account the simulation resource usage according to the rate set by the resource provider and charge the users. The module of user permission management provides mechanism for user management, such as account maintenance, identity authentication, role assignment, and access control.

4.4. Ubiquitous Portal Layer

The *Ubiquitous Portal Layer* provides interfaces supporting ubiquitous UI for users to access CSP services. This layer offers interface adapter for interactive terminal devices including PC, pad computer, and smart mobile phone, so users can acquire simulation services without time and space constrains.

The CSP portals include simulation service portal, virtual desktop portal, batch job processing portal, and collaborative simulation task portal. The simulation service portal is the homepage where users can search for simulation services and customize their own simulation applications. To better support typical simulation application patterns, three kinds of portal mentioned above are offered on the homepage.

In the virtual desktop portal, users can customize preferred system environment including hardware and software, then CSP find the matched virtualized

templates and instantiate the virtual machines. Finally CSP returns the remote virtual UI to users for interaction.

In the batch job processing portal, users submit batch job files to CSP, then CSP parses the file and find the needed simulation services and resources to deploy simulation environment. Once the environment is ready, the tasks are uploaded to the virtual machines, and CSP returns result to users when the job is done.

In the collaborative simulation task portal, users can establish the high-level model of simulation tasks. The formal description of the model is parsed by CSP, then CSP discover the needed services and resources to create simulation federation. In the simulation process, users can monitor the run-time status through the portal. In addition, users can start, pause, continue, and stop the simulation progress in the portal.

5. KEY TECHNOLOGIES IN CLOUD SIMULATION PLATFORM

5.1. Simulation Resource Virtualization and On-demand Use

To support high efficient sharing and flexible use of large-scale simulation resources, virtualization is one of the key techniques (Lei *et al.* 2011). Virtualization technology can decouple simulation applications from needed simulation resources, thereby allowing the fine-grained resources to integrate on demand flexibly.

The key issues in simulation resource virtualization include simulation resource taxonomy, unified formal description of simulation resource, virtualized simulation resource template, composition verification, management of large-scale virtualized resource pool, mapping approach from physical simulation resource to virtualized resource, and remote management of simulation resource status.

5.2. Service-oriented Simulation Resource Publication and Intelligent Match

Service computing technology plays a important role in CSP. The unified service encapsulation of simulation resource make it possible to shield the complexity derived from the distributed and heterogeneous resources. Moreover, standard service interfaces can implement effective inter-operability in collaborative simulation based on standard protocols. To realize efficient and intelligent search for simulation resources, semantics-based service match technique (Martin *et al.* 2007) is essential to CSP.

The key issues include formal description of service of simulation resource, semantics description of simulation service, semantic service encapsulation approach, simulation service publication, Ontology-based service match method, and semantic composition of simulation service.

5.3. Simulation System Dynamically Construction and Deployment

One of the advantages of CSP is the capability of constructing a simulation environment on demand dynamically and rapidly. This technique facilitates the time-consuming deployment work for complex simulation system to a large degree. In addition, it can

optimize the utilization of simulation resources in run-time simulation along with the fluctuating resource needs.

The key issues include formal description of simulation task requirement, automatic parsing of resource requirement, intelligent match of virtualized resource template, virtualized resource composition optimization, automatic deployment of simulation resource, and virtualized simulation resource instantiation and run-time management.

5.4. Fault Tolerance and Migration in Run-time Simulation

One of the most important targets of CSP is to achieve high reliability, because a variety of failures and errors are inevitable in collaborative simulation progress over unstable networks. CSP should have the ability of fault tolerance to ensure the simulation tasks proceed at the lowest cost once failures occurred. Migration technique offers a mechanism that enables fault tolerance in run-time simulation.

The key issues include run-time simulation monitoring, risk evaluation and fault prediction, run-time simulation failure detection, optimal selection for migration target, migration cost evaluation, live migration in run-time simulation, and post-migration simulation tasks synchronization.

5.5. Utilization Accounting

The distinguishing characteristic of CSP different from other networked M&S platform, such as Grid simulation platform, is that CSP gives support for business transaction between simulation resource consumers and providers. And the transaction billing accords with the actual usage of simulation resources, just like electricity utility charging. The simulation resource utilization accounting technology is the key to accomplish the target.

The key issues include multiple-level accounting model of simulation resource utilization, fine-grained simulation resource utilization statistics, transaction and rate standard management, and user account security management.

5.6. Ubiquitous UI and Remote Virtual Interface

User operating environment is no longer confined to PC desktop in CSP. CSP provides more powerful support for simulation in mobile environments relying on ubiquitous computing and virtualization technology. Users can access CSP by using mobile terminals such as pad computer and smart phone, and customize their virtual UIs to implement remote interaction.

The key issues include ubiquitous terminal adapter standard, interaction context perception and processing, adaptive visualization in small UIs, personalized portal customization, virtual desktop customization and automatic building, and remote interaction and visualization in virtual UIs.

6. CONCLUSION

Networked M&S platform gives powerful support for collaborative simulation applications by means of integrating simulation resources over networks. Facing the new challenge of increasing difficulty in making use

of massive simulation resources in distributed and heterogeneous network environment, users need a way to leverage the rich simulation resources to build complex simulation applications rapidly and simply. This paper introduced the idea of Cloud Computing into networked M&S platform and discussed the key issues in Cloud Simulation Platform. We discussed the operating principle of Cloud Simulation model, CSP architecture, and the key technologies in CSP. The proposed CSP provides a promising approach that users can get simulation services on demand and use simulation resources in a more efficient, transparent, and ubiquitous way. And this paper provides a research map for future research on CSP. We are developing a CSP prototype now and we plan to establish a M&S application of aircraft design on CSP in the future.

ACKNOWLEDGMENTS

The research is supported by the Fundamental Research Funds for the Central Universities in China, and the NSFC (National Science Foundation of China) Projects (No.61074144, No.51005012) in China.

REFERENCES

- Bohu, L., 2007. Research and Application on Virtual Prototyping Engineering Grid. *System Modeling and Simulation*, 2 (15): 304-308.
- Buyya, R., Yeo C.S., Venugopal S., Broberg J., and Brandic I., 2009. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, 25 (6), 599-616.
- Christopher, C., Keir, F., Steven, H., Jacob, G.H., Eric, J., Christian, L., Ian, P., and Andrew, W., 2005. Live Migration of Virtual Machines. *Proceedings of the 2nd ACM/USENIX Symposium on Networked Systems Design and Implementation*, pp. 273-286. Boston (CA, USA).
- Foster, I., Zhao, Y., Raicu, I., and Lu, S., 2008. Cloud Computing and Grid Computing 360-degree compared, *Proceedings of Grid Computing Environments Workshop*, pp. 1-10. Nov. 12-16, Austin (Texas, USA).
- Lei, R., and Lin, Z., 2010. VSim: A Virtual Simulation Framework for High Performance Simulation[C]. *Proceedings of Summer Simulation Multiconference*, pp. 38-44. July 11-15, Ottawa (Ontario, Canada).
- Lei, R., Lin Z., Fei T., Xiaolong, Z., Yongliang, L., and Yabin, Z., 2011. A methodology toward virtualization-based high performance simulation platform supporting multidisciplinary design of complex products. *Enterprise Information Systems*, 5 (3), 1-24.
- Martin, D., Burstein, M., Mcdermott, D., and Mcilraith, S., 2007. Bringing semantics to web services with owl-s, *World Wide Web*, 10 (3), 243-277.
- Mladen, A.V., 2008. Cloud Computing – Issues, Research and Implementations. *Journal of*

Computing and Information Technology, 16(4), 235-246.

- Papazoglou, M., Traverso, P., Dustdar, S., and Leymann F., 2007. Service-Oriented Computing: State of the Art and Research Challenges, *IEEE Computer*, 40(11), 38-45.
- Thomas, A., Larry, P., Scott, S., and Jonathan, T., 2005. Overcoming the Internet impasse through virtualization. *IEEE Computer*, 38 (4), 34-41.
- Zhang, H. and Chen, D., 2006. An approach of multidisciplinary collaborative design for virtual prototyping in distributed environments. *Proceedings of the 10th International Conference on Computer Supported Cooperative Work in Design*, pp. 1-6. May 3-5, Nanjing, China.

AUTHORS BIOGRAPHY

Lei Ren, Ph.D. Dr. Lei Ren received Ph.D. degree in 2009 at the Institute of Software, Chinese Academy of Sciences, China. From 2009 to 2010 he worked at the Engineering Research Center of Complex Product Advanced Manufacturing System, Ministry of Education of China. He is currently a researcher at the School of Automation Science and Electrical Engineering, BeiHang University. He is a member of SCS and SISO. His research interests include high performance simulation platform, integrated manufacturing systems, Cloud Simulation, Cloud Manufacturing and Cloud Computing.

Lin Zhang, Ph.D., Professor. Lin Zhang received M.S. degree and the Ph.D. degree in 1989 and 1992 from the Department of Automation at Tsinghua University, China, where he worked as an associate professor from 1994. From April 2002 to May 2005 he worked at the US Naval Postgraduate School as a senior research associate of the US National Research Council. Now he is a full professor in BeiHang University. His research interests include system modeling and simulation, integrated manufacturing systems, and software engineering.

Yabin Zhang is now a Ph.D. candidate at the School of Automation Science and Electrical Engineering, BeiHang University. His research interests include Cloud Simulation, and Cloud Computing.

Yongliang Luo is now a Ph.D. candidate at the School of Automation Science and Electrical Engineering, BeiHang University. His research interests include service-oriented manufacturing and integrated manufacturing systems.

Qian Li received the B.S. degree and the M.S. degree in 2002 and 2008 from the School of Computer Science and technology at Shandong University, China. She is now a researcher at the School of Information Technology, Shandong Institute of Commerce and Technology. Her research interests include ubiquitous user interface, VR and visualization.

RESEARCH ON KEY TECHNOLOGIES OF RESOURCE MANAGEMENT IN CLOUD SIMULATION PLATFORM

Ting Yu Lin^(a), Xu Dong Chai^(b), Bo Hu Li^(c)

^(a) School of Automatic Science and Electrical Engineering, BeiHang University, Beijing 100191 China

^(b) Beijing Simulation Center, Second Academy of Aerospace Science & Industry Co., Beijing 100854 China

^(c) School of Automatic Science and Electrical Engineering, BeiHang University, Beijing 100191 China

^(a) lintingyu2003@foxmail.com, ^(b) Xdchai@263.net, ^(c) bohuli@moon.bjnet.edu.cn

ABSTRACT

For the diversity and the life-cycle dynamics of the resources on the cloud simulation platform, this paper discuss how to cross and integrate all types of resources management systems to realize the centralized management of distributed resources and then supply services of unifying managed resources to distributed users efficiently. This paper proposed the guideline and architecture of the cloud simulation platform at first. Then, from the perspective of resource operation and running, present the unified simulation modeling to support the dynamic management for the life cycle of the resources. After that, based on the concept of the resource group, this paper proposes the resource selection technology to support the efficient allocation for types of the resources. Finally, this paper introduces an application in the collaborative design and simulation for the multi-disciplinary virtual prototype.

Keywords: cloud simulation platform, resource management, unified modeling, resource selection

1. INTRODUCTION

Cloud Simulation is a service-oriented, intelligent, agile, green, new networkized simulation paradigm. Combining with the emerged information technologies such as cloud computing, service-oriented, virtualization, high performance computing, and developing the existing networkized modeling and simulation (M&S) technology, cloud simulation encapsulates the simulation resource and capability as virtualization and service-oriented forms, and then constructs the cloud service pool of simulation resource and capability so as to implement unified, centralized management and operation, which will support users to access the services of simulation resource and capability on demand at any time through the network for their various activities during the life cycle of simulation. The paper (Bo Hu Li. *et al.* 2009) has fully introduced the technical content, application mode, the architecture and key technologies of the cloud simulation platform.

There are kinds of resources on cloud simulation platform, including physical machine resource (CPU, memory, storage), virtual machine resource, software

resource, licensing resource, model resource, equipment resource and capability resource. The distribute resources are accessed through virtualization middleware, service-oriented middleware and sensing middleware, and appeared as virtual resources, service resources and physical resources to the upper layer, shown in the Figure 1.

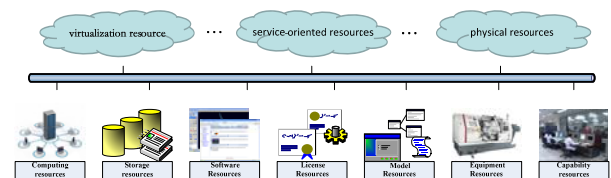


Figure 1: Resources in Cloud Simulation Platform

In terms of virtual resource, both Xen (Barham *et al.* 2003) and VMware (Anderson *et al.* 2005) are popular technology. The virtual machine cluster is managed through VMMs on each nodes and management center on the master node. In terms of service-oriented resources, Apache axis2⁴ is used as service container on each node to deploy and monitor the services, and Platform ego⁵ is used to manage the services of entire cluster. In terms of physical resources, the performance of physical cluster are monitored by Ganglia⁶, and the management of them are implemented by distribute agents.

However, the diversity of resources on the cloud simulation platform makes it difficult to aggregate resources on demand. We need to integrate all types of resource management systems, and take the resources' collaboration into account, to achieve the optimal selection and combination of various resources. Cloud simulation platform uses resource management middleware to achieve the unified management of various resources, and uses dynamic construction modular to achieve the dynamic combination of various resources, so as to solve the problem effectively.

Based on the analysis of related work in Section II, we propose the guideline and the architecture of the resource management in cloud simulation platform in Section III. The resource management of cloud simulation platform involves several key technologies,

including unified modeling techniques, selection techniques, combination technology, and operation/running technology, fault-tolerant technology, evaluation technology and so on. This paper focuses on the modeling technology and selection technology. In section IV, we present the unified formal description model of the simulation resource and resource instance from the perspectives of operation and running. In section V, we present the selection technology based on the resource group. Then, we introduce an application of the resource management in the field of multi-disciplinary virtual prototype. Finally, in section VII, we give a summary and future outlook.

2. RELATIVE WORK

Since 1983, the development of distributed simulation technology has experienced four periods which are Simulator Networking (SIMNET), Distributed Interactive Simulation (DIS), Aggregate Level Simulation Protocol (ALSP) and High Level Architecture (HLA) (Dahmann *et al.* 1998). HLA, whose latest progress is the HLA-Evolved⁸⁻¹⁰, is a popular technology of distributed simulation. It can provide a flexible general-purpose simulation framework for the M&S of complex systems, and can improve the Interoperability and reusability of the simulation model and simulation system. Therefore, it has been widely applied in M&S.

After making good solution of the Interoperability and reusability on the level of simulation model such as components, federates and services, in order to manage and use all kinds of simulation resources better, Grid technology (Foster 2002) is introduced into simulation, which is so called Simulation Grid (Bo Hu Li *et al.* 2006). Simulation Grid, such as Cosim Grid (Bo Hu Li *et al.* 2006), NessGrid (Pearlman *et al.* 2004) and FederationX Grid (Einstein 2005), is a new generation of M&S support system, combined with the new network technologies such as Internet technology, Web Service technology and Grid technology and the traditional M&S support technologies such as HLA technology. In the simulation grid, the simulation models are packaged as loose coupling simulation services with interoperability deployed and shared on the grid node, and are dynamically discovered and invoked in the runtime. As a result, the traditional pattern of chimney-like development and resource utilization are broken, and establish new resource management and utilization patterns of autonomous, dynamic sharing and on-demand collaborative.

However, from the view of application, current simulation grid services are almost fixed on some grid nodes, rather than created resource copy flexibly. In addition, current simulation grid services cannot penetrate the underlying hardware facilities so as to share fine-grained resources, including CPU, memory, software, etc, and it is difficult to satisfy multi-users to access all kinds of M&S services through the Internet anytime, anywhere. Again, the heterogeneous and loose environment of simulation grid cannot provide the

security and high availability of operating environment for the upper applications.

With the rapid development of virtualization technologies recently, we can shield the differences of the hardware architecture, operating environment and other simulation resource, and uncouple the coupling relationship between them, making the construction and operation of simulation systems more flexible; we can provide a unified encapsulating standard for kinds of the heterogeneous simulation resources, making share and transparent use of simulation resources as much as possible; we can migrate and expand simulation operating environment dynamically and efficiently, so as to enhance the reliability and stability of the simulation runtime environment. After merging the latest technology of information field, such as virtualization technology, into the simulation grid, we have developed Cloud Simulation, which makes a better solution to the shortcomings of the simulation grid.

It is known that the cloud simulation is the further development of the simulation grid, therefore, some technology of resource management in the simulation grid can be learned and continue to be adopted. Chang Feng Song, *et al.* (2009) proposes a resource selection model and algorithm in the environment of the simulation grid. For the "All-to-All" task model that any federate in the distribute interactive simulation possibly interacts with any other federate, the paper aims to select the grid node with CPU frequency as large as possible, communication delay as small as possible, communication speed as quick as possible from the grid nodes contained the required grid services. The work presents a Resource Selection Model (RSM) in matrix, and then uses various intelligent optimization algorithms to select a set of optimal resource instances from eligible simulation resources.

However, since there are subtle changes in the resource form, the resource management of cloud simulation platform needs to fine-tune. For example, in the past, the task was to select M optimal nodes from N candidate nodes contained the required resource services ($N > M$), however, one node can use virtualization technology to build multiple copies of resource services at present. In the past, the required resource services were just fixed on N (finite) candidate nodes, however, the number of candidate nodes now can be much larger than N, since the introduction of virtualization technology which can build virtual nodes dynamically. These differences are related to the adjustment of the unified modeling technology and the selection technology for the simulation resources, which is the focus of this paper.

3. GUIDELINE AND ARCHITECTURE OF RESOURCE MANAGEMENT

Before discussion the simulation resource management technology about unified modeling and selection of resources in detail, we need first to introduce the guideline and the architecture of cloud simulation platform.

The guideline of cloud simulation platform is shown as follow:

- Centralized management of distributed resources

Fully integration of networked and shared computing resources, storage resources, software resources, License resources, knowledge / model resources, equipment resources and capability resources conducts unified management;

- Supplying services of unifying managed resources to distributed users

The aim of system is to organize simulation resources and capabilities quickly and flexibly, so as to provide services transparently on demand at anytime and anywhere.

- Effective collaboration of resources on the life cycle of M&S

The resource management covers the whole life cycle of M&S to support four application patterns, which include multi-users complete the design and analysis tasks independently, multi-users complete the simulation tasks collaboratively, multi-users complete the muti-analysis process collaboratively and multi-users access simulation capability, which is the combination of the intellectual resource and the traditional simulation resource, on demand. This paper does not take the fourth pattern into consideration.

The architecture of cloud simulation platform is shown in the Figure2.

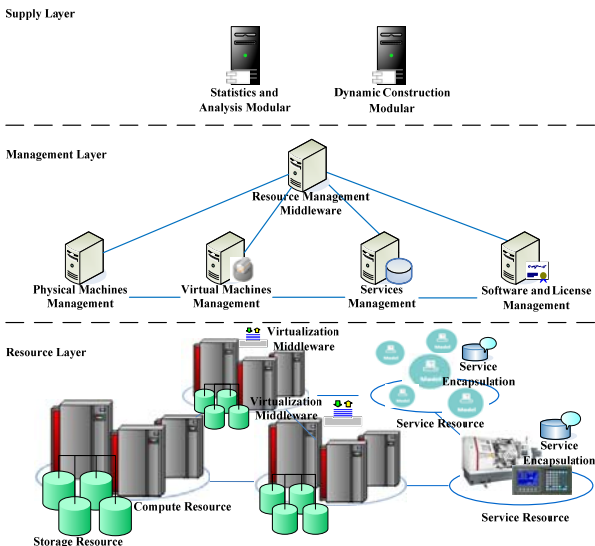


Figure 2: Architecture of Cloud Simulation Platform

There are three layers in the architecture, which are Supply Layer, Management Layer and Resource Layer.

Resource layer includes distributed computing resources, storage resources, service resources, which including service-oriented simulation software, simulation models, simulation equipments, and their license resources. Parts of the cluster is deployed virtualization middleware and become virtual computing resources so as to build virtual computing nodes on demand, other parts of the cluster is deployed

monitor agents and so called physical computing resources.

Management layer includes physical machine management, virtual machine management, service management, software and their License Management. They will be integrated by resource management middleware, which is deployed on a dedicated server, so as to manage resource uniformly. With the expansion of the resources' scale and terrain, the deployment of the resource management middleware will be distribute, however the pattern of management and allocation of the resources keeps still.

Supply layer includes statistics and analysis modular and dynamic construction modular, which both deploy on the dedicated servers respectively. Dynamic construction modular prepares hardware, software and service environment ready for the upper application, and statistics and analysis modular provides the reporting service about resources for the users.

From the architecture of cloud simulation platform mentioned above, we can summarize that the resource management of cloud simulation platform covers physical machine resource management, virtual machine resource management, service resource management, software resource management and license management, shown in the Figure 3. In essence, license resource management subordinates to software resource management.

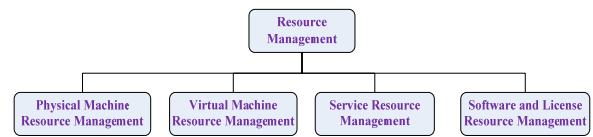


Figure 3: Types of Resource Management

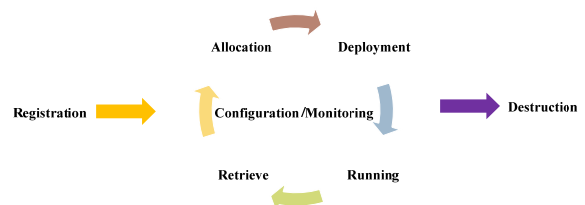


Figure 4: Life Cycle of Resource Management

In order to achieve the centralized management of distributed resources, as well as supply services of unifying managed resources for distributed users, resource management of cloud simulation platform covers the entire life cycle of resource, includes the unified registration, flexible configuration, real-time monitoring, on-demand allocation, efficient deployment, transparent running, timely retrieve and safe destruction for all kinds of simulation resources, shown in the Figure 4.

4. UNIFIED MODELING OF SIMULATION RESOURCES FOR OPERATION AND RUNNING

To support the unified management for the life cycle of various simulation resources, we first need to model all

kinds of simulation resources on the cloud simulation platform uniformly. Generally, there are two perspectives of unified modeling, the one is resource operation, and the other is resource running. The former concerns about the resource's grouping, ownership, availability and state of allocation; the later describes the static configuration, dynamic performance and running state of the whole life cycle for kinds of simulation resources.

Figure 5 shows the formal description model of simulation resource and resource instance, as well as the relationships between them, from the perspective of resource operation. The resource attributes include resource ID, resource name, resource type and set of resource instances. Resource type can be divided into physical machine resource (identified as HPC), virtual machine resource (identified as VM), service resource (identified as SVR), and software resource (identified as SFW), from the analysis mentioned above. The set of resource instances achieve the association between resource and resource instances. A simulation resource includes a number of resource instances, and can be reverse indexed by any resource instance.

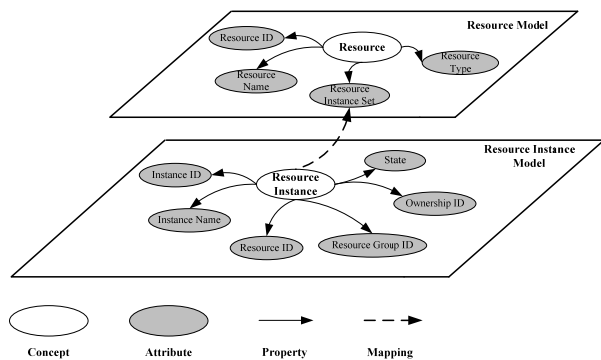


Figure 5: Formal Description Model from the perspective of Resource Operation

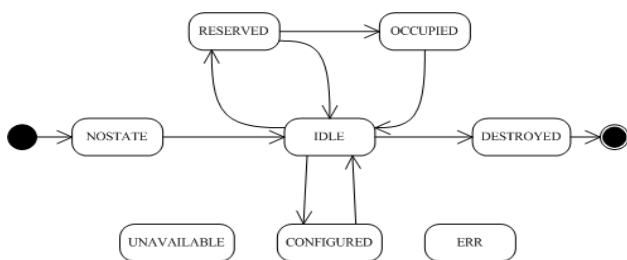


Figure 6: State Transition between Resource Instances

In addition, the resource instance attributes include instance ID, instance name, resource ID, resource group ID, ownership ID and state of allocation. Resource group ID is an important basis for the allocation of resources. Resource instances contacted with each other closely and executed collaboratively will be assigned to the same resource group. Ownership ID is an important symbol described the occupancy status and availability of the resource instance. From the perspective of resource operation, the state of allocation is an important property of the dynamic management for the

life cycle of resource instances, including NOSTATE, IDLE, CONFIGURED, RESERVED, OCCUPIED, UNAVAILABLE, ERROR and DESTROYED. The conversion between the various states is shown in Figure6.

Figure 7 shows the formal description model of simulation resource and resource instance, as well as the relationships between them, from the perspective of resource running. The resource attributes include basic information (such as resource ID), the character information (such as operating system type), resource type and set of resource instances. The resource instance attributes include static configuration (such as the number of CPU), dynamic performance (such as CPU utilization) and running state of the whole life cycle. Similarly, the set of resource instances achieves the association between resource and its instances. A simulation resource includes a number of resource instances, and can be reverse indexed by any resource instance.

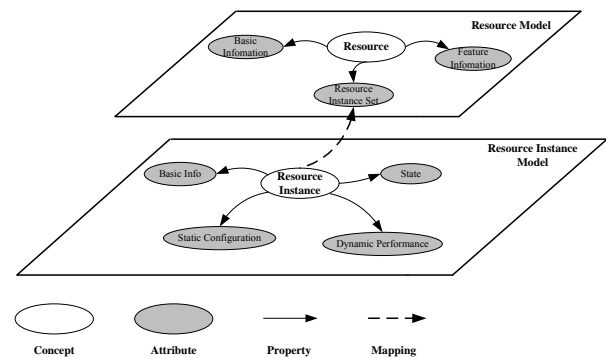


Figure 7: Formal Description Model from the perspective of Resource Running

In fact, there are subtle differences for different types of simulation resources and resource instances in their formal description models, from the perspective of resource running. There are detailed definitions of resource and resource instance for the physical machine, virtual machine, service and software respectively, shown as follow.

Definition I: Physical Machine Resource and Resource Instance

Physical Machine Resource
 < Basic Information <Resource ID, Resource Name>, Character Information <Is Cluster, OS Type> >
Physical Machine Node
 < Basic Information <Node ID, Node Name, Resource ID>, Static Configuration <Node IP, CPU Num, CPU Speed, Memory Size>, Dynamic Performance <CPU Utilization, Memory Utilization>, State {HPC_NOSTATE, HPC_INI, HPC_RUNNING, HPC_SHUTOFF, HPC_ERR}>

Definition II: Virtual Machine Resource and Resource Instance

Virtual Machine Resource

< Basic Information <Resource ID, Template Name>, Character Information <OS Type, Software List> >
Virtual Machine Node
< Basic Information <Node ID, Node Name, Resource ID>, Static Configuration <Node IP, CPU Num, Memory Size, Host ID>, Dynamic Performance <CPU Utilization, Memory Utilization>, State {VM_NOSTATE, VM_INI, VM_RUNNING, VM_PAUSE, VM_BLOCKED, VM_SHUTOFF, VM_ERR }>

Definition III: Service Resource and Resource Instance

Service Resource
< Basic Information < Resource ID, Resource Name >, Character Information < Is Parallel, OS Type > >
Service Instance
< Basic Information <Instance ID, Instance Name, Resource ID >, Static Configuration < UDDI, CPU Num, Memory Size, Host ID >, Dynamic Performance < >, State {SVR_NOSTATE, SVR_INI, SVR_IDLE, SVR_RUNNING, SVR_PAUSE, SVR_SHUTOFF, SVR_ERR}>

Definition IV: Software Resource and Resource Instance

Software Resource
< Basic Information < Resource ID, Resource Name >, Character Information < Is Parallel, OS Type > >
Software Instance
< Basic Information <Instance ID, Instance Name, Resource ID >, Static Configuration < UDDI, CPU Num, Memory Size, Host ID (List)>, Dynamic Performance < >, State {SFW_NOSTATE, SFW_INI, SFW_RUNNING, SFW_PAUSE, SFW_SHUTOFF, SFW_ERR}>

5. RESOURCE SELECTION TECHNOLOGY BASED ON RESOURCE GROUP

In the unified modeling of simulation resources mentioned above, we introduce the concept of resource group. Administrators can classify the simulation resource instances closely as a resource group. For example, a virtual machine template was deployed in a high performance cluster with shared storage, and then a virtual machine can be start up at any node of the cluster according to the mission requirement. Then the high performance cluster with the virtual machine template can be classified as a resource group. As another example, considering the security, the providers of some simulation models or services usually do not allow them migrating and being deployed free on the simulation platform, so as to isolate and publish them as the form of SOA. Then any of the simulation models or services can be classified as a separate resource group to reflect its isolated feature.

In general, there is high communication efficiency in a resource group, such as the high performance

cluster inside which the point to point communication is based on infiniband, so that the communication performance can almost always meet with the requirement of the collaborative simulation. Therefore, for the resource selection in a resource group, which usually refers to select suitable physical machines to deploy simulation models or startup virtual machines, we focus on considering the performance of each node from the perspective of CPU num and memory size. In the resource group, the algorithm of selecting and allocating the resources for collaborative simulation was shown as follow with the form of pseudo-code.

Variable Declaration

rCPUNum: the requirement of the cpu num, which is an integer

rMEMSize: the requirement of the memory size, which is an integer

rIndex: the requirement of the computing performance, which is a real number

pCPUNum, *pCPUSpeed*, *pCPUUtil*: the cpu number, its speed and its utilization in the physical compute node

pMEMSize, *pMEMUtil*: the memory size and its utilization in the physical compute node

pCPUIndex: the available performance of the cpu, which is a real number

pMEMIndex: the available performance of the memory, which is a real number

pIndex: the available performance of the computing, which is a real number

Selection Algorithm

For $i = 1$ to M // M is the requirement number of the computer node

$rIndex[i] = \alpha_1 * rCPUNum[i] + \alpha_2 * rMEMSize[i] // \alpha_1$ and α_2 are the weighting factors

EndFor

Rearrange the array *rIndex* in descending

For $j = 1$ to N // N is the candidate number of the computer node

$pCPUIndex[j] = pCPUNum[j] * pCPUSpeed[j] * (1 - pCPUUtil[j])$

$pMEMIndex[j] = pMEMSize[j] * (1 - pMEMUtil[j])$

End For

For $i = 1$ to M

For $j = 1$ to N

If $(rCPUNum[i] * pCPUSpeed[j] < pCPUIndex[j])$

Then $pIndex[j] = \beta_1 * pCPUIndex[j] // \beta_1$ is the weighting factor

```

ELSE pIndex[j] = β1 * pCPUIndex[j] * δ1 //δ1 is the
punishment factor
End If
If (rMEMNum[i] < pMEMIndex[j])
Then pIndex[j] = pIndex[j] + β2 * pMEMIndex[j] //β2
is the weighting factor
ELSE pIndex[j] = pIndex[j] + β2 * pMEMIndex[j] *
δ2 //δ2 is the punishment factor
End If
End For
Select the node j, the pIndex of which is max, as the target
node for the requirement i
pCPUIndex[j] = pCPUIndex[j] - rCPUNum[i] *
pCPUSpeed[j]
pMEMIndex[j] = pMEMIndex[j] - rMEMNum[i]
End For

```

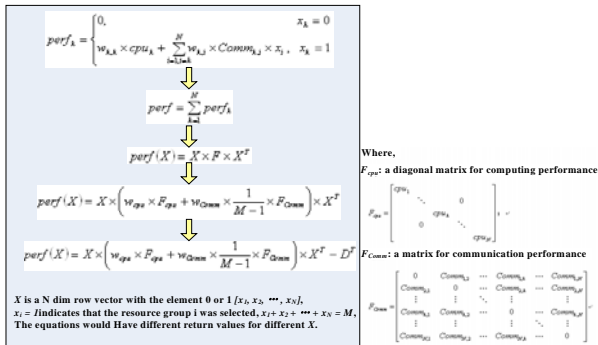


Figure 8: Resource Selection Model for Selecting Resource Groups

In the cloud simulation platform, every user corresponds to a prior resource group for utilization according to his identity, which is configured by the administrators. When the amount or performance of specific resource in the resource group can not satisfy the requirement of collaborative simulation task for a user, cloud simulation platform will find the specific resource in other resource groups. However, the cloud simulation platform has the feature of large-scale virtualization, the communication between resource groups may be on the WAN or on the Internet. At this situation, communication efficiency should be considered as an important indicator. If there are suitable resource instances in more than one resource groups to meet with the remaining simulation requirement, which cannot be satisfy in the prior resource group, both computing performance and communication performance need to be considered. The RSM (Chang Feng Song, et al. 2009) in simulation grid mentioned above may be learned and continue to be adopted to select the resource group, shown in the Figure 8. What we need to do is just setting the weights of communication performance between the resource

groups instead of the grid nodes. After selecting a resource group, we can use the above method to select specific resource instances in the resource group for the simulation task.

6. APPLICATION

The resource management technologies of cloud simulation platform described in this paper have played an important role in the field of multi-disciplinary virtual prototype engineering in their preliminary application. Based on the cloud simulation platform COSIM-CSP (Bo Hu Li *et al.* 2009), the application about collaborative design and simulation of the landing gear system of an aviation aircraft is shown as follow.

The multi-disciplinary virtual prototype of the landing gear system consists of several sub-system models, such as electronic control model, multi-body dynamics model and hydraulics model and so on. These models refer to various commercial software, such as control system design/simulation tools MATLAB / SIMULINK, dynamics system design/simulation tools MSC ADAMS, hydraulics system design/simulation tools MSC EASY5 and structure design tool CATIA. The top-level system model of the virtual prototype constructed by the COSIM-CSP was shown in Figure9.

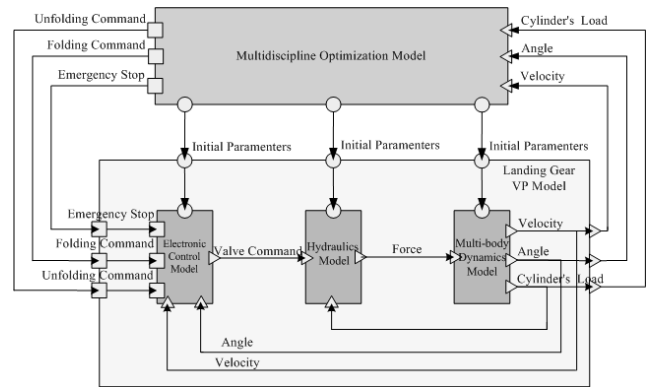


Figure 9: Top-level System Model of the Virtual Prototype Constructed by the COSIM-CSP

According to the formal description of unified modeling mentioned above, we can describe the requirements of various sub-system models of the virtual prototype as follow.

```

<?xml version="1.0" encoding="UTF-8" ?>
<requirement>
  <vmNode>
    <modelName> Electronic Control Model </modelName>
    <osType> Windows XP </osType>
  </vmNode>
  <softwareList>
    <softwareName> MATLAB/SI MULINK </softwareName>
  </softwareList>
  <cpuNum> 1 </cpuNum>
  <memSize> 1 </memSize>
</vmNode>
  <vmNode>
    <modelName> Hydraulics

```

```

    Model</modelName>
    <osType>CentOS5.4</osType>
    = <softwareList>
        <softwareName>MSC
            EASY5</softwareName>
        </softwareList>
    <cpuNum>1</cpuNum>
    <memSize>1</memSize>
</vmNode>
= <vmNode>
    <modelName>Multi-body Dynamics
        Model</modelName>
        <osType>Windows XP</osType>
        = <softwareList>
            <softwareName>MSC
                ADAMS</softwareName>
            </softwareList>
        <cpuNum>1</cpuNum>
        <memSize>1</memSize>
    </vmNode>
= <vmNode>
    <modelName>Structure
        Model</modelName>
        <osType>Windows XP</osType>
        = <softwareList>
            <softwareName>CATIA</softw
                areName>
            </softwareList>
        <cpuNum>2</cpuNum>
        <memSize>2</memSize>
    </vmNode>
</requirement>

```

service; model resource includes top-level model, the electronic control model, multi-body dynamics model, the hydraulics model, structure model.

In the process of implementation, the overall system designer and the sub-system designers complete the top-level system modeling and sub-system design respectively in their own virtual desktops, which are the research environment customized, and then upload the corresponding model files through the application portal of COSIM-CSP. Based on the resource management technologies mentioned above, COSIM-CSP constructs the running environment of the collaborative simulation dynamically, and builds the system-level virtual prototype automatically, so that achieving the aggregation and collaboration of the simulation resources. The process is shown in the Figure 11.

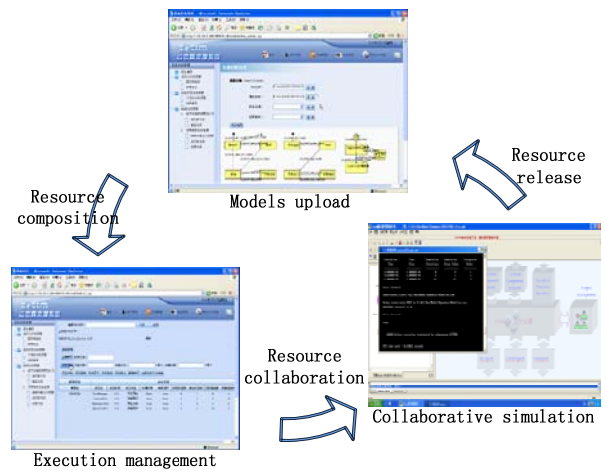


Figure 11: The Process of the Implementation

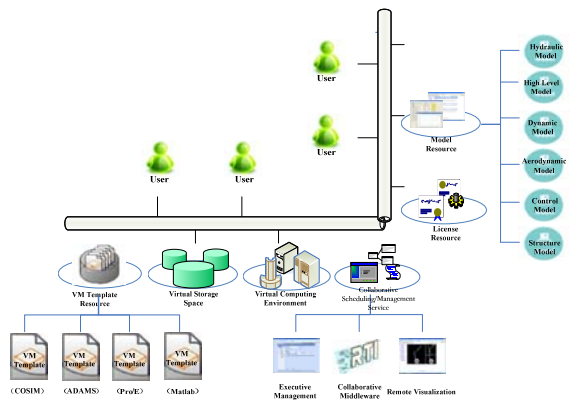


Figure 10: Virtual Resource Environment

In the COSIM-CSP, the overall system designer and the sub-system designers can see the virtual resource environment shown in the Figure 10, including virtual computing environment, virtual storage space, collaborative scheduling / management service, virtual machine template resource, licensing resource and model resource. Specifically, virtual machine template resource includes templates installed COSIM top-level modeling software, MATLAB / SIMULINK software, MSC ADAMS software, MSC EASY5 software, CATIA software respectively; cooperative scheduling / management service includes execution management tool, collaboration middleware, remote visualization

7. CONCLUSION AND FUTURE WORK

The research of this paper proposes the guideline and the architecture for the resource management of the cloud simulation platform. In this framework, in order to support the centralized management for the life cycle of all kinds of simulation resources, we present the formal description of the unified modeling from the perspectives of resource operation and resource running. In the unified modeling, we propose clearly the concept of resource group, so as to distinguish the selection methods in the resource group and inter-groups. The resource management technologies of cloud simulation platform described in this paper have been verified in the collaborative design and simulation process for the multi-disciplinary virtual prototype.

In future work, we will do more research on simulation performance evaluation, and enhancing the flexibility of resource selection, which will select the most appropriate resource instances instead of best performance ones. In addition, we need further research on the on-demand aggregation and high efficient collaboration for the simulation capabilities.

ACKNOWLEDGMENTS

This paper is supported by the National 973 plan (No. 2007CB310900).

REFERENCES

- LI Bo-hu, CHAI Xu-dong, HOU Bao-cun, et al., 2009. A Networked Modeling & Simulation Platform Based on the Concept of Cloud Computing – “Cloud Simulation Platform”. *Journal of System Simulation* 21(17): 5292-5299 . (in Chinese)
- Barham, P., et al., 2003. Xen and the art of virtualization. *Proceedings of the 19th ACM Symposium on Operating Systems Principles*, 164-177. Oct. 2003, New York USA.
- Anderson, T., et al., 2005. Overcoming the Internet impasse through virtualization. *IEEE Computer* 38 (4): 34-41.
- Apache, 2009. *Axis2/Java - Next Generation Web Services*. Available from: <http://ws.apache.org/axis2/>[2010].
- Platform, 2007. *Platform EGO white papers*. Available from: <http://platform.com/Products/platform-enterprise-grid-orchestrator/whitepapers>[2010].
- Ganglia, 2010. *Ganglia Monitoring System*. Available from: <http://ganglia.sourceforge.net/>[2010].
- Dahmann, J., et al., 1998. The DoD High Level Architecture: An Update. *Proceedings of the 1998 Winter Simulation Conference*, 797-804. 10 Dec. 1998, Washington, D. C., USA
- IEEE, 2008. *Draft Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)-Object Model Template (OMT) Specification*. Available from: <http://ieeexplore.ieee.org/servlet/opac?punumber=4478265> [2008].
- IEEE, 2009. *Draft Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)-Framework and Rules*. Available from: <http://ieeexplore.ieee.org/servlet/opac?punumber=5347324> [2009].
- IEEE, 2009. *Draft Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)-Federate Interface Specification*. Available from: <http://ieeexplore.ieee.org/servlet/opac?punumber=5347330> [2009].
- Foster, I., 2002. The Grid: A New Infrastructure for 21st Century Sciece. *Physical Today* 55 (2): 42-47.
- Bo Hu Li, Xudong Chai, Baocun Hou, et al., 2006. Research and Application on CoSim (Collaborative Simulation) Grid. *Proceedings of MS-MTSA2006*, 156-163. July 2006, Alberta, Canada.
- Pearlman, L., et al. 2004. Distributed Hybrid Earthquake Engineering Experiments: Experiences with a Ground-Shaking Grid Application. *Proceedings of 13th IEEE Symposium on High Performance Distributed Computing*, 14-23. June 2004, Hawaii, USA.
- Einstein, A. 2005. FederationX: A Technical Brief. Available from: <http://www.magnetargames.com/>.
- Chang Feng Song, Bo Hu Li and Xudong Chai, 2009. Node selection in simulation grid. *Journal of Beijing University of Aeronautics and Astronautics* 35(1): 56-60. (in Chinese)

AUTHORS BIOGRAPHY

Ting Yu Lin was born in 1984. He received his B.S.degree in BeiHang University. He is currently a Ph.D. candidate at the School of Automatic Science and Electrical Engineering, BeiHang University, Beijing, China. His research interests include multi-disciplinary virtual prototype and intelligent distributed simulation.

Xudong Chai was born in 1969. He is a researcher and deputy director at Beijing Simulation Center of Second Academy of Aerospace Science & Industry Co. and council members of Chinese System Simulation Association and National Standardization Technical Committee. His research interests include automatic control and simulation.

Bo Hu Li was born in 1938. He is a professor at School of Automatic Science and Electrical Engineering, BeiHang University, and Chinese Academy of Engineering, and the chief editor of “Int. J. Modeling, Simulation, and Scientific Computing”. His research interests include multi-disciplinary virtual prototype, intelligent distributed simulation and cloud manufacturing.

A FRAMEWORK FOR ENHANCED PROJECT SCHEDULE DESIGN TO AID PROJECT MANAGER'S DECISION MAKING PROCESSES

Sanja Lazarova-Molnar^(a), Rabeb Mizouni^(b), Nader Kesserwan^(a)

^(a) Faculty of Information Technology, United Arab Emirates University, United Arab Emirates

^(b) Department of Software Engineering, Khalifa University, United Arab Emirates

^(a) sanja.nkesserwan@uaeu.ac.ae, ^(b) rabeb.mizouni@kustar.ac.ae

ABSTRACT

Good schedule increases the chances of a project meeting its goals. The most popular formalisms for describing project schedules are very rigid and inflexible in modeling changes due to uncertainties. In this paper we describe a framework to support enhanced project schedule design. The proposed framework is based on the Enhanced Project Schedule (EPS) model. In addition to an initial Gantt Chart, EPS allows definition of Remedial Action Scenarios (RAS), which contain guidelines of actions to consider when uncertainties arise. This creates a dynamic and evolving schedule. It is meant to guide the project manager in the decision making processes throughout the project implementation. The process of selection of the remedial action scenario is an optimization one, based on simulation. We illustrate the dynamics of the EPS design framework by an example.

Keywords: project schedule, proxel-based simulation, remedial action scenarios, uncertainty

1. INTRODUCTION

During the past few decades project management has evolved into a discipline that studies planning, scheduling and controlling of activities that directly contribute to the achievement of project's objectives.

The pressure of time-to-market along with the increasing complexity of present-day projects, have contributed project management to become one of the main factors for projects success. A growing number of companies use various advanced project management tools and methods to ensure the project quality expected by customers, delivered within reasonable deadlines and at the lowest possible cost.

Many attempts have been conducted to improve the project scheduling prediction (Herroelen and Leus 2005; Arauzo, Galán et al. 2009; Huang, Ding et al. 2009; Jing-wen and Hui-fang 2009; Sobel, Szmerekovsky et al. 2009). Many of them are based on analytical models and simulation. Tools, such as Microsoft Project and Primavera Project Planner, are typically suggested to help managers in planning and controlling their projects. Existing frameworks and methods, however, fail, or, are insufficient; to answer

the real needs of a project. The models developed still suffer from many limitations that often make them not representative to real world situations. Typically, project schedules are described in very strict terms, using Gantt charts or PERT.

In real life, even small projects face risks and may, consequently, deviate from their original plans. As a consequence, even good projects can fail (Matta and Ashkenas 2003). For instance, in the software industry it has been reported (Denning and Riehle 2009) that approximately one-third of software projects fail to deliver anything, and another third deliver something workable but not satisfactory. In order to have a more realistic and effective project scheduling, management frameworks need to incorporate uncertainties on the one hand, and guide the managers to what actions to take when such uncertainties arise. This is the issue that we address in our paper, i.e. to lay out a strategy to create an *optimal enhanced project schedule*. Our objective is to answer the needs of managers by providing a framework that helps the generation of a *more realistic and insightful* project planning.

The proposed framework supports flexible and efficient project schedule modeling and simulation. It combines: (a) a novel model for describing project schedules in a more realistic way, accommodating uncertainties, and (b) facilities for model's simulation and assessment with respect to predetermined project goals. The objective of the framework is to provide managers with answers to the following types of questions:

- 1) What is the best Remedial Action Scenario (RAS) to adopt if some uncertainties arise during the implementation of the project?
- 2) What are the features that can be implemented within the deadline of the projects?
- 3) What are the best and robust deadlines to consider that take into consideration the deviation from the original scheduling because of uncertainties.

As shown in Figure 1, our framework is based on two main modules, and two supporting ones:

- 1) Multi-RAS EPS Proxel-Based Simulator is a simulator based on the proxel-based simulation method.
- 2) Result Visualization Module: responsible for visualizing and interpreting the results of the simulation with respect to the goals specified by the manager.
- 3) User Interface Module that supports and facilitates the input of project schedules.
- 4) A Data Storage Module that manages project schedule data.

The rest of the paper is organized as follows. In the next section we describe the EPS model. In Section 3 we present the framework that supports the generation of EPS models and we describe the different modules, focusing on the key ones. Section 4 demonstrates the idea of the framework by an example. Finally, Section 5 concludes the paper.

2. WHAT IS AN ENHANCED PROJECT SCHEDULE?

To illustrate the concept of an enhanced project schedule, we provide an example that displays side-by-

side a standard Gantt chart and an EPS, as shown in Figure 2. For comparison, Figure 2(a) illustrates a simple project schedule, modeled using a classical Gantt chart. The project schedule consists of four tasks (Task1, Taks2, Task3, and Task4) and two available teams (Team A and Team B). All tasks have predefined executors leading to one possible scenario of execution. Such model is in fact rigid and it is not able to anticipate the occurrence of any unpredictable events.

Figure 2(b) illustrates the EPS. While having the same number of tasks and teams, two majors features are added:

- 1) “floating task” (Task 2), which is a non-vital task that can be executed by any of the two teams, albeit with different duration distribution functions (based on teams’ expertise).
- 2) fuzzily described guidelines, provided below the schedule, which are meant to accompany the project schedule as RAS (remedial action scenario).

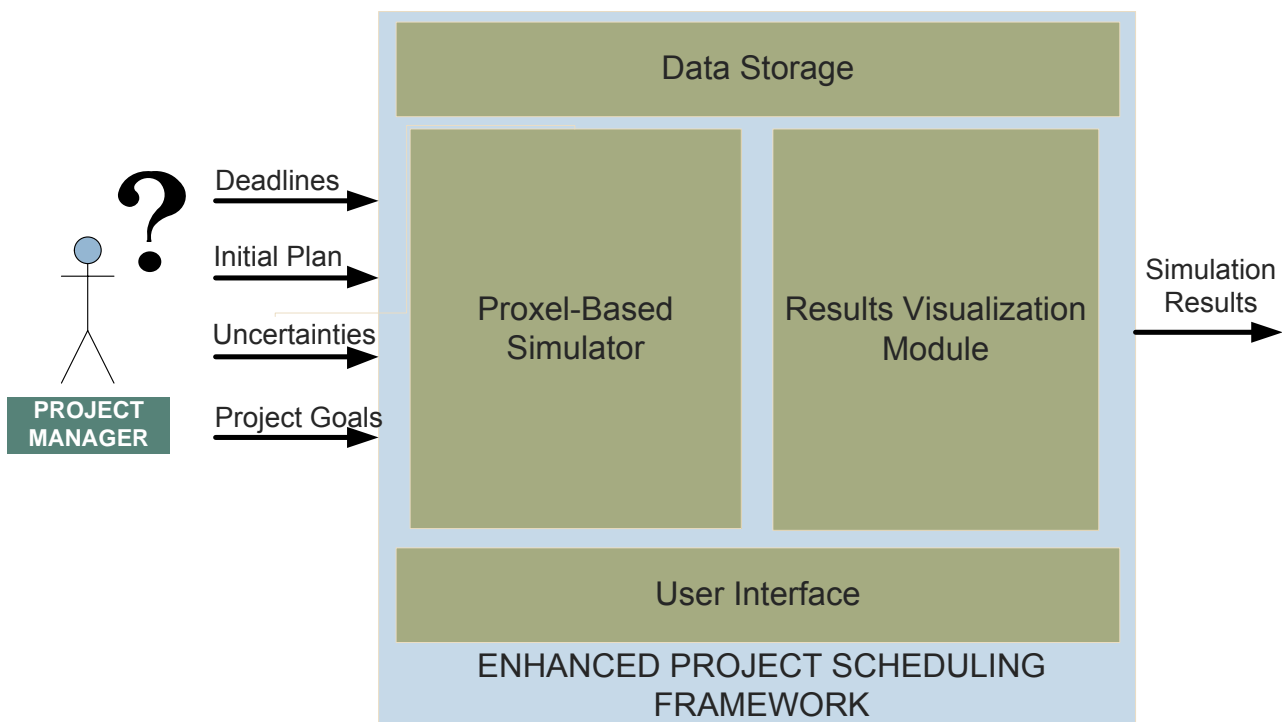
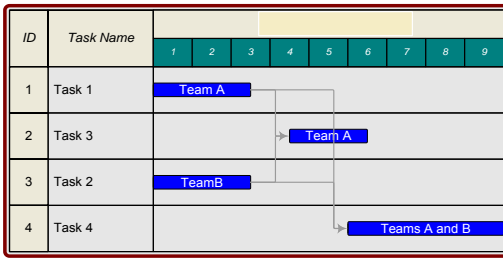
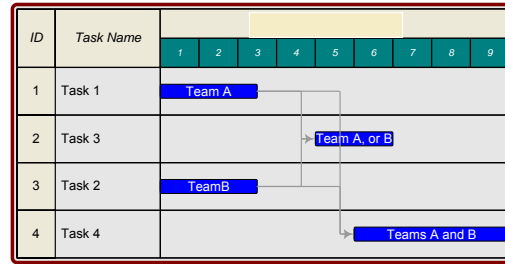


Figure 1: EPS Design Framework Architecture

(a) No additional rules



(b) EPS with a Remedial Action Scenario



1. If duration of Task 2 performed by team B is “very short” then start Task 3 by team B.
2. If duration of Task 1 is “too long” and it completes “shortly after” team B started to work on Task 3, then Task 3 is cancelled and both teams start working on Task 4.

Figure 2: Illustration of: a) Classical Gantt Chart and b) Enhanced Project Schedule with RAS

In our previous work (Lazarova-Molnar and Mizouni 2010; Lazarova-Molnar and Mizouni 2010) we successfully modeled and simulated the type of scenarios described in Figure 2(b). There, we also developed an approach to analyze and simulate the effects of the uncertainties and remedial actions on the duration of project. As expected, on-the-fly decisions make a significant difference in the duration of the project and need to be considered and, if possible, pre-determined. To account for resource re-allocations we have also defined a new type of tasks, which we termed as “floating task”. This task was a typically a non-crucial task for the success of the project, which could be implemented by a number of teams, albeit with different duration distribution functions, and based on their availabilities.

2.1. EPS Model Description

We propose the definition of a schedule to include the uncertainties that can arise and their quantification using statistical probability distributions. In addition to this, we formalize the remedial actions that managers can take. Every schedule along with the set of remedial actions (RAS) creates what we term as: *enhanced project schedule* (as shown in Figure 3). The RAS consists of a set of fuzzy *if-then* production rules. These rules make the project evolving and thus, the sequencing of tasks, dynamic and changing. Once the enhanced project schedule is designed, we simulate each RAS using the proxel-based method and pick the best one based on the success criteria for the project. “The probability that the project is delivered before deadline” and “the probability that the project is implemented within this budget” are examples of success criteria.

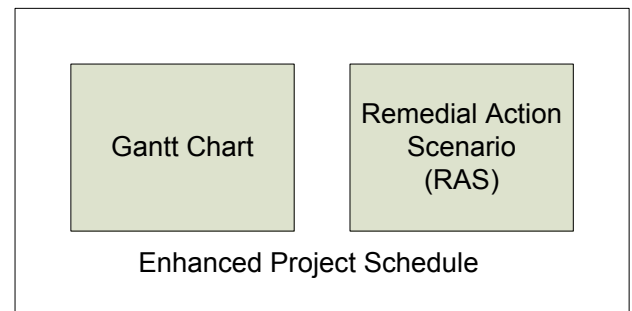


Figure 3: Enhanced Project Schedule Components

Let us take the example of a simple software development project schedule, subject to various uncertainties. The enhanced schedule would consist of a Gantt chart, where each task that corresponds to a requirement implementation, is associated with a probability distribution function for its duration, as well as a set of fuzzy rules that describe the remedial strategy under certain conditions (e. g. if task A finishes in a very short time than proceed to task B, else skip task B). The set of fuzzily specified guidelines are obtained by simulating a set of possible RAS, and accordingly selecting the most optimal one (similar to the simple example presented in Figure 2(b)).

Once the simulation of the chosen RAS provides good results with respect to project goals (e.g. complete as many tasks as possible, complete in as short time as possible, or minimize budget), the resulting EPS is communicated to the project manager to aid his/her decision making process. We see the fuzziness as a great advantage as it leaves a certain degree of freedom to the project manager as well, to involve his/her knowledge/perceptions he/she might have.

The proxel-based simulation method allows for a great flexibility in schedule description and provides solutions taking into account anticipated uncertainties. This helps us in picking the best RAS that specifies the

optimal set of recommended remedial actions when uncertainties occur.

2.2. EPS Formal Model

An Enhanced Project Schedule (EPS) is described as follows:

$$EPS = (A, P, T, D, W, F, IGC)$$

- $A = \{A_1, A_2, \dots, A_n\}$, set of tasks, where each task corresponds to a task in the project schedule,
- $P = \{P_1, P_2, \dots, P_m\}$, set of precedence constraints, that are actually tuples of two tasks where the completion of the first one is a pre-requirement for commencing the second one, e.g. (A_x, A_y) would mean that completing of A_x is a pre-requirement for beginning A_y ,
- $T = \{T_1, T_2, \dots, T_l\}$, set of teams available for the execution of the project,
- $D = \{d_1, d_2, \dots, d_s\}$, set of probability distribution functions that correspond to duration of tasks performed by the competent teams,
- $W = \{w_1, w_2, \dots, w_t\}$, set of mappings of distribution functions to competent teams and tasks,
- $F = \{f_1, f_2, \dots, f_r\}$, set of fuzzy rules that define the remedial action scenario,
- IGC – *Initial Gantt Chart*, initial sequencing of tasks that satisfies the set of precedence constraints provided by P ,

where $P \subseteq A \times A$, and $W \subseteq A \times M \times D$. Also, $A = A^c \cup A^n$, where A^c is the set of cancelable tasks and A^n is the set of non-cancelable tasks. Cancelable task is a task that is non-vital for the success of the project, and thus, not compulsory, however, useful for the value of the project. Non-cancelable tasks are the ones that are crucial for the success of the project. This differentiation is important for the realistic simulation of project schedules.

Each fuzzy rule is made up of two parts: *condition* and *action*, formally expressed as “*condition* \Rightarrow *action*”. Conditions can be described either by using strict terms, or fuzzy ones. An action can typically be canceling or interrupting some of the tasks, or one of the various types of rescheduling. This is the fact that makes our schedule description evolving, rather than rigid and inflexible. Two examples of fuzzy rules are:

$$A_x \text{ takes too long} \Rightarrow \text{cancel } A_y$$

or

$$A_x \text{ completes quickly after } A_y \Rightarrow \text{cancel } A_z.$$

Both are examples for typical proceedings during project execution. However, in our approach we formalize their modeling, assessment and quantitative

evaluation. This makes it straightforward to compare various RAS, as well as test for the best RAS to counteract uncertainties, as described by F . Note that F can be an empty set too, which would imply sticking to the original project schedule provided by IGC . Once an optimal remedial action scenario is selected, it is associated with the initial project schedule and handed to the project manager as a decision making aid. The goal of the proposed framework is to support the generation of the EPS. This process is further demonstrated by a simple example in Section 4.

2.3. Multi-RAS EPS

To facilitate the generation of the optimal EPS, the framework needs to analyze a number of various RAS that could potentially accompany a given EPS. For this purpose, we define the term Multi-RAS EPS (MEPS) as follows:

$$MEPS = (A, P, T, D, W, F', IGC),$$

where the only difference to the standard EPS is that $F' = \{F_1, F_2, \dots, F_q\}$ represents a set of RAS, where each $F_i, i = 1..q$, is a defined as a set of fuzzy rules. This defines the central input to the framework.

3. THE EPS DESIGN FRAMEWORK

Main functionalities of the proposed framework for EPS design are the following:

1. Support the expressive description of EPS,
2. Support definitions of project goals (e.g. minimize duration; maximize number of completed tasks; etc.),
3. Run simulations for a single project schedule in combination with a number of RASs (Multi-RAS EPS), and
4. Select the RAS that best meets the specified project goals to accompany the initial project schedule to yield the final EPS.

The simulation method of choice is the proxel-based simulation (Horton 2002; Isensee, Lazarova-Molnar et al. 2005) as it is highly flexible and provides high accuracy. Its additional advantage is that the simulation is carried out directly, based on the user model, i.e. EPS, without building the state space prior to that. The resulting, simulation-based calculated, optimal EPS will definitely take into account many of the uncertainty factors, thus reducing the risk in the project. In addition, it will provide managers with more insight and guidance when making decisions during project's implementation. A high-level diagram of the data-flow process that underlies the EPS supporting framework is presented in Figure 4. It shows that the inputs to the program are the Multi-RAS EPS and the Project Goals, and it produces an Optimal EPS as a final product.

In the following, we provide detailed description of the most complex module of the framework, i.e. the

EPS Proxel-Based Simulator as previously mentioned in Section 2, and a brief description of the remaining modules.

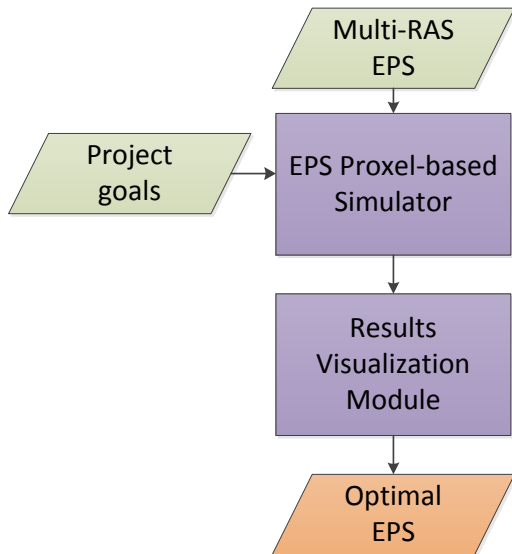


Figure 4: High-level diagram of the framework

3.1. EPS Proxel-Based Simulator Module

The Proxel-Based Simulator is the key-module of the framework. This is the module that performs simulation of the provided Multi-RAS EPS. During simulation, statistics that correspond to project's goals are collected. As previously stated, our simulation method of choice is the proxel-based simulation (Lazarova-Molnar 2005).

The proxel-based method is a simulation method based on the method of supplementary variables (Cox 1955). It was introduced and formalized in (Horton 2002; Lazarova-Molnar 2005). The advantages of the proxel-based method are its flexibility to analyze stochastic models that can have complex dependencies and the accuracy of results, which is comparable to the accuracy of numerical solvers (Stewart 1994).

The proxel-based method expands the definition of a state by including additional parameters which trace the relevant quantities in one model following a previously chosen time step. Typically this includes, but is not limited to, age intensities of the relevant transitions. The expansion implies that all parameters pertinent for calculating probabilities for future development of a model are identified and included in the state definition of the model.

In order to apply the proxel-based simulation algorithm, this module needs to process the information contained in the input file, i.e. the Multi-RAS EPS. In summary, it contains the following information:

- Maximum simulation time,
- Time step,
- Task information,

- Team information,
- Distribution functions in use,
- Mappings of distribution functions to teams and tasks,
- Multiple RAS,
- Deadline, and
- Initial state.

The proxel-based simulation of a given project schedule in combination with each provided RAS is the core element of the tool. Algorithm 1 provides more details of how this is performed. It describes the dynamics of the proxel-based simulation for a single-RAS EPS. This is further repeated for each provided RAS. The basic computational unit, i.e. the proxel, for each EPS is formed based on the information in the input file. The general simplified proxel format is the following:

$$Proxel = (State, t, Pr)$$

where:

$$State = (Task\ Vector, Age\ Vector, Completed\ Tasks), \text{ and}$$

- *Task Vector* is a vector whose size is equal to the number of teams available and records the task that each team is working on,
- *Age Vector* tracks the length that each team has been working on the task specified in the *Task Vector*, correspondingly,
- *Completed Tasks* stores the set of completed tasks,
- *t* is the time at which the afore-described state is observed, and
- *Pr* stores the probability that the schedule is in the afore-specified state at time *t*.

Algorithm 1 demonstrates the on-the-fly building of the state-space of the project schedule model. Thus, there is no need for any pre-processing to generate the state-space. It is directly derived from the input file specification. The initial state proxel is derived from the initial state that is specified in the input file as well.

The algorithm operates by using two interchangeable data structures, *Proxel_Tree[0]* and *Proxel_Tree[1]*, that store the proxels from two subsequent time steps (regulated by the *switch* variable). If two proxels represent the same state, there is only one proxel stored, and their corresponding probabilities are summed up.

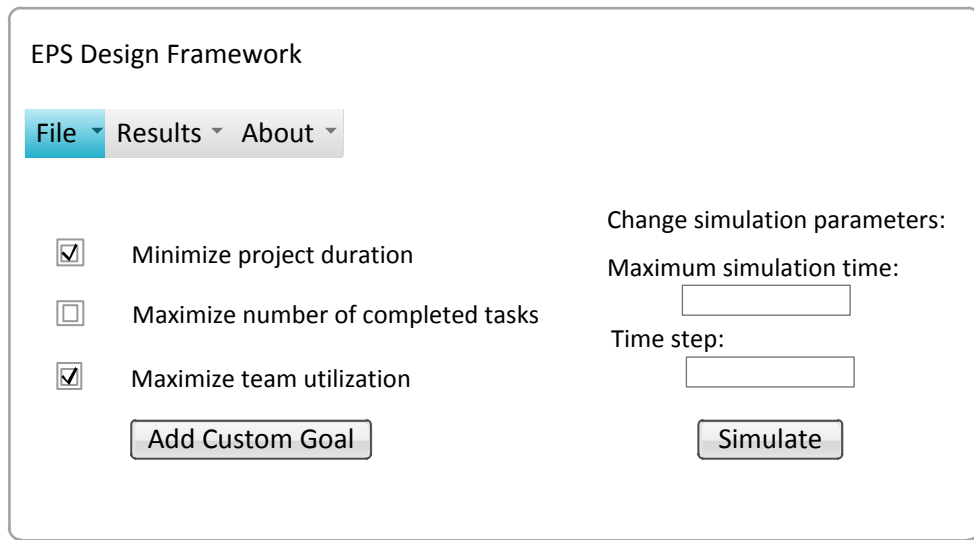


Figure 5: EPS Design Framework GUI prototype

Algorithm 1: Proxel-based simulation of enhanced project schedules

```

Input: EPS, Project Goals
Output: Simulation Results
switch = 0
insert Initial State Proxel in the Proxel_Tree[switch]
switch = 1 - switch
while (maximum simulation time has not been reached)
{
  px = get_proxel(Proxel_Tree[switch]);
  for (each task in the Task Vector(px))
  {
    check task precedence & team availability;
    generate next state S;
    compute probability for S in computed_prob
    search for the S in the Proxel_Tree[1-switch];
    if (S found)
    {
      px1 = found_proxel(S);
      probability(px1) = (probability(px1) ) +
        computed_prob;
    }
    else
    {
      generate new proxel px2(S);
      insert proxel in Proxel_Tree[1-switch];
    }
    delete px from Proxel_Tree[switch];
    increase simulation time by one time step;
    calculate statistics with respect to project goals;
    switch = 1- switch;
  }
}

```

3.2. Supporting Modules

The remaining modules, i.e. the Results Visualization Module and the two supporting ones, are trivial. The Visualization Module is charting the (transient or steady-state) solutions of the simulation with respect to project goals. An example of such solution is provided in the following Section 4.

The Graphical User Interface Module facilitates file-based and graphical input of Multi-RAS EPS, along with the set of project goals. The project goals are meant to be selected from a list of most commonly used ones. The list would include project goals as:

- minimize duration,
- maximize number of tasks completed, etc.

as well as allow the user to specify custom goal by using a scripting language. To illustrate our idea, a prototype of the framework GUI is shown in Figure 5.

The Data Storage Module ensures efficient memory manipulation and stores the statistics and intermediate solutions of the simulation experiments.

4. EXPERIMENTS

To demonstrate the proposed framework, we demonstrate the processing of an example Multi-RAS EPS. The example EPS contains 4 tasks, identified as: Task 1, Task 2, Task 3, and Task 4. Each task can be performed by one of the two teams: Team A or Team B. Tasks 1, 2 and 3 have fixed human resource allocation, i.e. performing team, and Task 4 is a *cancelable floating task*, and can be performed by either team A or B. The initial Gantt chart (*IGC*) of the sample project schedule is shown in Figure 6, where the green-colored tasks are cancelable and the team capable of carrying out task is labeled on the task itself. In addition to this the project schedule has a predefined deadline Δ .

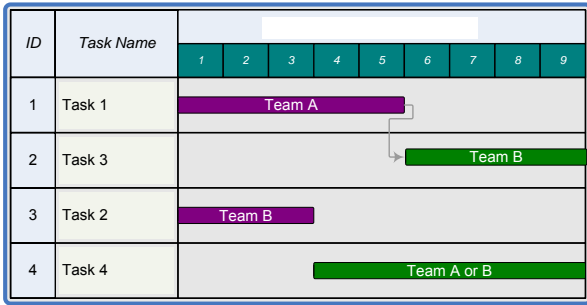


Figure 6: Example of an Initial Gantt chart of the example project schedule (the green-colored tasks are cancelable)

The multi-RAS enhanced project schedule features three RAS. In our case we choose among the two remedial action scenarios (a and b) and the default sequencing (seen as an empty set of fuzzy rules, c), which are defined as follows:

- If the duration of tasks 1 or 2 is close to the deadline Δ , then do not start working on any of the tasks 3 or 4 and do not interrupt the other team if they have already started to work on either of the latter two tasks.
- If the team assigned to a certain non-floating and cancelable task (Task 3 in our case) is unavailable at the time it can be initiated, then cancel the task.
- No guidelines are provided and the manager is instructed to follow the original schedule.

The performance measure, according to which we assess the three RAS, is:

- “the probability of completing the project before the deadline”

The project goal that is supported by this performance measure is defined as “Complete the project before the deadline”. The simulation targets to discover the RAS that yields the best performance, given the constraints of the initial project schedule. For this purpose, the EPS Proxel-Based Simulator Module runs the proxel-based simulation that collects the statistics that answer our question.

4.1. Input File Specifications

In the following we describe and explain the input file specification of the example model, which is shown in Figure 7.

```
#tasks
1 Task1 () 10 false false
2 Task2 () 05 false false
3 Task3 (1) 10 true false
4 Task4 () 10 true true
#teams
10 A
20 B
#distributions
100 U (2.0,10.0)
200 N (7.0,1.0)
300 U (2.0,8.0)
400 W (3.5,1.5)
500 U (2.0,5.0)
#ttd
1 10 100
2 20 200
3 20 300
4 10 400
4 20 500
#ras
(a)
1 fuzzy1 (11.25,15.0) C 3
1 fuzzy1 (11.25,15.0) C 4
2 fuzzy1 (11.25,15.0) C 3
2 fuzzy1 (11.25,15.0) C 4
(b)
3 non_avail (20) C 3
(c)
-
#deadline
20
#initial
(1,2)
```

Figure 7. Example Input File

The input file consists of:

- 1) Definition of all tasks and their duration probability distribution function,
- 2) Definition of all fuzzy functions in use,
- 3) Deadline of the project, and
- 4) Initial state(s)

The input file contains all parameters that are listed in Section 3.1. In the example case there are 4 tasks. Each task is specified by the following parameters: *Task ID*, *Task Name*, *Preceding Tasks*, *Cancelable*, and *Floating*, specified in the same order. Each team is specified by: *Team ID* and *Team Name*. Distributions are specified by: *Distribution ID*, *Distribution Type*, and *Parameters*. Each team-task-distribution mapping contains three values, i.e. the id's of the team, task and distribution that are connected to form the mapping. As specified in the input file, the values of the parameters of the duration distribution functions are:

- Duration of Task 1 ~ Uniform (2.0, 10.0)
- Duration of Task 2 ~ Normal (7.0, 1.0)
- Duration of Task 3 ~ Uniform (2.0, 8.0)
- Duration of Task 4, performed by:
 - Team A ~ Weibull (3.5, 1.5)
 - Team B ~ Uniform (2.0, 5.0)

Next, the definition of the RAS is provided. The fuzzy membership function that defines *fuzzy1* is defined in the framework as follows:

$$\mu(t, a, b) = \begin{cases} 0, & t < a \\ \frac{t-a}{b-a}, & a \leq t \leq b \\ 1, & t > b \end{cases}$$

As specified in the input file, the concrete fuzzy membership function is $\mu(t, 11.25, 15.0)$. The symbol *C* stands for “cancel” and the subsequent number specifies the task id of the task to be canceled. The *non_avail* function evaluates to true/false depending on the availability of team B (specified by its id, i.e. 20 as a parameter).

The framework allows custom specification of the fuzzy functions and actions. It is also extendable as to the type of actions that can be taken. Currently it features only “cancel”.

Finally the pre-determined deadline of 20 time units (as an important parameter) and the initial state of the EPS are provided. According to the latter one, the project begins by Team A working on Task 1 and Team B working on Task 2.

In addition to the EPS model specification, inputs to the framework are the simulation parameters (size of the time step and maximum simulation time) and project goals.

4.2. Proxel-Based Simulation Details

In the following we provide some insight in the proxel-based simulation of our example model to illustrate the simulation method.

The proxel-based simulation of the EPS commences with the initial state, as specified in the input file. This would create the following initial proxel:

$$(((1, 2), (0, 0), ()), 0, 1.0).$$

In the next time step, one of the three developments could be seen:

- 1) Team A completes working on Task 1,
 - implying that it can start working on Task 4, as the only possibility
- 2) Team B completes working on Task 2, and
 - implying that it can start working on Task 4, as the only possibility
- 3) Both teams continue working on the corresponding tasks.

This would create the following proxels, correspondingly:

- 1) (((4, 2), (0, Δt), (1)), 0, p1),
- 2) (((1, 4), (Δt, 0), (2)), 0, p2), and
- 3) (((1, 2), (Δt, Δt), ()), 0, 1-p1-p2).

Note that the age intensities for each task that is still being worked on in the next time step are updated accordingly.

4.3. Experimental Results

In the following we present the results of the simulation of our model. The proxel-based simulation provides complete results for any quantity of interest; in this case it is the probability function of the duration of the project (shown in Figure 8). The performance measures are not limited to this and it is provided for illustration only. In general, they can include any quantities that are relevant to project’s goals.

Simulation results provide us with an overview to aid the selection of the most suitable remedial action scenario with respect to project’s goals.

From the simulation results we can see that the best RAS that yields the highest probability of having the project completed before the deadline is RAS (a), closely followed by (b). Judging from this, we can conclude that the RAS (a) seems most favorable for this enhanced project schedule. Also, we can clearly see that the rigid RAS (c) which does not allow for any changes has the lowest probability of having the project completed before the deadline.

5. CONCLUSIONS

The framework presented in this paper demonstrates a vision and its implementation strategy of how to create a better project schedule, one that will provide more information and decision making guidance to project managers. This is what we term as EPS Design Framework. The output of the framework, i.e. the resulting EPSs is planned to be further used to support project managers in the decision making processes throughout project implementation. Decisions made in this way, based on the RAS recommendations that accompany the optimal EPS will not be solely based on human judgment, as it is the case in classical approach, but also based on sound models and their analysis.

In this paper we present the details of the framework and, in particular, the details of its core module, i.e. the Proxel-Based Simulator, for which we present the modified proxel-based simulation algorithm.

We believe that this is an effective way of designing schedules and it enhances the classical project schedule by allowing all available information to be utilized. Instead of having a static schedule, the remedial action scenarios make the schedule dynamic and evolving. In addition, our framework provides support for managers to incorporate their knowledge within the project schedules by simulating various possible RAS when new uncertainties occur. The

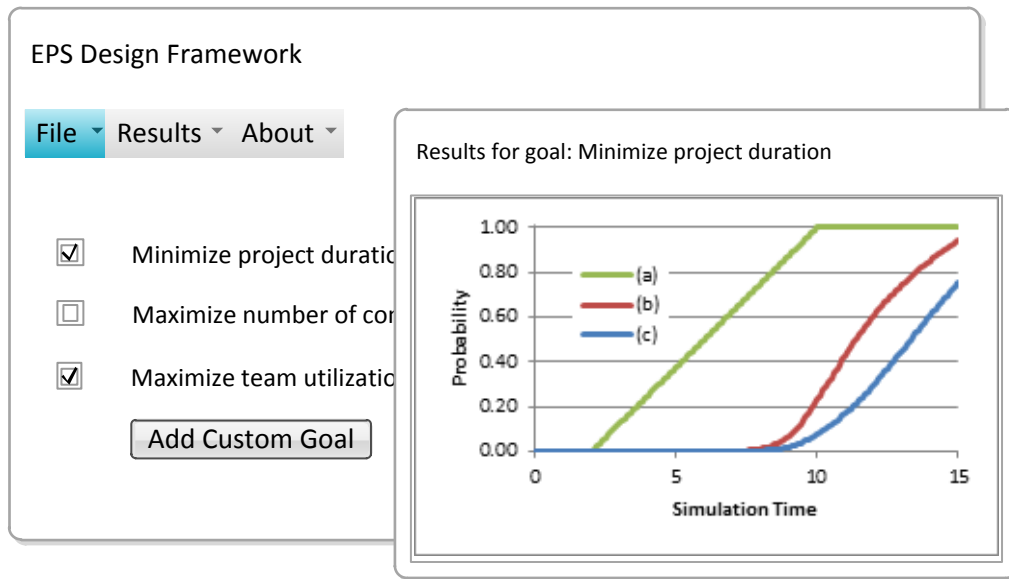


Figure 8: Probability of having the project completed for the three possible RAS

framework encourages analysis and deep thinking about project plans and, hence, supports the creation of a higher quality initial plan, identified as one of the key success factors for projects.

As part of our future work agenda, we plan to extend the capabilities of our simulation approach to handle multi-project resource sharing.

REFERENCES

- Arauzo, J. A., J. M. Galán, et al. (2009). "Multi-agent technology for scheduling and control projects in multi-project environments. An Auction based approach." *Inteligencia Artificial* **42**: 12-20.
- Cox, D. R. (1955). "The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables." *Proceedings of the Cambridge Philosophical Society* **51**(3): 433-441.
- Denning, P. J. and R. D. Riehle (2009). "The profession of IT Is software engineering engineering?" *Communications of the ACM* **52**(3): 24-26.
- Herroelen, W. and R. Leus (2005). "Project scheduling under uncertainty: Survey and research potentials." *European Journal of Operational Research* **165**(2): 289-306.
- Horton, G. (2002). "A new paradigm for the numerical simulation of stochastic Petri nets with general firing times." *Proceedings of the European Simulation Symposium*.
- Huang, W., L. Ding, et al. (2009). *Project Scheduling Problem for Software Development with Random Fuzzy Activity Duration Times*, Springer.
- Isensee, C., S. Lazarova-Molnar, et al. (2005). "Combining Proxels and Discrete Phases." *Proceedings of ICMSAO*.
- Jing-wen, Z. and S. Hui-fang (2009). *Multi-Mode Double Resource-Constrained Time/Cost Trade-Offs Project Scheduling Problems*. International Conference on Management and Service Science, 2009. MASS '09. .
- Lazarova-Molnar, S. (2005). The proxel-based method: Formalisation, analysis and applications, Otto-von-Guericke-Universität Magdeburg, Universitätsbibliothek.
- Lazarova-Molnar, S. (2005). The Proxel-Based Method: Formalisation, Analysis and Applications. *Faculty of Informatics*. Magdeburg, University of Magdeburg. **Ph.D.**
- Lazarova-Molnar, S. and R. Mizouni (2010). *Floating Task: Introducing and Simulating a Higher Degree of Uncertainty in Project Schedules*, IEEE.
- Lazarova-Molnar, S. and R. Mizouni (2010). "Modeling Human Decision Behaviors for Accurate Prediction of Project Schedule Duration." *Enterprise and Organizational Modeling and Simulation*: 179-195.
- Matta, N. F. and R. N. Ashkenas (2003). "Why good projects fail anyway." *Harvard Business Review* **81**(9): 109-116.
- Sobel, M. J., J. G. Szmerekovsky, et al. (2009). "Scheduling projects with stochastic activity duration to maximize expected net present value." *European Journal of Operational Research* **198**(3): 697-705.

Stewart, W. J. (1994). Introduction to the Numerical Solution of Markov Chains, Princeton University Press.

AUTHORS BIOGRAPHY

SANJA LAZAROVA-MOLNAR is an Assistant Professor in Computer Science at United Arab Emirates University. She received her Diploma in Computer Science from the “Sts. Cyril and Methodius” University in Macedonia and her M.Sc. in Computational Visualistics and Ph.D. in Computer Science from “Otto-von-Guericke” University of Magdeburg in Germany. Her main research interests are simulation and modeling, with the main focus on their methodology for achievement of more accurate and valid results. Her e-mail is sanja@uaeu.ac.ae.

RABEB MIZOUNI is an assistant professor in the Department Software engineering at Khalifa University, Abu Dhabi, UAE. She got her PhD and her MSc in Electrical and Computer Engineering from Concordia University, Montreal, Canada in 2007 and 2002 respectively. After her graduation, Rabeb joined SAP research labs in Montreal where she acquired experience in the development of mobile applications. Her research interests include modeling and simulation of software requirements to improve the development of robust project schedules, development of mobile applications, and web services. Her email is rabeb.mizouni@kustar.ac.ae.

NADER KESSERWAN got his Master Degree in Computer Science from McGill University, Montreal, Canada. He joined the college of Information Technology as Instructor in 2001. Before joining the academia, he acquired experience in diverse fields of computer science, simulation and real time systems. He spent two years at CAE Company, Canada working on a Flight Simulator, and on testing the integration of its software. Also, he had worked two years on games development for Disney Studio, Canada and USA. He taught programming courses and software engineering at McGill University, Canada. His experience includes team management, system analysis, software development and project management. His email is nkesserwan@uaeu.ac.ae.

A NOVEL APPROACH TO REALISTIC MODELING AND SIMULATION OF STATE-VARYING FAILURE RATES

Sanja Lazarova-Molnar

Faculty of Information Technology, United Arab Emirates University, United Arab Emirates

sanja@uaeu.ac.ae

ABSTRACT

There has been a lot of research on time-varying failure rates, which deems constant failure rates as inadequate to model failures accurately. However, besides time, failure rates can also be affected by the state of the system (or its history, in terms of sequences of states and events that it has been through). In our paper we define several classes of state-varying failure rates and extend the formalism of Petri nets to model them. We further use the flexibility of the proxel-based method to accurately analyze behavior of systems that incorporate these kinds of failures. To illustrate our approach and study the effect of such dependencies, we compare simulation results for two models: one that exhibits state-varying failure rates, and another that only contains predefined failure rate functions.

Keywords: state-varying failure rates, reliability, proxels, Petri nets

1. INTRODUCTION

There has been an intensive research on time-varying failure rates, including their significant impact on reliability (Hassett, Dietrich et al. 1995; Retterath, Venkata et al. 2005; Zhang, Cutright et al. 2010), which have been defined as such almost two decades ago (Billinton and Allan 1992). Recently, Xie developed an analytical model of unavailability due to aging failures too (Xie and Li 2009). Since long time ago it has been shown that constant failure rates are inadequate for describing systems' failures (Proschan 1963). Nevertheless, they are still widely used due to the fact that the methodology for their analysis is less complex and more accurate. The popular MTTF (meantime to failure) measure is still a widely used one (Coskun, Strong et al. 2009; Sharma, Kahlon et al. 2010), even though it has been deemed many times as inadequate (Schroeder and Gibson 2007). We go one step further as to claim that even time-varying failure rates are not sufficient, as in many cases the rates completely change their functions based on the occurrence of some events or based on the complete state of the system (e.g. a part has been replaced by a new one that is based on a new technology, or if a mechanical part has been physically broken, then it is logical that the failure rate would

increase with each time it breaks). This is what we term as a *state-varying* failure.

According to a study of medical equipment (Baker 2001) it was shown that there was a decreasing hazard of (first) failure after repair for some types of equipment. The interpretation was that it is a consequence of imperfect or hazardous repair, and also, because of differing failure rates among a population of machines.

Likewise, in (Liberopoulos and Tsarouhas 2005) a pizza production line is studied and it was found that most of the failures have a decreasing failure rate because proactive maintenance improves the operating conditions at different parts in the line, and a few failures have an almost constant failure rate. It was also concluded that the longer the time between two failures, the more problems accumulate, and therefore, it takes longer time to fix the latter failure. It also suggests that the more time the technicians spend fixing a failure, the more careful job they do, and therefore, the time period until the next failure is longer. This is a very interesting observation that calls for state-varying failure rates and it can be addressed using our approach.

These are some examples that show that failures need to be described more realistically to obtain accurate and useful simulation results. Unfortunately, this has very rarely been the case.

Our goal is to provide a deterministic approach to analyze systems that exhibit not only time-, but more importantly, state-varying failure rates. For this we use method of proxel-based simulation, which based on our previous experience, is highly adjustable to treat these complex activities. In (Lazarova-Molnar 2008) we have analyzed and described state-dependent transitions and used proxel-based simulation for their analysis. These are the types of transitions that correspond and can be used to describe state-varying failure rates. Thus, in addition to the simulation approach, this paper provides a concept of how to model this type of failure rates and what changes need to be undertaken in the standard stochastic Petri net (SPN) models to introduce them.

The paper is organized as follows. In the subsequent section we describe the state-varying failure rates, along with an introduction to the proxel-based simulation method. Further, we provide a concept for modeling state-varying failures using SPN. Next, we

present an example model which we use to demonstrate our approach and we run experiments based on it. Finally, we present the results of the experiments with a discussion and conclusions.

2. PRELIMINARIES

2.1. State-varying Failures

It is a common observation that a failure rate cannot simply be described by one function during its entire lifetime. Even more, failure rates in reality can change not solely based on time (Retterath, Venkata et al. 2005), but also based on the occurrence of certain events in the system (e.g. replacing the service person by another one which fixes them in a different manner, i.e. more thoroughly would influence the failure rate function). We refer to these types of failures as *state-varying failure rates*.

Description of failure rate functions of state-varying failure rates is a complex process and would require an algorithmic description to supplement the graphical model. To illustrate it, one such description may be:

If machine is repaired by repairman A
 Then the failure rate function \sim Normal(a, b)
 Else if machine is repaired by repairman B
 Then the failure rate function \sim Normal(c, d)

If we add another factor to this, i.e. the age of the machine, and then the description would change to:

If machine is repaired by repairman A
 Then the failure rate function \sim Normal(f(t), b)
 Else if machine is repaired by repairman B
 Then the failure rate function \sim Normal(g(t), d)

where t is the age of the machine (which can easily be exchanged to represent the number of failures or any other relevant quantity). This observation is more general than the one that uses fixed failure rate functions, and as such, more realistically models the phenomenon of a machine that exhibits failures.

Obviously, these models would need more advanced (or extended) modeling formalisms to be described. Thus, we extend stochastic Petri nets to account for the state-varying rates.

Finally, to show the difference and compare the effects of such (even very small) dependencies, we compare the simulation results for two models: one that exhibits state-varying failure rates, and another, similar and over-simplified one, that only contains predefined failure rate functions with fixed parameter values.

2.2. Proxel-based Simulation

The proxel-based method (Horton 2002; Lazarova-Molnar 2005) is a relatively novel simulation method, whose underlying stochastic process is a discrete-time Markov chain (Stewart 1994) and implements the method of supplementary variables (Cox 1955). The method, however, is not limited to Markovian models.

On the opposite, it allows for a general class of stochastic models to be analyzed regardless of the involved probability distribution functions. In other words, the proxel-based method combines the accuracy of numerical methods with the modeling power of discrete-event simulation.

The proxel-based method is based on expanding the definition of a state by including additional parameters which trace the relevant quantities in one model through a previously chosen time step. Typically this includes, but is not limited to, age intensities of the relevant transitions. The expansion implies that all parameters pertinent for calculating probabilities for the future development of a model are identified and included in the state definition of the model.

Proxels (stands for probability elements), as basic computational units of the algorithm, follow dynamically all possible expansions of one model. The state-space of the model is built on-the-fly, as illustrated in Figure 1, by observing every possible transiting state and assigning a probability value to it (Pr in the figure stands for the probability value of the proxel). Basically, the state space is built by observing all possible options of what can happen at the next time step. The first option is for the model to transit to another discrete state in the next time step, according to the associated transitions. The second option is that the model stays in the same discrete state, which results in a new proxel too. Zero-probability states are not stored and, as a result, no further investigated. This implies that only the truly reachable (i.e. tangible) states of the model are stored and consequently expanded. At the end of a proxel-based simulation run, a transient solution is obtained which outlines the probability of every state at every point in time, as discretized through the chosen size of the time step. It is important to notice that one source of error of the proxel-based method comes from the assumption that the model makes at most one state change within one time step. This error is elaborated in (Lazarova-Molnar 2005).

Each proxel carries the probability of the state that it describes. Probabilities are calculated using the instantaneous rate function (IRF), also known as hazard rate function. The IRF approximates the probability that an event will happen within a predetermined elementary time step, given that it has been pending for a certain amount of time τ (indicated as ‘age intensity’). It is calculated from the probability density function (f) and the cumulative distribution function (F) using the following formula:

$$\mu(\tau) = \frac{f(\tau)}{1 - F(\tau)} \quad (1)$$

As all state-space based methods, this method also suffers from the state-space explosion problem (Lin, Chu et al. 1987), but it can be predicted and controlled by calculating the lifetimes of discrete states in the model. In addition, its efficiency and accuracy can be

further improved by employing discrete phases and extrapolation of solutions (Isensee and Horton 2005). More on the proxel-based method can be found in (Lazarova-Molnar 2005).

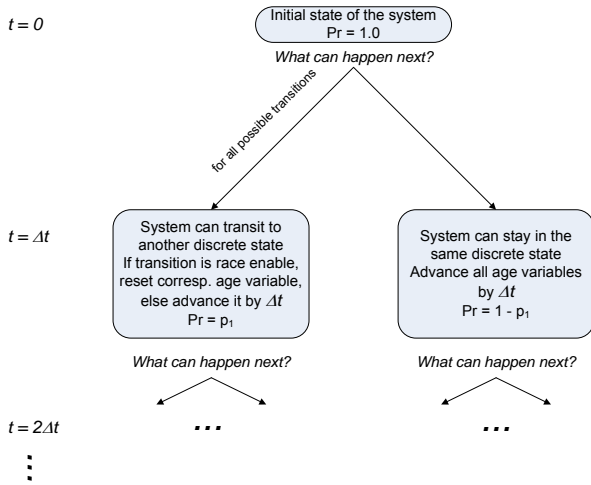


Figure 1: Illustration of the development of the proxel-based simulation algorithm

3. MODELING STATE-VARYING FAILURES

According to the performed observation, studies and research, we identify several classes of state-varying failure rates, i.e. failure rates that depend on:

- the number of failure occurrences up to the observed point in time,
- the age of the machine up to the observed point in time,
- the duration of the last repair,
- the time between the last two failures,
- the properties of the repair facilities, introduced as additional parameters, and
- the types of failures that have occurred.

We allow a combination of a number of these factors to occur in our sample model to illustrate their effects through the proxel-based simulation analysis. Proxel-based simulation can easily be applied to analyze a model that exhibits any combination of them, as well as other types of dependencies on quantities that are part of the model. In the following, we will provide the details of the formal classification of the state-varying failures and our simulation approach. This will be further demonstrated using an example model.

3.1. Formal Model of State-Varying Failure Rates

The underlying discrete stochastic model that exhibits state-varying failure rates is described using a stochastic Petri net (SPN) (Bause and Kritzinger 2002). Nevertheless, we further extend the basic description of SPN to allow the tracking of the *relevant rewards*. Those are the quantities that are in fact parameters of the distribution functions of the timed transitions, besides the age intensities of the relevant transitions.

Typically, they are introduced by extending the basic SPN with additional places and transitions that enable the tracking, as shown in Figure 2. However, to record quantities, such as the *duration of the last repair* (type (c)), we introduce a novel element which we term as *tracking variable* (TV), and it is represented by a hexagon in the SPN graphical model. TVs are connected by diamond-shaped arrows to the transitions for which they record the last firing time.

To summarize, the extension is at both the level of the SPN formalism, and at the Petri net model itself, which is enriched by a number of extra places and transitions to ensure the tracking of relevant rewards. As for the SPN formalism: we extend it by the new element TV, and, in order to account for the state-varying transitions, we allow distributions to have discrete states, i.e. markings, as parameters of the distribution functions that control firing of transitions. In the following, we show by example how a SPN can be extended to allow the tracking of the various relevant quantities.

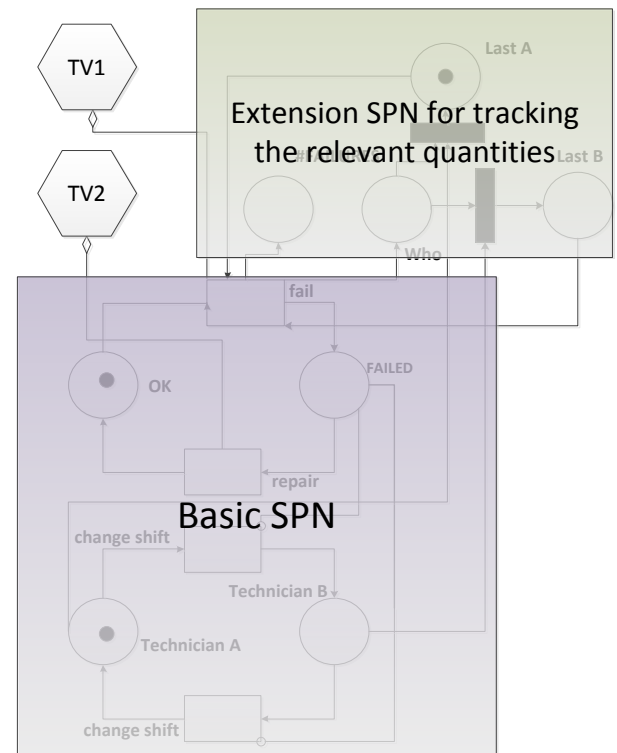


Figure 2: Illustration of the extended SPN model

3.2. Petri Net Specifications

In the following we provide the formal definition of the extension of the SPN to account for the state-varying failure rates. Each Petri net *SPN* is defined as:

$$SPN = (P, T, A, G, TV, m_0)$$

where:

- $P = \{P_1, P_2, \dots, P_n\}$, the set of places, drawn as circles
- $T = \{T_1, T_2, \dots, T_m\}$, the set of transitions along with their distribution functions or probability values, drawn as bars
- $A = A^I \cup A^O \cup A^H \cup A^T$, the set of arcs, where A^O is the set of output arcs, A^I is the set of input arcs, A^H is the set of inhibitor arcs, and A^T is the set of tracking arcs (connect transition to a tracking variable and are ended by a diamond-shape at the tracking variable end); each arc has a multiplicity assigned to it,
- $G = \{g_1, g_2, \dots, g_r\}$, the set of guard functions which are associated with different transitions,
- $TV = \{TV_1, TV_2, \dots, TV_m\}$, the set of tracking variables that store the last duration of the enabling time of a transition (drawn as hexagons),
- m_0 – the initial marking of the Petri net.

Each transition is defined as $T_i = (F, type)$, where $type \in \{enabling, age, immediate\}$ is the type of memory policy if it is a timed transition or “immediate” if the corresponding transition is an immediate one. F is a cumulative distribution function if the corresponding transition is a timed one. Immediate transitions have a constant value instead of a distribution function assigned to them, which is used for computing the probability of firing of an immediate transition if more than one are enabled at once. The sets of arcs are defined such that

$$A^O = \{a^o_1, a^o_2, \dots, a^o_k\}, A^I = \{a^i_1, a^i_2, \dots, a^i_j\}, A^H = \{a^h_1, a^h_2, \dots, a^h_i\}, \text{ and } A^T = \{a^t_1, a^t_2, \dots, a^t_l\},$$

where

$$A^H, A^I \subseteq P \times T \times \mathbb{N}, A^O \subseteq T \times P \times \mathbb{N}, A^T \subseteq T \times P \times \mathbb{R}.$$

The multiplicity of the tracking arcs can be a real number, unlike the others, where it is a non-negative integer number. We denote by $M = \{m_0, m_1, m_2, \dots\}$ the set of all reachable markings of the Petri net. Each marking is a vector made up of the number of tokens in each place in the Petri net along with the values of the tracking variables, $m_i = (\#P_1, \#P_2, \dots, \#P_n, val(TV_1), val(TV_2), \dots, val(TV_m))$. The set of all reachable markings is the discrete state space of the Petri net. The changes from one marking to another are consequences of the firing of enabled transitions which move (destroy and create) tokens; thus creating the dynamics in the Petri net. This makes the firing of a transition analogous to an event in a discrete-event system. The markings of a Petri net, viewed as nodes, and the possibilities of movement from one to another, viewed as arcs, form the reachability graph of the Petri net.

3.3. Adaptation of the Proxel-based method to Accommodate State-varying Failure Rates

The main adjustment of the proxel-based method to accommodate state-varying failure rates is the extension

of the definition of the proxel to introduce the notion of relevant rewards. They incorporate in the state definition all quantities, which in addition to age intensities; can be parameters of probability distribution functions of events in the model. This yields the following definition of a state:

$$\text{State} = (\text{Discrete state, Relevant Rewards, State Relevant Age Intensities})$$

Thus, all parameters required for computing transition distribution functions are contained in the state vector (making the model implicitly a non-homogeneous Markov chain). The discrete state typically corresponds to a marking in the SPN model. The relevant rewards are determined by the nature of the events in the model, i.e. what kind of dependencies the failure rates in the model exhibit. In Table 1 we provide the relevant rewards for the six classes of state-varying failures that we have identified.

Table 1: Relevant rewards for the six state-varying failure classes

State-varying failure class	Relevant reward
a) the number of failure occurrences up to the observed point in time,	Number of failures
b) the age of the machine up to the observed point in time,	Age of machine
c) the duration of the last repair,	Duration of last repair
d) the time between the last two failures,	Duration of operation of machine between two consecutive failures
e) the properties of the repair facilities, introduced as additional parameters,	Parameters of repair facilities (e.g. quantification of experience of repairman)
f) the types of failures that have occurred.	Types of failures that have occurred so far

4. EXPERIMENTS AND RESULTS

In this section we present an example model which we will use to illustrate the simulation of state-varying failure rates. We will describe the proxel-based simulation of this model and run it using various time steps.

4.1. The Model

The model that we use to demonstrate our approach is a simple model that describes a machine that incorporates both time- and state- varying failure rates, similar to the scenarios described in Section 2.1.

Using a Petri net, the model can be described as shown in Figure 2. It represents a machine that exhibits one of two possible states: OK and FAILED. When the machine has failed, one of the two repairmen arrives and repairs it, after what the machine's state becomes OK. Changing shifts during repair is not allowed. The two repairmen have different lengths of working experience. Thus, when the machine is fixed by the Repairman A, the time to the next failure is on average longer, than when it is repaired by Repairman B. This implies that the proxels will need to record the information of who performed the last repair as well.

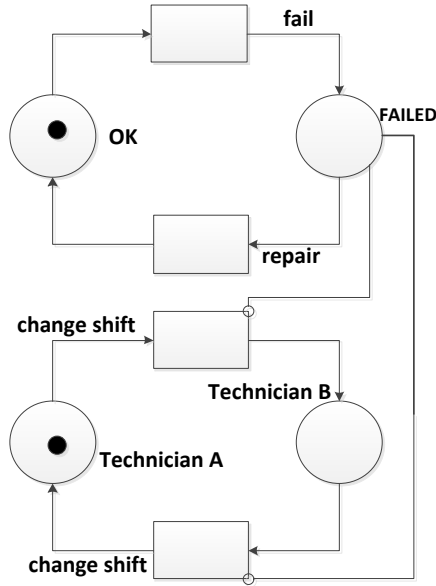


Figure 3: Basic Petri net model of the example

In addition to the afore-described scenario, the distribution of the time to next failure of the machine is also a function of the number of failures and the age of the machine. For instance, in our example model we use the following formula to describe the distribution of the repair time:

$$f_{repair}(R, age, n_f) \sim \begin{cases} N(10 + age * 0.01 + n_f * 0.1, age * n_f + 1.0), R = A \\ N(12 + age * 0.015 + n_f * 0.12, age * n_f + 1.5), R = B \end{cases} \quad (2)$$

where:

- *age* is the age of the machine, i.e. the global simulation time,
- n_f is the number of failures that have occurred, i.e. #FAILURES in the SPN,
- *R* is the repairman that did the serviced the last failure, i.e. can be obtained from the SPN by checking the token is in place *Last A* or *Last B*, and

- *N* stands for the normal distribution parameters, with the standard parameters: mean and variance, correspondingly.

In other words, repairs are normally distributed, where the mean and the variance are functions of the Repairman that completed the last repair, the total number of failures, and the age of the machine. This implies that we observe the dependences (a), (b), and (e), as pointed out in Section 3. This directly implies that, as described in Table 1, we need to add the following relevant rewards:

- Number of failures,
- Age of machine, and
- Parameters of repair facilities.

In order to illustrate the enhancements that the state-varying failures would require, we include the required information in the basic Petri net model, whose extended version is shown in Figure 4.

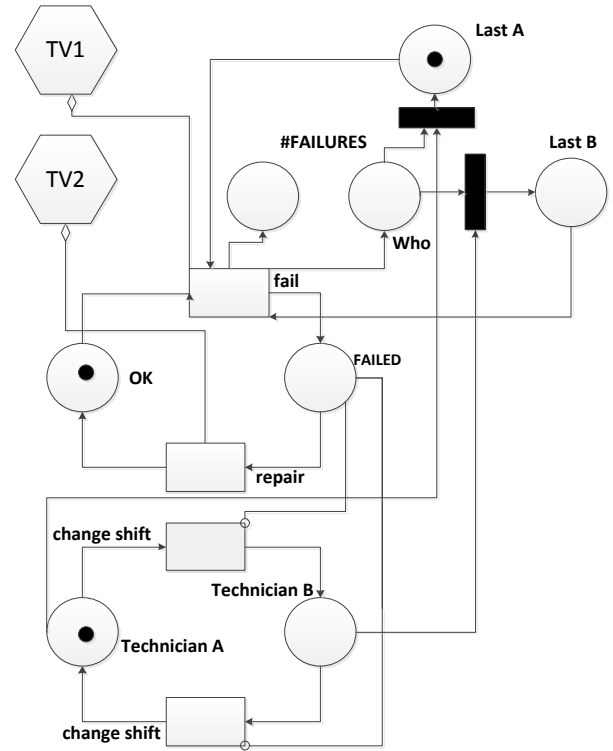


Figure 4: Extended Petri net model of the example

In Figure 5, the state-transition diagram of the Petri net model from Figure 4, is shown. Note that the model is an unbounded Petri net, i.e. it is practically impossible to accurately analyze it using numerical approaches. This, however, is not a limitation of the proxel-based method, as it dynamically builds the state space on-the-fly.

Besides the repair duration probability distribution function, as shown by the Equation (2), the remaining distribution functions that we used in our experiments are as follows:

- $f_{fail}(age, n_f) \sim E(150 + age * 0.05 + n_f * 0.1)$
- $f_{change\ shift}() \sim D(12)$

where $E()$ stands for the exponential distribution function, with the mean as its only parameter, and $D()$ is the deterministic probability distribution function.

As described in the following subsection, for this example we slightly modify the discrete state description to better explain our approach. In general, the proxel-based simulation can be directly performed on the enhanced Petri net model.

4.2. Insight in the Proxel-based Simulation

In the following we provide insight in the proxel-based simulation for the example model. The goal is show what exactly happens at lower level when simulating the state-varying failures. We begin by defining the state of the systems in the concrete example as:

$$((Machine\ State, Repairman), \\ (\#Failures, Last\ Repairman), \\ Age\ Intensity\ Vector),$$

which implies that the discrete state of the system is described by the state of the machine (*Machine State*) and the repairman on shift (*Repairman*). The relevant rewards are the number of failures of the machine (*#Failures*) and the repairman that completed the last repair of the machine (*Last Repairman*). Finally, the last element of the state of the system is the *Age Intensity Vector* that keeps track of the time that the machine has spent in the specified state, as well as the time during which the repairman has been on shift. This yields the initial proxel as:

$$((OK, A), (0, A), (0, 0)).$$

which shows that initially the machine is in state OK, and Repairman A is on shift. The number of failures up to simulation time $t = 0$ is zero, and the age intensities of both machine state OK and duration of Repairman A on shift are zero as well. Note that initially we assume that the last repair was completed by the more experienced repairman, i.e. Repairman A. The subsequent proxels which originate from the initial one at time $t = \Delta t$, along with the three potential events, are the following:

- Machine fails - $((F, A), (1, A), (0, \Delta t))$,
- Repairman shift change - $((F, B), (0, A), (\Delta t, 0))$,
- No events - $((OK, A), (0, A), (\Delta t, \Delta t))$.

where F stands for the machine's state *FAILED*. The age of the machine is implicitly recorded by the global simulation time variable. Note that we assume that the repairman on shift that has started to work on the repair also has to complete it, and thus, extend his shift. Further, for illustration purposes, we will develop the proxel for the case (a), i.e. when the machine has failed, which yields the following subsequent events and proxels:

- Machine is repaired - $((OK, A), (1, A), (0, 2\Delta t))$,
- No events - $((F, A), (1, A), (\Delta t, \Delta t))$.

The model description yields that the "change shift" transition is of *race age* policy, i.e. it needs to record the time spent on shift and not be restarted it when a failure occurs. During this processing, the required statistics that yield the simulation results are collected.

4.3. Experiments and Results

In the following we present the results of our simulation experiments, i.e. the statistics that were collected during the proxel-based simulation of the example model. The questions that our simulation model provides answers to are the following:

- What is the probability of having the machine running?, and
- What is the probability that the machine has 5 or more failures?

The question (a) is a classical reliability analysis case, and the most typical question for a model like this one. In Figure 6 and Figure 7 we present the answers to the questions (a) and (b), correspondingly. The simulation parameters that we used were: a maximum simulation time $t = 300$, and a time step $\Delta t = 0.5$. Apparently, in Figure 6, we can observe that the model has not reached a steady-state during the simulation time of 300 time units, thus it needs to be simulated for a longer period of time. We did this, i.e. we ran the simulation up to time $t = 10000$ using the same time step, and the obtained solution for the steady state reliability of the system is periodically oscillating, as shown in Figure 8.

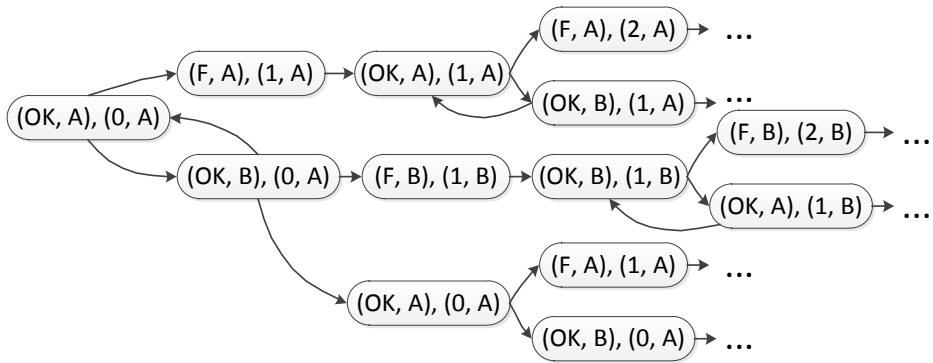


Figure 5: State-transition diagram of the unbounded Petri net model from Figure 4

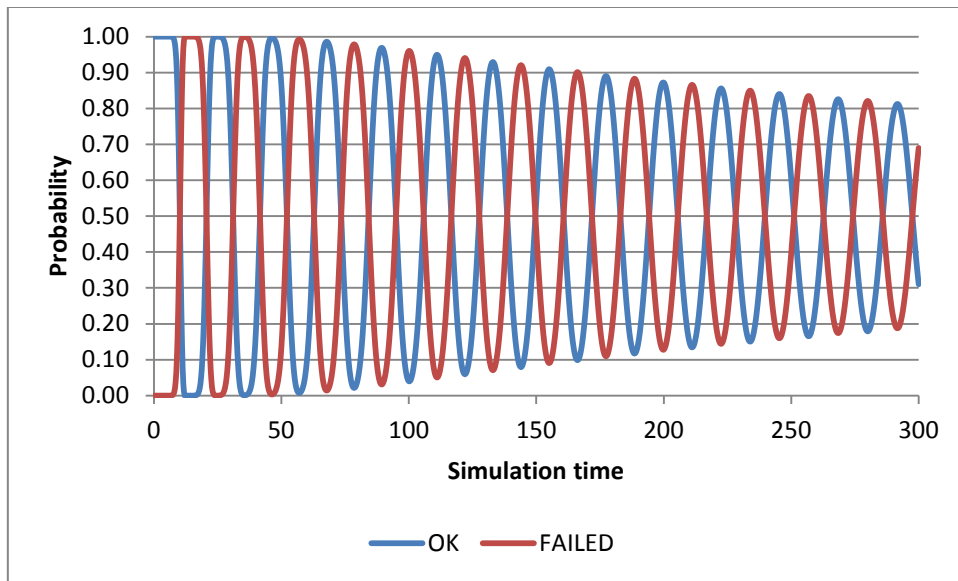


Figure 6: Transient solution for the 2 discrete states, neglecting the fact of what repairman is on shift

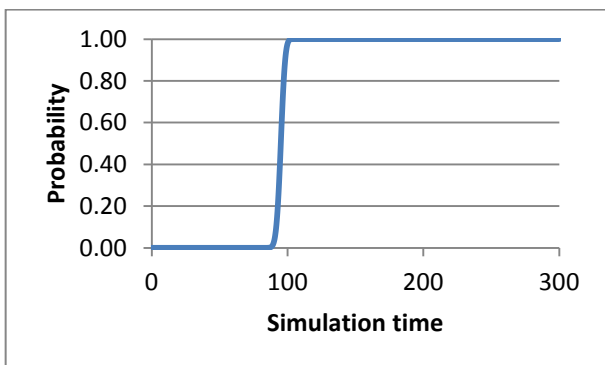


Figure 7: The probability of the machine having 5 or more failures

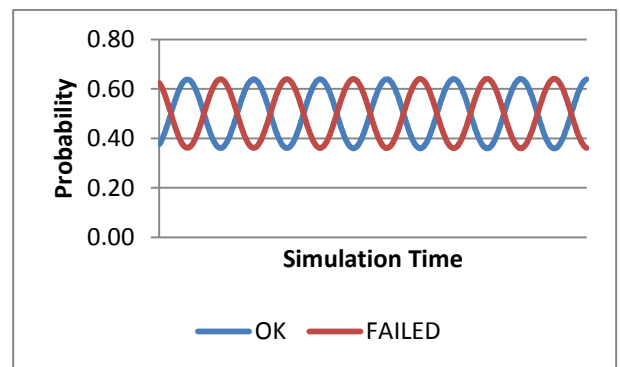


Figure 8: Steady-state solution of the discrete stochastic model

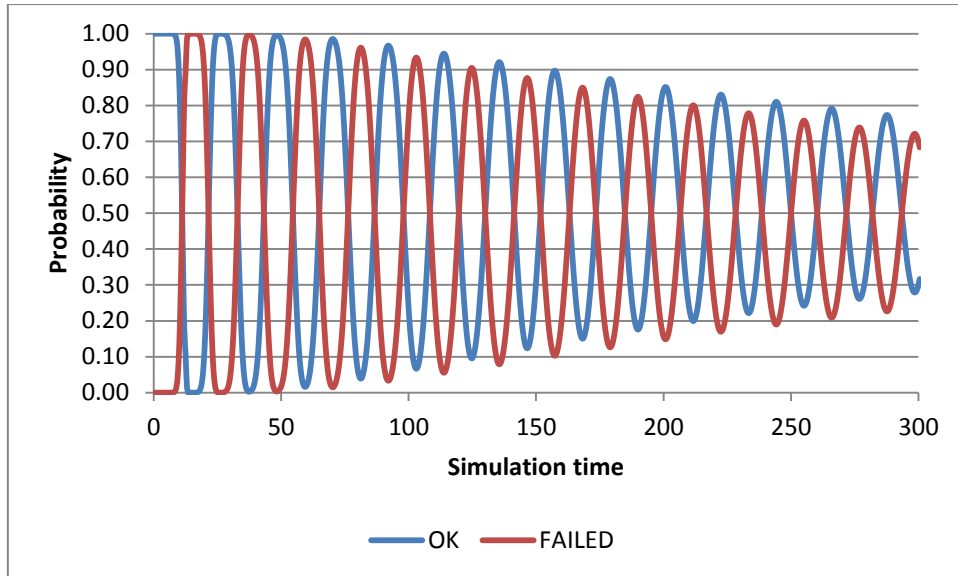


Figure 9: Transient solution for the 2 discrete states for the simplified model, neglecting the fact of what repairman is on shift

The simulation results were obtained in 0.5 seconds on an Intel Core i5 2.53GHz workstation with a 4GB of RAM. The extended computation for the steady-state solution took longer, i.e. 3.5 minutes, which is still an acceptable running time.

- $f_{repair}(\cdot) \sim N(11.0, 1.0)$,
- $f_{fail}(\cdot) \sim E(150)$

5. SUMMARY AND OUTLOOK

We presented an approach to more realistically model and simulate failures that exhibit a wide range of dependencies which are typically neglected. Their neglecting, however, can provide highly misleading results, and thus, it is imperative to avoid their oversimplification. The proxel-based method has shown to be very accurate and highly flexible in describing the complex types of dependencies that typically occur in stochastic models. We anticipate extending of the presented work to provide a tool that would facilitate reliability modeling considering state-varying failure rate functions.

REFERENCES

- Baker, R. D. (2001). "Data-based modeling of the failure rate of repairable equipment." *Lifetime Data Analysis* 7(1): 65-83.
- Bause, F. and P. S. Kritzinger (2002). *Stochastic Petri Nets*, Vieweg.
- Billinton, R. and R. N. Allan (1992). *Reliability evaluation of engineering systems*, Plenum Press New York.
- Coskun, A. K., R. Strong, et al. (2009). *Evaluating the impact of job scheduling and power management on processor lifetime for chip multiprocessors*, ACM.
- Cox, D. R. (1955). "The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables." *Proceedings of the Cambridge Philosophical Society* 51(3): 433-4
- Hassett, T. F., D. L. Dietrich, et al. (1995). "Time-varying failure rates in the availability and

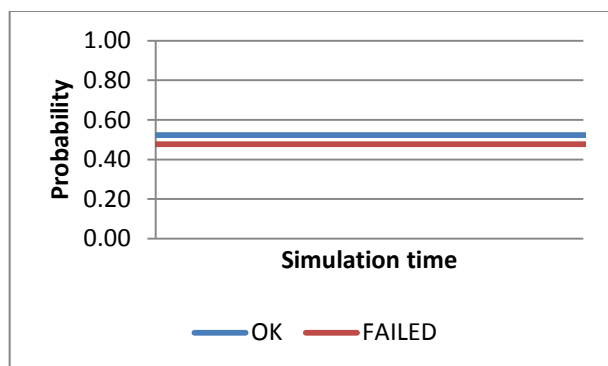


Figure 10: Steady-state solution of the simplified discrete stochastic model

For comparison, in Figure 9 and Figure 10, we provide the transient and steady state solutions of the same model, where the state-varying distributions are substituted with state-independent ones, i.e. with distributions with fixed parameters. Apparently, the results and their nature are quite different as the oscillating steady-state pattern is not present in the simplified model.

More specifically, the distribution functions that we used for the state-independent failure rates were the following:

- reliability analysis of repairable systems." Reliability, IEEE Transactions on 44(1): 155-160.
- Horton, G. (2002). "A new paradigm for the numerical simulation of stochastic Petri nets with general firing times." Proceedings of the European Simulation Symposium.
- Isensee, C. and G. Horton (2005). "Approximation of Discrete Phase-Type Distributions." Proceedings of the 38th annual Symposium on Simulation: 99-106.
- Lazarova-Molnar, S. (2005). The Proxel-Based Method: Formalisation, Analysis and Applications. Faculty of Informatics. Magdeburg, University of Magdeburg. Ph.D.
- Lazarova-Molnar, S. (2008). State-Dependent Transitions in Discrete Stochastic Models: Deterministic Simulation Approach. Summer Computer Simulation Conference 2008.
- Liberopoulos, G. and P. Tsarouhas (2005). "Reliability analysis of an automated pizza production line." Journal of Food Engineering 69(1): 79-96.
- Lin, F. J., P. M. Chu, et al. (1987). "Protocol verification using reachability analysis: the state space explosion problem and relief strategies." ACM SIGCOMM Computer Communication Review 17(5): 126-135.
- Proschan, F. (1963). "Theoretical explanation of observed decreasing failure rate." Technometrics 5(3): 375-383.
- Retterath, B., S. S. Venkata, et al. (2005). "Impact of time-varying failure rates on distribution reliability." International Journal of Electrical Power and Energy Systems 27(9-10): 682-688.
- Schroeder, B. and G. A. Gibson (2007). Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?, USENIX Association.
- Sharma, S., K. S. Kahlon, et al. (2010). "Reliability and path length analysis of irregular fault tolerant multistage interconnection network." ACM SIGARCH Computer Architecture News 37(5): 16-23.
- Stewart, W. J. (1994). Introduction to the Numerical Solution of Markov Chains., Princeton University Press.
- Xie, K. and W. Li (2009). "Analytical model for unavailability due to aging failures in power systems." International Journal of Electrical Power and Energy Systems 31(7-8): 345-350.
- Zhang, H., E. Cutright, et al. (2010). Time-varying failure rate for system reliability analysis in large-scale railway risk assessment simulation. Safety and Security in Railway Engineering: 29.

in Macedonia and her M.Sc. in Computational Visualistics and Ph.D. in Computer Science from "Otto-von-Guericke" University of Magdeburg in Germany. Her main research interests are simulation and modeling, with the main focus on their methodology for achievement of more accurate and valid results. Her e-mail is sanja@uaeu.ac.ae.

AUTHORS BIOGRAPHY

SANJA LAZAROVA-MOLNAR is an Assistant Professor in Computer Science at United Arab Emirates University. She received her Diploma in Computer Science from the "Sts. Cyril and Methodius" University

MODELLING AND SIMULATION OF ORDER SORTATION SYSTEMS

Fahrettin Eldemir^(a), Elif Karakaya^(b)

^(a)Industrial Engineering Department, Fatih University, Istanbul

^(b)Industrial Engineering Department, Fatih University, Istanbul

^(a)feldemir@fatih.edu.tr, ^(b)elif.krky@gmail.com

ABSTRACT

The Order Accumulation and Sortation Systems (OASS) are getting more important as distribution centers try to gain competitive advantages. The parameters that affect the sorting time in OASS are analyzed in this study. The length and speed of conveyors, sending the packages within a wave, the wave size, number of the sorting lanes and the sorting strategy are the main parameter in OASS. The time required to sort mixed items depends on the sortation strategy used as well. Different sorting strategies and different conveyor models are analyzed in this study. Available analytical models assume that all orders are at the same size (quantity). In this study this assumption also is relaxed. Simulation models have been developed to compare different design alternatives and design strategies. For different order combinations and for various design choices, simulation is used to compare sortation strategies. The results have been given in tables that show which strategy should be used under which scenario. AutoMod Software is used as the simulation tool.

Keywords: sortation strategies, inventory management, simulation

1. INTRODUCTION

In today's competitive world, it is desirable that a distribution center runs at its optimal settings to gain a competitive advantage. More efficient distribution centers are needed to respond to the increasing competition and to an increased emphasis placed on time-based service. In distribution centers, long list of orders are put together in an intensive way. Each customer order can be full of various items at different quantities. In classical order picking procedure, each order is collected by an assigned picker and the products in this list might be kept at different storage addresses. Therefore, picker may end up traveling to far distances in a warehouse in order to complete the list and searching the items all over the warehouse. This situation often causes unnecessary transportation costs and ineffective worker utilizations. To overcome shortages mentioned above, zone picking method widely used in warehouses. In this picking method, the

items from different orders are arranged over again (batch orders) and the same product types collected by the same workers. With this method, order pickers are assigned to a specific zone. In this way, unproductive travel time will disappear. However, although this situation saves time and speed, the items of accumulated orders completely mixed. Therefore the items collected by different pickers arrive to the packing area at different times. To wait the other items from the same order, the ready packages are accumulated in accumulation zone. There is no doubt that these products (items) have to be sorted according to the product type and quantity before shipment. At this point, sortation systems (these are often automated systems) are used.

The optimal condition for a given system studied would be one in which the rate of sortation (i.e., throughput rate) is maximized, so minimizing the wave sortation time without increasing the capital and operating costs. There is a trade-off between the rate and cost. Using more resources such as labor and machines can increase the rate of sortation; however, the cost of sortation thus increases. This study focuses on maximizing the throughput rate of a given system and assumes that the other variables, such as cost and operating design parameters, are held within satisfactory limits.

There are different sortation strategies available. Fixed Priority Rule, Next Available Rule, and Earliest Completion Rule. In the literature a few analytical models have been developed for these sortation strategies. However the sortation models are limited to the one induction lane and one sortation lane.

2. LITERATURE REVIEW

Order Accumulation and Sortation System (OASS) related publications are very few. The first example related sortation strategies comes from (Bozer et al., 1988) developed Fixed Priority Rule (FPR) for lane assignment by simulating different wave of orders. Johnson (1998) developed a dynamic sortation strategy which is called Next Available Rule (NAR) and compared it with "FPR". Eldemir (2006) developed an alternative sortation strategy called Earliest Completion Rule (ECR) by using order statistics.

Closed-loop simple conveyor design researches contain different number of induction lane and number of sortation lanes. Especially the first studies are related one induction and one sortation lane. However later on, because of the variability in products and order sizes, the conveyor designs seen in the literature adapt into many induction and sortation lanes. Following table summarizes the literature on closed-loop conveyor system analysis according to number of its induction and sortation lane.

Table 1: Literature Review about Conveyor Design

Sortation Literature Summary					
Citation	Method	Problem Setting			
		One Ind.	Many Ind.	One Sort.	Many Sort.
Bozer and Sharp (1985)	Simulation	√		√	
Bozer et al (1988)	Simulation	√			√
Johnson and Lofgren(1994)	Simulation	√			√
Johnson (1998)	Analytical	√			√
Meller (1997)	Analytical	√			√
Schmidt and Jackman(2000)	Analytical	√		√	
Johnson and Meller(2002)	Analytical		√		√
Russell and Meller(2003)	Descriptive Mdl		√		√
Bozer (2004)	Analytical		√		√
Eldemir (2006)	Analytical	√		√	

3. SORTATION SYSTEM DESIGN

3.1 One-One Model

In this design model, one induction lane and one sortation lane is available. When the literature is evaluated thoroughly, it is observed that this model is the first applied model to the re-circulating conveyor. For instance, Bozer and Sharp (1985) have carried out this model in order to develop sortation strategies.

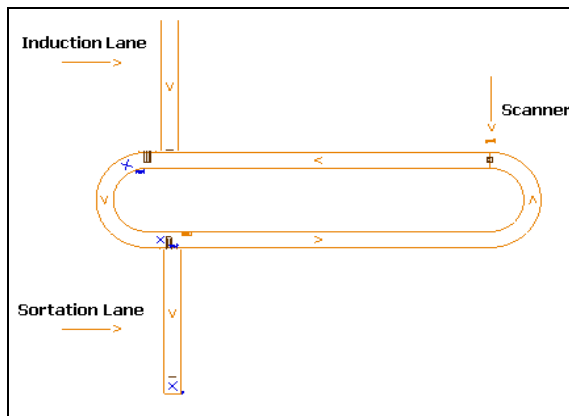


Figure 1: One- One Model Conveyor Design

3.2 One-Many Model

One –Many Model differs from previous model since it has more than one sortation lanes. When it is compared with others, this model is the most applied one. For instance, Johnson and Lofgren (1994), Johnson (1998), Meller (1997) have used this model in their studies.

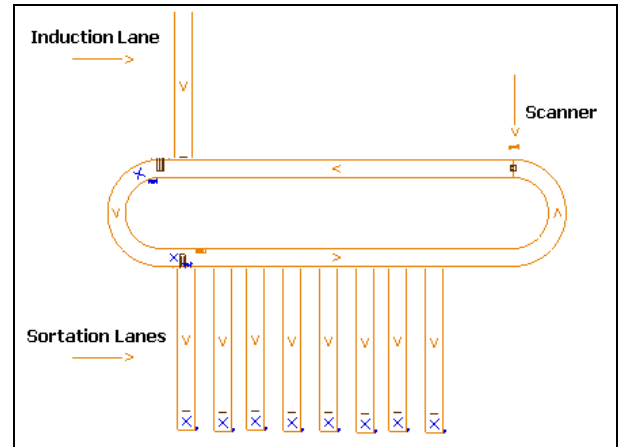


Figure 2: One- Many Model Conveyor Design

4. SORTATION STRATEGIES

Sortation strategies can be classified into two families, fixed priority rules (FPR) and dynamic assignment rules. In fixed priority rules, the orders are prioritized before sortation based on a certain rule. Dynamic assignment rules are assignment strategies that consider the item locations on the conveyor. The most common examples of this family are the next available rule (NAR) and the earliest completion rule (ECR). All parameters are determined below:

Table 2 : Notation

y	Number of items within an order
m	Number of orders within a wave
l	Length of the closed-loop conveyor
v	Speed of the conveyors
T	The time for an item to circulate around the main sortation line
n	Number of accumulation lanes
i	Item index within an order
j	Order index within a wave
q	The number of orders sorted thus far

4.1 Fixed Priority Rule (FPR)

Sortation time evaluation by using Fixed Priority Rule is given follows. The number of accumulation lane is accepted as one, and the number of items within the order is assumed to be constant.

Under FPR, The sorting time for all orders within the specific wave will be the summation of all the gaps and spreads as follows:

$$T_{FPR} = \frac{m \cdot T \cdot y}{y + 1} \quad (1)$$

4.2 Next Available Rule (NAR)

In this Next Available Rule, the expected sorting time each order depends on the number of orders which stays behind to be sorted. If it is supposed that the location of the items in the remaining orders are independent and uniformly distributed, and q the number of orders sorted thus further.

Under NAR, the sorting time for all orders within the specific wave will be as follows:

$$T_{NAR} = T \cdot \sum_{q=0}^{m-1} \left(1 - \frac{m-q}{y(m-q)+1} \right) \quad (2)$$

4.3 Earliest Completion Rule (ECR)

In dynamic assignment category, another sortation strategy model is Earliest Completion Rule (ECR). When sortation of an order is finished, the next order is determined based on the location of the last items. The order with the last item being closest to the accumulation lane is selected as next order to be sorted. Like NAR, the sortation time will be dependent on the number of orders which are going around on the main sortation lane. Assuming that all items are randomly and uniformly distributed and on the closed-loop conveyor and the item locations are independent of each other, from order statistics.

In Earliest Completion Rule, the total wave sortation time is given:

$$T_{ECR} = \sum_{q=0}^{m-1} \left(\frac{y(m-q)}{T^{y(m-q)}} \cdot \int_{l=0}^T [l^y \cdot (T^y - l^y)^{m-q-1}] dl \right) \quad (3)$$

where (l) is the location of last item on conveyor with the length of (T) .

5. EXPERIMENTATION

5.1 One-One Model

5.1.1 Analytical Model

To compare ECR, FPR and NAR, an empirical method is used. In developing the analytical models, several assumptions are made to facilitate the analysis. To illustrate the expressions for the three sorting strategies, the time to traverse the re-circulating conveyor is $T = 100$ seconds and there are $m = 10$ orders in each wave with $y = 5$ boxes per order. For analytical model experimentations MAPLE software is used.

Table 3: Sorting Times for Numerical Examples

Sorting Sequence (Order number)	Order Sorting time (seconds)		
	FPR	NAR	ECR
1	83,33	80,39	57,26
2	83,33	80,43	58,40
3	83,33	80,49	59,70
4	83,33	80,56	61,19
5	83,33	80,65	62,94
6	83,33	80,77	65,04
7	83,33	80,95	67,64
8	83,33	81,25	71,02
9	83,33	81,82	75,76
10	83,33	83,33	83,33
Total	833,30	810,64	662,29

5.1.2 Simulation Model

In order to compare ECR, FPR and NAR, a simulation method is used as well. Several assumptions are made to facilitate the simulation analysis. To illustrate the expressions for the three sorting strategies, the time to traverse the re-circulating conveyor is $T = 222$, 8 seconds and there are $m = 10$ orders in each wave with $y = 5$ boxes per order. A hundred repetitions are done for each simulation experiment. Then, the average of these repetitions is taken.

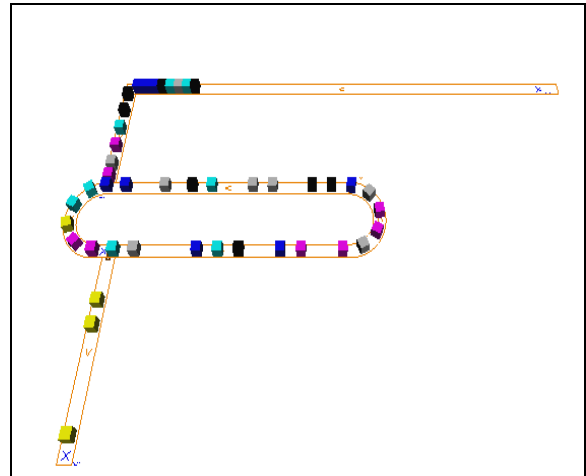


Figure 3: One-One Design Model Simulation Screenshot

For simulation model experimentations, AUTOMOD software is used. Figure 3 is the screenshot of the Automod software for One-One Design Model.

Table 4: Sorting Time Comparison for One-One Model by Using Simulation Model

Model	FPR	NAR	ECR
One-One Model	2.248,78	2.179,25	1.841,70

5.1.3 Simulation Model versus Analytic Model

Simulation model and Analytical model outputs, according to different scenarios are illustrated in following Table 5.

Table 5 : Sorting Time Comparison for One-One Model Both Simulation and Analytical Model

Orders/ Wave	Items/ Orders	Wave Sorting Time (seconds)			Wave Sorting Time (seconds)		
		<i>Analytical Model</i>			<i>Simulation Model</i>		
		FPR	NAR	ECR	FPR	NAR	ECR
24	1	2676	214	214	2915	442	442
12	2	1784	1478	992	2033	1724	1484
8	3	1338	1246	974	1574	1507	1268
6	4	1070	1033	878	1298	1279	1120
4	6	765	754	695	1003	991	920
3	8	595	591	563	823	829	792
2	12	412	411	403	640	643	634
1	24	214	214	214	442	442	442

It can be realized above Table 5 that Simulation Model's results are greater than Analytical Model in every case. The reason of this situation is that in simulation model, there are some additional spent times. The following shape points out spending time locations on the simulation system.

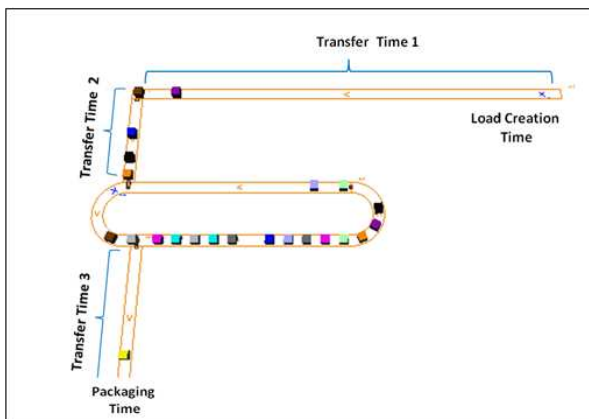


Figure 4: Extra Times Spending for Simulation

In Table 5 there are averages of extra time for each spending point which are shown in preceding shape. Besides, if subtraction is taken from simulation model to analytical model, the average difference is approximately 239 seconds. Also, summation of the extra spending time is 234.86 second. Thus, we can say that these two numbers are too close to each other.

Table 6: Sort of Spending Time for Simulation

Spending Time	Duration
Transfer Time 1	49,5
Transfer Time 2	29,76
Transfer Time 3	35,2
Load Creation time	69,7
Packaging Time	50,7
Total Time	234,86

5.2 One-Many Model

5.2.1 Simulation Model

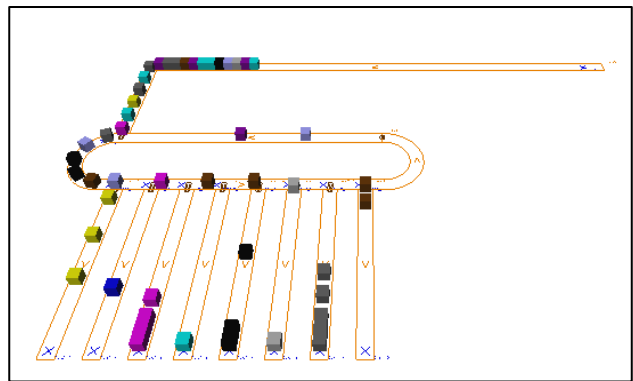


Figure 5: One- Many Design Model Simulation Screenshot

As it can be seen clearly, the best one is ECR model as One-Many Model. Since, the lowest value which emphasizes the average of the total sorting time is for ECR model.

Table 7: Sorting Time Comparison for One-Many Many by Using Simulation Model

Model	FPR	NAR	ECR
One-Many Model	690,61	678,72	651,21

5.3 Random and Equal Number of Items in the Order

Before studies assumed that number of item in an order are same. For Example, In Johnson (1998)'s article an accepted item number is $y=5$ for any event. In practice, it is known that it cannot be provided for every wave. Item number varies from one order to another order.

Table 8 : Sorting Time Comparison of Sorting Strategies According to Number of y

	Number of y	FPR	NAR	ECR
One-One Model	Random	2142,65	2.116,97	1.818,30
	Equal	2248,78	2.179,25	1.841,70
One-Many Model	Random	668,95	650,04	636,46
	Equal	690,61	678,72	651,21

From above shapes, random item size provides more time saving than equal item size in addition to, it does not reflect reality.

5.4 Number of Orders versus Number of Items

Different numbers of items and orders combinations are designed in order to comprehend the sortation strategies behavior for various situations. After preparing 8 combinations, for example, 24-1 means that there are 24 different orders within a wave and all orders have only one item, Table 5 represents the strategies' results:

Table 9: Total Sortation Time for Different Sortation Lane in O-OM

Orders/ Wave	Items/ Orders	Wave Sorting Time (seconds)		
		FPR	NAR	ECR
24	1	2.915,41	441,8	441,8
12	2	2.032,74	1.723,56	1.484,25
8	3	1.574,19	1.506,52	1.268,18
6	4	1.297,91	1.279,31	1.119,80
4	6	1.003,05	990,58	919,74
3	8	822,74	828,92	792,45
2	12	639,51	642,73	634,12
1	24	441,8	441,8	441,8

As can be seen from Figure 5, great savings can be accomplished in total sortation time for every experiment by using ECR

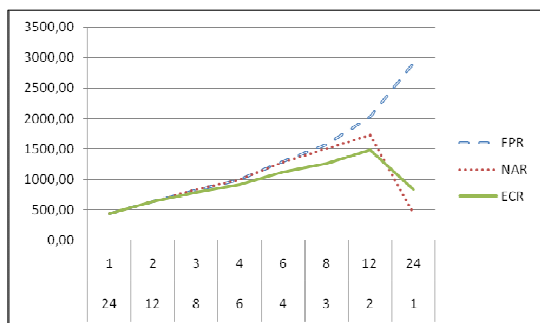


Figure 6: Total Sortation Time for Different Sortation Lane in O-OM

6. CONCLUSION

Available sortation strategies are compared and a set of modeling approach in simulation and in analytical is developed for the design and analysis of conveyor sortation system. Consequently, the following contributions are made:

Based on simulation models, FPR, NAR and ECR sortation strategies are compared. Overall outputs are represented as follows in Figure 7.

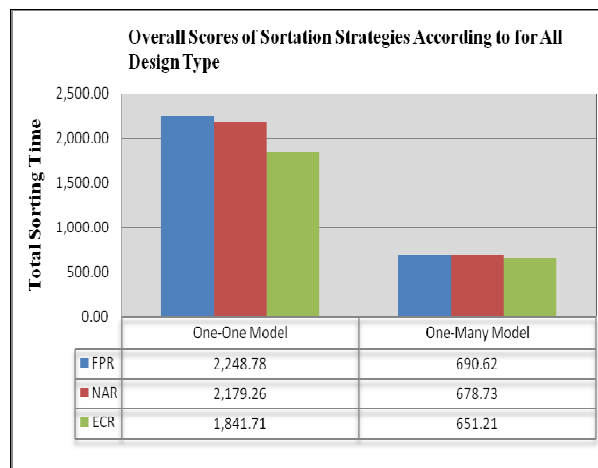


Figure 7: Effect of the Distance between Sortation Lanes

Simulation models are developed for all designs. Therefore, the results of simulation models are compared with analytical models and in this way, the validation of simulation is provided.

Different scenarios are simulated by varying design and operational parameters. For instance, despite of the literature, random item size in an order supports better results. Besides, it is more appropriate for a real case.

REFERENCES

Bozer, Y. A., and Hsieh, Y. J., 2004, Expected waiting times at loading stations in discrete-space closed-loop conveyors, *European Journal of Operational Research*, vol. 155(2), pp. 516-532

Bozer, Y.A., and Quiroz, M. and Sharp, G.P., 1988, An evaluation of alternate control strategies and design issues for automated order accumulation and sortation system, *Material Flow*, vol. 4, pp. 265-282

Demongodin, I. And Prunet, F., 1993, Simulation modeling for accumulation conveyors in transient behavior, *COMPEURO 93*, Paris, France, pp. 29-37

Dotoli M., Iacobellis, G, Stecco, G., Ukovich W., 2009, Performance Analysis and Management of an Automated Distribution Center” *IEEE*

Gagliardi, J, Angel R, Renauld, J. 2010, A simulation modeling framework of multiple-Aisles Automated Storage and Retrieval System” *CIRRELT -57*

Harit, S., Taylor G. D., 1995, Framework for the Design and Analysis of Large Scale Material Handling Systems” *Winter Simulation Conference*

Jayaraman, A., R. Narayanaswamy, et al. 1997, A sortation system model, *Winter Simulation Conference*

Jing, G. and Kelton, W. D. and Arantes, J.C., 1998, Modeling a controlled conveyor network with merging configuration, *Proceeding of the Winter Simulation Conference*, pp. 1041-1048

Johnson, M. E., 1998, The impact of sorting strategies on automated sortation system performance”, *IIE Transactions*, vol. 30(1), pp. 67-77

- Johnson, M. E. and Russell, M., 2002, Performance analysis of split-case sorting systems, *Manufacturing & Service Operations Management*, vol. 4(4), pp 258-274
- Johnson, M. E. and Lofgren, T., 1994 Model Decomposition speeds distribution center design”, *Interfaces*, vol. 24(5), pp. 95-106
- Kale, N., Zottolo, M., Ülgen, O.M., &Williams, E.J. 2007, Simulation improves end-of line sortation and material handling pickup scheduling at appliance manufacturer. *Proceedings of the 2007 winter simulation conference* pp.1863–1868, Washington, D.C., USA
- Koster, R., Le-Duc., T., Roordbergen, K., J., 2007 Design and control of warehouse order picking: A literature review, *European Journal of Operational Research* 182, 481–501
- Le-Duc, T., and de Koster, R. 2005, Determining Number of Zones in a Pick-and-pack Order picking System, *ERIM Report Series Research in Management*, Rotterdam
- Maxwell, W. and Wilson, R., 1981, Deterministic models of accumulation conveyor dynamics, *International Journal of Production Research*, vol. 19(6), pp. 645-655
- Meller, R. D. 1997, Optimal order-to-lane assignments in an order accumulation/sortation system, *IIE Transactions* 29: 293-301
- Roodbergen, K.J., and Vis, I.F.A., 2009, A survey of literature on automated storage and retrieval system. *European Journal of Operational Research*, Vol.194, pp.343, 362
- Russell, M. L. and Meller, R. D., 2003, Cost and throughput modeling of manual and automated order fulfillment systems, *IIE Transactions*, 35, 589-603
- Schmidt, L. C. and Jackman, J., 2000, Modeling recirculating conveyors with blocking, *European Journal of Operational Research*, vol. 124, pp. 422-436
- Sonderman, D., 1982, An analytic model for recirculating conveyors with stochastic inputs and outputs, *Internal Journal of Production Research*, vol. 20(5), pp.591-605

AUTHORS BIOGRAPHY

Fahrettin Eldemir is an Assistant Professor in the Department of Industrial Engineering at Fatih University, Istanbul, Turkey. He has a Ph.D. in Decision Sciences and Engineering Systems, an M.E. in Operations Research and Statistics B.S in Industrial and Management Engineering from Rensselaer Polytechnic Institute. Before joining to Fatih University, he served as an engineer and a consultant at Omega Advanced Solutions.

DEVELOPMENT OF THE SURFACE TO AIR MISSILE SIMULATOR THROUGH THE PROCESS OF COMPONENT COMPOSITION AND DYNAMIC RECONFIGURATION OF WEAPON SYSTEM

J.B. Suk^(a), J.O. Lee^(b), Y.H. Seo^(c)

^(a)Department of Industrial Management Engineering, Korea University, Seoul, Korea

^(b)Department of Industrial Management Engineering, Korea University, Seoul, Korea

^(c)Department of Industrial Management Engineering, Korea University, Seoul, Korea

^(a)sjb1010@korea.ac.kr, ^(b)etoztt@korea.ac.kr, ^(c)yoohoseo@korea.ac.kr

ABSTRACT

The concerns of the technology of the reusable component are increasing for producing the software effectively and developing the products quickly on demand of customers. Technologies of composition and dynamic reconfiguration of the component are needed to develop the simulator of the weapon system in field of national defense. It is needed to develop the surface to air missile simulator as the example of the weapon system through component-based development. In order to develop the simulator, creation of components of the simulator is required and the process of reconfiguration is defined and realized. The process of reconfiguration consists of four parts: management, semantic test, composition and performance evaluation. In first part, the developer can easily create lots of component models including characteristics of the product of the weapon system and quickly manage them by using the management tool. In second part, semantic tester tests that whether the reconfiguration and import in the simulator are possible or not. In next part, the composer constructs the product models configuring surface to air simulator by composing existing component and newly created component. Then the results of construction can be entities of detecting radar, the surface to air missile, the launcher of the missile and aircrafts of the simulator of the weapon system. In last part, the existing 3D-based simulator for evaluating the performance of the component of the weapon system confirms the effects of reconfigured components by importing it. Therefore, this study provides basic framework to simulate common weapon systems through the technology of composition and the process of dynamic reconfiguration.

Keywords: surface to air missile, modeling and simulation, reusability, component

1. INTRODUCTION

Developing a simulation model will require a lot of time and money in the field of modeling and simulation. Developed model is hard to be used as other applications by its closed architecture. To overcome this

hardness, we require developing a basic component model that can give flexibility, scalability, reusability in modeling and simulation. Also, we require assembling and composing component model and need a way of component based development for new simulation model. This way is to assemble developed components and create new applications. It has advantages such cost and time of development and maintenance of software. Ultimate goals of this development are maximizing reusability of components and increasing productivity of the software.

New software model of the weapon system is developed by applying the technology of software product line in the field of modeling and simulation of weapon system that has a lot of similarities for each product. Therefore, in order to effectively reuse, pre-built, core assets of the software, studies are needed for this processes; management, selection, assembly and composition of component models.

The component reconfiguration of weapon system and related works are handled in second chapter. In third chapter, the development of surface to air missile simulator and its application example are organized. In the last chapter, it ends the study with conclusion and future work.

2. COMPONENT RECONFIGURATION OF WEAPON SYSTEM

Through software product line based development, it designs the dynamic component reconfiguration framework which can reuse and compose quickly the component of weapon system. Software product line based development is the way of keeping common elements of the product and only changing the distinguishing characteristics of it for reusing components. In order to realize the framework, components will be defined and developed as the unit of reusable model. The component model consists of physical part and behavior part. It is reconfigured through considering characteristics of two parts of it and being selected and composed.

The component model can be created, modified and deleted by graphic-based management tool of the

component. Also, it can be approved to reuse for constructing the reconfiguration framework of weapon system by semantic tester of the component after different developers created it. In order to develop the simulator of surface to air missile of weapon system, it is reconfigured as composing the reusable components. For composing existing component and newly created component by the tool, the composers of the component model are developed. They also consist of physical composer and behavior composer. Physical composer constructs physical model of the product by composing components grouped by physical characteristic of component model of weapon system like the performance of the component. Furthermore, behavior composer constructs behavior model of product by including logics and rules in product. Then we can reconfigure product model of the simulator of weapon system through use of two composers.

2.1. The Process of Component Composition

2.1.1. Software Product Line Engineering

The recent interest in the field of modeling and simulation is development of reusable and configurable simulation model. It is generalized small quantity batch production from rise of importance of personalization of the customers and issued software product line engineering for corresponding the demand of customers and environment of the market. Software product line engineering is paradigm that reusing the core asset from similar product and inevitable choice for development of new product that satisfying time and economic restriction of the development (Chen, Yu, Gannold, Gerald C., Collofello and James 2006).

The definition of software product line engineering is paradigm for developing the applications of software sing platform and mass customization (Pohl, Klaus and Bockle 2005). The objective of software product line is increasing efficiency of the development by reusing the product strategically as analysis of similarities and differences between the products of software set. So it is to choice the option of product according to the user's intention and to produce new software on one of basic platforms along the choice. The basic concepts of it are understanding the similarity among characteristics of the product and supporting the variability among the application programs by distinguishing that of the product complicity (Clements, Paul and Northrop 2002).

The structure of software product line engineering is as below Figure 1. It is classified domain engineering and application engineering. Domain engineering is process of development by setting up core assets from common function through analysis of similarity and the variability. And it specifies requirements and characteristic of components and grafts them onto the basic framework. On the other hand, application engineering classifies the variability from characteristics of components and develops the target application. These two engineering help developing target application by reconfiguring pre-modeling

components as specification of the variability from the basic framework.

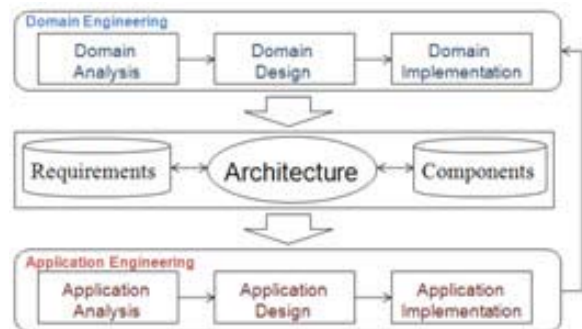


Figure 1: The Structure of Software product Line Engineering

2.1.2. Development of Basic Component Model

By applying software product line engineering, it is needed to develop basic component model of weapon system that is possible to reconfigure as unit of the component. This study develops basic component model that is possible to create various products efficiently by considering scalability of weapon system. Basic component model consists of physical part and behavior part. Physical component model contains representing physical characteristics of the weapon model. In order to develop physical component, it is needed to define criteria for description of weapon system. It is decided by requirements of the domain. This study decided criteria limits at entity level.

Next, this study considers fidelity of physical component model that represents correspondence with real entity model like the missile. High fidelity of entity model causes low reality. So it is needed to decide at optimum level. As example, the missile is composed 4 parts by analyzing its functionality and structure in figure 2.

Behavior component model also represents tactical behavior and decision. And it is developed by dividing 3 parts; basic behavior, composition behavior and judgment.

- Basic behavior: The behavior that performs alone fundamentally designed
- Composition behavior: Combined behavior except basic and original function of entity
- Judgment: It represents the state of the behavior by deciding changing behavior or not.

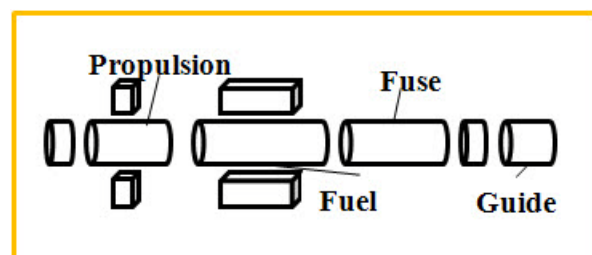


Figure 2: Structure of Missile Considering Fidelity

2.2. The Process of Dynamic Reconfiguration

For dynamic reconfiguration of component model, developed basic component models are configured onto product model that means the entity of weapon system. Next, the product models are also reconfigured to system model as application program by following the designed DCRF (Dynamic Component Reconfiguration Framework).

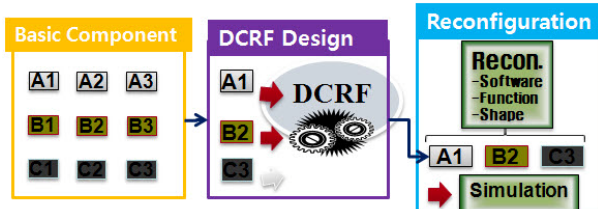


Figure 3: The Process of Dynamic Reconfiguration

2.2.1. Description of Dynamic Component Reconfiguration Framework

The framework of dynamic component reconfiguration is proposed based on requirement of customers and reusable basic components. It consists of system layer, product layer, component layer and supporting tools. At component layer, the components are reconfigured to form product models and they also are reconfigured to develop the target application system.

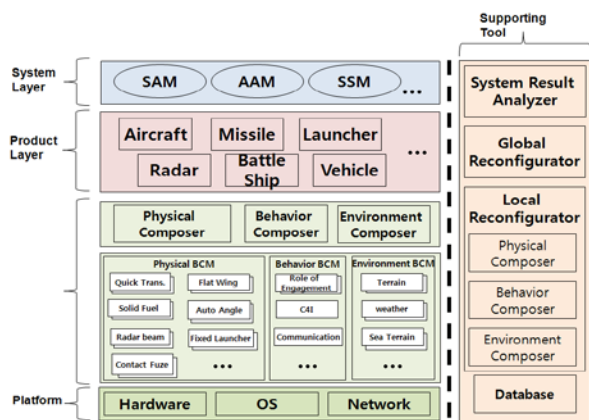


Figure 4: Dynamic Component Reconfiguration Framework

2.3. Related works

Typical national defense simulation system by reusing components and configuring is OneSAF (One Semi-Automated Forces) model. OneSAF is comprehensive model that used in ACR (Advanced Concepts and Requirements), RDA (Research, Development and Acquisition) and TEMO (training, exercise and military operation) (Giampapa JA, Sycara K, Owens S, Ginton R, Seo YW, Yu B 2004). OneSAF model has the concept of assembled software product line. This concept is based on completed system that consists of many products, and they also consist of many components. OneSAF is possible to comprise

various and differenced product needed to customers as other composition of products.

3. DEVELOPMENT OF THE SURFACE TO AIR MISSILE SIMULATOR

By applying designed the framework, the surface to air missile simulator is developed as one of software application. By using this simulator, reusability of developed component model is confirmed and performance evaluation of it is also possible. The user develops surface to air missile simulator by importing composed physical part and behavior part of the product model in existing 3D-based performance evaluation simulator of the weapon system. In addition, the user reconfigures various products and evaluates performance of it through the use of management tool and semantic tester.

3.1. Framework of Surface to Air Missile Simulator

As application from reconfiguration framework, SAM simulator is one of systems in system layer. Basically, it needs 4 product models such as missile, radar, launcher, and aircraft. With these as the center, there exist various components in these. For completing the product, core assets are existed and distinguishing components are composed with relative function independently.

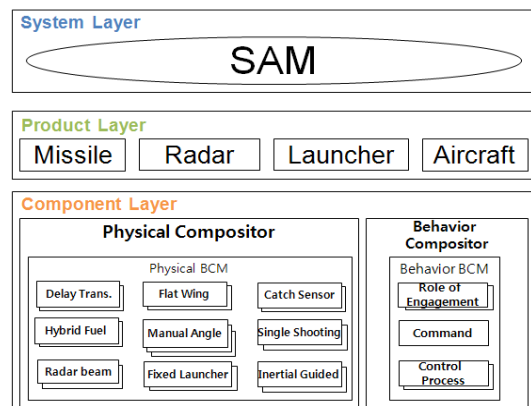


Figure 5: Framework of SAM Simulator

3.2. Application example

In this study, it develops SAM simulator that comprised of 4 product models based on the framework. Objective of this simulator is to bring down aircraft of enemy that has the mission of destroying our core facilities by shooting our missiles. The roles of each product model are as below in SAM simulator.

1. Radar: Detect aircraft and delivery information and order.
2. Launcher: Receive information and order of aircraft and fire the missile.
3. Missile: Hit the aircraft of enemy.
4. Aircraft: Move and destroy the core facilities.

After the execution of simulation, the simulator provides accuracy rate and information of product models in real time.

Also the process of reconfiguration as application is realized the developments of management tool, semantic, physical composer and behavior composer.

- Management tool: It manages so many various the basic components that have distinguishing performance models such that add, delete and modify them.
- Semantic Tester: It tests whether newly added or modified component model is possible to be used at simulator of weapon system.
- Physical Composer and Behavior Composer: By composing the basic component models, both composer creates new product models that having characteristics that are choiced by user

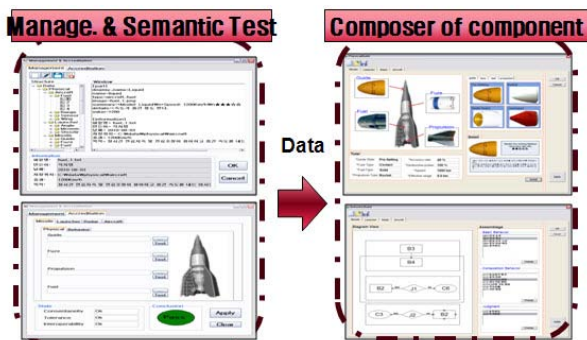


Figure 6: Application of Reconfiguration

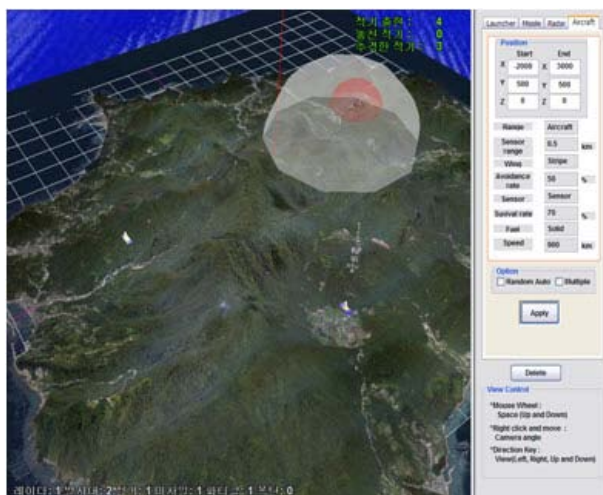


Figure 7: Surface to Air Missile Simulator

4. CONCLUSION

In this study, dynamic component reconfiguration framework is designed and realized through software product line-based development to produce the simulator easy and fast. Developed components of modeling and simulation of the weapon system increase productivity of simulation model and reduce the development cost and time. It is possible to reconfigure products as the user intended by composing reusable

components that developed by the management tool and semantic tester of the component. Product models are entities of missiles, detecting radar, launchers and aircrafts for constructing surface to air missile simulator through the use of the composers. The entities include not only the physical model, but also include the behavior model to express developer's logical and specific intentions. It is possible to evaluate the characteristics of the component of the system quickly and easily through realization of the surface to air simulator. In addition, more studies are needed to store and manage the components

ACKNOWLEDGMENTS

This work was supported by Defense Acquisition Program Administration and Agency for Defense Development under the contract UD110006MD, Korea.

REFERENCES

Chou SC, Chen YC. 2006. Retrieving reusable components with variation points from software product lines, *Information Processing Letters*, 99 (3), 106-110.

Giampapa JA, Sycara K, Owens S, Ginton R, Seo YW, Yu B, 2004. Extending the OneSAF Testbed into a C4ISR Testbed. *Simulation*, 80(12), 681-691.

Henderson C, Rodriguez A, 2002. Modeling in OneSAF. *Computer generated forces and behavioral representation*, 337-348.

Chen Y, Gerald C, Gannod, Collofello JS, 2006. A software product line process simulator. *Software process improvement and practice*, 11(4),385-409.

Clements, Paul and Northrop, Linda, 2002. *Software Product Lines: Practices and Patterns*, 5-50.

Pohl, Klaus and Bockle, Gunter, 2005. *Software Product line engineering : Foundation, Principles, and Techniques*, Springer, 159-370.

MODELLING AND SIMULATING A BENCHMARK ON DYNAMIC RELIABILITY AS A STOCHASTIC ACTIVITY NETWORK

Daniele Codetta-Raiteri^(a)

^(a)Dipartimento di Informatica, Università del Piemonte Orientale, Alessandria, Italy

^(a)raiteri@mfn.unipmn.it

ABSTRACT

Several versions of a benchmark on dynamic reliability taken from the literature are examined: each version deals with particular aspects such as state dependent failure rates, failures on demand, and the repair of components. The benchmark was modelled in the past, using two types of Petri Nets; in this paper, we exploit another Petri Net based modelling formalism called Stochastic Activity Network (SAN). This allows a more compact model of the system by exploiting input and output gates, together with the possibility to represent float variables by means of extended places. The SAN model of the system undergoes simulation in order to compute the system unreliability: the results are coherent with those obtained with other methods, and this confirms that Petri Net based models can be a valid approach to dynamic reliability evaluation.

Keywords: dynamic reliability, benchmark, modelling, simulation, Stochastic Activity Network, Petri Net.

1. INTRODUCTION

We talk about safety critical systems when their incorrect behaviour may cause undesirable consequences to the system itself, the operators, the population or the environment. This definition fits categories of systems such as industrial production plants, electric power plants, and transportation systems. *Dependability* is a fundamental requirement for this class of systems. The Dependability level of a system is the degree of confidence that the system will provide its service correctly during its life cycle.

There are two main methods to evaluate the Dependability: the *Measurement-based* method and the *Model-based* method. The first one requires the observation of the behaviour of the physical objects composing the system, in the operational environment. This method is more believable, but it may be impractical or too expensive. Therefore the model-based method is preferable and consists of the construction of a model representing the behaviour of the system in terms of modelling primitives defined in a formalism. The model of the system must be a convenient abstraction of the system; this means that the level of accuracy of the model must be high enough to represent correctly the aspects of the system

behaviour which are relevant to Dependability evaluation. The mechanisms that lead to the failure of a technological item are very complex and depend on physical, technical, human and environmental factors which may not obey deterministic laws; so, the model-based method follows the probabilistic approach.

The concept of Dependability is quite general; in order to quantify the Dependability, we need particular measures: the *Reliability* $R(t)$ of an item (component or system) is the probability that the item performs the required function in the time interval $(0, t)$, given the stress and environmental conditions in which the item operates; the *Unreliability* $U(t)=1-R(t)$ is the probability that the item is in the failed state at time t (Sahner, Trivedi, and Puliafito 1996).

We talk about dynamic reliability (Marseguerra, Zio, Devooght, and Labeau 1998) when the reliability parameters of the system change according to the current configuration of the system. For instance, the failure rate of a component may be expressed as a function of one or more variables describing the current behaviour or state of the system. In dynamic reliability, considering only the combinations of failure events is not sufficient to evaluate the system (un)reliability, but we actually have to take into account the complete behaviour of the system. This means modelling the normal functioning of the system, the occurrence of failure events and their effect on the system functions. For this reason, combinatorial models such as *Fault Trees* and *Reliability Block Diagrams* (Sahner, Trivedi, and Puliafito 1996) are not suitable to deal with cases of dynamic reliability because such kinds of model can only represent combinations of component failure events. Their extensions such as *Dynamic Fault Trees* (Dugan, Bavuso, and Boyd 1992) and *Dynamic Reliability Block Diagrams* (Distefano and Xing 2006) introduce the possibility to represent dependencies among the events, but they still only focus on the failure propagation ignoring the other aspects of the system behaviour.

Such aspects could be represented instead by means of state space based models, such as *Markov Chains* and *Stochastic Petri Nets* (Sahner, Trivedi, and Puliafito 1996), but their use typically leads to the state space explosion because the complete dynamics of the system has to be modelled. Therefore the model analysis

becomes impractical because of the high computing cost (and time). For these reasons, dynamic reliability cases are typically evaluated by means of simulation.

In this paper, we take into account a benchmark on dynamic reliability taken from the literature (Marseguerra and Zio 1996). The system consists of a tank containing some liquid whose level is influenced by two pumps and one valve managed by a controller, with the aim of avoiding the failure of the system occurring in case of the dry out or the overflow of the liquid. Such events are consequences of the pumps or valve failure because in such condition the components ignore the orders coming from the controller. The dry out or the overflow does not occur as soon as a particular combination of component failures occurs, but such basic failures may influence the liquid level, possibly leading the system to the failure after that some time has elapsed or another event has happened. Because of this, not only the component failure combinations have to be modelled, but also any variation in the liquid level caused by the components action or failure.

In Marseguerra and Zio (1996), several versions of the benchmark are proposed: the initial case of state independent failure rates (that we call Version 1), the case of state dependent failure rates (Version 2), the case with possible failure on demand of the controller (Version 3), the case with repairable components (Version 4), and finally the case with temperature dependent failure rates (Version 5). All the versions are described in Sec. 3. In Sec. 5, each version of the system is modelled as a *Stochastic Activity Network* (SAN) (Sanders and Meyer 2001), a particular form of Stochastic Petri Net; the SAN formalism is described in Sec. 4. The SAN models are designed and simulated by means of the *Möbius* tool (Deavours, Clark, Courtney, Daly, Derisavi, Doyle, Sanders, and Webster 2002); the aim is to compute the system unreliability in each version of the benchmark (Sec. 6). The advantages of SAN with respect to other forms of Petri Nets (Sec. 2) are presented in Sec. 7.

2. RELATED WORK

In Marseguerra and Zio (1996) the unreliability of the system in Versions 1, 2, and 3 is evaluated first in an analytical way by computing the probabilities of the minimal cut sets of component failure events leading to the dry out or the overflow. Then, the system unreliability is evaluated by means of Monte Carlo simulation. The cut set analysis only considers the combinations of events, while the Monte Carlo simulation deals with the complete dynamics of the system: therefore there is a relevant difference between the unreliability values returned by the two approaches, in particular in Versions 2 and 3. Such difference highlights the necessity to take into account the complete behaviour in order to evaluate the system in an accurate way. Versions 4 and 5 are only evaluated by means of Monte Carlo simulation in Marseguerra and Zio (1996).

In Codetta and Bobbio (2005a), Versions from 1 to 4 have been modelled as *Generalized Stochastic Petri Nets* (GSPN) (Ajmone, Balbo, Conte, Donatelli, Franceschinis 1995) by means of the *GreatSPN* tool (Chiola, Franceschinis, Gaeta, and Ribaudo 1995). The GSPN model can undergo analysis, but this requires the liquid level to be discretized in several intermediate integer levels. This is because in GSPN models, only discrete variables can be represented as the number of tokens (*marking*) inside places. The number of such intermediate levels must not be high; otherwise the state space dimensions may explode, with the consequent increase of the computing cost. Moreover, in the GSPN model, some deterministic timed events such as the action of the pumps or the valve on the liquid level, have to be approximated by stochastic events, in order to allow the model analysis. So, despite of the advantages given by the model analysis instead of simulation, the GSPN model suffers from some approximation about the liquid level and its variations during the time.

In Codetta and Bobbio (2005a), Versions from 1 to 4 are modelled also as *Fluid Stochastic Petri Net* (FSPN) (Gribaudo, Sereno, Horvath, and Bobbio 2001), a particular form of Petri Net including fluid places containing a continuous amount of fluid instead of a discrete number of tokens. Fluid places directly represent continuous variables, such as the liquid level or temperature. FSPN models are typically simulated. This can be done by means of the *FSPNedit* tool (Gribaudo 2001).

Version 5 of the benchmark could not be modelled as a GSPN because the temperature would have been approximated in several intermediate integer values leading to an unacceptable approximation of the current temperature, and consequently to the approximation of the current failure rates. Besides this, the expression of the failure rate as a function of the liquid temperature (Eq. 3 in Sec. 3) cannot be represented in the GSPN model. Moreover, the combination of the possible temperature values together with the possible values of the other parameters describing the system state, would have led to the explosion of the underlying state space dimensions. For this reason, in Codetta and Bobbio (2005b), Version 5 of the system has been modelled and simulated only as a FSPN.

In this paper, the benchmark is modelled and simulated using the SAN formalism. The SAN models can undergo analysis as well, but to this aim, the deterministic activities (transitions) have to be replaced by stochastic activities reducing the accuracy of the model with respect to the real behaviour of the system.

3. THE BENCHMARK

The system (Fig. 1.a) is composed by a tank containing some liquid, two pumps (P1 and P2) to fill the tank, one valve (V) to remove liquid from the tank, and a controller (C) monitoring the liquid level (H) and acting on P1, P2 and V.

Initially H is equal to 0, with P1 and V in state ON, and P2 in state OFF; since P1, P2 and V have the same level variation rate ($Q=0.6 \text{ m/h}$), the liquid level does not change while the initial configuration holds. The cause of a variation of H is the occurrence of a failure of one of the components consisting of turning to the state Stuck-ON or Stuck-OFF. The time to failure is random and obeys the negative exponential distribution; the failure rate (Tab. 1) does not depend on the current state of the component, so the effect of the failure is the stuck condition, while the state transitions toward the Stuck-ON or the Stuck-OFF state are uniformly distributed (Fig. 2.a).

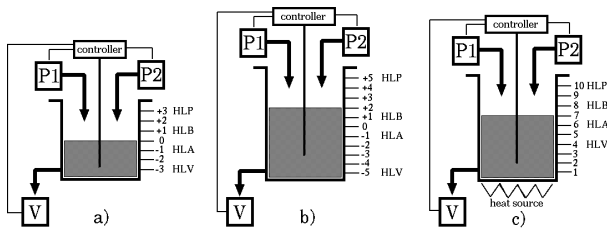


Figure 1. The System Schemes in Versions 1, 2, 3 (a), in Version 4 (b), in Version 5 (c)

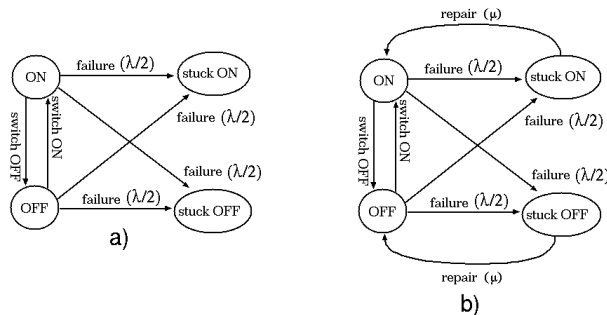


Figure 2: The States of P1, P2 and V in Versions 1, 3, 5 (a). The States of P1, P2 and V in Version 4 (b)

Tab. 2 shows how H changes with respect to the current configuration of the component states; the controller believes that the system is correctly functioning while H is inside the region between the levels HLA (-1 m) and HLB (+1 m). If H exceeds HLB, the controller orders both pumps to switch OFF, and the valve to switch ON, with the aim of decreasing H (Tab. 3) and avoiding the liquid overflow; this event occurs when H exceeds the level HLP (+3 m). If a component is stuck, it does not obey the controller order and maintains its current state.

The other undesired situation is the tank dry out; this happens when H is below HLV (-3 m); if H goes below HLA, the controller orders both pumps to switch ON, and the valve to switch OFF, with the aim of increasing H (Tab. 3) and avoiding the dry out.

The failure of the whole system happens when the dry out or the overflow occurs.

We denote such configuration of the system as **Version 1**. In this paper, we deal with several versions of the benchmark, still proposed in Marseguerra and Zio (1996).

Table 1: Failure Rates in Versions 1, 3, 4

component	failure rate (λ)
P1	0.004566 h^{-1}
P2	0.005714 h^{-1}
V	0.003125 h^{-1}

Table 2: The Level Variation in each State Configuration

configuration	P1	P2	V	effect on L
1	ON	OFF	OFF	\uparrow
2	ON	ON	OFF	$\uparrow\uparrow$
3	ON	OFF	ON	$=$
4	ON	ON	ON	\uparrow
5	OFF	OFF	OFF	$=$
6	OFF	ON	OFF	\uparrow
7	OFF	OFF	ON	\downarrow
8	OFF	ON	ON	$=$

Table 3: Control Boundaries and Laws

boundary	P1	P2	V
$H < \text{HLA}$	ON	ON	OFF
$H > \text{HLB}$	OFF	OFF	ON

3.1. Version 2: state dependent failure rates

In this version, the failure rate of a component changes according to its current state and the state reached as a consequence of the failure (Tab. 4).

Table 4: Failure Rates for each Component in each State, in Version 2

component	from	to	failure rate (λ)
P1	ON	Stuck-ON	$0.004566/2 \text{ h}^{-1}$
P1	ON	Stuck-OFF	$0.004566/2 \text{ h}^{-1}$
P1	OFF	Stuck-ON	0.045662 h^{-1}
P1	OFF	Stuck-OFF	0.456621 h^{-1}
P2	ON	Stuck-ON	0.057142 h^{-1}
P2	ON	Stuck-OFF	0.571429 h^{-1}
P2	OFF	Stuck-ON	$0.005714/2 \text{ h}^{-1}$
P2	OFF	Stuck-OFF	$0.005714/2 \text{ h}^{-1}$
V	ON	Stuck-ON	$0.003125/2 \text{ h}^{-1}$
V	ON	Stuck-OFF	$0.003125/2 \text{ h}^{-1}$
V	OFF	Stuck-ON	0.031250 h^{-1}
V	OFF	Stuck-OFF	0.312500 h^{-1}

3.2. Version 3: controller failure on demand

Here, the controller has a probability of failure on demand equal to 0.1. This means that each time H exceeds the region of correct functioning ($\text{HLA} < H < \text{HLB}$), the controller may not send the command to P1, P2 and V, so they maintain their current state.

3.3. Version 4: repairable components

In this version, a stuck (failed) component can be repaired during the grace period which begins when the region of correct functioning is exceeded for the first time, and ends when the dry out or the overflow occurs. The time to repair of a component is a random variable obeying the negative exponential distribution with the repair rate equal to 0.2 h^{-1} . The effect of the repair

consists of removing the stuck condition of a component (Fig. 2.b). As soon as the repair is completed, the component is set to the state ON or OFF if H is currently out of the region of correct functioning (Tab. 3). Moreover, the repaired component can respond to future orders from the controller, changing its state again if necessary. After the repair, a component may fail and undergo repair again.

In this version of the benchmark, HLV and HLP are set to $-5 m$ and $+5 m$ respectively (Fig. 1.b).

3.4. Version 5: temperature dependent failure rates

The current temperature (T) of the liquid in the tank is taken into account, and T influences the failure rate of the components P1, P2 and V. A heat source increases T according to the heating power $w = 1m^{\circ}C/h$. There is no heat released outside the tank, and the heat is uniformly distributed on the liquid. The initial temperature of the liquid inside the tank is $15.6667^{\circ}C$; the temperature of the liquid introduced in the tank by the pumps is $T_{in} = 15^{\circ}C$, and it gets mixed instantaneously with the liquid in the tank.

The level variation rate for P1, P2 and V is now $Q=1.5 m/h$. Assuming that a pump is activated at time t_0 and is still active at time $t > t_0$, we use the Eq. 1 and 2 to provide the liquid level and temperature respectively, at time $t > t_0$. In Eq. 1, L_0 is the liquid level at time t_0 ; in Eq. 2, T_0 is the liquid temperature at time t_0 .

The failure rates of P1, P2 and V are temperature dependent according to Eq. 3 where λ_0 is the failure rate of the component for a temperature equal to $20^{\circ}C$ (Tab. 5). Besides the dry out and the overflow, another condition determines the failure of the system: T reaches $100^{\circ}C$.

The initial level of the liquid in the tank is $7 m$; HLA and HLB are set to $6 m$ and $8 m$ respectively; HLV and HLP are equal to $4 m$ and $10 m$ respectively (Fig. 1.c).

Table 5: Failure Rates for $T = 20^{\circ}C$ in Version 5

component	λ_0
P1	$0.004566 h^{-1}$
P2	$0.005714 h^{-1}$
V	$0.003125 h^{-1}$

$$L(t) = L_0 + Q \cdot (t - t_0) \quad (1)$$

$$T(t) = T_0 \cdot L_0/L(t) + T_{in} \cdot Q \cdot (t - t_0) / L(t) \quad (2)$$

$$\lambda(T) = \lambda_0 \cdot (0.2e^{0.005756(T-20)} + 0.8e^{-0.2301(T-20)}) \quad (3)$$

Three versions of the benchmark are characterized by the aspects described above:

- **Version 5.1:** the controller cannot fail.
- **Version 5.2:** the controller has a probability of failure on demand equal to 0.2.
- **Version 5.3:** initially the controller has a probability of failure on demand equal to 0.2; due to the wear of the controller, such probability is increased of 50% every time that the controller has to act (at each demand).

4. BASIC NOTIONS ABOUT SAN

A SAN model can contain two kinds of *places*. A *standard* place contains a certain number of *tokens* (*marking*) corresponding to an integer variable. The marking of an *extended* place corresponds to a variable whose type is not integer, but it can be a float number, a character, an array, etc. A place graphically appears as a circle, while *activities* (transitions) graphically appear as vertical bars.

The completion (firing) of an activity is enabled by a particular condition on the marking of a set of places. This marking can be expressed by connecting the activity to the standard places by means of oriented arcs, as it is possible in Petri Nets. The effect of the activity completion on the standard places can be specified in the same way. Another way to express the enabling condition consists of using *input gates*. An input gate is connected to an activity and to a set of standard or extended places; the input gate is characterized by two expressions:

- a *predicate* consists of a Boolean condition expressed in terms of the marking of the places; if this condition holds, then the activity is enabled to complete.
- A *function* expresses the effect of the activity completion on the marking of the places.

A SAN model can contain also *output gates*. The role of an output gate is specifying only the effect of the activity completion on the marking of the places. Therefore an output gate is characterized only by a function. The marking enabling the same activity can be expressed by means of oriented arcs, or by means of an input gate. Gates graphically appear as triangles (input gate: ◀ - output gate: ▶).

In a SAN model, it is possible to set several completion cases for an activity; each case corresponds to a certain effect of the completion and has a certain probability: when the activity completes, one of the cases happens. A case graphically appears as a small circle close to the activity; from the case an arc is directed to a gate or a place.

The completion of an activity can be immediate or timed. In the second case, the completion time can be constant or random. A random completion time has to be ruled by a probability distribution; in this paper, we always resort to the negative exponential one, but several other distributions are available in the SAN formalism. In this paper (and in Codetta (2011)), we call “immediate activity” an activity completing as soon as it is enabled; we call “deterministic activity” an activity whose completion time is deterministic and not immediate; finally, we call “stochastic activity” an activity whose completion time is random.

5. MODELING THE SYSTEM

Each version of the benchmark (Sec. 3) has been modelled as a SAN where each aspect of the system behaviour is represented: the state of components, the

failure events, the current liquid level and its variations, the orders by the controller, the liquid dry out or overflow, etc.

5.1. Modelling Version 1

The SAN model of Version 1 is depicted in Fig. 3 where the current state of the pump P1 is represented by means of the places $P1_on$ and $P1_stuck$: if $P1_on$ is empty, this means that P1 is off; if instead the place $P1_on$ is marked with one token, then P1 is on. The place $P1_stuck$ is used to represent the stuck condition of P1: if such place is empty, then P1 is not stuck; if instead the place $P1_stuck$ contains one token, this means that P1 is currently stuck. According to the marking combinations of the places $P1_on$ and $P1_stuck$, we can model all the possible states of P1: ON, OFF, Stuck-ON, Stuck-OFF (Sec. 3).

Initially the place $P1_on$ is marked with one token, and the place $P1_stuck$ is empty, in order to model that P1 is initially in state ON. The state transitions of P1 caused by its failure are modelled by the stochastic activity $P1_fail$ whose completion rate is equal to the failure rate of P1 (Tab. 1). The completion of such activity is ruled by the input gate I_{P1_fail} : $P1_fail$ may complete only if the place $P1_stuck$ is empty (P1 is not stuck), while there is no condition about the place $P1_on$ (the failure may occur during both the ON state and the OFF state). The same gate partially specifies the effect of the completion of $P1_fail$: the gate sets the marking of the place $P1_stuck$ to 1 (P1 becomes stuck), and sets the marking of $P1_on$ to 0. The effect of the activity $P1_fail$ is ruled also by two completion cases: in one case the marking of the place $P1_on$ is not changed, and in this way we model the state transition toward the state Stuck-OFF; in the other case, one token appears in $P1_on$, in order to represent the state transition toward the state Stuck-ON. These two completion cases have the same probability to occur: 0.5.

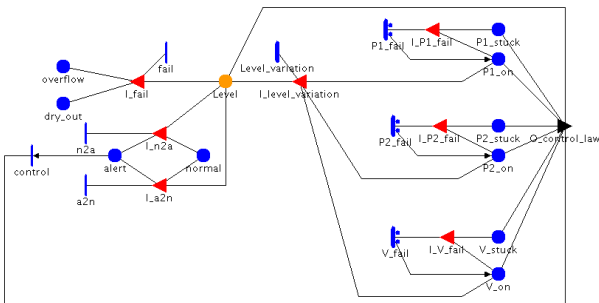


Figure 3: The SAN Model of Version 1

The current state of the pump P2 and the state transitions due to a failure of P2 are modelled in the same way by the places $P2_on$ and $P2_stuck$, the stochastic activity $P2_fail$ and the input gate I_{P2_fail} . Initially both $P2_on$ and $P2_stuck$ are empty in order to model that P2 is initially in state OFF. The state evolution of the valve V is modelled by the places V_on and V_stuck , the stochastic activity V_fail and the input

gate I_{V_fail} . The initial ON state of V is modelled by the presence of one token inside V_on and no tokens inside V_stuck .

The current level (H) of the liquid in the tank expressed in meters, is represented by the extended place $Level$ whose marking is a float variable initially set to 0 corresponding to the initial level of the liquid (Sec. 3). In the SAN model, we model any variation of H by 0.01 m; this is done by increasing or decreasing the marking of $Level$ by 0.01. The action of P1, P2 and V on H are modelled by the deterministic activity $Level_variation$ and in particular by the corresponding input gate $I_{Level_variation}$. Such gate enables $Level_variation$ to complete only when the state configuration 1, 2, 4, 6, or 7 holds (Tab. 2). The action of P1, P2 or V on H is ruled by a level variation rate equal to 0.6 m/h (Sec. 3); this means that the action of a pump (valve) increases (decreases) the liquid level by 0.01 m every 0.016667 h. Since we are interested in representing any variation of H by 0.01 m, $Level_variation$ completes every 0.016667 h in state configurations 1, 4, 6, 7, or every 0.016667/2 h in state configuration 2 (Tab. 2). The gate $I_{Level_variation}$ specifies also the effect of the completion of $Level_variation$: each time such activity completes, the marking of the place $Level$ is increased by 0.01 in the state configurations 1, 2, 4, 6, or is decreased by 0.01 in the state configuration 7 (Tab. 2).

The place $normal$ is initially marked with one token in order to represent that H is inside the region of correct functioning (Sec. 3). The completion of the immediate activity $n2a$ removes the token inside the place $normal$; according to the input gate I_{n2a} , this happens if the marking of the extended place $Level$ is less than HLA or more than HLB (Tab. 3). The same gate sets the marking of the place $alert$ to 1. In this way, we model that the liquid level in the tank is outside the correct region. The presence of one token inside $alert$ enables the immediate activity $control$ to complete. The effect of its completion is ruled by the output gate $O_{control_law}$ executing the control laws in Tab. 3: such gate acts on the marking of the places $P1_on$, $P2_on$ and V_on , and consequently on the state of P1, P2 and V. So, $control$ together with $O_{control_law}$, models the orders given by the controller. If the place $P1_stuck$, $P2_stuck$ or V_stuck is marked, then the output gate $O_{control_law}$ has no effect on the place $P1_on$, $P2_on$ or V_on respectively. In this way we model that the controller cannot act on the state of a stuck component.

The controller action on the component states may lead H back to the region of correct functioning. In this case, the immediate activity $a2n$ is enabled to complete by the input gate I_{a2n} checking that the marking of the extended place $Level$ is equal or greater than HLA and less or equal to HLB. The effect of the completion of $a2n$ is the presence of one token inside the place $normal$ in order to represent that H is inside the region of correct functioning.

The dry out and the overflow condition determining the system failure, are detected by the immediate activity *fail* and in particular by the corresponding input gate *I_fail*: if the marking of the extended place *Level* is less than HLV, then one token appears in the place *dry_out* in order to model the occurrence of the dry out. If instead the marking of *Level* is greater than HLP, then the effect of the completion of *fail* is the presence of one token inside the place *overflow* modelling the occurrence of the overflow.

Further details of the SAN model in Fig. 3 are reported in Codetta (2011).

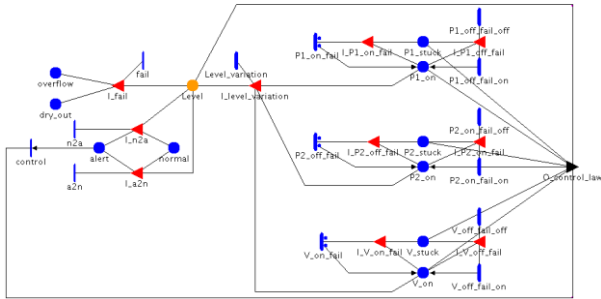


Figure 4: The SAN Model of Version 2

5.2. Modelling Version 2

In Version 2 (Sec 3.1), the failure rates of P1, P2 and V are state dependent (Tab. 4). The SAN model of Version 2 appears in Fig. 4 where the current state of P1 is still modelled by the marking of the places *P1_on* and *P1_stuck*, but the state transitions caused by the failure are now modelled by three stochastic activities: *P1_on_fail* models the failure of P1 during the state ON; *P1_off_fail_off* represents the failure during the state OFF and leading to the state Stuck-OFF; *P1_off_fail_on* models the failure during the state OFF, but leading to the state Stuck-ON. The state evolution of P2 and V is modelled in a similar way. The other parts of the SAN model in Fig. 5 are the same as in the model of Version 1 (Fig. 3).

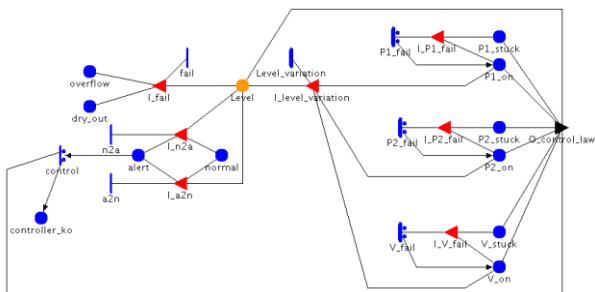


Figure 5: The SAN Model of Version 3

5.3. Modelling Version 3

In Version 3, the controller failure on demand is introduced (Sec. 3.2); this aspect is represented in the SAN model in Fig. 5 by the presence of two completion cases for the immediate activity *control* modelling the action of the controller. In one case, the effect of the completion of *control* is ruled by the output gate *O_control* executing the control laws (Tab. 3). In the

other case, the failure on demand occurs and the only effect is the addition of one token to the marking of the new place *controller_ko* counting the number of failures of the controller.

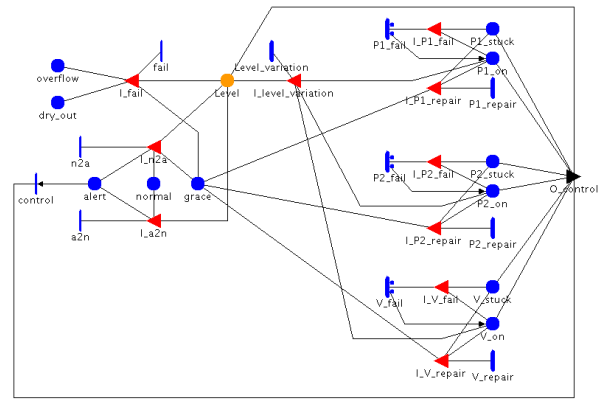


Figure 6: The SAN Model of Version 4

5.4. Modelling Version 4

Version 4 is modelled in Fig. 6 where the immediate activity *n2a* still completes when H exceeds the region of correct functioning, but now *n2a* inserts also one token inside the new place *grace* in order to represent that the grace period (Sec. 3.3) has begun. Such marking enables the new stochastic activities *P1_repair*, *P2_repair* and *V_repair* ruled by the input gates *I_P1_repair*, *I_P2_repair* and *I_V_repair*, and modelling the repair of P1, P2 and V respectively. The effect of the completion of such activities is the removal of the token inside the place representing the stuck condition of the component (*P1_stuck*, *P2_stuck* and *V_stuck* respectively). The immediate activity *fail* still models the system failure, but now it also removes the token inside the place *grace*, in order to represent the end of the grace period.

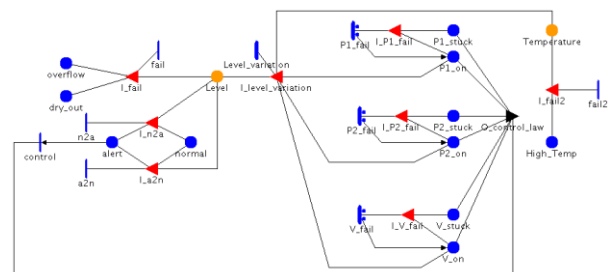


Figure 7: The SAN Model of Version 5.1

5.5. Modelling Version 5

The SAN models representing Versions 5.1, 5.2, 5.3 are depicted in Figures 7, 8, 9, respectively. Such models are characterized by the presence of a new extended place called *Temperature* representing the current liquid temperature (T). The deterministic activity *Level_variation*, together with the input gate *I_level_variation*, models the variation of H and T, as a consequence of the heat source (Sec. 3.4) and the injection of new liquid in the tank by the pumps (Eq. 2). The rates of the stochastic activities *P1_fail*, *P2_fail*

and V_{fail} modelling the failure of P1, P2 and V respectively, are expressed as a function of the marking of $Temperature$ according to Eq. 3.

In Versions 5.1, 5.2, 5.3, the system failure condition due to the high temperature of the liquid is introduced (Sec. 3.4). In the SAN models in Fig. 7, 8, 9, such condition is detected by the new immediate activity $fail2$ ruled by the input gate I_{fail2} : when the marking of $Temperature$ reaches the value of 100, one token appears inside the new place $High_Temp$.

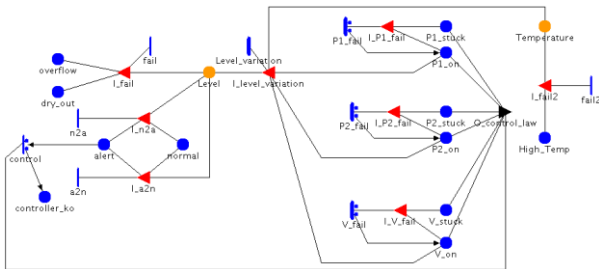


Figure 8: The SAN Model of Version 5.2

The Version 5.2 is characterized by the possible failure on demand of the controller (Sec. 3.4). In the SAN model in Fig. 8, such aspect is represented in the same way as in the SAN model of Version 3 (Fig. 5).

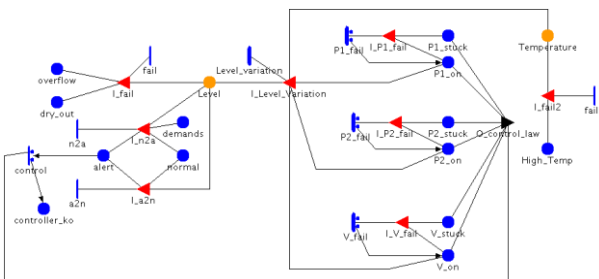


Figure 9: The SAN Model of Version 5.3

In Version 5.3, the probability of failure on demand of the controller is increased by 50% after each demand (Sec. 3.4). In Fig. 9, the marking of the new place $demands$ indicates the number of demands: every time that the immediate activity $n2a$ completes (H reaches the control boundaries), the marking of $demands$ is increased by one. The immediate activity $control$ still has two completion cases, but now their probabilities are a function of the marking of $demands$.

The full details of all the SAN models can be found in Codetta (2011).

6. SIMULATION RESULTS

The SAN models presented in the previous section have been simulated. In particular, for each model, 100'000 simulation batches have been performed by means of the *Möbius* tool, requiring a confidence level equal to 0.95, and a relative confidence interval equal to 0.1. The measures computed by the simulation are the *cumulative distribution function* (cdf) of the probability of each system failure condition (Sec. 3) for a mission time varying between 0 and 1000 h (or between 0 and

500 h in Version 4, as in Marseguerra and Zio (2006)). The cdf provides the system unreliability (Sec. 1) according to a specific failure condition. For instance, the value of the dry out cdf at time $t > 0$ is the probability that the system has failed because of the dry out, during the time period $(0, t)$.

The cdf of the dry out probability is computed as the mean value over the 100'000 simulation batches, of the marking of the place dry_out present in all the SAN models (Sec. 5). In each simulation batch and at a certain time, the number of tokens inside the place dry_out is equal to 0 if the dry out has not occurred, or it is equal to 1 if the dry out condition holds (Sec. 5). So, the mean value of its marking at a certain time, over the 100'000 simulation batches, provides the probability that the dry out condition holds at that time.

The cdf of the overflow probability is computed in the same way, but with reference to the place $overflow$ present in all the SAN models (Sec. 5). In Versions 5.1, 5.2 and 5.3, another system failure condition is taken into account: the temperature of the liquid reaching 100°C (Sec. 3.4). The cdf of such condition is computed as the mean number of tokens inside the place $High_Temp$ present in the SAN models in Fig. 7, 8 and 9 (Sec. 5.5).

6.1. Results for Versions 1, 2, 3

The values of the cdf of the dry out in Versions 1, 2 and 3 (SAN model in Fig. 4, 5 and 6 respectively) are reported in Tab. 6 and are graphically compared in Fig. 10. The values of the cdf of the overflow are reported in Tab. 7 and are graphically compared in Fig. 11. The results returned by the SAN model simulation are similar to the values returned by Monte Carlo simulation, GSPN analysis and FSPN simulation.

Table 6: The cdf of the Dry Out in Versions 1, 2, 3

time	Version 1	Version 2	Version 3
100 h	4.5900E-03	2.0240E-02	4.9090E-02
200 h	2.2390E-02	4.0400E-02	8.6710E-02
300 h	4.4890E-02	5.4090E-02	1.0952E-01
400 h	6.5990E-02	6.3360E-02	1.2664E-01
500 h	8.2600E-02	6.9870E-02	1.3844E-01
600 h	9.5290E-02	7.3750E-02	1.4707E-01
700 h	1.0393E-01	7.6650E-02	1.5313E-01
800 h	1.1003E-01	7.8340E-02	1.5739E-01
900 h	1.1435E-01	7.9440E-02	1.6024E-01
1000 h	1.1747E-01	8.0240E-02	1.6220E-01

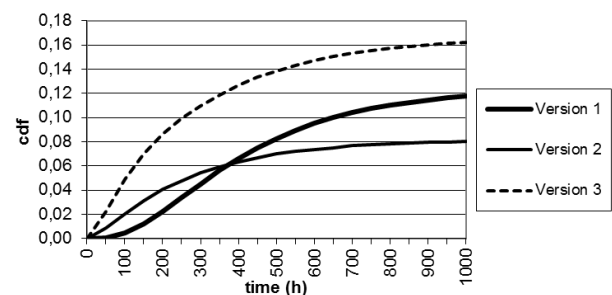


Figure 10: The cdf of the Dry Out in Versions 1, 2, 3

Table 7: The cdf of the Overflow in Versions 1, 2, 3

time	Version 1	Version 2	Version 3
100 h	7.8880E-02	7.9370E-02	1.3517E-01
200 h	1.9914E-01	1.6852E-01	2.7244E-01
300 h	2.9386E-01	2.3411E-01	3.6541E-01
400 h	3.6207E-01	2.7882E-01	4.2492E-01
500 h	4.0667E-01	3.0943E-01	4.6332E-01
600 h	4.3665E-01	3.2938E-01	4.8808E-01
700 h	4.5683E-01	3.4310E-01	5.0444E-01
800 h	4.7063E-01	3.5284E-01	5.1537E-01
900 h	4.7929E-01	3.6009E-01	5.2298E-01
1000 h	4.8572E-01	3.6500E-01	5.2797E-01

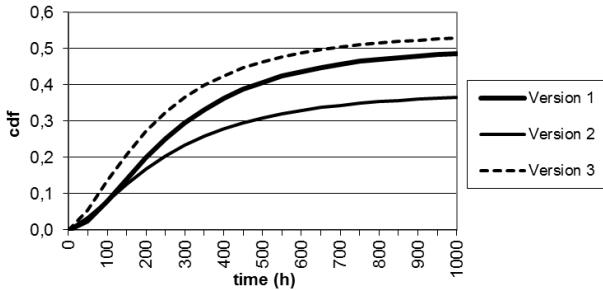


Figure 11: The cdf of the Overflow in Versions 1, 2, 3

6.2. Results for Version 4

The results obtained by simulating the model in Fig. 6, are reported in Tab. 8, in Fig. 12 (dry out) and in Fig. 13 (overflow). They differ from the results returned by Monte Carlo simulation in Marseguerra and Zio (1996), even though they are in the same order of magnitude. Moreover, they differ from the results obtained by means of GSPN analysis and FSPN simulation in Codetta and Bobbio (2005a) where the repair is erroneously assumed to be allowed only while the level is outside the region of correct functioning, instead of during the grace period (Sec. 3.3).

Table 8: The cdf of the Dry Out and the Overflow in Version 4

time	dry out	overflow
50 h	0.000E+00	8.000E-04
100 h	6.000E-05	2.430E-03
150 h	1.500E-04	4.230E-03
200 h	2.200E-04	6.090E-03
250 h	3.300E-04	7.920E-03
300 h	3.700E-04	9.460E-03
350 h	4.500E-04	1.069E-02
400 h	4.500E-04	1.197E-02
450 h	4.700E-04	1.298E-02
500 h	5.100E-04	1.363E-02

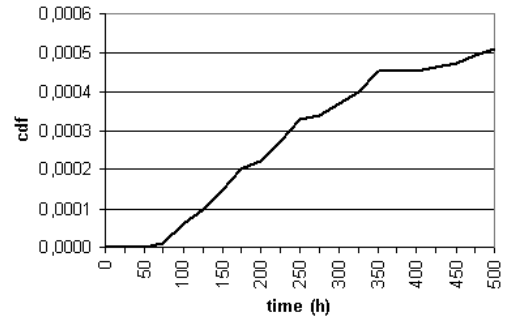


Figure 12: The cdf of the Dry Out in Version 4

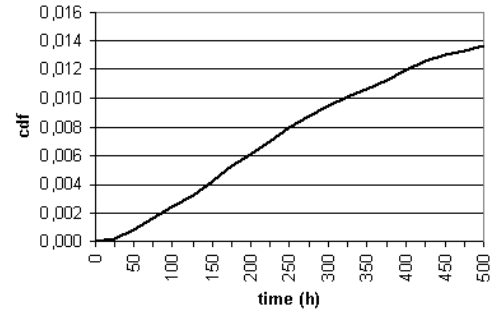


Figure 13: The cdf of the Overflow in Version 4

6.3. Results for Version 5

The results for Versions 5.1, 5.2 and 5.3 are reported in Tables 9, 10 and 11 respectively. In particular, the results in Version 5.1 (Fig. 14) and Version 5.2 (Fig. 15) are similar to the values returned by Monte Carlo simulation in Marseguerra and Zio (1996) and FSPN simulation in Codetta and Bobbio (2005b). Version 5.3 was not modelled as a FSPN in the past. According to the results for such version (Fig. 16), the wear of the controller (Sec. 3.4) does not seem to have a relevant impact on the cdf values, with respect to Version 5.2. In Marseguerra and Zio (1996) instead, the controller wear determines a slight increase of the dry out cdf values.

Table 9: The cdf of the Dry Out, the Overflow and the High Temperature in Version 5.1

time	dry out	overflow	high temp.
100 h	3.1650E-02	2.4007E-01	0.0000E+00
200 h	7.9330E-02	3.9531E-01	0.0000E+00
300 h	1.0517E-01	4.5631E-01	0.0000E+00
400 h	1.1706E-01	4.8133E-01	0.0000E+00
500 h	1.2200E-01	4.9161E-01	1.3000E-04
600 h	1.2376E-01	4.9588E-01	3.8200E-02
700 h	1.2424E-01	4.9750E-01	7.3850E-02
800 h	1.2436E-01	4.9826E-01	1.1855E-01
900 h	1.2438E-01	4.9864E-01	1.2614E-01
1000 h	1.2438E-01	4.9884E-01	1.2724E-01

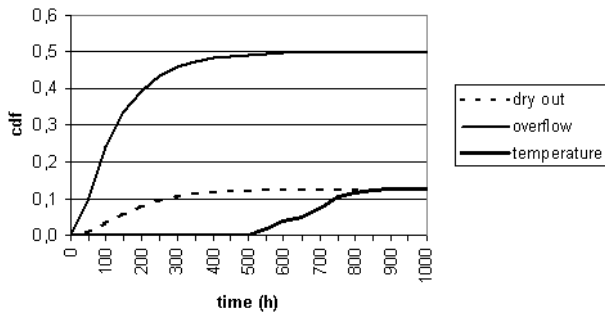


Figure 14: The cdf of the Dry Out, the Overflow and the High Temperature in Version 5.1

Table 10: The cdf of the Dry Out, the Overflow and the High Temperature in Version 5.2

time	dry out	overflow	high temp.
100 h	1.1102E-01	3.3740E-01	0.0000E+00
200 h	1.4934E-01	4.7526E-01	0.0000E+00
300 h	1.6601E-01	5.2228E-01	0.0000E+00
400 h	1.7271E-01	5.4072E-01	2.0000E-05
500 h	1.7559E-01	5.4828E-01	2.2000E-04
600 h	1.7650E-01	5.5178E-01	2.0370E-02
700 h	1.7677E-01	5.5325E-01	4.0250E-02
800 h	1.7685E-01	5.5387E-01	6.7460E-02
900 h	1.7687E-01	5.5416E-01	7.1930E-02
1000 h	1.7687E-01	5.5428E-01	7.2550E-02

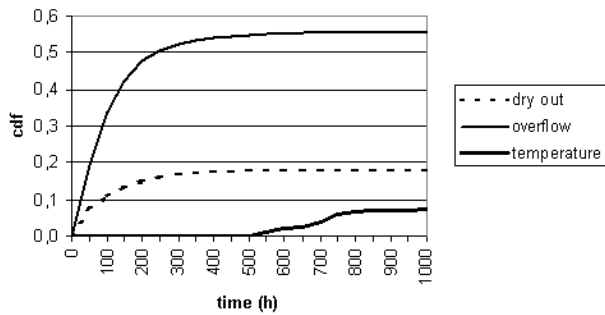


Figure 15: The cdf of the Dry Out, the Overflow and the High Temperature in Version 5.2

Table 11: The cdf of the Dry Out, the Overflow and the High Temperature in Version 5.3

time	dry out	overflow	high temp.
100 h	1.1427E-01	3.4165E-01	0.0000E+00
200 h	1.5201E-01	4.7762E-01	0.0000E+00
300 h	1.6799E-01	5.2503E-01	0.0000E00
400 h	1.7470E-01	5.4360E-01	1.0000E-05
500 h	1.7748E-01	5.5109E-01	2.0000E-04
600 h	1.7849E-01	5.5462E-01	1.9230E-02
700 h	1.7878E-01	5.5604E-01	3.7920E-02
800 h	1.7886E-01	5.5663E-01	6.5320E-02
900 h	1.7889E-01	5.5694E-01	6.9780E-02
1000 h	1.7889E-01	5.5708E-01	7.0460E-02

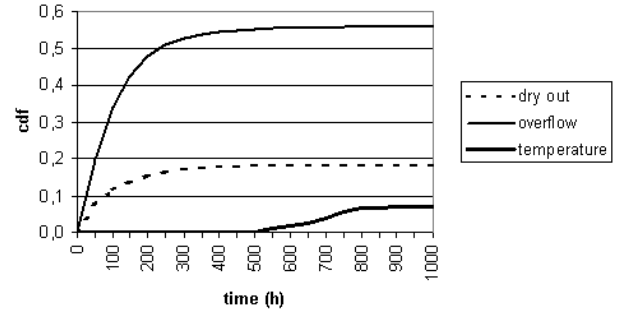


Figure 16: The cdf of the Dry Out, the Overflow and the High Temperature in Version 5.3

7. CONCLUSIONS

A benchmark on dynamic reliability taken from the literature has been examined. Each version focuses on a particular aspect of the dynamic behaviour of the system, such as state or temperature dependent failure rates, repairable components, failures on demand. The benchmark was originally evaluated in terms of system unreliability by means of Monte Carlo simulation. In this paper, the benchmark versions have been modelled and simulated using SAN, a particular form of Petri Net. The results in this paper are in general coherent to the original ones and those obtained by means of other Petri Net based formalisms such as GSPN and FSPN. This confirms that Petri Net models are a valid approach to deal with dynamic reliability cases because of the possibility to model the stochastic, timed or immediate events characterizing the complete behaviour of the system.

In particular, the use of SAN has several advantages. Gates make the SAN model more compact: many predicates and functions (Sec. 4) that are incorporated into the input or output gates, would have required more transitions (activities) and arcs in order to be represented in a GSPN or FSPN. For instance, in the GSPN models, the orders by the controller are represented by 6 transitions, while in the SAN model, only the activity (transition) *control*, together with its output gate, is necessary. In the GSPN model, the variations to the liquid level are represented by 5 transitions, while the activity *Level_variation* is enough in the SAN. The failure of a pump or valve is represented by four transitions in the GSPN; in the SAN instead, one activity is necessary (Sec. 5).

The SAN formalism can represent float variables, as in FSPN, by means of extended places. An example is the place *Level* modelling the liquid level. This avoids the discretization into integer values of float variables, required in GSPN. The negative values of variables can be directly mapped into the marking of SAN places. For instance, the liquid level in Versions 1, 2, 3 varies between $-3 m$ and $+3 m$ (Sec. 3), just like the marking of the place *Level* (Sec. 5). In the GSPN and FSPN model instead, the liquid thresholds had to be redefined in order to avoid negative values.

AUTHOR BIOGRAPHY

Daniele Codetta-Raiteri received the Ph.D. in Computer Science from the University of Turin, Italy, in 2006. Now he is a researcher at the University of Eastern Piedmont, Italy. His research focuses on stochastic models for Reliability evaluation, with a particular experience in Fault Trees, Petri Nets and Bayesian Networks. He is the (co-)author of more than thirty papers published in proceedings or journals.

REFERENCES

- Ajmone-Marsan, M., Balbo, G., Conte, G., Donatelli, S., Franceschinis, G., 1995. *Modelling with Generalized Stochastic Petri Nets*. Wiley Series in Parallel Computing.
- Chiola, G., Franceschinis, G., Gaeta, R., Ribaud, M., 1995. GreatSPN 1.7: Graphical Editor and Analyzer for Timed and Stochastic Petri Nets. *Performance Evaluation, special issue on Performance Modeling Tools*, 24(1&2):47–68.
- Codetta-Raiteri, D., 2011, *SAN models of a benchmark on dynamic reliability*, Università del Piemonte Orientale. Available from: <http://people.unipmn.it/dcr>
- Codetta-Raiteri, D., Bobbio, A., 2005a. Solving Dynamic Reliability Problems by means of Ordinary and Fluid Stochastic Petri Nets. *Proceedings of the European Safety and Reliability Conference (ESREL)*, pp. 381–389. June, Gdansk (Poland).
- Codetta-Raiteri, D., Bobbio, A., 2005b. Evaluation of a benchmark on dynamic reliability via Fluid Stochastic Petri Nets. *Proceedings of the International Workshop on Performability Modeling of Computer and Communication Systems (PMCCS)*, pp. 52–55. September, Turin (Italy).
- Deavours, D., Clark, G., Courtney, T., Daly, D., Derisavi, S., Doyle, J., Sanders, W., Webster, P.G., 2002. The Möbius Framework and its Implementation. *IEEE Transactions on Software Engineering*, 28(10):956–969.
- Distefano, S., Xing, L., 2006. A New Approach to Model the System Reliability: Dynamic Reliability Block Diagrams. *Proceedings of the Annual Reliability and Maintainability Symposium (RAMS)*, pp. 189–195. January, Newport Beach (California, USA).
- Dugan, J.B., Bavuso, S.J., Boyd, M.A., 1992. Dynamic Fault-Tree Models for Fault-Tolerant Computer Systems. *IEEE Transactions on Reliability*, 41:363–377.
- Gribaudo, M., 2001. FSPNedit: A fluid stochastic Petri net modeling and analysis tool. *Proceedings of Tools of International Multiconference on Measurements Modelling and Evaluation of computer Communication Systems*, pp. 24–28. September, Aachen (Germany).
- Gribaudo, M., Sereno, M., Horvath, A., Bobbio, A., 2001. Fluid Stochastic Petri Nets augmented with flush-out arcs: Modelling and analysis. *Discrete Event Dynamic Systems*, 11(1&2):97–117.
- Marseguerra, M., Zio, E., 1996. Monte Carlo Approach to PSA for dynamic process system. *Reliability Engineering and System Safety*, 52:227–241.
- Marseguerra, M., Zio, E., Devooight, J., Labeau, P.E., 1998. A concept paper on dynamic reliability via Monte Carlo simulation. *Mathematics and Computers in Simulation*, 47:371–382.
- Sahner, R.A., Trivedi, K.S., Puliafito, A. 1996. *Performance and Reliability Analysis of Computer Systems; An Example based Approach Using the SHARPE Software Package*. Kluwer Academic Publisher.
- Sanders, W.H., Meyer, J.F., 2001. Stochastic activity networks: Formal definitions and concepts. *Lecture Notes in Computer Science*, 2090:315–343.

A STOCHASTIC APPROACH TO RISK MODELING FOR SOLVENCY II

Vojo Bubevski

Bubevski Systems & Consulting™
TATA Consultancy Services™

vojo.bubevski@landg.com

ABSTRACT

Solvency II establishes EU-wide capital requirements and risk management standards for (re)insurers. The capital requirements are defined by the Solvency Capital Requirement (SCR), which should deliver a level of capital that enables the (re)insurer to absorb significant unforeseen losses over a specified time horizon. It should cover insurance, market, credit and operational risks, corresponding to the Value-at-Risk (VAR) subject to a confidence level of 99.95% over one year. Standard models are deterministic, scenario-based or covariance-based, i.e. non-stochastic. They don't optimise the investment portfolios. These are two major deficiencies. A stochastic approach is proposed, which combines Monte Carlo Simulation and Optimisation. This method determines minimal variance portfolios and calculates VAR/SCR using the optimal portfolios' simulation distributions, which ultimately eliminates the standard models' deficiencies. It offers (re)insurers internal model options, which can help them to reduce their VAR/SCR providing higher underwriting capabilities and increasing their competitive position, which is their ultimate objective.

Keywords: Solvency II stochastic model, VAR/SCR reduction, portfolio optimisation – minimal variance, Monte Carlo simulation

1. INTRODUCTION

The Solvency II regulations are fundamentally redesigning the capital adequacy regime for European (re)insurers and will be effective from 1st January 2013.

Solvency II establishes two levels of capital requirements: i) Minimal Capital Requirement (MCR), i.e. the threshold below which the authorization of the (re)insurer shall be withdrawn; and ii) Solvency Capital Requirement (SCR), i.e. the threshold below which the (re)insurer will be subject to a much higher supervision. The SCR should deliver a level of capital that enables the (re)insurer to absorb significant unforeseen losses over a specified time horizon. It should cover, at a minimum, insurance, market, credit and operational risks, corresponding to the VAR of the (re)insurer's own basic funds, subject to a confidence level of 99.95% over a one-year period.

Solvency II offers two options for calculating VAR/SCR, i.e. by applying either: i) a standard model, which will be provided by the regulator; or ii) an internal model developed by the (re)insurer's risk department.

The standard models are non-stochastic risk models. They are rather deterministic, scenario-based or covariance models. They are also conservative by nature and generic across the EU so they cannot consider the company's specific factors. Moreover, they do not use optimisation to determine the minimal variance investment portfolios in order to minimise the financial risk for the (re)insurers. Thus the calculated VAR/SCR will be higher. These are apparent most important limitations of the standard models.

For example, the deterministic model applies analytically calculated or estimated input parameters to calculate the results. However, the likelihood of the outcome is not considered at all. Also, the scenario-based models consider the worst, most likely and best case scenarios. However, they fail to answer the two basic questions: i) how likely are the worst and best case scenarios? And more importantly, ii) how likely is the most likely case?

Solvency II offers capital-reduction incentives to insurers that invest in developing advanced internal models, which apply a stochastic approach for risk management and control. Thus, insurers will benefit from using internal models. A very good explanation of developing the Enterprise Risk Management (ERM) frameworks in (re)insurance companies for Solvency II is presented in a handbook edited by Cruz (2009).

There are a number of published examples of recommended internal model, which could be used for Solvency II (Cruz 2009). These suggested internal model examples don't consider optimisation to determine the minimal variance investment portfolios in order to minimise the financial risk, which is a significant deficiency.

The stochastic models usually apply the Monte Carlo Simulation method, which assigns distributions of random variables to the input parameters and the calculated results are presented in the form of a histogram. This allows statistical and probabilistic tools to be used to analyse the results. A comprehensive

elaboration of Monte Carlo Simulation in Finance is given by Glasserman (2004).

An investment portfolio is defined by the fraction of the capital put in each investment. The problem of determining the minimum variance portfolio that yields a desired expected return was solved by Markowitz in the 1950's. He received the 1991 Nobel Prize for his work in Economics (Markowitz 1987). Mostly, the Optimisation methodology is used to find the minimum variance portfolio in order to minimise the financial risk.

VAR is a widely used financial risk measure. The approach to calculate VAR is well summarised by Jorion (2011). This approach includes the VAR Parametric method and VAR Monte Carlo Simulation method.

This paper proposes a stochastic approach to risk modelling for Solvency II. This method applies combined Monte Carlo Simulation and Optimisation methodologies. The method uses Optimisation to calculate the minimal variance portfolios that yield desired expected returns to determine the Efficient Frontier of optimal portfolios. Monte Carlo Simulation is used to calculate VAR/SCR for every portfolio on the Efficient Frontier by using the respective portfolios' simulation distributions. Therefore, by using the synergy of Monte Carlo Simulation and Optimisation, the method eliminates the deficiencies and limitations, which are identified above.

This approach can help (re)insurers to develop and improve their internal risk models in order to reduce their VAR/SCR. Consequently, this will provide insurers with higher underwriting capabilities and increase their competitive position, which is their ultimate objective.

According to research by Mercer Oliver Wyman, the impact of the four quantifiable risks on the economic capital of insurance companies is: i) 64% Investment Asset Liability Management (ALM) Risk, i.e. Market Risk; ii) 27% Operational Risk; iii) 5% Credit Risk; and iv) 4% Insurance Risk. Considering that the Market (ALM) Risk is the top contributing risk factor, the method is demonstrated by using an example of Market (ALM) Risk Management. Also, in order to facilitate the presentation, a simple Market (ALM) Risk model is demonstrated.

Only the practical aspects of the Market (ALM) Risk modelling are discussed. Microsoft™ Excel® and Palisade™ @RISK® and RISKOptimizer® were used in these experiments.

1.1. Related Work

The following is a summary of some published works related to Market (ALM) Risk modelling for Solvency II.

1.1.1. Market Risk in the GDV Model

The GDV (Gesamtverband der Deutschen Versicherungswirtschaft) Model is the standard model of the German Insurance Association for Solvency II (GDV 2005). This model is to some extent a Static

Factor deterministic model, where the risk capital calculation is based on linear combinations of static risk factors. Actually, the model is mostly a Covariance (or VAR) Model, which is a very simplified version of Stochastic Risk Models.

1.1.2. Market Risk in the Swiss Solvency Test (SST) Model

This is the standard model of Swiss Federal Office of Private Insurance. The Market Risk in the SST model is handled by the ALM model. The SST ALM model is a Risk Factor Covariance model complemented with Scenario-Based models (SST 2004).

1.1.3. Bourdeau's Example of Internal Market Risk Model

Michele Bourdeau published an example of an internal model for Market Risk. This model calculates VAR using Monte Carlo Simulation. This is an example of a true Stochastic Risk Model (Bourdeau 2009).

2. ALM RISK MODELLING PROCEDURE

The following sections demonstrate the ALM Risk modelling procedure for Solvency II step-by-step. Actual financial market data are used in the presentation.

2.1. Problem Statement

The following is a simplified problem statement for the demonstrated investment ALM risk model under Solvency II.

Determine the minimum variance investment portfolio that yields a desired expected annual return to cover the liabilities of the insurance company. Calculate the VAR considering all the company's specific factors including their risk appetite. The model should allow the insurer to reduce their VAR (SCR) providing for higher underwriting capabilities and increasing their competitive position. The model should help the company to achieve their ultimate objective.

2.2. Calculating Compounded Monthly Return

The monthly returns of four investment funds are available for a period of seven years, i.e. 1990-1996 (Table 1). Note that the data for the period July/1990-June/1996 are not shown.

Table 1: Monthly Return

Month	Fund 1	Fund 2	Fund 3	Fund 4
Jan/1990	0.048	-0.01	-0.06	-0.01
Feb/1990	0.066	0.096	0.037	0.038
Mar/1990	0.022	0.022	0.12	0.015
Apr/1990	0.027	-0.04	-0.02	-0.04
May/1990	0.112	0.116	0.123	0.075
Jun/1990	-0.02	-0.02	-0.04	-0.01
Jul/1996	0.086	-0.07	-0.12	-0.02
Aug/1996	0.067	0.026	0.146	0.018
Sep/1996	0.089	-0.03	-0.04	0.092
Oct/1996	0.036	0.117	0.049	0.039

The Compounded Monthly Return (CMR) is calculated for each month and each investment fund from the given Monthly Return (MR) fund using the following formula (Table 2):

$$CMR = \ln(1 + MR)$$

Table 2: Compounded Monthly Return (CMR)

Month	CMR1	CMR2	CMR3	CMR4
Jan/1990	0.047	-0.01	-0.06	-0.01
Feb/1990	0.063	0.092	0.036	0.038
Mar/1990	0.021	0.022	0.113	0.015
Apr/1990	0.027	-0.04	-0.02	-0.04
May/1990	0.106	0.11	0.116	0.073
Jun/1990	-0.02	-0.02	-0.04	-0.01
Jul/1996	0.082	-0.07	-0.13	-0.02
Aug/1996	0.065	0.026	0.136	0.018
Sep/1996	0.085	-0.03	-0.04	0.088
Oct/1996	0.036	0.111	0.048	0.038

2.3. Fitting Distributions to Compounded Monthly Return

For the Monte Carlo method, we need the distribution of the compounded monthly return for each investment fund. Thus, for each investment fund, we determine the best fit distribution based on the Chi-Square measure. For example, the best fit distribution for the compounded monthly return of Fund 4 (i.e. CMR4) is the normal distribution presented in Figure 1.

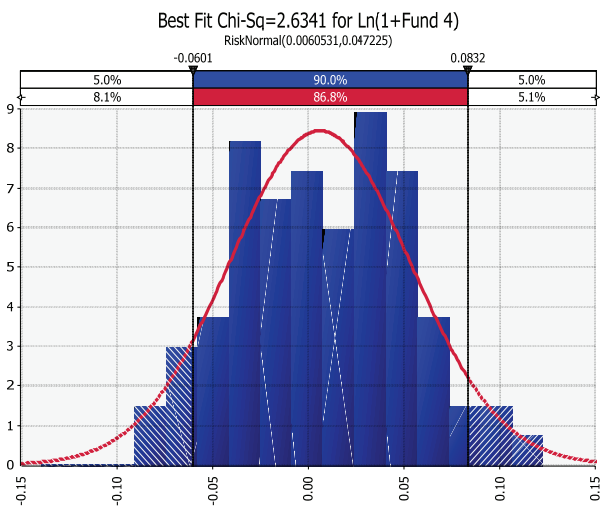


Figure 1: Fund 4 Best Fit Distribution

2.4. Finding Compounded Monthly Return Correlations

The compounded monthly returns of the investment funds are correlated. We need to find the correlation to allow the Monte Carlo method to generate correlated random values for the compounded monthly returns. The correlation matrix is presented in Table 3.

Table 3: Correlation Matrix

	CMR1	CMR2	CMR3	CMR4
CRM 1	1	0.263	0.038	0.0868
CRM2	0.263	1	0.244	0.0895
CRM3	0.038	0.244	1	0.095
CRM4	0.087	0.089	0.095	1

2.5. Generating Compounded Monthly Return

The Compounded Monthly Return (CMR) is randomly generated for each investment fund from the best fit distribution considering the correlations. The following distribution functions of the Palisade™ @RISK® are used:

$$CMR1 = RiskLogistic(0.0091429, 0.044596)$$

$$CMR2 = RiskLognorm(1.1261, 0.077433, Shift(-1.1203))$$

$$CMR3 = RiskWeibull(6.9531, 0.46395, Shift(-0.42581))$$

$$CMR4 = RiskNormal(0.0060531, 0.047225)$$

The correlation is applied by using the “RiskCorrmat” function of the Palisade™ @RISK®.

2.6. Calculating Compounded Annual Return by Fund

The Compounded Annual Return (CAR) is calculated for each investment fund from the respective Compounded Monthly Return (CMR), using the following formula:

$$CAR = 12 * CMR$$

2.7. Calculating Expected Annual Mean Return on the Portfolio

The expected annual mean return on the portfolio (EAR-Mean) is calculated from the asset allocation weights vector (Weights-V) and the vector of compounded annual returns of funds (CAR-V) by using the following Excel® formula:

$$EAR-Mean = SumProduct(Weights-V, CAR-V)$$

2.8. Calculating Variance, Standard Deviation and VAR of the Portfolio

The variance, standard deviation and VAR of the portfolio are calculated from the distribution of the expected annual mean return on the portfolio (EAR-Mean) by using the following Palisade™ @RISK® functions:

$$Variance = RiskVariance(EAR-Mean)$$

$$Standard-Deviation = RiskStdDev(EAR-Mean)$$

$$VAR = RiskPercentile(EAR-Mean, 0.005)$$

2.8.1. Portfolio Simulation and Optimisation #1

Palisade™ RISKOptimizer® is used to solve the portfolio simulation and optimisation problem. That is to find the minimal variance portfolio of investments, which yields sufficient return to cover the liabilities. Thus, the aim of the simulation and optimisation model is to minimise the variance of the portfolio subject to the following specific constraints:

- The expected portfolio return is at least 8.2%, which is sufficient to cover the liabilities;
- All the money is invested, i.e. 100% of the available funds is invested; and
- No short selling is allowed so all the fractions of the capital placed in each investment fund should be non-negative.

The model should also calculate the Standard Deviation and VAR of the portfolio.

2.8.2. Finding the Efficient Frontier of Portfolios

Palisade™ RISKOptimizer® is used repetitively to solve the portfolio simulation and optimisation problem in order to find the Efficient Frontier of investment portfolios. That is to find the minimal variance portfolios of investments, which yield expected portfolio returns of at least 8.4%, 8.6%, ..., 10% and 10.2%. Thus, the aim of the simulation and optimisation models is to find in ten iterations, the ten minimal variance portfolios subject to the following specific constrains:

- The expected portfolio return is at least 8.4%, 8.6%, ..., 10% and 10.2% respectively;
- All the money is invested, i.e. 100% of the available funds is invested; and
- No short selling is allowed so all the fractions of the capital placed in each investment fund should be non-negative.

The model should also calculate the Standard Deviation and VAR of these ten portfolios.

3. RESULTS AND DISCUSSION

3.1. Simulation and Optimisation #1

The optimal portfolio found by this model has the following investment fractions: 14.6% in Fund 1; 11.6% in Fund 2; 18.6% in Fund 3; and 55.2% in Fund 4. The Portfolio Return is 8.2% with Variance of 19.9%, Standard Deviation of 44.6% and VAR of -7%.

The probability distribution of this optimal portfolio is given in Figure 2. From the graph, we can read the confidence levels as follows.

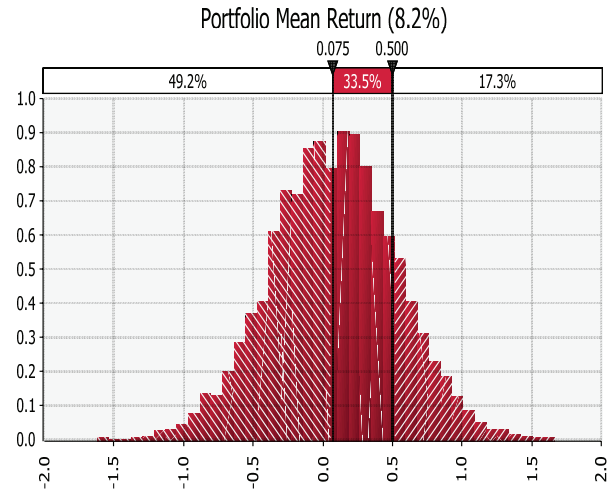


Figure 2: Probability Distribution #1

The probability that the portfolio return is below 7.5% is 49.2%. There is a 33.5% probability that the return is in the range of 7.5%-50%. From the simulation statistics we also find that there is a 43.2% probability that the portfolio return is negative, and 51.4% probability that the return is below 10%.

From the correlation graph (Figure 3), we can conclude that the portfolio return is most dependent on the return of Fund 4 with a correlation coefficient of 77%. The other three funds, i.e. Fund 3, Fund 2 and Fund 1, are less influential with correlation coefficients of 48%, 46% and 44% respectively.

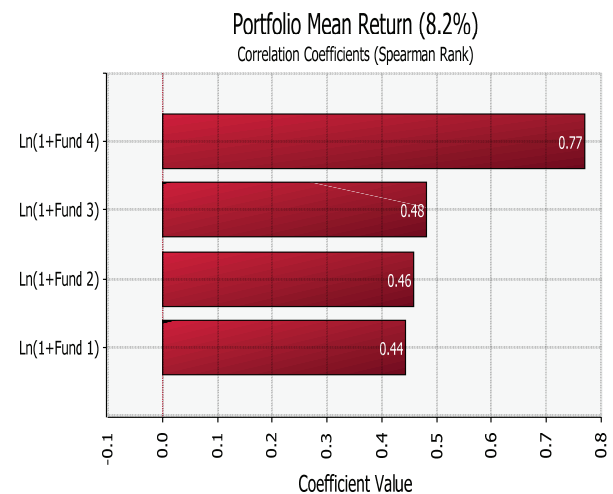


Figure 3: Correlation Sensitivity

The regression sensitivity graph is given in Figure 5. This graph shows how the portfolio mean return is changed in terms of Standard Deviation, if the return of a particular fund is changed by one Standard Deviation.

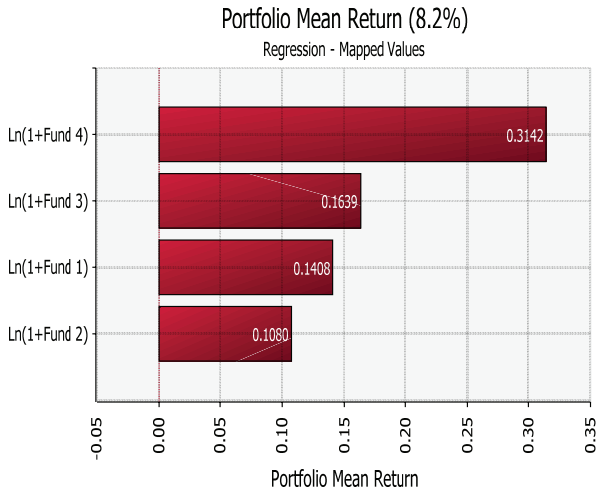


Figure 4: Regression Mapped Values Sensitivity

Therefore, we can read from the graph for example, that if Fund 4 return is changed by one Standard Deviation, the portfolio return will be changed by 0.3142 Standard Deviations (as shown by the regression coefficient of 0.3142). Again, the other three funds, i.e. Fund 3, Fund 2 and Fund 1, are less influential as their regression coefficients are 0.1639, 0.1408 and 0.1080 respectively.

3.2. Overall Simulation & Optimisation Results

The overall results of all the eleven simulation and optimisations are presented in Table 4 showing the Mean Return, Variance, Standard Deviation and VAR of the optimal portfolios.

Table 4: The overall results

Portfolio No.	Mean Return	Variance	Standard Deviation	VAR
1	0.082	0.199	0.446	-0.067
2	0.084	0.202	0.45	-0.088
3	0.086	0.204	0.452	-0.11
4	0.088	0.215	0.464	-0.147
5	0.09	0.222	0.471	-0.223
6	0.092	0.246	0.496	-0.257
7	0.094	0.266	0.516	-0.308
8	0.096	0.295	0.543	-0.403
9	0.098	0.33	0.574	-0.498
10	0.1	0.37	0.608	-0.608
11	0.102	0.42	0.648	-0.753

3.3. Efficient Frontier of Portfolios

Efficient Frontier of the optimal portfolios is presented in Figure 5.

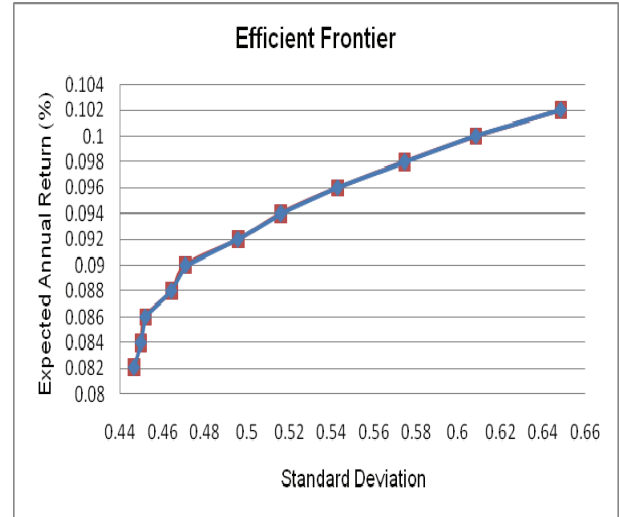


Figure 5: Efficient Frontier of Optimal Portfolios

The Efficient Frontier shows that an increase in expected return of the portfolio causes an increase in portfolio Standard Deviation. Also, the Efficient Frontier gets flatter as expected. This tells us that each additional unit of Standard Deviation allowed, increases the portfolio mean return by less and less.

3.4. Portfolio Expected Mean Return versus VAR

Figure 6 shows the dependency of the expected portfolio returns against VAR. From the graph we can see that an increase in expected return of the portfolio causes an increase in portfolio VAR in terms of money. (It should be noted that mathematically, VAR is a negative number, which actually decreases when the return increases.) Also, the curve on the graph gets flatter, again as expected. This tells us that each additional unit of VAR allowed, increases the portfolio mean return by less and less.

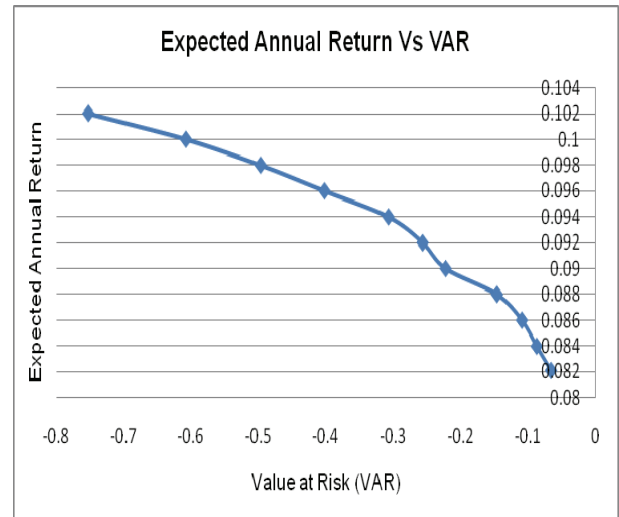


Figure 6: Portfolios Return versus VAR

3.5. Portfolio Standard Deviations versus VAR

Figure 7 shows the dependency of the portfolio Standard Deviation against VAR.

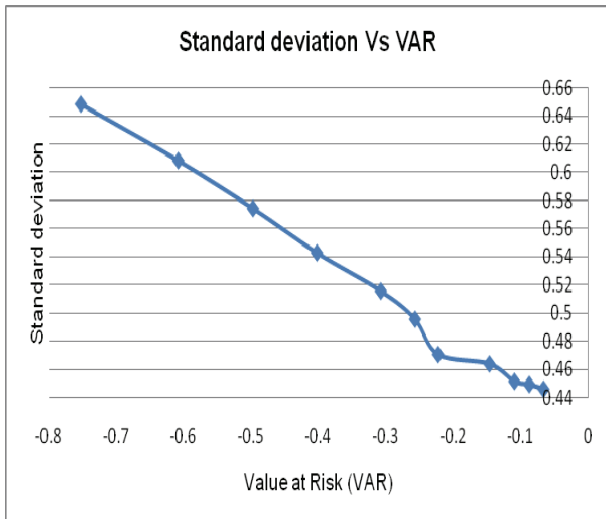


Figure 7: Standard Deviation versus VAR

From the graph we can see that the portfolio VAR is almost linearly proportional to the Standard Deviation. This is also as expected because a higher Standard Deviation translates to a higher risk, thus the VAR also increases in money terms (decreases mathematically).

3.6. Decision Support

The results presented above provide for comprehensive and reliable decision support for the decision makers, i.e. the financial risk executives of the insurance company. In particular, considering the Efficient Frontier of portfolios and the dependencies between portfolio expected return, Standard Deviation and VAR (shown in Figure 5, Figure 6 and Figure 7), the decision maker can decide in which assets to invest according to the desired expected return, risk appetite (i.e. standard deviation) and VAR. These results can help to reduce the SCR as required.

3.7. The Simulation & Optimisation Approach Comparison with the Related Work Examples

A comparison of the Simulation and Optimisation method proposed in this paper with the related work examples summarized in Sec. 1.1 is given below.

3.7.1. The Simulation & Optimisation Method versus the GDV Model

The GDV Model inherits its limitations from the Static Factor deterministic model. Also, this model is a very simplified Stochastic Model, which is an additional limitation. Moreover, the model doesn't use Optimisation to minimise the variance of the investment portfolios of the insurer, which is another major limitation.

In contrast, the proposed method does not have these two major limitations because they are resolved by using the Monte Carlo Simulation and Optimisation methodologies. Thus, the proposed approach is superior to the GDV Model.

3.7.2. The Simulation & Optimisation Method versus the Swiss Solvency Test (SST) Model

The SST ALM model is a Risk Factor Covariance model complemented with Scenario-Based models. Therefore, The SST ALM Model has the same deficiencies as the Scenario-Based models and the Covariance Models, which are not true Stochastic Models. In addition, the SST ALM Model does not apply optimisation to minimise the variance of the investment portfolios, which is another major deficiency.

The proposed method has eliminated these deficiencies by using the synergy of the Monte Carlo Simulation and Optimisation methodologies. Therefore, the proposed approach is also superior to the SST ALM Model.

3.7.3. The Simulation & Optimisation Model versus Bourdeau's Internal Market Risk Model Example

The Market Risk internal model proposed by Michele Bourdeau is a true Stochastic Risk Model. However, it does not use optimisation to minimise the variance of the investment portfolio, which is a main limitation. In this sense, the proposed method has a significant advantage versus this example because it minimises the variance of the investment portfolio, which ultimately minimises the risk and VAR.

4. CONCLUSION

This paper proposed a stochastic method for risk modelling under Solvency II. The method combines Monte Carlo Simulation and Optimisation methodologies in order to manage financial risk. The Optimisation methodology is used to calculate the minimal variance portfolios that yield desired expected returns in order to determine the Efficient Frontier of portfolios. The Monte Carlo methodology is used in order to calculate VAR/SCR for every portfolio on the Efficient Frontier by using the respective portfolios' simulation distributions. Consequently, the synergy of the Monte Carlo Simulation and Optimisation methodologies, which are used by the method, eliminates the identified significant limitations of the standard models. Also, the method has a significant advantage against the internal models, which do not use simulation and optimisation methodologies.

This stochastic approach can help the insurance and reinsurance companies to develop or improve their Solvency II internal risk models in order to reduce their VAR/SCR. Reducing the VAR and SCR will ultimately provide the insurance and reinsurance firms with higher underwriting capabilities, which will increase their competitive position on the market. Moreover, the proposed method can significantly assist the insurance and reinsurance companies to achieve their business objectives.

ACKNOWLEDGMENTS

I would like to thank my daughter, Ivana Bubevska, for reviewing the paper and suggesting relevant

improvements. She has also substantially helped with the editing and formatting of the paper. Her contribution has been essential to the successful publication of this work.

REFERENCES

- Bourdeau, M., 2009. Market Risk Measurement Under Solvency II. In: Cruz, M., ed. *The Solvency II Handbook*, London, Risk Books – Incisive Media, 193–226.
- Cruz, M., 2009. *The Solvency II Handbook*, London, Risk Books – Incisive Media.
- GDV, 2005. *Discussion Paper for a Solvency II Compatible Standard Approach*, Gesamtverband der Deutschen Versicherungswirtschaft. Available from:
http://www.gdv.de/Downloads/English/Documentation_Sol_II.pdf [Accessed 20 June 2011]
- Glasserman, P., 2004. *Monte Carlo Methods in Financial Engineering*, New York, Springer Science.
- Jorion, P., 2011. *Financial Risk Manager Handbook*, New Jersey, John Wiley & Sons.
- Markowitz, H.M., 1987. *Mean-Variance Analysis in Portfolio Choice and Capital Markets*, Oxford, UK, Basil Blackwell.
- SST, 2004. *White Paper of the Swiss Solvency Test*, Swiss Federal Office of Private Insurance. Available from:
http://www.naic.org/documents/committees_smi_int_solvency_switzerland_sst_wp.pdf [Accessed 20 June 2011]

AUTHORS BIOGRAPHY

Vojo Bubevski comes from Berovo, Macedonia. He graduated from the University of Zagreb, Croatia in 1977, with a degree in Electrical Engineering - Computer Science. He started his professional career in 1978 as an Analyst Programmer in Alkaloid Pharmaceuticals, Skopje, Macedonia. At Alkaloid, he worked on applying Operations Research methods to solve commercial and pharmaceutical technology problems from 1982 to 1986.

In 1987 Vojo immigrated to Australia. He worked for IBM™ Australia from 1988 to 1997. For the first five years he worked in IBM™ Australia Programming Center developing systems software. The rest of his IBM™ career was spent working in IBM™ Core Banking Solution Centre.

In 1997, he immigrated to the United Kingdom where his IT consulting career started. As an IT consultant, Vojo has worked for Lloyds TSB Bank in London, Svenska Handelsbanken in Stockholm, and Legal & General Insurance in London. In June 2008, he joined TATA Consultancy Services Ltd.

Vojo has a very strong background in Mathematics, Operations Research, Modeling and Simulation, Risk & Decision Analysis, Six Sigma and Software Engineering, and a proven track record of delivered solutions applying these methodologies in practice. He

is also a specialist in Business Systems Analysis & Design (Banking & Insurance) and has delivered major business solutions across several organizations. He has received several formal awards and published a number of written works, including a couple of textbooks. Vojo has also been featured as a guest speaker at several prominent conferences internationally.

DESIGN AND IMPLEMENTATION OF A FUZZY COGNITIVE MAPS EXPERT SYSTEM FOR OIL PRICE ESTIMATION

A. Azadeh ^(a), Z.S. Ghaemmohamadi ^(b)

^(a) Department of Industrial Engineering and Center of Excellence for Intelligent Based Experimental Mechanics, College of Engineering, University of Tehran, Tehran, Iran

^(b) Department of Industrial Engineering, College of Engineering, University of Tehran, Iran

^(a) aazadeh@ut.ac.ir, ^(b) ghaem@ut.ac.ir

ABSTRACT

The objective of this study is to design a fuzzy cognitive maps expert system for estimation of monthly oil price based on intelligent approaches and meta heuristics. Oil price is influenced by several elements, such as politic and social factors. In this paper a fuzzy cognitive maps (FCMs) approach is presented in order to explore the importance of these factors in oil price estimation. To this purpose, causal relationship between affective factors and oil price are depicted and relationship values between them are computed. The proposed expert system utilizes Genetic Algorithm (GA), Artificial Neural Network (ANN) and Adaptive Neural Fuzzy Inference System (ANFIS). The system is coded in .net environment by C# and Matlab and Excel are also used linked for data processing and evaluation. The expert system identifies the preferred method (from GA, ANN, ANFIS) through mean absolute percentage error (MAPE).

Keywords: Expert system, Fuzzy cognitive maps, oil price, Estimation

1. INTRODUCTION

Crude oil, sometimes called the blood of industries, plays an important role in any economies (Fan et al. 2008). The role of oil in the world economy becomes more and more significant because nearly two-thirds of the world's energy consumption comes from the crude oil and natural gas (Alvarez-Ramirez et al. 2003). The crude oil price is basically determined by its supply and demand, and is strongly influenced by many events like the weather, inventory, GDP growth, refinery operable capacity, political aspects and people's expectation. Sharp oil price movements are likely to disturb aggregate economic activity, volatile oil prices have been considerable interest to many researchers and institutions. Therefore, forecasting oil prices is an important and very hard topic due to its intrinsic difficulty and practical applications (Wang et al. 2004).

There is an array of methods that are available today for forecasting energy price. An appropriate method is chosen based on the nature of the data available and the desired nature and level of detail of

the forecasts (Azadeh et al. 2010). For crude oil price forecasting, Mirmirani and Li (2004) applied VRA and ANN techniques to make ex-post forecast of US oil price movement. Lagged oil price, lagged oil supply, and lagged energy consumption were used as three endogenous variables for VAR-based forecast. Ye et al. (2006) provided a model to forecast crude oil spot prices in the short-run using high- and low-inventory variables. They showed that the non-linear-term model better captures price responses at very high- or very low-inventory levels and improves forecasting capability. Wang et al. (2005) proposed a new integrated methodology-TEI@I methodology and showed a good performance in crude oil price forecasting with back propagation neural network (BPNN) as the integrated technique. Xie et al. (2006) proposed a support vector regression (SVR) model to predict crude oil price. Similarly, Shambora and Rossiter (2007) and Yu et al. (2007) also used the ANN model to predict crude oil price. Yousefi et al. (2005) introduces a wavelet-based prediction procedure and market data on crude oil is used to provide forecasts over different forecasting horizons. Sadorsky (2006) uses several different univariate and multivariate statistical models such as TGARCH and GARCH to estimate forecasts of daily volatility in petroleum futures price returns. Amin-Naseri and Gharacheh (2007) proposed a hybrid AI approach integrating feed-forward neural networks, genetic algorithm, and k-means clustering, to predict the monthly crude oil price and obtain better results.

In this paper, we develop a fuzzy cognitive maps expert system model for forecasting monthly crude oil spot prices using readily available data. The objective of this model is to provide a forecast of monthly West Texas Intermediate (WTI) prices using readily available data. In addition, this paper examines the feasibility of applying fuzzy cognitive maps expert system in crude oil price forecasting through the contrast with ANN, ANFIS and GA models.

The rest of the paper is organized as follows: Section 2 describes fuzzy cognitive maps expert system method for crude oil price prediction. To evaluate the fuzzy cognitive maps expert system, a main crude oil

price series, West Texas Intermediate (WTI) crude oil spot price is used to test the effectiveness of the proposed methodology, and its comparable results with ANN, ANFIS and GA methods. Some concluding remarks are made in section 4.

2. FUZZY COGNITIVE MAPS EXPERT SYSTEM FOR CRUDE OIL PRICE

In this section, a fuzzy cognitive maps expert system method for time series forecasting and its application in crude oil price prediction are presented. We apply ANN, ANFIS and GA in this fuzzy cognitive maps expert system model. Then present the fuzzy cognitive maps expert system method for oil price forecasting.

2.1. Fuzzy Cognitive Maps Expert System

Expert system technology has proven to benefit decision making process in businesses and accounting management of corporations. Most applications are developed in production/operations management area with lowest number of applications in the human resources area (Mearns et al. 2003). There are several applications in the area of diagnosis. They include defects diagnostic system for tire production and service (Prez-Carretero et al. 2002). Benefits of an expert system approach to productivity analysis include cost reductions due to the reduced need for manpower, faster analysis of pressing productivity problems, and more consistent appraisals and interpretation of productivity performance (Azadeh et al. 2008).

2.1.1. Fuzzy Cognitive Maps (FCM)

Cognitive maps (CMs) were introduced by Axelrod (1976) in the 1970s. CMs are signed diagraphs designed to represent the causal assertions and belief system of a person (or group of experts) with respect to a specific domain, and use that statement in order to analyze the effects of a certain choice on particular objectives. Two elements are used when realizing CMs: concepts and causal belief. Concepts represent the variables that describe the belief system of a person, while the causal belief consists in the causal dependencies between variables. Such variables can be continuous, ordinal or dichotomous (Kardaras and Karakostas, 1999). In signed cognitive maps, each relationship is linked to a sign that represents the sense of casual influence of the cause variable on the effect variable. Fuzzy cognitive map is a well-established artificial intelligence technique that incorporates ideas from artificial neural networks and fuzzy logic. FCMs were introduced by Kosko (1986) to extend the idea of cognitive maps by allowing the concepts to be represented linguistically with an associated fuzzy set rather than requiring them to be precise. In order to describe the degree of the relationship between concepts it is possible to use a number between [0,1] and [-1, 1], or use linguistic terms, such as “often”, “always”, “some”, “a lot”, etc. Figure 1 shows an example of FCM used by Kosko to define the indirect and the total effects for an FCM

(Kosko 1986). Three paths connect C_1 to C_5 , so there are three indirect effects of C_1 on C_5 :

along path $P_1(C_1, C_2, C_4, C_5)$:

$$I_1(C_1, C_5) = \min \{e_{12}, e_{24}, e_{45}\} = \text{some}$$

along path $P_2(C_1, C_3, C_5)$:

$$I_2(C_1, C_5) = \min \{e_{13}, e_{35}\} = \text{much}$$

along path $P_3(C_1, C_3, C_4, C_5)$:

$$I_3(C_1, C_5) = \min \{e_{13}, e_{34}, e_{45}\} = \text{some}$$

Thus, the total effect of C_1 to C_5 is:

$$T(C_1, C_5) = \max \{I_1(C_1, C_5), I_2(C_1, C_5), I_3(C_1, C_5)\} = \text{much}$$

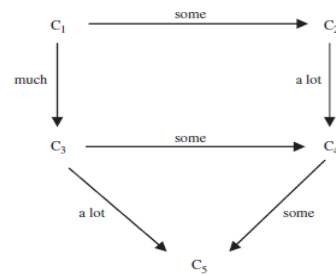


Figure 1: Example of FCM

2.1.2. Data Fuzzification

Fuzzification is a process in which the input data, precise or imprecise is converted into linguistic formation, which is easily perceptible by the human minds (Wagner et al. 2001). All relationship between concepts (indicators of the proposed oil price volatility estimation) are linguistic variables. The most typical fuzzy set membership function has the graph of a triangle. The fuzzy set membership function of our model is also a triangle. This approach translates the point (x_1^*, \dots, x_n^*) in set A to a fuzzy set A' as shown in (1). Fuzzy sets for relationship between concepts are shown in Figure 2.

$$\mu_{A'}(x) = \begin{cases} 1 - \frac{|x_1 - x_1^*|}{b_1}, \dots, 1 - \frac{|x_n - x_n^*|}{b_n}, & |x_i - x_i^*| \leq b_i \\ 0, & \text{else} \end{cases} \quad (1)$$

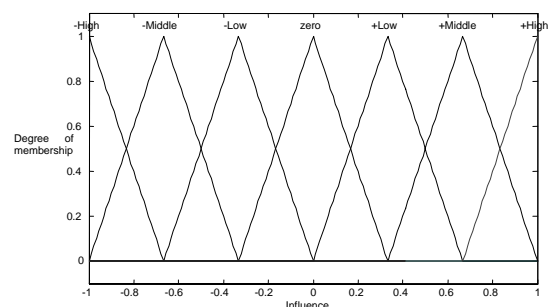


Figure 2: The fuzzy set for relationship between concepts

2.1.3. Data Defuzzification

Defuzzification is the process of producing a quantifiable result in fuzzy logic (Wilson et al. 1992). There are some methods for defuzzification such as centroid average (CA), center of gravity (CG), maximum center average (MCA), mean of maximum (MOM), smallest of maximum (SOM), largest of maximum (LOM) (Wong et al. 1995).

Our fuzzy cognitive maps expert system uses center of gravity (CG). This is because the approach provides better solution than other methods via data engine. We have the center of gravity (CG) as shown in (2):

$$y' = \frac{\int_A y \mu_{A'}(y) dy}{\int_A \mu_{A'}(y) dy} \quad (2)$$

2.2. The Proposed Fuzzy Cognitive Maps Expert System

An excellent approach is fuzzy cognitive maps expert system that can implement crude oil price forecasting in the volatile crude oil market. The flow chart of the fuzzy cognitive maps expert system is shown in Figure 3.

From Figure 3, the fuzzy cognitive maps expert system for crude oil price forecasting consists of some main components, i.e., graphical user interface module (GUI), oil price forecasting with ANN, ANFIS, GA module, oil price volatility correction with fuzzy cognitive maps module and integration module.

GUI: It is a graphical window through which users can exchange information with the fuzzy cognitive maps expert system and also users enter necessary data in system. In details, it handles all input/output between users and the fuzzy cognitive maps expert system.

Oil price forecasting with ANN, ANFIS and GA module: in this study ANN, ANFIS and GA predict the future value of oil price using the historical data. The crude oil prices data are used in this paper are monthly spot prices of West Texas Intermediate (WTI) crude oil. For a univariate time-series forecasting problem, the inputs of the network are the past lagged observations of the data series and the outputs are the future values. According to Pierson coefficient of correlation, oil price of a month before and oil price of two months before are chosen for oil price forecasting. Oil price forecasting accomplish with three methods, ANN, ANFIS and GA, then according to MAPE best forecasting is chosen between three methods. The parameters in these methods are chosen based on previous studies and also using Trial and error.

Oil price volatility correction with fuzzy cognitive maps module: Crude oil market is an unstable market with high volatility and oil price is often affected by many related factors (Wang et al. 2004). In this paper we used from nine factors that they

are premier than other factors, according to experts' ideas. Eight experts draw casual graphs between these factors and volatility oil price. The experts use linguistic terms for expressing relationships between factors. Taber (1991) suggested a relation to unify different judgments of experts, based on the credibility weight of each expert. In our applications, each expert had the same credibility. At the end of the process the result of different factors effects on volatility oil price appears as in Table 1.

Table 1: The effects of different factors on volatility oil price

Effective factors	The effect on volatility oil price
World oil demand	0.78
Reduce excess capacity	0.67
Agiotage	0.42
Devaluation dollar	0.83
Financial Crisis	0.25
Government changes and internal turmoil	- 0.16
Environmental policy	0.29
OPEC cuts oil production	0.4
Natural events	0.25

Integration module: Crude oil price forecasting obtains by implementing, oil price forecasting with ANN, ANFIS, GA module and oil price volatility correction with fuzzy cognitive maps module. Indeed crude oil price forecasting obtains by adding two values of two modules.

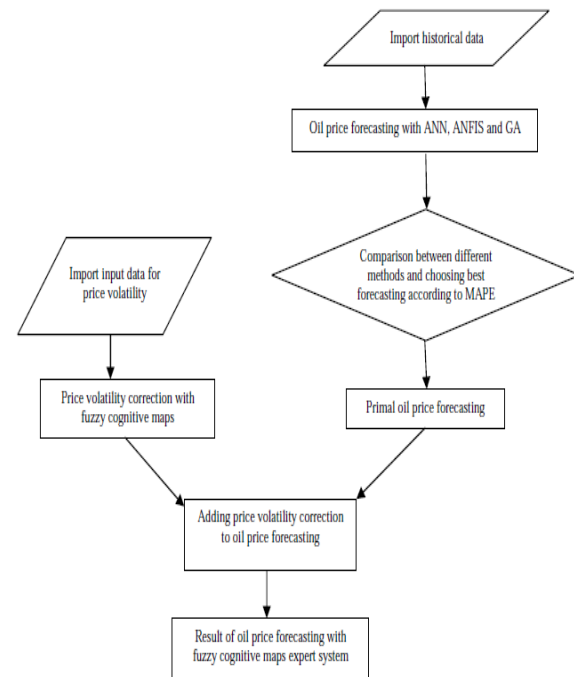


Figure 3: The overall computational flow chart of the fuzzy cognitive maps expert system

3. A CASE STUDY

In this section, we first describe the data, and then define some evaluation criteria for prediction purposes. Finally, the empirical results are presented.

3.1. Data

The crude oil price data used in this study are monthly spot prices of West Texas Intermediate (WTI) crude oil from January 1991 to December 2010 with a total of $n = 240$ observations. These data include train data and test data. The train data are 216 (90 percent of total data as usual) observations and the test data are 24 (10 percent of total data as usual) observations. The main reason of selecting this oil price indicator is that this crude oil price is the most famous benchmark price, which is used widely as the basis of many crude oil price formulae (Yu et al. 2008). The crude oil price data used in this study are obtainable from the energy information administration (EIA) website of Department of Energy of USA (<http://www.eia.doe.gov>).

3.2. Data Preprocessing

As in time-series methods making the process, covariance stationary is one of the basic assumptions and also using preprocessed data is more useful in most heuristic methods (Zhang et al. 2005), and so the stationary assumption should be studied for the models. In time series forecasting, the appropriate preprocessing method should have two main properties. It should make the process stationary and have post processing capability. The most useful preprocessed methods are presented in the sections.

The first difference method: The difference method was proposed by Box et al. (1994) In this method, transformation should be applied:

$$y_t = x_t - x_{t-1} \tag{3}$$

However, for the first difference of the logarithm method the transformation is adjusted as follows:

$$y_t = \log(x_t) - \log(x_{t-1}) \tag{4}$$

Normalization: There are different normalization algorithms which are Min-Max Normalization, Z-Score Normalization and Sigmoid Normalization.

We used these methods to estimate time series functions and finally according to mean absolute percentage error (MAPE) we didn't apply any methods for preprocessing.

3.3. Evaluation Criteria

There are four basic error estimation methods which are listed: Mean absolute error (MAE), Mean square error (MSE), Root mean square error (RMSE) and Mean absolute percentage error (MAPE). They can be calculated by the following equations, respectively:

$$\begin{aligned} MAE &= \frac{\sum_{t=1}^n |x_t - x'_t|}{n} \\ MSE &= \frac{\sum_{t=1}^n (x_t - x'_t)^2}{n}, \\ RMSE &= \sqrt{\frac{\sum_{t=1}^n (x_t - x'_t)^2}{n}}, \\ MAPE &= \frac{\sum_{t=1}^n \left| \frac{x_t - x'_t}{x_t} \right|}{n} \end{aligned} \tag{5}$$

All methods, except MAPE have scaled output. MAPE method is the most suitable method to estimate the relative error because input data used for the model estimation, preprocessed data and raw data have different scales (Azadeh et al. 2011).

3.4. Results and Analysis

Result of Each of the forecasting method described in the last section is presented in Table 2 in which a comparison among them is performed. Each of methods is estimated and validated by train data. The model estimation selection process is then followed by an empirical evaluation which is based on the test data.

Table 2 shows the detailed results of the simulated experiment via the four methods. It can be seen that the fuzzy cognitive maps expert system method outperforms other models in term of MAPE. Focusing on the MAPE indicators, the values of fuzzy cognitive maps expert system model are explicitly lower than those of ANN, ANFIS and GA except in the third sub-period.

The main reasons for the above conclusions are as follows. As Panas et al. (2000) reported, the crude oil market is one of the most volatile markets in the world and shows strong evidence of chaos. All methods can in principle describe the nonlinear dynamics of crude oil price. Best method between ANN, ANFIS and GA for different periods is different, so, a hybrid method that it uses from all methods is necessary and useful. Fuzzy cognitive maps expert system is resistant to the over-fitting problem and can model nonlinear relations in an efficient and stable way.

Table 2: Crude oil forecast results according to mape

Method	Full period (1991-2010)	Sub-period 1 (1991-1995)	Sub-period 2 (1996-2000)	Sub-period 3 (2001-2005)	Sub-period 4 (2006-2010)
ANN	0.0594	0.0256	0.0588	0.0525	0.0222
ANFIS	0.1534	0.0392	0.0658	0.334	0.0328
GA	0.0654	0.047	0.0799	0.0457	0.0285
FCM expert system	0.0474	0.0245	0.0423	0.0386	0.0216

Actual values and forecasting values with fuzzy cognitive maps expert system for test data are shown in Figure 3.

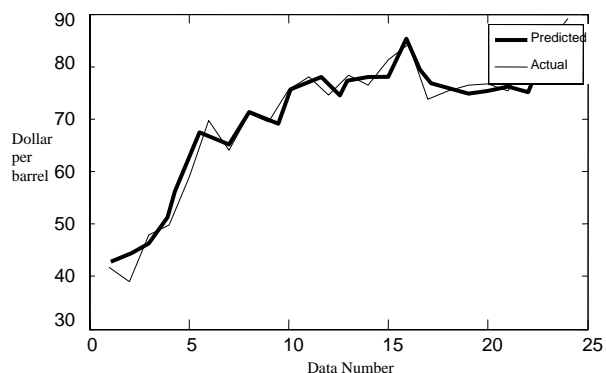


Figure 3: WTI crude oil price forecast based on fuzzy cognitive maps expert system for test data

REFERENCES

- Alvarez-Ramirez, J., Soriano, A., Cisneros, M., Suarez, R., 2003. Symmetry/anti-symmetry phase transitions in crude oil markets, *Physica A*, 322, 583-596.
- Amin-Naseri, M. R., Gharacheh, E. A., 2007. A hybrid artificial intelligence approach to monthly forecasting of crude oil price time series. *The Proceedings of the 10th International Conference on Engineering Applications of Neural Networks*, CEUR-WS 284, 160-167.
- Axelrod, 1976. *Structure of Decision*. Princeton University Press, Princeton, NJ.
- Azadeh, A., Asadzadeh, S. M., Ghanbari, A., 2010. An adaptive network-based fuzzy inference system for short-term natural gas demand estimation: Uncertain and complex environments. *Energy Policy*, 38, 1529-1536.
- Azadeh, A., Fam, I. M., Khoshnoud, M., Nikafrouz, M., 2008. Design and implementation of a fuzzy expert system for performance assessment of an integrated health, safety, environment (HSE) and ergonomics system The case of a gas refinery. *Information Sciences*, 178, 4280-4300.
- Azadeh, A., Saberi, M., Asadzadeh, S.M., 2011. An adaptive network based fuzzy inference system-auto regression-analysis of variance algorithm for improvement of oil consumption estimation and policy making The cases of Canada, United Kingdom, and South Korea. *Applied Mathematical Modelling*, 35, 581-593.
- Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 1994. *Time Series Analysis: Forecasting and Control*. Prentice-Hall, Englewood Cliffs, NJ.
- Fan, Y., Liang, Q., Wei, Y. M., 2008. A generalized pattern matching approach for multi-step prediction of crude oil price. *Energy Economics*, 30, 889-904.
- Kardaras, D., Karakostas, B., 1999. The use of cognitive maps to simulate the information systems strategic planning process. *Information and Software Technology*, 41, 197-210.
- Kosko, B., 1986. Fuzzy cognitive maps. *International Journal of Man-Machine Studies*, 24, 65-75.
- Taber, R., 1991. Knowledge processing with fuzzy cognitive maps. *Expert Systems with Applications*, 2, 83-87.
- Wang, Sh., Yu, L., Lai, K. K., 2004. A Novel Hybrid AI System Framework for Crude Oil Price Forecasting. *CASDMKM, LNAI 3327*, 233-242.
- Mearns, M., Whitaker, S. M., Flin, R., 2003. Safety climate, safety management practice and safety performance in offshore environments. *Safety Science*, 41, 641-680.
- Mirmirani, S., Li, H.C., 2004. A comparison of VAR and neural networks with genetic algorithm in forecasting price of oil. *Advances in Econometrics*, 19, 203-223.
- Panas, E., Ninni, V., 2000. Are oil markets chaotic? A non-linear dynamic analysis. *Energy Economics*, 22, 549-568.
- Prez-Carretero, C., Laita, L. M., Roanes-Lozano, E., Lazaro, L., Gonzalez-Cajal, J., Laita, L., 2002. Logic and computer algebra based expert system for diagnosis of anorexia. *Mathematics and Computers in Simulation*, 58, 183-202.
- Sadorsky, P., 2006. Modeling and forecasting petroleum futures volatility. *Energy Economics*, 28, 467-48.
- Shambora, W. E., Rossiter, R., 2007. Are there exploitable inefficiencies in the futures market for oil?. *Energy Economics*, 29, 18-27.
- Wang, S. Y., Yu, L.A., Lai, K. K., 2005. Crude oil price forecasting with TEI@I methodology. *Journal of Systems Science and Complexity*, 18, 145-166.
- Wagner, W.P., Najdawi, M.K., Chung, Q.B., 2001. Selection of knowledge acquisition techniques based upon the problem domain characteristics of production and operations management expert systems. *Expert Systems*, 18, 76-87.
- Wilson, J.R., Corlett, E.N., 1992. *Evaluation of Human Work: A Practical Ergonomics Methodology*. Taylor and Francis, USA, 141-169.
- Wong, B.K., Monaco, J.A., 1995. A bibliography of expert system applications for business (1984-1992). *European Journal of Operational Research* 85, 416-432.
- Xie, W., Yu, L., Xu, S.Y., Wang, S.Y., 2006. A new method for crude oil price forecasting based on support vector machines. *Lecture Notes in Computer Science*, 3994, 441-451.
- Ye, M., Zyren, J., Shore, J., 2006. Forecasting short-run crude oil price using high- and low-inventory variables. *Energy Policy*, 34, 2736-2743.
- Yousefi, S., Weinreich, I., Reinartz, D., 2005. Wavelet-based prediction of oil prices. *Chaos, Solitons and Fractals*, 25, 265-275.
- Yu, L., Lai, K.K., Wang, S.Y., He, K.J., 2007. Oil price forecasting with an EMD-based multi scale neural

network learning paradigm. *Lecture Notes in Computer Science*, 4489, 925–932.

Yu, L., Wang, Sh., Lai, K. K., 2008. Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. *Energy Economics*, 30, 2623–2635.

Zhang, G.P., Oi, M., 2005. Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research*, 160, 501–514.

AUTHORS BIOGRAPHY

Ali Azadeh is an Associate Professor and the founder Department of Industrial Engineering and co-founder of RIEMP at the University of Tehran. He graduated with first class honor Degree (BS) in Applied Mathematics from the University of San Francisco and obtained his MS and PhD in Industrial and Systems Engineering from the San Jose State University and the University of Southern California. He received the 1992 Phi Beta Kappa Alumni Award for excellence in research and innovation of doctoral dissertation in USA. He is the recipient of the 1999–2000 Applied Research Award and has published more than 300 academic papers.

Zeinab Sadat Ghaemmohamadi is currently a Graduate student in Industrial Engineering at the University of Tehran. He earned his BS in statistics from Isfahan University of Technology and his MS in Industrial Engineering from Tehran University, Iran. His current research interests include systems modeling and forecasting of supply, demand and price of energy via artificial intelligence tools.

HOW TO BENEFIT MORE FROM INUITIVE POWER AND EXPERIENCE OF THE HUMAN SIMULATION KNOWLEDGE STAKEHOLDER

Gaby Neumann

Faculty of Engineering / Industrial Engineering
Technical University of Applied Sciences Wildau

gaby.neumann@th-wildau.de

ABSTRACT

Generally, it is pretty clear and widely accepted that the human actor plays a significant role in any simulation project – although in recent years some authors proclaimed a revival of human-free simulation at least related to distinct parts of a simulation study. Therefore, the paper aims to provide an overview on needs and challenges in model-user interaction as well as on approaches, methods and tools to support the user in bringing in his/her knowledge in all phases of a simulation project from model building via understanding a model and using it for experimentation to correctly interpreting simulation outcome. Furthermore, barriers and problems hindering a simulation stakeholder in sharing his/her knowledge are identified and approaches to access and extract such knowledge are discussed in order to avoid inefficiency and failure in future projects.

Keywords: knowledge-based simulation, simulation knowledge, discrete event simulation, knowledge management

1. INTRODUCTION AND MOTIVATION

The impact of a person's knowledge and background on the design, level-of-detail and focus of the simulation model, i.e. on the way a simulation model appears and functions, was demonstrated, for example, by Neumann and Page (2006). Here, two groups of students with different background (computing vs. logistics) but the same level of simulation knowledge and experience were assigned with the same problem to be investigated. In the end both student projects produced valid and usable simulation models, but efforts for model implementation, model modification in the course of experimentation and visualization of results were quite different. Results achieved from either model equally allowed responding to the initial questions addressed to the simulation project; from this it was possible to conclude that despite of different modeling approaches simulation results are comparable and of similar quality. This way, the case study gave proof of the fact that different persons with different background might produce different but in the same way correct and usable simulation models of the same problem and

situation just because of their individual knowledge and experience. Consequently, the individual background significantly impacts the whole range of a simulation project from model building till interpretation of results (see Figure 1).

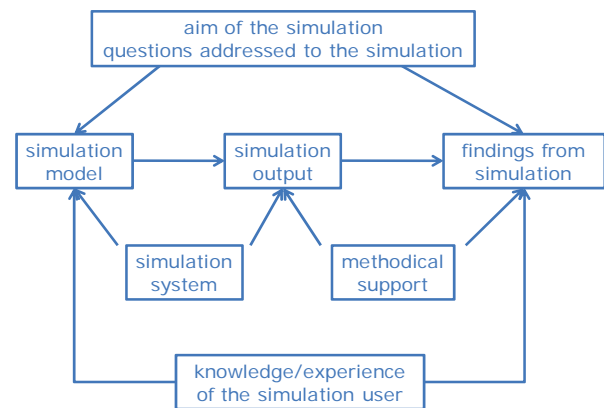


Figure 1: Impact of the simulation user on the outcome of a simulation project

Neumann and Ziems (1997) went into detail with identifying human simulation knowledge stakeholders' impact on certain simulation project stages. According to this, simulation experts are primarily responsible for model building and implementation steps, whereas domain experts mainly provide application-specific knowledge for problem description, identification of input data and evaluation of results. This corresponds to the type of knowledge and experience brought into a simulation project and gained from a simulation project by the different actors. Therefore, simulation needs to be understood in its entire characterization as complex problem-solving, knowledge-generation and learning process at the same time.

This view is in line with literature characterizing modeling and simulation in general as both, knowledge-processing activity and goal-directed knowledge-generation activity (Ören 1990). Based upon this, advanced methodologists and technologists were expected to be allowed to integrate simulation with several other knowledge techniques. But looking at today's situation in simulation projects it still has to be considered that a sound application of a knowledge

management perspective to modeling and simulation is still missing. Instead, the term ‘knowledge-based simulation’ is typically used for applying AI approaches to automatically create simulation models from expert knowledge. Research focuses, for example, on developing efficient and robust models and formats to capture, represent and organize the knowledge for developing conceptual simulation models that can be generalized and interfaced with different applications and implementation tools (Zhou, Son and Chen 2004). Other work aims to develop concepts for modeling human decisions, e. g. in manufacturing systems (Zülch 2006), or to model and simulate human behavior to support workplace design by use of digital human models (Monteil et al. 2010) and especially by incorporating human characteristics like fear, aggressiveness, fatigue and stress in particularly challenging situations (Bruzzone et al. 2010).

In contrast to this, the fact that non-formalized expert knowledge finds its way into the simulation model on one hand or is created throughout the simulation lifecycle and needs to be externalized on the other is not in the focus of research in this field. That is why, information about decisions taken when building the model or running experiments as well as really new knowledge about the particular application or even about the simulation methodology gained in the course of a simulation project quite often stays in the heads of the people involved in the project. Furthermore, the simulation model itself also forms a kind of dynamic repository containing knowledge about parameters, causal relations and decision rules gathered through purposeful experiments. This knowledge is being somewhat hidden as long as not being discovered, understood and interpreted by another person.

Against this background, research on implementing a knowledge management perspective in simulation projects should address the following questions:

- Which information and knowledge is needed by whom at what stage of a simulation project?
- Which knowledge and information is provided by whom in which step of a simulation project?
- Which knowledge is generated with whom in which step of the simulation project?
- Which knowledge is “stored” in the conceptual and simulation models, evolves from simulation experiments, and is “hidden” in the input/output data of simulation runs?
- How simulation knowledge with the different stakeholders or repositories can be accessed, extracted, externalized and distributed, shared, applied?

To generalize research needs, the biggest challenge for properly handling modeling and simulation knowledge by applying knowledge management methods and tools consists in providing the right knowledge of the right

quality and with the right costs at the right place and time. In other words, it is essential not to focus on the introduction of knowledge management technology and integration of software tools for storing and retrieving knowledge and information only, but to put the human resources running model building and simulation projects into the centre of gravity and to try to give them that kind and amount of support which is needed in a particular situation.

Therefore, the paper aims to provide an overview on needs and challenges in model-user interaction (Section 2) as well as approaches, methods and tools to support the user in bringing in his/her knowledge in all phases of a simulation project from model building via understanding a model and using it for experimentation to correctly interpreting simulation outcome (Section 3). Barriers and problems hindering a simulation stakeholder in sharing his/her knowledge are identified and approaches to access and extract such knowledge are discussed (Section 4). Findings are summarized and conclusions are drawn in Section 5.

2. NEEDS AND CHALLENGES IN MODEL-USER INTERACTION

Once a valid simulation model is available it serves as tool for different types of studies:

- In a *what-if analysis* the user discovers how a system reacts on changing conditions or performance requirements, i.e. system loads. During experimentation a particular type of changes is introduced to the model in a systematic way in order to understand sensitivity of a certain parameter, design or strategy.
- A *what-to-do-to-achieve investigation* aims to answer questions like how to set system parameters or how to improve process control in order to reach a certain behavior or performance level. Experimentation might be multidimensional including different types of changes to the model; it is strongly oriented towards identifying modification strategies for reaching a particular performance objective or target behavior.
- *Performance optimization experiments* serve to solve a particular target function such as minimizing job orders’ time in system or stock level, maximizing service level or resources’ utilization, etc. Here, the limits of typical performance characteristics are to be identified with the respective limit value itself forming the goal of the investigation.

No matter which type of investigation is on the agenda the user always needs to interact with the model in order to implement the intended experimentation strategy and to gain simulation results.

Interaction prior to the simulation run (or a batch of simulation runs) might consist in adjusting the structure

of the simulation model, in purposefully changing one or more model or simulation parameters, or even simply in starting the simulation in order to produce and collect simulation output data that are expected to be of use for the investigation. Post-run interaction focuses on accessing and dealing with simulation output data in the form of dynamic visualization (i.e. watching animations) or statistical analysis (i.e. checking original or condensed data, viewing diagrams or other types of graphical representation) in order to achieve findings with regard to the focus and aim of the investigation. Consequently, the entire interaction cycle can be characterized as a user-model dialogue: any pre-run interaction with the model corresponds to the concept of asking questions; post-run interaction is adequate to the concept of responding to questions. Pre-condition for a successful user-model dialogue is true understanding in both directions. The simulation model needs to “understand” what the user is interested in and looking for. This requires the ability to ask the right questions from the user. Those questions might either be very specific and clearly matching “the language of the model” (i.e. directly addressing input/output data of a simulation) or they are of more principle, general, eventually even fuzzy nature requiring a kind of translation for being understandable to the model. When it comes to the responding part of the dialogue the user needs to understand the simulation output for getting the answers s/he was looking for.

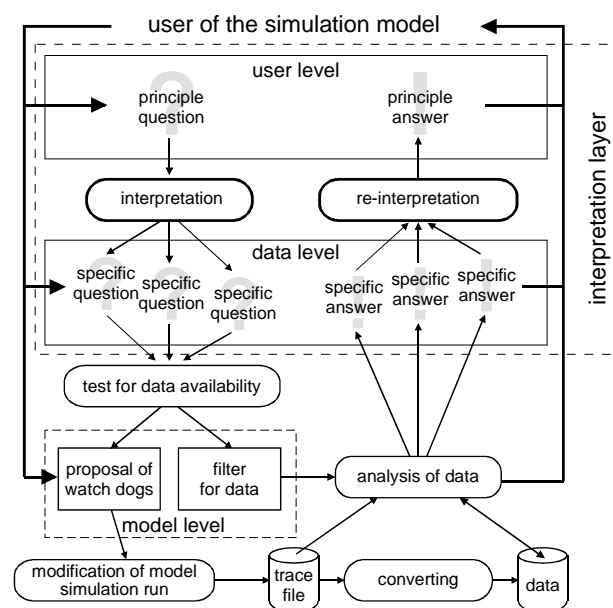


Figure 2: User-data interaction for simulation output analysis

When the potential interests a simulation user might have in a simulation study are compared, one significant difference emerges: specific questions formulated by the user might directly be answered with concrete simulation output at data level; those usually fuzzy questions of principle from the more global user’s point of view require interpretation and re-interpretation steps before being answered. Here, any question of

principle has to be transferred to the data level by explaining it in detail and putting it in terms of concrete data (see Figure 2). As result of this process of interpretation a set of specific questions is defined with each of them providing a specific part of the overall answer in which the user is interested. Questions at data level correspond to results that can be delivered directly by the simulation even if minor modifications to the simulation model should be required (Tolujew 1997). This is the kind of study also current approaches for automatic trace file analysis in order to better cope with large amounts of simulation output data support (Kemper and Tepper 2009, Wustmann et al. 2009). Those approaches mainly focus on formalizing simulation outcome in the context of a certain application area. With this they remain at data level, whereas deriving answers of principle to questions of principle requires processing further the respective set of specific answers. These steps of additional analysis and condensing can be understood as a process of re-interpretation to transfer results from data to user level.

All steps of interpretation and re-interpretation aim to link the user’s point of view to that of the simulation model. They not only require an appropriate procedure, but, even more importantly, an interpretative model representing the application area in which simulation takes place. This model needs to be based on knowledge and rules expressed in the user’s individual expertise, but also in generalized knowledge of the problem environment regarding design constraints or system behavior and the experience of the model building expert derived from prior simulations. This knowledge might not only be of explicit nature, i.e. existing independent of a person and suitable to be articulated, codified, stored, and accessed by other persons, but also comprises implicit or tacit knowledge carried by a person in his or her mind often not being aware of it. Whereas explicit knowledge might be transferred into rules and algorithms, tacit knowledge cannot be separated from its owner and therefore requires direct involvement of the knowledge holder in the interpretation process.

In the end, knowledge stored in the simulation model can be considered proven, independently of whether it was developed by the domain expert him- or herself or by a consultant simulation expert (Neumann and Ziems 1997). Unfortunately, this knowledge is usually not very well documented and therefore does exist implicitly only inside the simulation model. To be used when the results of the simulation project are put into practice, it needs to be explained in such a way as to be accessible to the domain expert in the subject-specific terminology and to be applicable without any loss of information or misrepresentation. Otherwise the technical or organizational solution in the real world cannot be expected to work in the way demonstrated by the respective simulation model or knowledge important for the realization of simulated functionality needs to be re-developed by renewed implementation and testing.

3. METHODS AND TOOLS FOR BRINGING IN SIMULATION KNOWLEDGE

Human resources involved in a simulation project are the key factors for its success and efficiency. As discussed in the previous section it is always up to the simulation user to define objectives of any simulation and target functions of any experimentation. For this detailed knowledge and understanding on the particular system/process to be investigated and problem to be solved is needed as well as sound background knowledge on the domain and experiences in simulation-based problem solving. As this individual knowledge and experience belongs to the person carrying it and continuously develops and grows over time with each new simulation project, it can be separated from the person, i.e. externalized, to some extent only. Therefore, a mix of methods and tools for bringing in a user's knowledge and experience into the simulation project is needed:

- Formalize what can be formalized and incorporate this into simulation tools completed by a rule-based supporting system and an interface for its continuous improvement.
- Apply algorithms to routine problem-solving (Kemper and Tepper 2009, Wustmann et al. 2009).
- Enable a structured dialogue between the user and the tool by applying the concept of oracle-based simulation model validation (Helms and Strothotte 1992).
- Provide support in structured documentation of problem, model, experiments, solution/findings and lessons learned (Neumann 2006).
- Use human intuition and tacit knowledge for all that cannot be formalized (yet).
- Allow the user to bring in his/her ability of flexible thinking for problems and questions that unexpectedly pop-up in the course of a simulation study (Tolujew et al. 2007).

Here, it is crucial to initiate an ongoing learning and improvement process as basis of structured knowledge explication and gathering of experiences similar to what has been proposed by Brandt et al. (2001) for software engineering projects. Applying this approach to learning from simulation projects, a well-defined and well-structured documentation of both simulation model and simulation runs and the simulation project with all its assumptions, agreements, and decisions has to be established (seamlessly and continuously). Procedures help to identify who knows what about the system and process, but also about the simulation project behind it, why something was decided in which way, which system configuration and which set of parameters work well together, what is in the simulation model and what the limitations of its validity and usability are. With this the process of a simulation project becomes a process of knowledge

creation and acquisition at the same time without too much additional effort for all involved (see Figure 3).

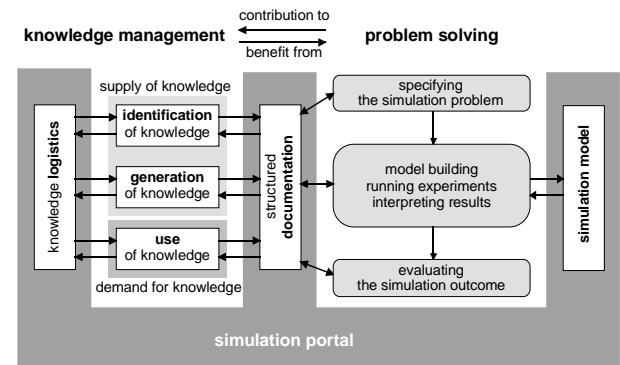


Figure 3: Problem solving and knowledge acquisition in the course of a simulation project

The clue to the successful implementation of those knowledge management procedures is often an appropriate (supporting) environment and climate in the organization. Concerning this, there is a greater need for a cultural shift than for additional software tools and IT solutions. Adopting a statement on human needs for computer technology by Shneiderman (2002) the link between knowledge management and (simulation-based) problem solving can generally be described as follows: the old discussion about how to support problem solving is about what (software) tools can do; the new discussion about how to support problem solving is (and must be) about what kind of problem-solving support people really need.

4. PROBLEMS AND BARRIERS IN SIMULATION KNOWLEDGE SHARING

In the course of a simulation project there are bidirectional links between activities for problem solving and knowledge management. On one hand knowledge available with persons, inside organizations and in the form of technology is (re-)used to build a model, plan and run experiments, analyze and understand simulation output. On the other hand knowledge about the problem's final solution and the chosen mode of action for its generation characterizes the increased scientific basis and additional experience of the problem-solving person, team or organization. Usually these links are based upon the persons directly involved in the simulation project. It's quite common to make use of own experience, but to benefit from knowledge, experience and lessons learned from other parts of the organization that is still not the usual procedure yet. To overcome this and to make knowledge of a successful or even unsuccessful problem solving process available to future simulation projects that is the challenge for knowledge management and its integration into personalized problem solving.

Being aware of this, organizations invest a large amount of money in technology to better leverage information, but often the deeper knowledge and

expertise that exists within the organization remains untapped. The sharing of knowledge remains limited in most respects, and at least, strained. APQC (2004) sees major reasons for this in technology that is too complicated and the human nature that poses barriers to knowledge sharing. Cultural aspects can enhance an open knowledge transfer or inhibit a positive attitude towards knowledge sharing. Taking cultural aspects into consideration requires letting the knowledge management approach – and with this the knowledge sharing process in particular – fit the culture, instead of making an organization's culture fitting the knowledge management approach (McDermott and O'Dell 2000).

In a perfect world the benefits of accessing and contributing knowledge would be intrinsic: people who share knowledge are better able to achieve their work objectives, can do their jobs more quickly and thoroughly, and receive recognition from their peers and mentors as key contributors and experts. Nevertheless, knowledge is often not shared. O'Dell and Grayson (1998) identified four common reasons for this:

- *Ignorance.* Those who have knowledge don't realize others may find it useful and at the same time someone who could benefit from the knowledge may not know another person in the company already has it.
- *No absorptive capacity.* Many times, an employee lacks the money, time, and management resources to seek out information they need.
- *Lack of pre-existing relationship.* People often absorb knowledge from other people they know, respect, and like. If two managers don't know each other, they are less likely to incorporate each other's experiences into their own work.
- *Lack of motivation.* People do not see a clear business reason for pursuing the transfer of knowledge.

To meet these challenges, the discipline of knowledge sharing should continuously be reinforced. For this, there are two different approaches: the organization might host visible knowledge-sharing events to reward people directly for contributing to knowledge or the organization might rely on the link between knowledge sharing and everyday work processes by embedding knowledge sharing into "routine" work processes. Here, initiating of a close, interpersonal link between a mentor or coach (the expert) and the novice is a promising way not to rely on enthusiasm only, but to bring in a personal commitment to the process of developing another person's simulation competence.

Those expert-novice links might also be part of learning processes to improve an individual's simulation competence in a learning-by-doing scenario. The pedagogical framework for this is formulated by the cognitive apprenticeship theory (see Collins et al.

1989): in general an apprentice is a learner who is coached by a master to perform a specific task. Based on this, the theory transfers the traditional apprenticeship model as known from crafts, trade and industry to the cognitive domain. More precise, cognitive apprenticeship aims at externalizing processes that are usually carried out internally. This approach works with methods like modeling, coaching, scaffolding, articulation, reflection and exploration. Coaching, for example, is to be understood as helping a person in actively creating and successfully passing individual learning processes through guidance-on-demand. In the end, the coach (i.e. the expert) offers support in case of difficulties (i.e. scaffolding), provides hints, feedback and recommendations, and eventually takes over certain steps for solving the given problem. However, the coach only appears when explicitly being called by the person to be coached (i.e. like a help system) and the scaffolding is gradually fading as the learning novice proceeds. So, coaching seems to be a very useful concept for sharing and developing simulation knowledge in practice as it aims to develop heuristic strategies through establishing a culture of expertise and with this goes far beyond pure learning as typically provided in workplace learning environments.

5. SUMMARY AND CONCLUSIONS

To the same extent as a new simulation project provides another challenge to model building, experimentation and interpretation of results it rarely can be planned comprehensively and in all details. Therefore, the simulation knowledge stakeholder cannot be fully replaced by algorithms in a simulation project. Instead his/her intuitive power and experience is needed to appropriately and creatively cope with the unexpected. Here, challenges typically consist in enabling or strengthening purposeful interaction between the simulation model and its user, supporting the user in bringing in his/her simulation knowledge, and overcoming barriers hindering in distributing and sharing knowledge and experience for extending the organizational simulation knowledge base and speeding up the learning curve in human resource development.

The paper presents approaches for dealing with those challenges from a knowledge management perspective. Here, the focus is clearly put on the methodological aspect, whereas implementation into simulation tools or supportive systems remains an open task.

Against this background the main message of the paper consists in underlining the key role a human resources play in simulation projects – no matter if we talk about simulation experts, experts from the application area or even novices to those fields. Despite of this, there are many useful methods, concepts and algorithms even coming from other areas that should be applied to simulation-based investigations in order to support the simulation knowledge stakeholder in more efficient and effective problem-solving and sustainable knowledge explication.

REFERENCES

- APQC, 2004. *Failures in knowledge sharing*. APQC - American Productivity and Quality Center.
- Brandt, M., Ehrenberg, D., Althoff, K., Nick, M., 2001. Ein fallbasierter Ansatz für die computergestützte Nutzung von Erfahrungswissen bei der Projektarbeit. *Proceedings 5. Internationale Tagung Wirtschaftsinformatik*. September 19-21, Augsburg (Germany). (A case-based approach for computer-based use of experience-based knowledge in projects, in German)
- Bruzzone, A., Madeo, F., Tarone, F., 2010. Modelling country reconstruction based on civil military cooperation. *Proc. of European Modeling and Simulation Symposium*, pp. 315-322. October 13-15, Fes (Morocco).
- Collins, A., Brown, J. S., Newman, S. E., 1989. Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. Resnick, L. B., ed. *Knowing, learning, and instruction*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 453-494.
- Helms, C., Strothotte, T., 1992. Oracles and Viewpoint Descriptions for Object Flow Investigation. *Proc. of EUROSIM Congress on Modelling and Simulation*, pp. 47-53, September 30 – October 3, Capri (Italy).
- Kemper, P., Tepper, C., 2009. Automated trace analysis of discrete-event system models. *IEEE Transactions on Software Engineering*, 2 (35), pp. 195-208.
- McDermott, R., O'Dell, C., 2000. *Overcoming the 'Cultural Barriers' to Sharing Knowledge*. APQC - American Productivity and Quality Center.
- Neumann, G., 2006. Projektwissen in der Logistiksimulation erschließen und bewahren: Auf dem Weg zu einer neuen Dokumentationskultur. *Proc. of Simulation in Production and Logistics*, pp. 341-350. September 26-27, Kassel (Germany). (Gaining and storing project knowledge in logistics simulation: on the way towards a new documentation culture, in German).
- Neumann, G., Page, B., 2006. Case study to compare modelling and simulation approaches of different domain experts. *Proc. of the International Mediterranean Modelling Multiconference*, pp. 517-522. October 4-6, Barcelona (Spain).
- Neumann, G., Ziems, D., 1997. Transparente Modell-dokumentation und Resultatpräsentation schafft Vertrauen. *Proc. of Simulation and Animation*, pp. 237-250. March 6-7, Magdeburg (Germany). (Transparent model documentation and presentation of results increases trust, in German)
- O'Dell, C., Grayson, J., 1998. *If only we knew what we know*. Free Press.
- Ören, T.I., 1990. A paradigm for artificial intelligence in software engineering. *Advances in Artificial Intelligence in Software Engineering*, vol. 1, pp. 1-55.
- Rego Monteil, N., del Rio Vilas, D., Crespo Pereira, D., Rios Prado, R., 2010. A simulation-based ergonomic evaluation for the operational improvement of the slate splitters work. *Proc. of European Modeling and Simulation Symposium*, pp. 191-200. October 13-15, Fes (Morocco).
- Shneiderman, B., 2002. *Leonardo's Laptop: Human Needs and the New Computing Technologies*. Cambridge: The MIT Press.
- Tolujew, J., 1997. Werkzeuge des Simulationsexperten von morgen. *Proc. of Simulation and Animation*, pp. 201-210. March 6-7, Magdeburg (Germany). (Tools of the simulation expert of tomorrow, in German)
- Tolujew, J., Reggelin, T., Sermpetzoglou, C., 2007. Simulation und Interpretation von Datenströmen in logistischen Echtzeitsystemen. Engelhardt-Nowitzki, C., Nowitzki, O., Krenn, B., eds. *Management komplexer Materialflüsse mittels Simulation. State-of-the-Art und innovative Konzepte*. Wiesbaden: Deutscher Universitäts-Verlag, pp. 215-232. (Simulation and interpretation of data streams in logistics real-time systems, in German)
- Wustmann, D., Vasyutynskyy, V., Schmidt, T., 2009. Ansätze zur automatischen Analyse und Diagnose von komplexen Materialflusssystemen. *Proc. Expert Colloquium of WGTL*, pp. 1-19. October 1-2, Ilmenau (Germany). (Approaches for automatic analysis and diagnosis of complex material flow systems, in German).
- Zhou, M., Son, Y.J., Chen, Z., 2004. Knowledge representation for conceptual simulation modeling. *Proc. of Winter Simulation Conference*, pp. 450-458. December 5-8, Washington (D.C., USA).
- Zülch, G., 2006. Modelling and simulation of human decision-making in manufacturing systems. *Proc. of Winter Simulation Conference*, pp. 947-953. December 3-6, Monterey (California, USA).

AUTHOR BIOGRAPHY

Gaby Neumann received a Diploma in Materials Handling Technology from the Otto-von-Guericke-University of Technology in Magdeburg and a PhD in Logistics from the University of Magdeburg for her dissertation on "Knowledge-Based Support for the Planner of Crane-Operated Materials Flow Solutions". Between December 2002 and June 2009 she was Junior Professor in Logistics Knowledge Management at the Faculty of Mechanical Engineering there. In December 2009 she became Professor on Engineering Logistics at the Technical University of Applied Sciences Wildau. Since 1991 she also has been working as part-time consultant in material handling simulation, logistics planning and specification of professional competences in certain fields of logistics. Her current activities and research interests are linked amongst others to fields like problem solving and knowledge management in logistics simulation. She has been or is being involved in a couple of research projects in these fields. Gaby Neumann has widely published and regularly presents related research papers at national and international conferences.

OBJECT-ORIENTED MODELLING AND VERIFICATION AIDED BY MODEL SIMPLIFICATION TECHNIQUES

Anton Sodja Borut Zupančič

Faculty of Electrical Engineering, University of Ljubljana
Tržaška 25, 1000 Ljubljana, Slovenia
anton.sodja@fe.uni-lj.si borut.zupancic@fe.uni-lj.si

ABSTRACT

Object-oriented modeling approach brought efficient model reuse and thus possibility to create rich model libraries which enable rapid development of large heterogeneous models. However, verification and debugging of large complex models is becoming and increasingly challenging task. Furthermore, model should not be more complicated as needed for a given purpose. A suitable component describing a subsystem in sufficient detail should be selected from a library which might contain several components describing the same system but with different level of detail.

Benefits of model simplification techniques for object-oriented model development are discussed in this paper. A modeler may help himself with them in a decision making process of how detailed components should be used (e.g., how complicated model should be) and also as an assistance for verifying model by some informal verification methods. Simplified model should be represented in the same way as original, therefore, two simplification techniques are discussed, simplification of object-diagrams and simplification of equations, which are the usual representation of models in Modelica, one of the commonly used object-oriented modeling language today.

Keywords: nonlinear model simplification, model verification, Modelica

1. INTRODUCTION

Dynamic models are an important part of many engineering applications. Various purposes which models have in engineering applications include assistance for system and control design, explaining complex-system behavior and help in operators training.

The purpose of the model also determines a desirable complexity of the model. Modeling is an iterative process and the complexity of the model usually increases during the process as well as the model purpose tends to change. In the early stages of model development relatively simple conceptual models are used which help examine some

general design situations. The error bounds on model prediction are relatively large and validation with experimental data is usually not possible due to the lack of measurements. Therefore, the model is only verified with general design experience and comparison with earlier models of systems with similar characteristics (Murray-Smith, 2009).

Later in the model building process, as more data becomes available, more complex description of the system are integrated into the model. However, more thorough validation may even indicate that some parts of the model are unnecessarily detailed and model is consequentially simplified.

The model building process may be also carried out in reverse: all known details about the systems are included into a model and in the further modeling phases the unnecessary parts of the model are identified and removed (Murray-Smith, 2009).

The emergence of the object-oriented modeling exacerbated the difficulty of assuring proper model complexity and model-verification in some respects. Object-oriented models are built of inter-connected components (models of subsystems). Large collections of various prepared components are available in designated model libraries (Tummescheit, 2002; Andres et al., 2009; Cellier, 1991). Therefore, a very complex model can be built easily. However, this kind of modeling paradigm has some pitfalls, for example, when the modeler is not familiar with the assumptions made at formulation of the components, they can be used incorrectly, e.g., resulting model is invalid due to incompatible components (especially when they come from different libraries). Furthermore, at different modeling stages, different level of model complexity is required and hence a set of components describing the same subsystem but with different level of detail must be present. Another option is to include switches into the components which enables to turn off unnecessary level of details (Casella et al., 2006).

When a model consists of many components (as it is the case with large models) it can be quite intricate to determine how complex each component should be. The complexity of the components with low impact on overall

model behavior should be of course kept low to bound overall model complexity.

2. MODEL SIMPLIFICATION TECHNIQUES

Engineers use experience and intuition to determine important parts of the model which have the highest impact on system's dominant dynamics or on the model's simulation response in specific scenario. In an attempt to diminish reliance on subjective factors such as experience, numerous model simplification and order-reduction methods have been developed (Schwarz et al., 2007; Sommer et al., 2008; Lall et al., 2003; Louca, 1998; Chang et al., 2001). Model simplification techniques consist of running a series of simulations, ranking the individual coordinates or elements by the appropriate metric and removing those that fall below a certain threshold (Chang et al., 2001). The choice of the ranking metric and simplification steps depends on modeling formalism used and may be limited by modeling domains. A special class of model simplification methods are those that produce *proper models*, i.e., simplified models have physically meaningful variables and parameters.

It is obvious how model simplification tools can help us determine if the model is too complicated: if model could be simplified a lot without losing too much accuracy, it is clearly too complicated.

Model simplification methods can also facilitate model verification when less formal approaches such as desk-checking are used (Sodja and Zupančič, 2010). Even rather small models and their simulation results can be too large to be human interpretable and understandable. Hence, physically interpretable simplified models could be used instead to provide a deeper insight into the system's behavior needed for model verification. In some cases when only a part of the model (a submodel) is under consideration, it is desired that only this submodel could be simplified.

3. SIMPLIFICATION OF MODELICA MODELS

Modelica modeling language was designed for efficient modeling of large, complex and heterogeneous physical systems (Modelica Association). Model is usually decomposed into several hierarchical levels. On the bottom of the hierarchy there are submodels of basic physical phenomena which are most commonly stated as a set of (acausal) differential-algebraic equations. It is thus most conveniently that these equations can be entered directly (e.g., without a need for any kind of manipulation or even transformation to some other description formalism). On higher hierarchical levels, the model is described graphically by schematics (i.e., *object diagrams*) and the obtained scheme efficiently reflects the topology of the system. Such model representation in Modelica is thus understandable also to domain specialists who do not have a profound knowledge about computer simulation.

Prior to simulation, model must be translated. First, model is *flattened* (i.e., hierarchical structure of a model is mapped into a set of differential, algebraic and discrete equations together with the corresponding variable declarations (Modelica Association)), then some further modifications (e.g., tearing of algebraic loops and DAE index reduction) are performed so that model can be brought into the form required by numerical solvers.

There are no tools known to the authors which support simplification of Modelica models directly. Model must be thus exported to external tools using either *Functional Mockup Interface* (Blochwitz et al., 2011) or by reparsing flattened model (if this feature is supported by modeling environment). This kind of an approach have many downsides, because model is flattened or even other translation steps are performed (algebraic-loop tearing and index reduction) before the export. Therefore much information about the organization of the model (for example, hierarchical structure) is lost and the simplified models may not be convenient for verification purposes. Furthermore, it may be very intricate and laborious to simplify only certain submodel and evaluate it together with whole model.

We believe that model simplification tool should be closely integrated into modeling environment and representation of a simplified model should be represented in the same way as original model. Models are in Modelica generally represented graphically (i.e., object diagrams) on a higher levels and textually (i.e., equations) on the lowest. Therefore, there are needed two classes of simplification techniques, simplification of object-diagrams and simplification of equation sets. In many cases only rankings of elements might be sufficient and simplification of the model could be done manually if needed. Simplification of submodels should be supported as only subset of model might be under consideration.

4. SIMPLIFICATION OF MODELICA OBJECT-DIAGRAMS

Object diagrams consist of connected components (submodels). A connection defines interactions which are determined by types of connectors (i.e., ports) which are used in components. Connector are rather loosely defined in Modelica. In general, a list of variables with some qualifications (e.g., causality, type of variable: intensive – extensive, etc.) is defined, but it can also have a hierarchical structure (Modelica Association).

4.1. Choice of ranking metric

Although different domains are modeled with rather different schemes and connections, acausal connections for modeling physical interactions are of special interest. Each (dynamic) interaction between physical systems results in an energy exchange between the systems. So it is very intuitive to choose the energy-based metrics for the simplification of a physical systems models.

Metric we chose Louca (1998) is defined by Eq. 1. It is the integral of absolute net energy flow that element exchanges with environment in time interval $[t_1, t_2]$.

$$\text{Activity} = \int_{t_1}^{t_2} \left| \sum_i p_i(t) \right| \cdot dt \quad (1)$$

In Eq. 1, $p_i(t)$ designates an energy flow through the boundary of the element (in Modelica usually modeled with *connection*).

4.2. Determination of energy-flows in object diagrams

Modelica object diagrams, when modeling physical systems, share many similarities with bond graphs, which can be efficiently used for object-oriented acausal modeling. Therefore it is possible to adapt most of bond-graph simplification techniques to Modelica's object diagrams. Of course the energy concept in bond graphs is much more unified in comparison with different Modelica libraries. So we analyzed the energy interactions between components in different Modelica libraries at the beginning.

Connectors usually contain a pair of effort and flow variables. However, their product is not necessarily an energy flow like in bond graph formalisms. This can be seen by inspecting Modelica Standard Library (Mod, 2010) where elementary connector definitions for almost all physical domains are gathered:

- Interaction between components in analog circuits (*Modelica.electric*) is determined by voltage v and current i , the latter is a flow variable, and the power of the interaction is the product of both variables: $p = v \cdot i$.
- Similar features has also the connector in *Modelica.Magnetic*, which is composed of variables for magnetic potential difference V_m and magnetic flux Φ , an effort and flow variables respectively. The power of the connection is the product of both variables: $p = V_m \cdot \Phi$.
- Connectors used for modeling of 1-D translational and rotational mechanics consist of position s and angle ϕ respectively, and force f and torque τ respectively. However the product of connector's effort and flow variables is no longer the power. For determination of the power of the connection, displacement variable has to be differentiated: $p = \frac{d}{dt} s \cdot f$ and $p = \frac{d}{dt} \phi \cdot \tau$ for translational and rotational mechanics respectively.
- In *Modelica Multibody* library, which deals with 3-D mechanics, effort and flow variables are no longer scalars, they are 6-dimensional vectors, so a state of a free-body (having 6 degree-of-freedom) can be determined. Furthermore, due to computational restrictions, implementation of connector takes into

account also a suitable selection of a frame of reference (forces, torques and orientation are expressed in local, while position is in global frame of reference). A definition of the connector is the following:

```
connector Frame
SI.Position r_0[3];
Frames.Orientation R;
flow SI.Force f[3];
flow SI.Torque t[3];
end Frame;
```

The position is determined with the variable r_0 , while the orientation R is a structure containing the transformation matrix T from global to local frame of the reference and the vector of angular velocities ω in the local frame of reference. Forces and torques are given by vectors f and t respectively. The power of the connection can be calculated by the expression: $p = \frac{d}{dt} (\mathbf{T} \cdot r_0) \cdot f + \omega \cdot t$, where again, there is a need to differentiate the position after transformation to local frame.

- Connector for modeling the heat transfer in 1-D consists of the effort variable temperature T and the flow variable for heat-flow rate Q_{flow} . The energy transfer is in this case equal to flow variable $p = Q_{flow}$.
- Library *Modelica.Fluid* deals with modeling of heat and mass transfer. The connector used in library's components which covers also mass transfer is implemented as following:

```
connector FluidPort
replaceable package Medium =
Modelica.Media.Interfaces.PartialMedium;

flow Medium.MassFlowRate m_flow;
Medium.AbsolutePressure p;
stream Medium.SpecificEnthalpy
h_outflow;
stream Medium.MassFraction
Xi_outflow[Medium.nXi];
end FluidPort;
```

Besides effort and flow variables, pressure p and mass-flow rate m_{flow} respectively, the connector includes also additional information about properties of the substance which is being exchanged in the interaction modeled by a connection of type *FluidPort*: specific enthalpy h and composition of substance (vector of mass fractions X_i if substance is a mixture). The thermodynamic state of the substance is uniquely determined by the variables of connector and all the other (thermodynamic) properties can be calculated by using functions provided by package *Medium* which is a parameter of the connector. However, thermal diffusion is not covered by this connector (it is neglected).

Energy flow associated with the connector is composed of thermal, hydraulic and chemical term and could be calculated as following (in Modelica By

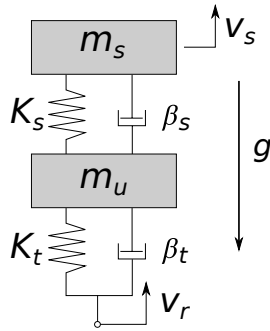


Figure 1: Scheme of a car suspension system.

Element	Activity [J]	Relative [%]	Accumulated [%]
gravityForce_s	2,270.06	37.06	37.06
spring_s	1,763.33	28.79	65.85
ground	795.02	12.98	78.82
mass_s	787.65	12.86	91.68
damper_s	198.82	3.25	94.93
spring_t	192.57	3.14	98.07
gravityForce_t	92.98	1.52	99.59
mass_t	24.53	0.40	99.99
damper_t	0.53	0.01	100.00
displacement_s	0.00	0.00	100.00
displacement_t	0.00	0.00	100.00

Table 1: Ranking of components when model from Fig. 2 is fed by input shown in Fig. 3.

Use of Chemo-bonds, 2009): $p = \dot{m} \cdot s \cdot T + \dot{m} \cdot p/\rho + \sum \mu_i \cdot \dot{N}_i$. Quantities specific entropy s , temperature T , density ρ , chemical potential μ_i and molar flow \dot{N}_i can be calculated from thermodynamical state equations provided by package *Medium*.

4.3. Ranking of elements

Although it is possible to calculate energy flow of the connector from the variables of the connector, this is not always possible to do from simulation results, because some variables can not be available. For example, the derivative of a position or an angle in the connector of the library for 1-D mechanics may not be available if this variable is not chosen as a state variable. This implies instrumentation of the model, i.e., additional auxiliary variables and equations are inserted into the model.

Ranking is done as post-processing of simulation results. *Activity* of each element required for ranking is calculated with Eq. 1. Each hierarchical level is considered separately.

After the ranking of the elements is available, model can be simplified by removing all the elements that fall below certain threshold (value of *activity* in our case). However, our current implementation provides only results of ranking in a printed form (a table). The ranking table can be then used to simplify the model manually. Automatic simplification is the matter of future investigations.

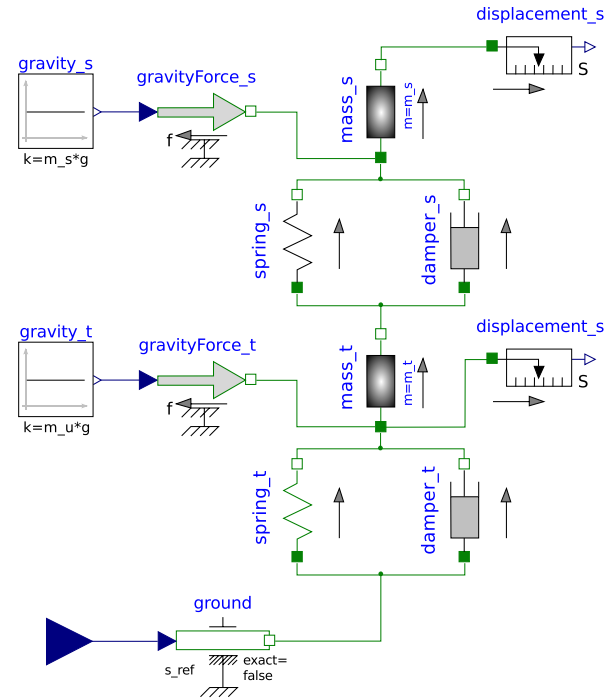


Figure 2: Car suspension system: model represented by a Modelica object diagram.

4.4. Example

The model from Fig. 2 is excited by signal depicted in Fig. 3. Components of the model are ranked with activity metrics (Fig. 1) and results are shown in table 1. The second column of table 1 consists of activities of all components calculated with Eq. 1. The third column contains relative activities components (the sum equals 100%) and the last column shows the accumulated relative activities. This column very illustratively shows how many components has to be taken into account for a reasonable accuracy.

The aim of the ranking tables is to simplify the model in Fig. 2 by removing components from the bottom of the tables 1. However, a high accumulated relative activity of the remaining components (e.g., components not removed) do not necessary imply high similarity of original and simplified model responses. It is necessary to validate the simplification comparing the original and simplified responses.

5. SIMPLIFICATION OF MODEL'S EQUATIONS

From a mathematical point of view, models in Modelica are systems of hybrid differential-algebraic equations (DAE). Therefore, in some cases we might want to investigate these equations directly. In all modeling environments supporting Modelica, models can be printed in *flat* form (i.e., as a system of equations). However, this can yield complex expression even for relatively small models, so symbolic model reduction techniques are applicable to achieve favorable representation.

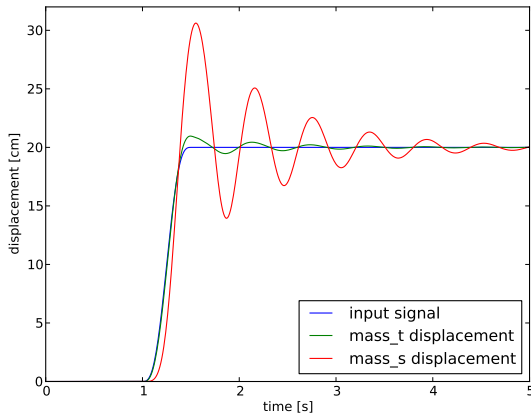


Figure 3: A car hits a smooth curb: low-frequency excitation signal is given as an input to the model in Fig.2. Also responses – displacement of an unsprung (*mass_t*) and sprung mass (*mass_s*) are depicted.

5.1. Simplification strategies

There are many mature simplification methods available for linear systems (Eitelberg, 1981; Fodor, 2002; Innis and Rexstad, 1983), while there is a lack of more general simplification methods that could be applied effectively on multi-domain models described as a set of nonlinear DAE.

Most commonly employed simplification strategies for nonlinear DAE combine model order reduction techniques (i.e., deletion of variables and variables' time derivatives or substitution of variables with constant values) and an approximation of single terms, for example, linearization, deletion or substitution of the term with constant value, etc. (Sommer et al., 2008; Wichmann et al., 1999).

The order of the applied simplification steps is determined by ranking. The most apparent ranking metric is estimation of reduced-model error for selected variables (variables of interest). One possible approach is to repeat simulation after for each possible simulation step which would yield a perfect ranking. However, this method is too time consuming to have any practical meaning.

Another option is to use energy-based ranking metric as in case of object-diagram simplification. In method suggested by (Chang et al., 2001) it is required that Lyapunov function of the system is known which is a rather harsh restriction.

The metric (Wichmann et al., 1999) suggest for simplification of nonlinear DAE systems obtained in analog circuit design estimates the error caused by simplification step (e.g., term deletion, substitution of the term with a constant value, term linearization) and it is done in two parts: the DC analysis and AC analysis. The former requires calculation of several *design points*, i.e. steady-states of the system at different inputs.

$$F(\mathbf{x}, \mathbf{0}, \mathbf{y}, \mathbf{u}, t) = 0 \quad (2)$$

A set of nonlinear algebraic equations (3) is obtained for each *design point* by solving original equation system (2) with $\dot{\mathbf{x}} = 0$ and given \mathbf{u} .

$$F(\mathbf{x}, \dot{\mathbf{x}}, \mathbf{y}, \mathbf{u}) = 0 \quad (3)$$

Values of variables at steady-state of the modified system $\tilde{F}(\mathbf{x}, \dot{\mathbf{x}}, \mathbf{y}, \mathbf{u}, t)$ (where a single term in one of equations is changed) are estimated by performing single Newton-Raphson iteration (4) and the solution of the original system is used as a guess value.

$$\begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{y}^{(1)} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^* \\ \mathbf{y}^* \end{bmatrix} - \mathbf{J}_{\tilde{F}}^{-1}(\mathbf{x}^*, \dot{\mathbf{x}}^*, \mathbf{y}^*) \cdot \tilde{F}(\mathbf{x}^*, \dot{\mathbf{x}}^*, \mathbf{y}^*) \quad (4)$$

The error estimation ε_i is then calculated by equation (5).

$$\begin{aligned} \varepsilon_i &= \| [\mathbf{x}^*, \mathbf{y}^*]^T - [\mathbf{x}^{(1)}, \mathbf{y}^{(1)}]^T \| \\ &= \| \mathbf{J}_{\tilde{F}}^{-1}(\mathbf{x}^*, \dot{\mathbf{x}}^*, \mathbf{y}^*) \cdot \tilde{F}(\mathbf{x}^*, \dot{\mathbf{x}}^*, \mathbf{y}^*) \| \end{aligned} \quad (5)$$

To further reduced computational costs, inverse Jacobian matrix is computed only once for original system at each design point and inverse Jacobians of the modified system are obtained by Sherman-Morrisson formula (6).

$$\mathbf{J}_{\tilde{F}}^{-1} = \mathbf{J}_F^{-1} - (1 + \mathbf{v}^T \mathbf{J}_F^{-1} \mathbf{e}_l)^{-1} \mathbf{J}_F^{-1} \mathbf{e}_l \mathbf{v}^T \mathbf{J}_F^{-1} \quad (6)$$

Finally, error estimations ε_i are combined in equation (7).

$$\varepsilon = \| [\varepsilon_1, \dots, \varepsilon_n]^T \| \quad (7)$$

The latter part of the metric, the AC analysis, requires linearization of the DAE system at selected design points and then transfer functions are computed. The resulting transfer functions are then simplified using methods for linear-systems simplification.

However, the simplification method of (Wichmann et al., 1999) is limited on DAE systems representing analog electrical circuits and analogue systems. The most impractical limitation preventing its use for more general multi-domain models in Modelica is that it requires calculation of *design points*, i.e. solving usually large nonlinear system of equations.

Could be this method extended to handle simplify transients of nonlinear DAE system directly (i.e., without linearization at selected *design points*)? Influence of a simplification on the equations' transient solution could be predicted by performing single Newton-Raphson iteration of equation system (2) at different time instants of the transients and then combine the obtained error estimates as suggested by the (Wichmann, 2003). However, this kind of error estimation is much more difficult as in case of purely algebraic nonlinear system (system at steady-state), because only local integration error is estimated and the elimination of low-ranked terms often results in an unstable system. (Wichmann, 2003) does not report how this problem was solved.

6. CONCLUSION

As the complexity of the models continuously increases, it is important than contemporary modeling environments include tools which help understand and cope with these models effectively. One class of those tools are also model simplification techniques.

Models which can be built in Modelica can be very heterogeneous and include submodels from different physical domains. On contrary, most model simplification techniques require strict modeling formalism and are limited on certain physical domains. They are thus not easily applicable on most models in Modelica. As it was shown in the paper, simplification methods developed for bond-graphs can be easily adapted for simplification of Modelica's object-diagrams. However, the simplification of the model on equation level is much more difficult and there are no publicly published simplification methods known to the author which could be efficiently used for a simplification of wide variety of Modelica models.

REFERENCES

- M. Andres, D. Zimmer, and F. E. Cellier. Object-oriented decomposition of tire characteristics based on semi-empirical models. In *Proceedings of the 7th Modelica Conference*, pages 9–18, Como, Italy, 2009.
- T. Blochwitz et al. The functional mockup interface for tool independent exchange of simulation models. In *Proceedings of 8th International Modelica Conference*, Dresden, Germany, 2011.
- F. Casella et al. The modelica fluid and media library for modeling of incompressible and compressible thermo-fluid pipe networks. In *Proceedings of the 5th Modelica Conference*, pages 631–640, Vienna, Austria, 2006.
- F. E. Cellier. *Continuous System Modeling*. Springer - Verlag, New York, 1991.
- Samuel Y. Chang, Christopher R. Carlson, and J. Christian Gerdes. A Lyapunov function approach to energy based model reduction. In *Proceedings of the ASME Dynamic Systems and Control Division – 2001 IMECE*, pages 363–370, New York, USA, 2001.
- E. Eitelberg. Model reduction by minimizing the weighted equation error. *International Journal of Control*, 34(6):1113–1123, 1981.
- I. K. Fodor. A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Laboratory, 2002. Technical Report UCRL-ID-148494.
- Modeling Chemical Reactions in Modelica By Use of Chemo-bonds. Cellier, f. e. and greifeneder, j. In *Proceedings of the 7th Modelica Conference*, pages 142–150, Como, Italy, 2009.
- G. Innis and E. Rexstad. Simulation model simplification techniques. *Simulation*, 41(1):7–15, 1983.
- Sanjay Lall, Petr Krysl, et al. Structure-preserving model reduction for mechanical systems. *Physica D*, 284: 304–318, 2003.
- Loucas Sotiri Louca. *An Energy-based Model Reduction Methodology for Automated Modeling*. PhD thesis, University of Michigan, 1998.
- Modelica Standard Library 3.1, User's Guide*. Modelica Association, 2010. <https://www.modelica.org/libraries/Modelica>.
- Modelica Association. *Modelica Specification, version 3.2*, 2010. <http://www.modelica.org/documents/ModelicaSpec32.pdf>.
- J. D. Murray-Smith. Simulation model quality issues in product engineering: A review. *Simulation News Europe*, 19(2):47–57, 2009.
- P. Schwarz et al. A tool-box approach to computer-aided generation of reduced-order models. In *Proceedings EUROSIM 2007*, Ljubljana, Slovenia, 2007.
- A. Sodja and B. Zupančič. Model verification and debugging of eoo models aided by model reduction techniques. In *Proceedings of the EOLT 2010*, pages 117–120, Oslo, Norway, 2010.
- Ralf Sommer, Thomas Halfmann, and Jochen Broz. Automated behavioral modeling and analytical model-order reduction by application of symbolic circuit analysis for multi-physical systems. *Simulation Modelling Practice and Theory*, 16:1024–1039, 2008.
- Hubertus Tummescheit. *Design and Implementation of Object-Oriented Model Libraries using Modelica*. PhD thesis, Lund Institute of Technology, 2002.
- T. Wichmann. Transient ranking methods for the simplification of nonlinear dae systems in analog circuit design. *Proceedings in Applied Mathematics and Mechanics*, 2:448–449, 2003.
- T. Wichmann et al. On the simplification of nonlinear dae systems in analog circuit design. In *Proceedings of CASC'99*, Munich, Germany, 1999.

THE EXCLUSIVE ENTITIES IN THE FORMALIZATION OF A DECISION PROBLEM BASED ON A DISCRETE EVENT SYSTEM BY MEANS OF PETRI NETS

Juan Ignacio Latorre-Biel^(a), Emilio Jiménez-Macías^(b)

^(a) Public University of Navarre. Department of Mechanical Engineering, Energetics and Materials.
Campus of Tudela, Spain

^(b) University of La Rioja. Industrial Engineering Technical School. Department of Electrical Engineering.
Logroño, Spain

^(a)juanignacio.latorre@unavarra.es, ^(b)emilio.jimenez@unirioja.es

ABSTRACT

The design of discrete event systems (DES) can be seen as a sequence of decisions, which allows obtaining a final product that comply with a set of specifications and operates with efficiency. A decision support system can alleviate the decision making as well as provide with more information and tools to make the best choice to the decision-maker.

The decisions related to the design of a DES may include the choice among a set of alternative structural configurations. These alternatives may be defined by the designer by mere combinations of subsystems that solve subproblems associated to the specifications and behaviour of the DES. As a consequence, it is possible that the alternative configurations share redundant information that lead to improvements in the classical approaches to solve this type of decision problems.

In this paper, the formalization of a decision problem based on a DES, underlining the characteristic feature of exclusivity between alternative configurations is presented as a tool that broadens the classical approach with new ideas and techniques to improve the efficiency in the solving of decision problems.

Keywords: decision support system, discrete event systems, Petri nets, exclusive entity.

1. INTRODUCTION

Many technological systems can be described as discrete event systems (DES) due, in a large number of cases, to the presence of digital computers in their control. Mobile phones, computer networks, manufacturing facilities or logistic systems constitute examples of DES whose presence is common in our technological society (Cassandras and Lafortune 2008). The efficiency and correctness in the operation of these systems can save important quantities of

money to companies and users. The design of discrete event systems is likely to influence in a decisive way the performance in their operation. For this reason to define adequately the design process of a discrete event system to be manufactured or constructed is a very productive activity with clear consequences in the whole life of the system (Balbo and Silva 1998). On the other hand, to consider the performance of the operation of the system is an adequate approach to afford a design process that aims to achieve a desired behaviour for the system once it is in a running stage. Nevertheless, to forecast the future behaviour of a system in process of being designed and that is not a reality yet presents several handicaps.

On the first hand, it is necessary to approximate the results, since a model of the system and not the real one should be used. On the other hand, it is needed to select the type of model to be used. In certain DES it is possible to develop physical prototypes to test certain properties of the behaviour of the systems. More commonly, formal models are developed to apply algorithmic methodologies to analyze the behaviour of the original system. Sometimes it is possible and productive to combine the construction of several models of different nature in the design process of a system: physical systems that model specific characteristics of the real system can be combined with formal models developed on a computer to forecast the behaviour of the real system.

The formal models developed to forecast the performance of a discrete event system in process of being designed are usually complemented with simulation in order to evaluate the behaviour of the model (Piera *et al.* 2004). In fact, simulation allows exploring the region of the state space of the system under specific conditions. On the other hand, it is

common that modelling and simulation present a rate between the information obtained from the system and cost associated to the modelling process, which is more favourable than other techniques that imply the physical implementation of the model by means of prototyping.

Furthermore, the design process of a DES requires stating and solving several decision problems. In particular, it is usual that it is necessary to choose among a set of alternative configurations or structures for the system (Latorre *et al.* 2009). This is the case, for example, when different layouts for the material or products conveying can be chosen to define a final configuration for a chain supply in process of being designed. A classical approach to this type of problems is intensive in computer resources, when the solution is searched by means of formal algorithms. This fact is due to the analysis by simulation of every one of the alternative configurations, knowing that this analysis requires launching not only one but sometimes an undetermined number of simulations.

An analysis of the design process described in the previous paragraph, characterizes the discrete event system to be designed among a set of alternative structures with the property of mutual exclusive evolution between the structures. This idea allows developing methodologies that reduces the computational cost of performing simulations to the different alternative configurations by the removal of the redundant information present in the models of every one of them (Latorre *et al.* 2010b).

This concept of exclusive entity is abstracted in a more general idea defined by the exclusive entities, associated to an undefined model. On the other hand, this new idea can be particularized in a variety of formalisms to be able to represent the model of the system in a compact way able to develop fast sets of simulations to support the decision making process in the design of discrete event systems.

The general approach given by a set of exclusive entities to the exclusiveness associated to the different alternative structural configurations for a discrete event system to be designed, constitutes a characteristic feature of a model defined as a disjunctive constraint in the formalization of a decision problem based on a DES. A model of this kind can be called undefined Petri net since it contains certain parameters whose values should be chosen among a domain set as result of decisions.

The model of the system is only a part of the formalization process of a decision problem stated on a discrete event system. There are other elements that can be included in the resulting formal problem as the type of solution expected for the problem, the solution space and, depending on the decision problem, the objective function that evaluates the cost or the performance of the DES after the selection of a certain solution from the solution space.

In this paper, an overview on the statement and formalization process of a decision problem based on a discrete event system is given underlying the exclusiveness feature in the different alternative structural configurations for the DES that can be particularized by different formalisms. On the other hand, this exclusiveness can be abstracted into the concept of set of exclusive entities that leads to an interesting property that will be defined in this paper.

2. DEFINITIONS

A discrete event system can be defined in the following way:

Definition 1. Discrete event system.

Dynamic system whose behaviour can be described by discrete state variables and is governed by asynchronous and instantaneous incidences, called events, which are solely responsible for the state changes.

A discrete event system may be defined in a more or less ambiguous way by a set of specifications and some constraints and expectations in its dynamic behaviour. The ambiguity in the definition of a DES can be interpreted as freedom degrees, some of which should be particularized in the design process, while others can be specified in the operation processes. The mentioned freedom degrees may be called undefined characteristics of the discrete event system.

The previous paragraph allows to define a particular type of DES.

Definition 2. Undefined DES.

A discrete event system with at least one undefined characteristic is said to be an undefined DES.

The type of decisions stated in the design process of an undefined DES try to reduce the

ambiguity in the description of the discrete event system, by the transformation of undefined characteristics in defined ones. As a consequence, it is possible to state a decision problem in the way described in the following.

Definition 3. Decision problem based on a DES.

Let D be an undefined discrete event system. A **decision problem based on D** is a choice, among several alternatives, in response to a question posed on any set of the undefined characteristics of D or its evolution.

Once a decision problem has been stated, it is necessary to solve it. For this purpose, it is convenient to represent it in a formal language.

A formal language shows important advantages from a natural language to state a decision problem. On the one hand, it provides with precision to the description of the problem, removing ambiguity and allowing the application of an algorithmic solving methodology. On the other hand, the consequence of the successful application of a solving methodology to a decision problem expressed in a formal language is one or several quantitative results, which can easily be compared with numerical references or the results of other methodologies.

A first element that is convenient to include in the formal statement of the decision problem is the discrete event system itself. There are a number of formal languages that can cope with the modelling of a generic discrete event system. However, the decision of the formal language to be considered in this paper, the Petri nets (Petri 1962), is based in the versatility and double representation that may be matrix-based or graphical. On the other hand complex behaviours of collaboration and competence may be modelled in an easy and natural way (David and Alla 2005), (Jiménez *et al.* 2005).

In particular, it is possible to define an autonomous unmarked Petri net in the following way (Cassandras and Lafortune 2008) and (Silva 1993), where an introduction to the Petri net paradigm can be found as well as in (Peterson 1981) or (David and Alla 2005).

Definition 4. Petri net graph

A (generalized) *Petri net graph* (or *Petri net structure*) is a weighted bipartite graph

$$N = \langle P, T, F, w \rangle$$

where

$P = \{p_1, p_2, \dots, p_n\}$ is the finite, non-empty, set of places (one type of node in the graph).

$T = \{t_1, t_2, \dots, t_m\}$ is the finite, non-empty, set of transitions (the other type of node in the graph)

$F \subseteq (P \times T) \cup (T \times P)$ is the set of directed arcs (from places to transitions and from transitions to places) in the graph, called flow relation.

$w : F \rightarrow \mathbb{N}^*$ is the weight function on the arcs.

In this paper, a new approach in the definition of a Petri net will be defined.

Definition 5. Unmarked Petri net.

A (generalized) *unmarked Petri net* (or *Petri net structure*) is a triple

$$N = \langle n_p, S_\gamma, S_{val_\gamma} \rangle$$

where

$n_p \in \mathbb{N}^*$ is the number of places.

$S_\gamma = \{ \gamma_1, \gamma_2, \dots, \gamma_n \}$ is a set of structural parameters.

$S_{val_\gamma} = \{ cv_1, cv_2, \dots, cv_m \}$ is the set of feasible combinations of values for the parameters of S_γ .

It is verified that $n = k \cdot n_p$, where $k \in \mathbb{N}^*$ and $\forall cv_i \in S_{val_\gamma}, cv_i = (v_1, v_2, \dots, v_n)$.

This new definition of unmarked Petri net allows constructing the incidence matrices of the formalism and underlines the approach of this paper, focussed on the formalization process of a decision problem based on a discrete event system.

Reducing the concept of Petri net to a collection of parameters and their feasible values, the formalization process from a discrete event system can be considered as the translation of a subset of characteristics of the DES into a set of parameters of the Petri net. The characteristics of the DES translated and included in the Petri net will depend on the degree of detail of the model.

Subsequently, an undefined characteristic of the discrete event system will be modelled by means of one or several undefined parameters in the Petri net.

The process of obtaining a formal model from an original system, the modelling process, can be interpreted by means of the translation of the undefined characteristics of the DES into a set of undefined parameters. As a consequence, an

undefined parameter can be defined as indicated below.

Definition 6. Undefined parameter

Any numerical variable of a Petri net model or its evolution that has not a known value but it has to be assigned as a consequence of a decision from a set of at least two different feasible values. The value assigned to the undefined parameter must be unique.

A parameter of a Petri net may belong to the set of structural parameters; nevertheless, it is possible to define other types of parameters according to the role they play in the model. For example, there is a category of marking parameters that includes the initial marking of all the places of the Petri net. As a consequence, an autonomous marked Petri net can be defined in the following way.

Definition 7. Marked Petri net.

A (generalized) *marked Petri net* (or *Petri net system*) is a triple

$$N = \langle n_p, S_\gamma, S_{val\gamma} \rangle$$

where

$n_p \in \mathbb{N}^*$ is the number of places.

$S_\gamma = \{ \gamma_1, \gamma_2, \dots, \gamma_n \}$ is a set of structural parameters.

$S_{val\gamma} = \{ cv_1, cv_2, \dots, cv_m \}$ is the set of feasible combinations of values for the parameters of S_γ .

It is verified that $n = (k+1) \cdot n_p$, where $k \in \mathbb{N}^*$ and $\forall cv_i \in S_{val\gamma}, cv_i = (v_1, v_2, \dots, v_n)$.

It is possible to notice from the comparison of the **definition 5** and the **definition 7** that the addition of the n_p marking parameters have modified the definition of the unmarked Petri net by increasing the size of the set S_γ and hence the number of values in the feasible combinations of values belonging to S_γ .

Furthermore, it is easy to deduce that this parametric definition of a Petri net allows easily to be extended to Petri nets with extended features as interpreted Petri nets, including timed Petri nets, and other nets that include priorities, colours, etc.

As it has already been explained previously, the design process of a discrete event system is usually associated to several alternative structural configurations for the DES. A classical approach for the modelling of such a

system is associated to so many different Petri nets as alternative structural configurations for the DES can be found. These Petri nets can be called alternative Petri nets and belonging to the same model for the original DES should comply with a property of exclusiveness (Latorre *et al.* 2011). Of course it is not possible that several of them can be chosen as solution for the DES design process. The only option for the model to be coherent with the reality of the decision problem is to comply with a property of exclusiveness. This property can be imposed by means of the concept of mutually exclusive evolution defined below.

Definition 8. Mutually exclusive evolution

Given two Petri nets R and R' . They are said to have mutually exclusive evolutions if it is verified:

- i) If $\mathbf{m}(R) \neq \mathbf{m}_0(R) \Rightarrow \mathbf{m}(R') = \mathbf{m}_0(R')$
- ii) If $\mathbf{m}(R') \neq \mathbf{m}_0(R') \Rightarrow \mathbf{m}(R) = \mathbf{m}_0(R)$

As a consequence, a set of alternative Petri nets can be described as:

Definition 9. Set of alternative Petri nets.

Given a set of Petri nets $S_R = \{ R_1, \dots, R_n \}$, S_R is said to be a set of alternative Petri nets if $n > 1$ and $\forall i, j$ such that $1 \leq i, j \leq n$, R_i and R_j verify:

- i) R and R' have mutually exclusive evolution.
- ii) $\mathbf{W}(R) \neq \mathbf{W}(R')$.

R_i is called the i -th alternative Petri net of S_R .

This classical approach of modelling an undefined DES with alternative structural configurations by means of a set of alternative Petri nets is not the only option. Even it is not necessarily the most efficient option (Latorre *et al.* 2009) for posing and solving a formal statement of the decision problem based on the DES.

In this search for new formalisms, it is interesting to abstract the representation of the undefined DES performed with the set of alternative Petri nets. On the first hand, it is possible to obtain a general abstraction for the mutually exclusive evolution of the alternative Petri nets by means of the concept of the exclusive entities. The alternative Petri nets can be considered exclusive entities since only one of them can be chosen at a time. In fact, a set of exclusive entities can be defined in the following way:

Definition 10. Set of exclusive entities.

Given a discrete event system, a set of exclusive entities associated to it is a set $S_x = \{ X_1, \dots, X_n \}$, which verifies that

i) The elements of S_x are exclusive, that is to say, only one of them can be chosen as a consequence of a decision.

ii) $\forall i, j \in \mathbb{N}^*, 1 \leq i, j \leq n$ it is verified that $X_i \neq X_j$.

iii) $\exists f: S_x \rightarrow S_R$, where

$S_R = \{ R_1, \dots, R_n \}$ is a set of alternative Petri nets, feasible models of D .

f is a bijection $\Rightarrow \forall X_i \in S_x \exists! f(X_i) = R_i \in S_R$ such that R_i is a feasible model for D and $\forall R_i \in S_R \exists! f^{-1}(R_i) = X_i \in S_x$.

Definition 11. Undefined Petri net.

An undefined Petri net is a 4-tuple

$$N = \langle n_p, S_\gamma, S_{val\gamma}, S_x \rangle$$

where

$n_p \in \mathbb{N}^*$ is the number of places.

$S_\gamma = \{ \gamma_1, \gamma_2, \dots, \gamma_n \}$ is a set of structural parameters.

$S_{val\gamma} = \{ cv_1, cv_2, \dots, cv_m \}$ is the set of feasible combinations of values for the parameters of S_γ .

$S_x = \{ X_1, X_2, \dots, X_q \}$, where $q > 1$, is a set of exclusive entities.

It is verified that $n = (k+1) \cdot n_p$, where $k \in \mathbb{N}^*$ and $\forall cv_i \in S_{val\gamma}, cv_i = (v_1, v_2, \dots, v_n)$.

In fact, the set of exclusive entities S_x does not provide with more structural or marking parameters to the model. It simply organizes or classifies the parameters into exclusive subsets. The specific representation of this undefined Petri net can be made according to different formalisms that should include a set of exclusive entities.

3. PROPERTIES AND APPLICATIONS

Several properties can be stated in relation with the idea of an undefined Petri net:

Proposition 1. The feasible combinations of values for the undefined structural parameters of a compound Petri net is a set of exclusive entities.

Proposition 2. A set of choice variables is a set of exclusive entities.

Proposition 3. A natural choice colour is a set of exclusive entities.

Proposition 4. A set of alternative Petri nets is a set of exclusive entities.

All these valid representations of a set of exclusive entities lead to different formalisms able to model a discrete event system with alternative structural configurations. For more details on the elements that appear in the statements of the propositions see (Latorre *et al.* 2010a) and (Latorre *et al.* 2010c).

An additional property should be guaranteed for any representation of a set of exclusive entities.

Theorem. Given an undefined Petri net R^U associated to a set of exclusive entities S_x , any representation of the set of exclusive entities S_z verifies that

$$\text{card}(S_x) = \text{card}(S_z)$$

This last property implies that no matter which representation is chosen for the set of exclusive entities of an undefined Petri net, the cardinality of its representation is the same than its abstraction S_x . In other words, the number of alternative structural configurations of the original discrete event system is constant.

The applications of the concept of set of exclusive entities can be found in the decision field that is associated to the discrete event systems.

Each exclusive entity can be related to a feasible configuration of the original discrete event system. The set of all these configurations determine the complete set of possible choices to define univocally the controllable parameters of the associated Petri net model.

One interesting application of the concept of set of exclusive entities consists of restricting the association to the structural configurations of the original DES to the exclusive entities. In this case it is possible to develop a methodology to choose among different structures to design, modify or control certain systems modelled by Petri nets. Every exclusive entity may be associated to several undefined or controllable parameters that lead to diverse behaviours,

however, the exclusive entities are reserved, in this approach, to the structural parameters. In this methodology, it is possible to state the following theorem:

The search for an efficient representation of an undefined Petri net, can lead to formalisms that profit from the search methodology, from the similarities between the different structural configurations of the DES, from a single solution space, etc, enhancing the performance of the optimization algorithms aimed to take the best decisions.

Some formalisms that can be mentioned are the sets of alternative Petri nets, the compound Petri nets, the alternatives aggregation Petri nets and the disjunctive coloured Petri nets. Their suitability for representing an undefined Petri net can be deduced from the propositions 1 to 4.

4. CONCLUSIONS AND FUTURE RESEARCH

In this paper, a decision problem based on a discrete event system is analysed. Moreover, some important topics in the formalization process of this problem are considered. In particular, it has been underlined a relevant type of decision problem stated in the design process of a discrete event system. The new approach of considering a Petri net from a parametric point of view leads to an abstraction of a model of a discrete event system with alternative structural parameters. Some properties allow relating the set of exclusive entities with different representations that comply with the invariance of their number of elements. This property is related with the fact that the number of alternative structural configurations for the DES being designed is constant.

The topic presented and summarized in this paper is an important part in the theory that affords the solution of the decision problems based on DES with different alternative structural configurations by means of the removal of redundant information and obtaining compact formalisms that behave efficiently in the algorithms to solve the associated problems.

Open research lines so far are the search for new formalisms to represent undefined Petri nets, as well as to develop criteria and algorithms to choose the best formalism to solve a given decision problem.

REFERENCES

- Balbo, G. y Silva, M. (eds.), "Performance Models for Discrete Event Systems with Synchronizations: Formalisms and Analysis Techniques", Editorial Kronos, Zaragoza, Spain, 1998
- Cassandras, Christos G., Lafortune, S., "Introduction to Discrete Event Systems". Second Edition, Springer, 2008
- David, R., Alla. H., Discrete, Continuous and Hybrid Petri nets, Springer, 2005
- Jiménez, E., Pérez, M., Latorre, J.I., "On deterministic modelling and simulation of manufacturing systems with Petri nets". Proceedings of European Modelling Simulation Symposium. Marseille, pp. 129-136. Marseille. 2005
- Latorre, J.I., Jiménez, E., Pérez, M., Martínez, E., "The design of a manufacturing facility. An efficient approach based on alternatives aggregation Petri," Proceedings of the 21st European Modelling and Simulation Symposium (EMSS 09). Puerto de la Cruz, Spain, vol. 2, pp. 33-39, September 2009.
- Latorre, J.I., Jiménez, E., Pérez, M., "The alternatives aggregation Petri nets as a formalism to design discrete event systems." International Journal of Simulation and Process Modeling, Special Issue. 2010
- Latorre, J.I., Jiménez, E., Pérez, M., "On the Solution of Optimization Problems with Disjunctive Constraints Based on Petri Nets" Proceedings of the 22nd European Modelling and Simulation Symposium (EMSS 10). Fez, Morocco, pp. 259-264, October 2010.
- Latorre, J.I., Jiménez, E., Pérez, M., "Coloured Petri Nets as a Formalism to Represent Alternative Models for a Discrete Event System". Proceedings of the 22nd European Modelling and Simulation Symposium (EMSS 10). Fez, Morocco, pp. 247-252, October 2010.
- Latorre, J.I., Jiménez, E., Pérez, M., "Efficient Representations Of Alternative Models Of Discrete Event Systems Based On Petri Nets". Proceedings of the UKSim 13th International Conference on Computer

Modelling and Simulation. Cambridge, United Kingdom, March 2011.

Peterson, J.L. "Petri Net Theory and the Modelling of Systems", Prentice Hall, Englewood Cliffs, 1981.

Petri, Carl A. (1962). "Kommunikation mit Automaten". Ph. D. Thesis. University of Bonn (German).

Piera, M.À., Narciso, M., Guasch, A., Riera, D., "Optimization of logistic and manufacturing system through simulation: A colored Petri net-based methodology," Simulation, vol. 80, number 3, pp 121-129, May 2004

Silva, M. "Introducing Petri nets", In Practice of Petri Nets in Manufacturing", Di Cesare, F., (eds.), pp. 1-62. Ed. Chapman&Hall. 1993

MATRIX-BASED OPERATIONS AND EQUIVALENTE CLASSES IN ALTERNATIVE PETRI NETS.

Juan Ignacio Latorre-Biel^(a), Emilio Jiménez-Macías^(b)

^(a) Public University of Navarre. Department of Mechanical Engineering, Energetics and Materials. Campus of Tudela, Spain

^(b) University of La Rioja. Industrial Engineering Technical School. Department of Electrical Engineering. Logroño, Spain

^(a) juanignacio.latorre@unavarra.es, ^(b) emilio.jimenez@unirioja.es

ABSTRACT

Petri net is a modelling paradigm used for discrete event systems (David and Alla 2005), (Cassandras *et al.* 2008). Their transformations are a powerful tool for the validation and verification of Petri nets as models of discrete event systems. If a simplified Petri net verifies a certain set of properties, the original Petri net will verify them if the applied net transformation preserves these properties. On the other hand, the control of Petri nets may also improve with the application of Petri net transformations. If the unobservable transitions are removed by the application of transformation rules, then the state of the simplified net evolves under the firing of the observable transitions.

In this paper, the transformation of nets will be applied to modify the formalism that represents a disjunctive constraint of a decision problem stated on a Petri net model. Some formalisms may be more suitable for the modelling process than others (alternative Petri nets), while others are more compact and suitable for the development of optimization processes in an efficient way (compound Petri nets, alternatives aggregation Petri nets or disjunctive coloured Petri nets).

A set of transformation rules that preserve the structure of the associated graph or reachable markings are provided in this paper, as well as an example of application in a net transformation between two formalisms. As a consequence of the application of these rules, both the original and resulting formalisms will show an equivalent behaviour. Hence, any of them can be used to state a decision problem but the efficiency of the algorithm to solve the problem may be different when considering the required computer resources and the quality of the obtained solution.

Keywords: equivalence operations, Petri net transformations, decision support system, compound Petri nets, alternative Petri nets.

1. INTRODUCTION

One of the stages that can be considered in the modeling process of a discrete event system is the validation and verification (Peterson 1981), (Jimenez *et al.* 2005). According to (Silva, 1993), it is possible to reduce the cost and the duration of the design process of a SED by checking if certain properties are verified by the model. One of the techniques of qualitative analysis is based in the transformation of the Petri net structure. Some important issues of these techniques are described in (Berthelot 1987), (Silva 1993) and (Haddad and Pradat-Peyre 2006).

These early developments in the theory of the Petri nets have led to other transformation techniques in the static structure and to new formalisms based on PN that are aimed to simplify the modeling of DES whose structure varies with time. For example (Van der Aalst 1997) provides with eight transformation rules, based in the previously mentioned techniques, applied to systems that experience the frequent changes in their structure.

An undefined Petri net can be interpreted as a model of a discrete event system that includes freedom degrees in its structural characteristics (Latorre *et al.*, 2009b). The undefined Petri net is an abstraction that can be particularized in a specific formalism. A classical approach to obtain a model of an undefined discrete event system is a set of alternative Petri nets (Latorre *et al.*, 2007). This type of nets verifies the property of mutually exclusive evolutions, hence in the same Petri net alternative structural configurations of an original discrete event system can be included (Latorre *et al.* 2011).

In this paper the concept of equivalence class is applied to every alternative Petri net. This concept allows substituting any Petri net belonging to a set of alternative Petri nets by another one whose behaviour and properties are the same than the original one. Hence, the resulting set of alternative Petri nets will verify the same properties and show equivalent behaviour. Every equivalence class will be composed by Petri nets with the same behaviour and all of them will be said to be equivalent.

Given an alternative Petri net, the methodology to obtain equivalent Petri nets to the first one will be based in matrix-based operations, applied to the incidence matrix of the net. These matrix-based operations will lead to new Petri nets but the graphs of reachable markings will be isomorphous in the original and the resulting net. This fact ensures that the properties and the behaviour of both Petri nets is the same and, hence, that they can be considered as equivalent ones.

The matrix based operations can be applied with the aim of transforming the set of alternative Petri nets into a compound Petri net (Latorre *et al.* 2010). This process requires the merging of the sets of parameters of all the nets belonging to the set of alternative Petri nets. In particular, it is necessary to merge the structural parameters that are the elements of the incidence matrices. In order to obtain a compound Petri net with the smallest set of undefined structural parameters and hence to obtain a compact model that requires a reduced computation resources to simulate the evolution of the original DES, it is convenient to apply the matrix-based operations to the set of alternative Petri nets.

These operations might lead to alternative Petri nets whose incidence matrices have more similarities in the same positions (common elements). This fact imply that when the element of all the incidence matrices in a given position is the same, there is not an associated undefined parameter and when most of the elements are the same in a certain position, then the set of feasible values for the undefined structural parameter that appears is reduced.

2. TRANSFORMATION OPERATIONS

Once the objectives of the transformations are clear, the matrix-based operations will be presented and some examples will be given.

Definition 1. Operation of swapping two rows of a matrix.

The operation of swapping two rows of a matrix is defined as the following function:

$$\text{swapr: } \mathbf{M}_{m \times n} \times \{1, 2, \dots, m\} \times \{1, 2, \dots, m\} \rightarrow \mathbf{M}_{m \times n} \\ (\mathbf{A}, i, j) \rightarrow \mathbf{B} \in \mathbf{M}_{m \times n}$$

$$\text{where, } \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots \\ a_{i1} & a_{i2} & \dots & a_{in} \\ \dots & \dots & \dots & \dots \\ a_{j1} & a_{j2} & \dots & a_{jn} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \in \mathbf{M}_{m \times n},$$

$$\mathbf{B} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots \\ a_{j1} & a_{j2} & \dots & a_{jn} \\ \dots & \dots & \dots & \dots \\ a_{i1} & a_{i2} & \dots & a_{in} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \in \mathbf{M}_{m \times n}.$$

In other words, **definition 1**, describes the swapping of the i th and j th rows in a matrix \mathbf{A} . This operation is denoted by $\text{swapr}(\mathbf{A}, i, j)$

Remark 1. When applying this operation to the incidence matrix of a Petri net it has to be taken into consideration that the i th row represents the input and output arcs of a place of the PN. Let us call this place p_i . For this reason, swapping two rows, the i th and the j th, implies that the arcs associated to p_i are no longer present in the i th row of the incidence matrix but in the j th. As a natural consequence, if this new incidence matrix is to be included in the characteristic equation of the Petri net it has to be considered that the i th element of \mathbf{m} , the marking of the Petri net, does not represent $\mathbf{m}(p_i)$ the marking of the place p_i anymore. The same considerations can be made for p_j , the j th row of the incidence matrix and $\mathbf{m}(p_j)$. Therefore, the $\text{swapr}(\mathbf{A}, i, j)$ operation implies the swapping of the i th and j th elements in the marking vector \mathbf{m} of the Petri net. This statement is also true for the particular case of \mathbf{m}_0 . It is clear then that it is necessary to apply the same swapping operation that is applied to the incidence matrix, to the marking of the Petri net. In a subsequent section, it will be mentioned the reference name and the alias of any place of a Petri net. With these concepts it will be generalized the previous considerations on the swapping of

rows of an incidence matrix and the application of the same operation in the marking of the associated Petri net.

Definition 2 . Operation of swapping two columns of a matrix.

The operation of swapping two columns of a matrix is defined as the following function:

$$\text{swapc: } \mathbf{M}_{m \times n} \times \{1, 2, \dots, n\} \times \{1, 2, \dots, n\} \rightarrow \mathbf{M}_{m \times n}$$

$$(\mathbf{A}, i, j) \rightarrow \mathbf{B} \in \mathbf{M}_{m \times n}$$

where,

$$\mathbf{A} = \begin{pmatrix} a_{11} & \dots & a_{1i} & \dots & a_{1j} & \dots & a_{1n} \\ a_{21} & \dots & a_{2i} & \dots & a_{2j} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{m1} & \dots & a_{mi} & \dots & a_{mj} & \dots & a_{mn} \end{pmatrix} \in \mathbf{M}_{m \times n},$$

and

$$\mathbf{B} = \begin{pmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1i} & \dots & a_{1n} \\ a_{21} & \dots & a_{2j} & \dots & a_{2i} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{m1} & \dots & a_{mj} & \dots & a_{mi} & \dots & a_{mn} \end{pmatrix} \in \mathbf{M}_{m \times n}$$

In other words, **definition 2**, describes the swapping of the columns i and j in matrix \mathbf{A} , which is denoted by $\text{swapc}(\mathbf{A}, i, j)$

Remark 2. The state equation of a Petri net requires representing the characteristic vector that summarizes the information contained in the sequence of transitions fired. The characteristic vector (also called firing count vector) contains elements that are different to zero in the positions that correspond to the transitions fired. If an operation swapc is applied to an incidence matrix and the state equation is represented, the characteristic vector should be modified according to this same swapc operation.

Definition 3. Operation of adding a row of zeros to a matrix.

The operation of adding a row of zeros to a matrix is defined as the following function:

$$\text{addr: } \mathbf{M}_{m \times n} \rightarrow \mathbf{M}_{(m+1) \times n}$$

$$\mathbf{A} \rightarrow \mathbf{B}, \text{ such that}$$

$$\text{Given } \mathbf{A} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} \in \mathbf{M}_{m \times n} \Rightarrow$$

$$\text{addr}(\mathbf{A}) = \mathbf{B} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{m1} & \dots & a_{mn} \\ 0 & \dots & 0 \end{pmatrix} \in \mathbf{M}_{(m+1) \times n}$$

The operation described in the **definition 3** is denoted by $\text{addr}(\mathbf{A})$ and adds a row of zeros to the matrix \mathbf{A} .

Remark 3. The operation addr applied to the incidence matrix of a Petri net implies the addition of a new place with a particular property: every input and output arc has weight zero. In other words, this new place is an isolated node of the Petri net.

The marking of the Petri net that results from the application of this operation should include the marking of the new place, which will occupy the last position of the vector. However, being isolated, the place cannot experience any variation of its initial marking in the evolution of the Petri net. Furthermore, the marking of other places will not be influenced by the added place, hence the marking of the new Petri net, excluding the added place, will be the same to the original one. If the new place is considered in this comparison it is possible to say that the significant marking (the one that varies in at least an evolution of the PN) is the same in both Petri nets, hence the graphs of reachable markings are isomorphic.

Definition 4. Operation of removing a row of zeros of a matrix.

The operation of removing a row of zeros of a matrix is defined as the following function:

$$\text{removr: } S \rightarrow \mathbf{M}_{(m-1) \times n}$$

$$\mathbf{A} \rightarrow \mathbf{B}, \text{ such that}$$

$S = \{\mathbf{A} \in \mathbf{M}_{m \times n} \mid a_{m*} = (0 \ 0 \ \dots \ 0)\}$, in other words, S is the set of matrices whose m th (last) row is a row of zeros.

$$\text{Given } \mathbf{A} = \begin{pmatrix} a_{1,1} & \dots & a_{1,n} \\ \dots & \dots & \dots \\ a_{m-1,1} & \dots & a_{m-1,n} \\ 0 & \dots & 0 \end{pmatrix} \in \mathbf{M}_{m \times n} \Rightarrow$$

$$\text{removr}(\mathbf{A}) = \mathbf{B}$$

$$\mathbf{B} = \begin{pmatrix} a_{1,1} & \dots & a_{1,n} \\ \dots & \dots & \dots \\ a_{m-1,1} & \dots & a_{m-1,n} \end{pmatrix} \in \mathbf{M}_{(m-1) \times n}$$

The operation described in the **definition 4** is denoted by $\text{removr}(\mathbf{A})$ and removes the last row of a matrix \mathbf{A} , which should contain only zeros.

Remark 4. The operation *removr* applied to the incidence matrix of a Petri net implies the removal of a place with a particular property: every input and output arc has weight zero. In other words, this new place is an isolated node of the Petri net. Moreover, the place should be associated to the last row of the incidence matrix (if this last condition is not verified it is always possible to apply an operation *swapr* to guarantee this fact).

The marking of the Petri net that results from the application of this operation should not include the marking of the removed place (which occupied the last position of the vector before the operation). However, being isolated, the place could not experience any variation of its initial marking in the evolution of the Petri net. Furthermore, the marking of other places will not be influenced by the removed place, hence the marking of the new Petri net, will be the same to the original one (excluding the added place). If the removed place is included in this comparison it is possible to say, as it was mentioned in the **remark 4**, that the reachable significant markings are the same in both Petri nets, hence the graphs of reachable markings (including the non-significant markings) are isomorphous.

The swapping of rows and columns of the incidence matrices simply locates in a different place of the matrices the information (weights) related to the arcs that link a certain place with the transitions of the PN or a certain transition with the places of the Petri net. Furthermore, if the parameters associated to a place or a transition that changes its position in the incidence matrices do not remain attached to the position in the matrix but move with the place or transition, the behaviour of the net, and its structure, will be the same. For this reason, to apply such a transformation as the swapping of rows and columns of the incidence matrices, it is necessary to ensure that the appropriate amount of parameters are associated to the moving places and transitions. In order to facilitate this operation it is convenient to define a reference name for every place and transition, to which its parameters will be also referred. This reference name will be attached to the information (weights of arcs) of the rows and columns of the incidence matrices (that can change its position). On the other hand, it is also convenient to define an alias for every place and transition. The alias will be attached to the position in the incidence matrices, in the way that the first row of the incidence matrix will always be associated to the alias p_1 , the second

to the alias p_2 and so on. The same may happen with the transitions: the first column will always be associated to t_1 , the second with t_2 and so on.

3. APPLICATION OF THE TRANSFORMATIONS

Example 1.

Let $\mathbf{A} \in \mathbf{M}_{m \times n}$ be the incidence matrix of a Petri net R .

The names of places and transitions of the Petri net can be shown in the following representation of \mathbf{A} :

$$\mathbf{A} = \begin{matrix} & \begin{matrix} t_1^r & \dots & t_u^r & \dots & t_v^r & \dots & t_n^r \end{matrix} \\ \begin{pmatrix} a_{11} & \dots & a_{1u} & \dots & a_{1v} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{i1} & \dots & a_{iu} & \dots & a_{iv} & \dots & a_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{j1} & \dots & a_{ju} & \dots & a_{jv} & \dots & a_{jn} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{m1} & \dots & a_{mu} & \dots & a_{mv} & \dots & a_{mn} \end{pmatrix} & \begin{matrix} p_1^r \\ \dots \\ p_i^r \\ \dots \\ p_j^r \\ \dots \\ p_m^r \end{matrix} \end{matrix}$$

Let us now apply two operations to the incidence matrix \mathbf{A} :

$$\mathbf{B} = \text{swapr}(\mathbf{A}, i, j); \mathbf{C} = \text{swapc}(\mathbf{B}, u, v)$$

In other words: $\mathbf{C} = \text{swapc}(\text{swapr}(\mathbf{A}, i, j), u, v)$

The resulting incidence matrix, with the new alias for the places and transitions is presented below:

$$\mathbf{C} = \begin{matrix} & \begin{matrix} t_1^r & \dots & t_v^r & \dots & t_u^r & \dots & t_n^r \end{matrix} \\ \begin{pmatrix} t_1 & \dots & t_u & \dots & t_v & \dots & t_n \\ a_{11} & \dots & a_{1v} & \dots & a_{1u} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{j1} & \dots & a_{jv} & \dots & a_{ju} & \dots & a_{jn} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{i1} & \dots & a_{iv} & \dots & a_{iu} & \dots & a_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{m1} & \dots & a_{mv} & \dots & a_{mu} & \dots & a_{mn} \end{pmatrix} & \begin{matrix} p_1 & p_1^r \\ \dots & \dots \\ p_i & p_j^r \\ \dots & \dots \\ p_j & p_i^r \\ \dots & \dots \\ p_m & p_m^r \end{matrix} \end{matrix}$$

In this representation of the incidence matrix \mathbf{C} , it can be seen that the reference names of the swapped rows and columns have changed the position according to these swaps. This change is a consequence of the fact that the reference names are related to structural and marking parameters (among others) such as the elements of the incidence matrices and not to positions in the matrix. If the Petri net is the model of a real

system, the reference names are likely to be associated to a physical meaning as well.

However, in certain applications it is very useful to define an alias for every place and transition. This alias is associated to the position that the elements of the incidence matrix occupy. The aliases do not bear the superindex “r”. For example, in the matrix $C = \text{swapr}(A, i, j)$, the alias of the place whose input and output arcs are stated in the j th row is p_j , whereas its reference name is p_i^r .

Example 2.

Let us consider the simple alternative Petri nets presented in the figure 1 and their incidence matrices shown in the figure 2. Some equivalence operations will be applied to transform the simple alternative Petri nets into matching ones able to be merged. The result of this merging is obtaining an equivalent compound Petri net with the smallest size of the set of undefined structural parameters and the smallest size of the feasible combination of values of these parameters. In this example it is not intended to obtain the optimal compound Petri net, just to illustrate the application of some equivalence operations.

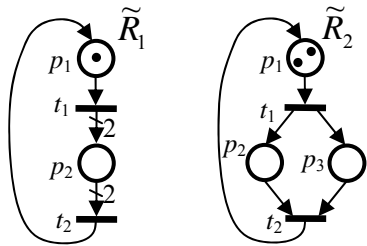


Fig. 1. Simple alternative Petri nets.

$$W(\tilde{R}_1) = \begin{pmatrix} t_1 & t_2 \\ -1 & 1 \\ 2 & -2 \end{pmatrix} \begin{matrix} p_1 \\ p_2 \end{matrix}$$

$$W(\tilde{R}_2) = \begin{pmatrix} t_1 & t_2 \\ -1 & 1 \\ 1 & -1 \\ 1 & -1 \end{pmatrix} \begin{matrix} p_1 \\ p_2 \\ p_3 \end{matrix}$$

Fig. 2. Incidence matrices of the simple alternative Petri nets.

The first equivalence operation consists of increasing the size of the incidence matrix to reach the dimensions 4×3 . This process requires the addition of isolated places and transitions as it can be seen in the figure 3.

The operations that have been applied are the following:

$$W(R_1^m) = \text{addc}(\text{addc}(\text{addr}(W(\tilde{R}_1))))$$

$$W(R_2^m) = \text{addc}(\text{addc}(\text{addr}(W(\tilde{R}_2))))$$

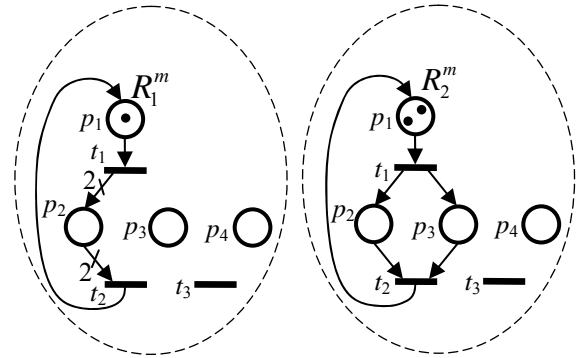


Fig. 3. Addition of isolated places and transitions to increase the size of the incidence matrices.

The new incidence matrices are shown in the figure 4.

$$W(R_1^m) = \begin{pmatrix} t_1 & t_2 & t_3 \\ -1 & 1 & 0 \\ 2 & -2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{matrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{matrix}$$

$$W(R_2^m) = \begin{pmatrix} t_1 & t_2 & t_3 \\ -1 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{matrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{matrix}$$

Fig. 4. Incidence matrices of the alternative Petri nets after increasing their size.

The second set of equivalence operations will be the swapping of one row and one column in $W(R_1^m)$. The purpose of this operation may be to make the largest number of elements placed in the same position of both incidence matrices to coincide in order to reduce the number of undefined structural parameters of the resulting Petri net.

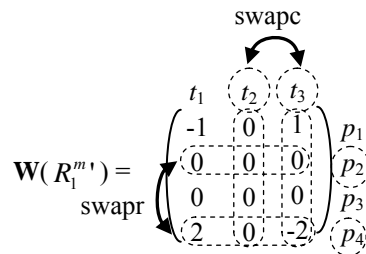


Fig. 5. Incidence matrix of $R_1^{m'}$ after the swapping operations.

The process may be seen in the figure 5 and correspond to the operations:

$$W(R_1^{m'}) = \text{swapr}(\text{swapc}(\text{addr}(W(\tilde{R}_1), 2, 3), 2, 4))$$

After this last operation it is possible to merge the incidence matrices of both alternative Petri nets to obtain a single compound Petri net.

4. CONCLUSIONS AND FURTHER RESEARCH

In this paper it has been shown how it is possible to apply matrix-based operations to the incidence matrices of a Petri net that preserve the structure of the graph of reachable markings. As a consequence, the properties and behaviour of the original and resulting Petri nets are equivalent. This idea constitutes a powerful tool that allows, in the application described in this document, to transform a set of alternative Petri nets into a compound Petri net with a set of undefined parameters is expected to be reduced, comparing with a case where the matrix-based operations are not applied.

As future research it is expected to extend the application of these ideas to other fields in the modeling of discrete event systems, as well as to develop new matrix-based operations that allow obtaining equivalent Petri nets.

REFERENCES

- Berthelot, G., 1987 *Transformations and decompositions of nets* in "Petri Nets: Central Models and Their Properties, Advances in Petri Nets". Lecture Notes in Computer Science, Brauer, W., Reisig, W., and Rozenberg, G. (eds.), vol. 254-I, pp. 359–376. Springer.
- Cassandras, Christos G., Lafortune, S., 2008 *Introduction to Discrete Event Systems*. Second Edition, Springer.
- David, R., Alla, H., , 2005 *Discrete, Continuous and Hybrid Petri nets*, Springer
- Haddad, S. and Pradat-Peyre, J.F., 2006 *New Efficient Petri Nets Reductions for Parallel Programs Verification*. Parallel Processing Letters, pages 101-116, World Scientific Publishing Company.
- Jiménez, E., Pérez, M., Latorre, J.I., 2005 *On deterministic modelling and simulation of manufacturing systems with Petri nets*. Proceedings of European Modelling Simulation Symposium. Marseille, pp. 129-136. Marseille.
- Latorre, J.I., Jiménez, E., Pérez, M., 2007 *Macro-Reachability Tree Exploration for D.E.S. Design Optimization*. Proceedings of the 6th EUROSIM Congress on Modelling and Simulation (Eurosime 2007). Ljubljana, Slovenia.
- Latorre, J.I., Jiménez, E., Pérez, M., Martínez, E., 2009 *The design of a manufacturing facility. An efficient approach based on alternatives aggregation Petri* Proceedings of the 21st European Modelling and Simulation Symposium (EMSS 09). Puerto de la Cruz, Spain, vol. 2, pp. 33-39.
- Latorre, J.I., Jiménez, E., Pérez, M., Blanco, J., 2009 *The problem of designing discrete event systems. A new methodological approach*. Proceedings of the 21st European Modelling and Simulation Symposium (EMSS 09). Puerto de la Cruz, Spain, vol. 2, pp. 40-46.
- Latorre, J.I., Jiménez, E., Pérez, M., 2010 *Control of Discrete Event Systems Modelled by Petri Nets*. Proceedings of the 7th EUROSIM Congress. Prague.
- Latorre, J.I., Jiménez, E., Pérez, M., 2011 *Efficient Representations Of Alternative Models Of Discrete Event Systems Based On Petri Nets*. Proceedings of the UKSim 13th International Conference on Computer Modelling and Simulation. Cambridge, United Kingdom.
- Peterson, J.L., 1981 *Petri Net Theory and the Modelling of Systems*. Prentice Hall, Englewood Cliffs.
- Silva, M., 1993 *Introducing Petri nets*. In Practice of Petri Nets in Manufacturing", Di Cesare, F., (eds.), pp. 1-62. Ed. Chapman&Hall.
- Van der Aalst, W.M.P., 1997 *Verification of Workflow Nets*. In "Application and Theory of Petri Nets 1997", P. Azéma and G. Balbo (eds.), volume 1248 of Lecture Notes in Computer Science, pages 407-426. Springer-Verlag, Berlin.

SYNTHESIS OF FEEDBACK CONTROLLER FOR STABILIZATION OF CHAOTIC HÉNON MAP OSCILLATIONS BY MEANS OF ANALYTIC PROGRAMMING

Roman Senkerik^(a), Zuzana Oplatkova^(a), Ivan Zelinka^(b), Donald Davendra^(b), Roman Jasek^(a)

^(a)Tomas Bata University in Zlin , Faculty of Applied Informatics, Nam T.G. Masaryka 5555, 760 01 Zlin, Czech Republic

^(b) Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science, 17. listopadu 15, 708 33 Ostrava-Poruba, Czech Republic

^(a)senkerik@fai.utb.cz, ^(b)ivan.zelinka@vsb.cz

ABSTRACT

This research deals with a synthesis of control law by means of analytic programming for selected discrete chaotic system – Hénon Map. The novelty of the approach is that analytic programming as a tool for symbolic regression is used for the purpose of stabilization of higher periodic orbits, which represent the oscillations between several values of chaotic system. The paper consists of the descriptions of analytic programming as well as used chaotic system and detailed proposal of cost function used in optimization process. For experimentation, Self-Organizing Migrating Algorithm (SOMA) with analytic programming and Differential evolution (DE) as second algorithm for meta-evolution were used.

Keywords: Chaos Control, Analytic programming, optimization, evolutionary algorithms.

1. INTRODUCTION

The interest about the interconnection between evolutionary techniques and control of chaotic systems is spread daily. First steps were done in (Senkerik et al. 2010a; 2010b), (Zelinka et al. 2009) where the control law was based on Pyragas method: Extended delay feedback control – ETDAS (Pyragas 1995). These papers were concerned to tune several parameters inside the control technique for chaotic system. Compared to this, current presented research shows a possibility as to how to generate the whole control law (not only to optimize several parameters) for the purpose of stabilization of a chaotic system. The synthesis of control law is inspired by the Pyragas's delayed feedback control technique (Just 1999, Pyragas 1992). Unlike the original OGY (Ott, Grebogi and York) control method (Ott et al.1990), it can be simply considered as a targeting and stabilizing algorithm together in one package (Kwon 1999). Another big advantage of the Pyragas method for evolutionary computation is the amount of accessible control parameters, which can be easily tuned by means of evolutionary algorithms (EA).

Instead of EA utilization, analytic programming (AP) is used in this research. AP is a superstructure of EAs and is used for synthesis of analytic solution according to the required behaviour. Control law from the proposed system can be viewed as a symbolic structure, which can be synthesized according to the requirements for the stabilization of the chaotic system. The advantage is that it is not necessary to have some “preliminary” control law and to estimate its parameters only. This system will generate the whole structure of the law even with suitable parameter values.

This work is focused on the expansion of AP application for synthesis of a whole control law instead of parameters tuning for existing and commonly used method control law to stabilize desired Unstable Periodic Orbits (UPO) of chaotic systems.

This research is an extension of previous research (Oplatkova et al. 2010a; 2010b, Senkerik et al., 2010c) focused on stabilization of simple p-1 orbit – stable state. In general, this research is concerned to stabilize p-2 UPO – higher periodic orbits (oscillations between two values).

Firstly, AP is explained, and then a problem design is proposed. The next sections are focused on the description of used cost function and evolutionary algorithms. Results and conclusion follow afterwards.

2. ANALYTIC PROGRAMMING

Basic principles of the AP were developed in 2001 (Zelinka et al. 2005). Until that time only Genetic Programming (GP) and Grammatical Evolution (GE) had existed. GP uses Genetic Algorithms (GA) while AP can be used with any EA, independently on individual representation. To avoid any confusion, based on the nomenclature according to the used algorithm, the name - Analytic Programming was chosen, since AP represents synthesis of analytical solution by means of EA.

The core of AP is based on a special set of mathematical objects and operations. The set of mathematical objects is a set of functions, operators and so-called terminals (as well as in GP), which are usually constants or independent variables. This set of variables

is usually mixed together and consists of functions with different number of arguments. Because of a variability of the content of this set, it is termed the “general functional set” – GFS. The structure of GFS is created by subsets of functions according to the number of their arguments. For example GFS_{all} is a set of all functions, operators and terminals, GFS_{3arg} is a subset containing functions with only three arguments, GFS_{0arg} represents only terminals, etc. The subset structure presence in GFS is vitally important for AP. It is used to avoid synthesis of pathological programs, i.e. programs containing functions without arguments, etc. The content of GFS is dependent only on the user. Various functions and terminals can be mixed together (Zelinka et al. 2005, Zelinka et al. 2008, Oplatkova et al. 2009).

The second part of the AP core is a sequence of mathematical operations, which are used for the program synthesis. These operations are used to transform an individual of a population into a suitable program. Mathematically stated, it is a mapping from an individual domain into a program domain. This mapping consists of two main parts. The first part is called Discrete Set Handling (DSH) (See Figure 1) (Zelinka et al. 2005, Lampinen and Zelinka 1999) and the second one stands for security procedures which do not allow synthesizing pathological programs. The method of DSH, when used, allows handling arbitrary objects including nonnumeric objects like linguistic terms {hot, cold, dark...}, logic terms (True, False) or other user defined functions. In the AP, DSH is used to map an individual into GFS and together with security procedures creates the above-mentioned mapping, which transforms arbitrary individual into a program.

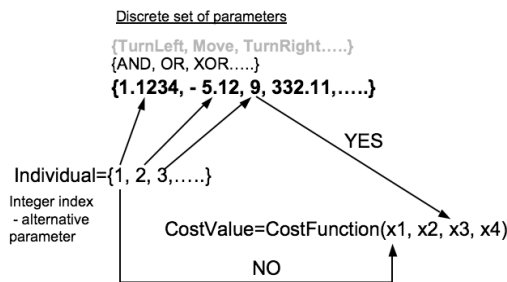


Figure 1: Discrete set handling

AP needs some EA (Zelinka et al. 2005) that consists of a population of individuals for its run. Individuals in the population consist of integer parameters, i.e. an individual is an integer index pointing into GFS. The creation of the program can be schematically observed in Figure 2. The individual contains numbers which are indices into GFS. The detailed description is represented in (Zelinka et al. 2005, Zelinka et al. 2008, Oplatkova et al. 2009).

AP exists in 3 versions – basic without constant estimation, AP_{nf} – estimation by means of nonlinear fitting package in *Mathematica* environment and AP_{meta} – constant estimation by means of another evolutionary algorithms; meta implies metaevolution.

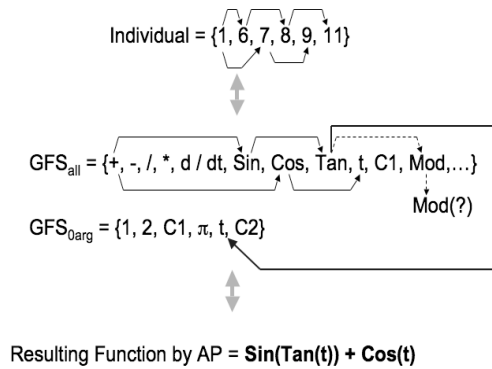


Figure 2: The main principles of AP

3. PROBLEM DESIGN

The brief description of used chaotic system and original feedback chaos control method ETDAS (Pyragas 1995) is given here. The ETDAS control technique was used in this research as an inspiration for synthesizing a new feedback control law by means of evolutionary techniques.

3.1. Selected Chaotic System

The chosen example of chaotic system was the two dimensional Hénon map in form (1):

$$\begin{aligned} x_{n+1} &= a - x_n^2 + by_n \\ y_{n+1} &= x_n \end{aligned} \quad (1)$$

This is a model invented with a mathematical motivation to investigate chaos. The Hénon map is a discrete-time dynamical system, which was introduced as a simplified model of the Poincaré map for the Lorenz system. It is one of the most studied examples of dynamical systems that exhibit chaotic behavior. The map depends on two parameters, a and b , which for the canonical Hénon map have values of $a = 1.4$ and $b = 0.3$. For these canonical values the Hénon map is chaotic (Hilborn 2000). The example of this chaotic behavior can be clearly seen from bifurcation diagram – Figure 3.

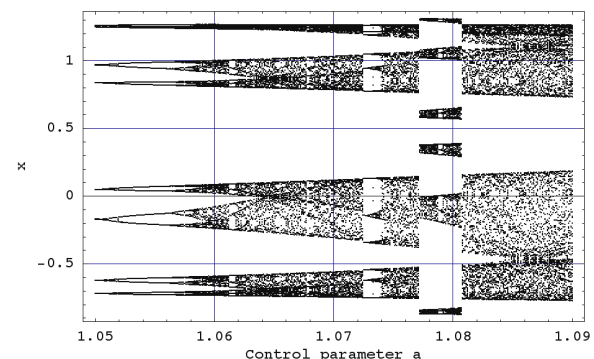


Figure 3: Bifurcation diagram of Hénon Map

Figure 3 shows the bifurcation diagram for the Hénon map created by plotting of a variable x as a function of

the one control parameter for the fixed second parameter.

3.2. ETDAS Control Method

This work is focused on explanation of application of AP for synthesis of a whole control law instead of demanding tuning of EDTAS (Pyragas 1995) method control law to stabilize desired Unstable Periodic Orbits (UPO). In this research desired UPO is only p-2 (higher periodic orbit – oscillation between two values). ETDAS method was obviously an inspiration for preparation of sets of basic functions and operators for AP. The original control method – ETDAS has form (2):

$$\begin{aligned} F(t) &= K[(1-R)S(t-\tau_d)-x(t)], \\ S(t) &= x(t) + RS(t-\tau_d), \end{aligned} \quad (2)$$

where: K and R are adjustable constants, F is the perturbation; S is given by a delay equation utilizing previous states of the system and τ_d is a time delay.

The original control method – ETDAS in the discrete form suitable for two-dimensional Hénon Map has the form (3):

$$\begin{aligned} x_{n+1} &= a - x_n^2 + by_n + F_n, \\ F_n &= K[(1-R)S_{n-m} - x_n], \\ S_n &= x_n + RS_{n-m}, \end{aligned} \quad (3)$$

where: m is the period of m -periodic orbit to be stabilized. The perturbation F_n in equations (3) may have arbitrarily large value, which can cause diverging of the system outside the interval $\{0, 1.0\}$. Therefore, F_n should have a value between $-F_{\max}$, F_{\max} . In this preliminary study a suitable F_{\max} value was taken from the previous research. To find the optimal value also for this parameter is in future plans.

Previous research concentrated on synthesis of control law only for p-1 orbit (a fixed point). An inspiration for preparation of sets of basic functions and operators for AP was simpler TDAS (Pyragas 1992) control method (4) and its discrete form given in (5):

$$F(t) = K[x(t-\tau) - x(t)], \quad (4)$$

$$F_n = K(x_{n-m} - x_n). \quad (5)$$

Compared to this work, the data set for AP presented in the previous research required only constants, operators like plus, minus, power and output values x_n and x_{n-1} . Due to the recursive attributes of delay equation S utilizing previous states of the system in discrete ETDAS (3), the data set for AP had to be expanded and cover longer system output history (x_n to x_{n-9}), thus to imitate inspiring control method for the successful synthesis of control law securing the stabilization of higher periodic orbits

3.3. Cost Function

Proposal for the cost function comes from the simplest Cost Function (CF). The core of CF could be used only for the stabilization of p-1 orbit. The idea was to minimize the area created by the difference between the required state and the real system output on the whole simulation interval – τ_i .

But another universal cost function had to be used for stabilizing of higher periodic orbit and having the possibility of adding penalization rules. It was synthesized from the simple CF and other terms were added. In this case, it is not possible to use the simple rule of minimizing the area created by the difference between the required and actual state on the whole simulation interval – τ_i , due to many serious reasons, for example: degrading of the possible best solution by phase shift of periodic orbit.

This CF is in general based on searching for desired stabilized periodic orbit and thereafter calculation of the difference between desired and found actual periodic orbit on the short time interval – τ_s (40 iterations) from the point, where the first min. value of difference between desired and actual system output is found. Such a design of CF should secure the successful stabilization of either p-1 orbit (stable state) or higher periodic orbit anyway phase shifted. The CF_{Basic} has the form (6).

$$CF_{Basic} = pen_1 + \sum_{t=\tau_1}^{\tau_2} |TS_t - AS_t|, \quad (6)$$

where:

TS - target state, AS - actual state

τ_1 - the first min value of difference between TS and AS

τ_2 – the end of optimization interval ($\tau_1 + \tau_s$)

$pen_1 = 0$ if $\tau_1 - \tau_2 \geq \tau_s$;

$pen_1 = 10 * (\tau_1 - \tau_2)$ if $\tau_1 - \tau_2 < \tau_s$ (i.e. late stabilization).

4. USED EVOLUTIONARY ALGORITHMS

This research used two evolutionary algorithms: Self-Organizing Migrating Algorithm (Zelinka 2004) and Differential Evolution (Price and Storn 2001, Price 2005). Future simulations expect a usage of soft computing GAHC algorithm (modification of HC12) (Matousek 2007) and a CUDA implementation of HC12 algorithm (Matousek 2010).

4.1. Self Organizing Migrating Algorithm – SOMA

SOMA is a stochastic optimization algorithm that is modelled on the social behaviour of cooperating individuals (Zelinka 2004). It was chosen because it has been proven that the algorithm has the ability to converge towards the global optimum (Zelinka 2004). SOMA works with groups of individuals (population) whose behavior can be described as a competitive – cooperative strategy. The construction of a new population of individuals is not based on evolutionary principles (two parents produce offspring) but on the

behavior of social group, e.g. a herd of animals looking for food. This algorithm can be classified as an algorithm of a social environment. To the same group of algorithms, Particle Swarm Optimization (PSO) algorithm can also be classified sometimes called swarm intelligence. In the case of SOMA, there is no velocity vector as in PSO, only the position of individuals in the search space is changed during one generation, referred to as ‘migration loop’.

The rules are as follows: In every migration loop the best individual is chosen, i.e. individual with the minimum cost value, which is called the Leader. An active individual from the population moves in the direction towards the Leader in the search space. At the end of the crossover, the position of the individual with minimum cost value is chosen. If the cost value of the new position is better than the cost value of an individual from the old population, the new one appears in new population. Otherwise the old one remains there. The main principle is depicted in Figures 4 and 5.

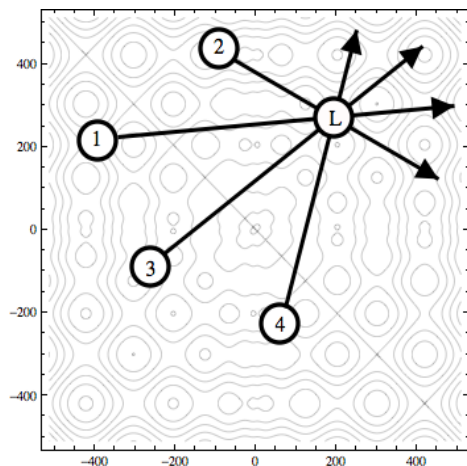


Figure 4: Principle of SOMA, movement in the direction towards the Leader

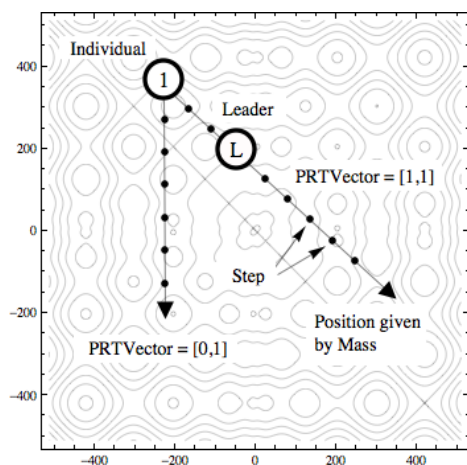


Figure 5: Basic principle of crossover in SOMA, PathLength is replaced here by Mass

4.2. Differential Evolution

DE is a population-based optimization method that works on real-number-coded individuals (Price 2005). For each individual $\bar{x}_{i,G}$ in the current generation G , DE generates a new trial individual $\bar{x}'_{i,G}$ by adding the weighted difference between two randomly selected individuals $\bar{x}_{r1,G}$ and $\bar{x}_{r2,G}$ to a randomly selected third individual $\bar{x}_{r3,G}$. The resulting individual $\bar{x}'_{i,G}$ is crossed-over with the original individual $\bar{x}_{i,G}$. The fitness of the resulting individual, referred to as a perturbed vector $\bar{u}_{i,G+1}$, is then compared with the fitness of $\bar{x}_{i,G}$. If the fitness of $\bar{u}_{i,G+1}$ is greater than the fitness of $\bar{x}_{i,G}$, then $\bar{x}_{i,G}$ is replaced with $\bar{u}_{i,G+1}$; otherwise, $\bar{x}_{i,G}$ remains in the population as $\bar{x}_{i,G+1}$. DE is quite robust, fast, and effective, with global optimization ability. It does not require the objective function to be differentiable, and it works well even with noisy and time-dependent objective functions. Description of used DERand1Bin strategy is presented in (7). Please refer to (Price and Storn 2001, Price 2005) for the description of all other strategies.

$$u_{i,G+1} = x_{r1,G} + F \cdot (x_{r2,G} - x_{r3,G}) \quad (7)$$

5. SIMULATION RESULTS

As described in section 2 about Analytic Programming, AP requires some EA for its run. In this paper AP_{meta} version was used. Meta-evolutionary approach means usage of one main evolutionary algorithm for AP process and second algorithm for coefficient estimation, thus to find optimal values of constants in the evolutionary synthesized control law.

SOMA algorithm was used for main AP process and DE was used in the second evolutionary process. Settings of EA parameters for both processes were based on performed numerous experiments with chaotic systems and simulations with AP_{meta} (See Table 1 and Table 2).

Table 1. SOMA settings for AP

Parameter	Value
PathLength	3
Step	0.11
PRT	0.1
PopSize	50
Migrations	4
Max. CF Evaluations (CFE)	5345

Table 2. DE settings for meta-evolution

Parameter	Value
PopSize	40
F	0.8
CR	0.8
Generations	150
Max. CF Evaluations (CFE)	6000

The Analytic Programming used following setting-up:
 Basic set of elementary functions for AP:
 $GFS_{2arg} = +, -, /, *, ^$
 $GFS_{0arg} = data_{n-9} \text{ to } data_n, K$

Total number of cost function evaluations for AP was 5345, for the second EA it was 6000, together 32.07 millions per each simulation. The novelty of this approach represents the synthesis of feedback control law F_n (8) (perturbation) for the Hénon Map inspired by original ETDAS control method.

$$x_{n+1} = a - x_n^2 + by_n + F_n \quad (8)$$

Following two presented simulation results represent the best examples of synthesized control laws. Based on the mathematical analysis, the real p-2 UPO for unperturbed logistic equation has following values: $x_1 = -0.5624, x_2 = 1.2624$.

Description of the two selected simulation results covers direct output from AP – synthesized control law without coefficients estimated; further the notation with simplification after estimation by means of second algorithm DE and corresponding CF value.

5.1. Example 1

The first example of a new synthesized feedback control law F_n (perturbation) for the controlled Hénon map (8) inspired by original ETDAS control method (3) has the form (9) – direct output from AP and form (10) – with estimated coefficients by means of the second EA.

$$F_n = -\frac{x_{n-1}(K_2 - x_{n-3} - x_n)}{K_1} \quad (9)$$

$$F_n = 0.342699x_{n-1}(0.7 - x_{n-3} - x_n) \quad (10)$$

Simulation depicted in Figure 6 lends weight to the argument, that AP is able to synthesize a new control law securing very quick and very precise stabilization. The CF Value was $3.8495 \cdot 10^{-12}$, which means that average error between actual and required system output was $9.6237 \cdot 10^{-14}$ per iteration.

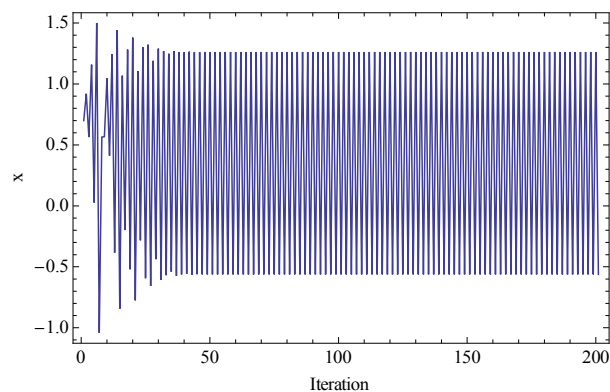


Figure 6: Simulation results - the first example

5.2. Example 2

The second example of a new synthesized feedback control law F_n (perturbation) for the controlled Hénon map (8) inspired by original ETDAS control method (3) has the form (11) - direct output from AP and form (12) – with estimated coefficients.

$$F_n = \frac{x_{n-5}x_{n-1}(K_1 + x_{n-3})\left(-\frac{x_{n-7} - x_n}{K_3}\right)}{K_2\left(-\frac{x_{n-7}}{x_{n-5}x_{n-2}x_n} + x_{n-6} + x_{n-4} - x_{n-1}\right)x_n} \quad (11)$$

$$F_n = -\frac{x_{n-5}x_{n-1}(x_{n-3} - 25.168)(-0.5402x_{n-7} - x_n)}{4.4124\left(-\frac{x_{n-7}}{x_{n-5}x_{n-2}x_n} + x_{n-6} + x_{n-4} - x_{n-1}\right)x_n} \quad (12)$$

Simulation output representing successful and quick stabilization of Hénon map is depicted in Figure 7. The CF Value was 0.7540 (average error 0.01885 per iteration).

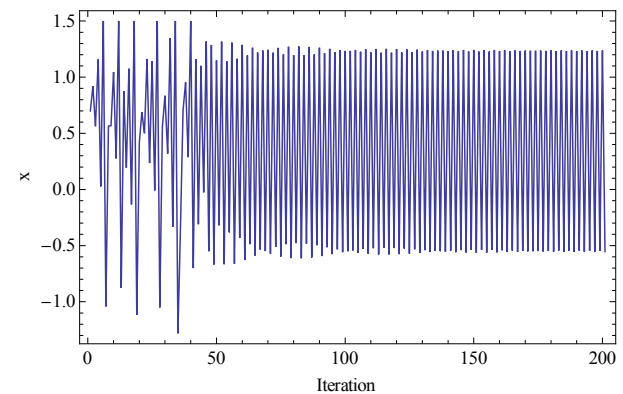


Figure 7: Simulation results - the second example

6. CONCLUSION

This paper deals with a synthesis of a control law by means of AP for stabilization of selected chaotic system at higher periodic orbit. Hénon map as an example of two-dimensional discrete chaotic system was used in this research.

In this presented approach, the analytic programming was used instead of tuning of parameters for existing control technique by means of EA's as in the previous research.

Obtained results reinforce the argument that AP is able to solve this kind of difficult problems and to produce a new synthesized control law in a symbolic way securing desired behaviour of chaotic system and stabilization.

Presented two simulation examples show two different results. Extremely precise stabilization and simple control law in the first case and not very precise and slow stabilization and relatively complex notation of chaotic controller in the second case. This fact lends

weight to the argument, that AP is a powerful symbolic regression tool, which is able to strictly and precisely follow the rules given by cost function and synthesize any symbolic formula, in the case of this research – the feedback controller for chaotic system. The question of energy costs and more precise stabilization will be included into future research together with development of better cost functions, different AP data set, and performing of numerous simulations to obtain more results and produce better statistics, thus to confirm the robustness of this approach.

ACKNOWLEDGMENTS

This work was supported by the grant NO. MSM 7088352101 of the Ministry of Education of the Czech Republic and by grants of Grant Agency of Czech Republic GACR 102/09/1680 and by European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089.

REFERENCES

- Hilborn R.C., 2000. *Chaos and Nonlinear Dynamics: An Introduction for Scientists and Engineers*, Oxford University Press, 2000, ISBN: 0-19-850723-2.
- Just W., 1999, *Principles of Time Delayed Feedback Control*, In: Schuster H.G., Handbook of Chaos Control, Wiley-Vch, ISBN 3-527-29436-8.
- Kwon O. J., 1999. *Targeting and Stabilizing Chaotic Trajectories in the Standard Map*, Physics Letters A. vol. 258, 1999, pp. 229-236.
- Lampinen J., Zelinka I., 1999, *New Ideas in Optimization – Mechanical Engineering Design Optimization by Differential Evolution*, Volume 1, London: McGraw-hill, 1999, 20 p., ISBN 007-709506-5.
- Matousek R., 2007, *GAHC: Improved GA with HC station*, In WCECS 2007, San Francisco, pp. 915-920. ISBN: 978-988-98671-6-4.
- Matousek R., 2010, *HC12: The Principle of CUDA Implementation*. In MENDEL 2010, Mendel Journal series, pp. 303-308. ISBN: 978-80-214-4120- 0. ISSN: 1803- 3814.
- Oplatková, Z., Zelinka, I.: 2009. *Investigation on Evolutionary Synthesis of Movement Commands*, Modelling and Simulation in Engineering, Vol. 2009.
- Oplatkova Z., Senkerik R., Zelinka I., Holoska J., 2010a, *Synthesis of Control Law for Chaotic Henon System - Preliminary study*, ECMS 2010, Kuala Lumpur, Malaysia, p. 277-282, ISBN 978-0-9564944-0-5.
- Oplatkova Z., Senkerik R., Belaskova S., Zelinka I., 2010b, *Synthesis of Control Rule for Synthesized Chaotic System by means of Evolutionary Techniques*, Mendel 2010, Brno, Czech Republic, p. 91 - 98, ISBN 978-80-214-4120-0.
- Ott E., C. Greboki, J.A. Yorke, 1990. *Controlling Chaos*, Phys. Rev. Lett. vol. 64, 1990, pp. 1196-1199.
- Price K., Storn R. M., Lampinen J. A., 2005, *Differential Evolution : A Practical Approach to Global Optimization*, Natural Computing Series, Springer.
- Price, K. and Storn, R. (2001), *Differential evolution homepage*, [Online]: <http://www.icsi.berkeley.edu/~storn/code.html>. [Accessed 30/04/2011].
- Pyragas K., 1992, *Continuous control of chaos by self-controlling feedback*, Physics Letters A, 170, 421-428.
- Pyragas K., 1995. *Control of chaos via extended delay feedback*, Physics Letters A, vol. 206, 1995, pp. 323-330.
- Senkerik, R., Zelinka, I., Davendra, D., Oplatkova, Z., 2010a, *Evolutionary Design of Chaos Control in 1D*, In. Zelinka I., Celikovski S., Richter H., Chen G.: Evolutionary Algorithms and Chaotic Systems, SpringerVerlag Berlin, pp.165 - 190.
- Senkerik R., Zelinka I., Davendra D., Oplatkova Z., 2010b, *Utilization of SOMA and differential evolution for robust stabilization of chaotic Logistic equation*, Computers & Mathematics with Applications, Volume 60, Issue 4, pp. 1026-1037.
- Senkerik R., Oplatkova Z., Zelinka I., Davendra D., Jasek R., 2010c, *Synthesis Of Feedback Controller For Chaotic Systems By Means Of Evolutionary Techniques*, Proceeding of Fourth Global Conference on Power Control and Optimization, Sarawak, Borneo, 2010..
- Zelinka I., 2004. "SOMA – Self Organizing Migrating Algorithm", In: *New Optimization Techniques in Engineering*, (B.V. Babu, G. Onwubolu (eds)), chapter 7, 33, Springer-Verlag, 2004, ISBN 3-540-20167X.
- Zelinka I., Oplatkova Z., Nolle L., 2005. *Boolean Symmetry Function Synthesis by Means of Arbitrary Evolutionary Algorithms-Comparative Study*, International Journal of Simulation Systems, Science and Technology, Volume 6, Number 9, August 2005, pages 44 - 56, ISSN: 1473-8031.
- Zelinka I., Senkerik R., Navratil E., 2009, *Investigation on evolutionary optimization of chaos control*, Chaos, Solitons & Fractals, Volume 40, Issue 1, pp. 111-129.
- Zelinka, I., Guanrong Ch., Celikovskiy S., 2008. *Chaos Synthesis by Means of Evolutionary algorithms*, International Journal of Bifurcation and Chaos, Vol. 18, No. 4, pp. 911-942

A SIMULATION STUDY OF THE INTERDEPENDENCE OF SCALABILITY AND CANNIBALIZATION IN THE SOFTWARE INDUSTRY

Francesco Novelli

SAP Research Darmstadt, Germany

francesco.novelli@sap.com

ABSTRACT

Dominant software vendors, whose applications have been predominantly delivered on-premises so far (i.e., installed, maintained and operated at customers' premises) are challenged by the rising adoption of software as a service (SaaS) solutions, outsourced applications delivered through the web under subscription or usage-based pricing terms. As a response, these incumbent vendors extend their product portfolios with SaaS offerings, but thus risk to engender revenue cannibalization, as a newly introduced SaaS application may attract their own on-premises customers instead of expanding the market or drawing from a competitor's customer base. At the same time they face the novel, severe scalability requirements of the technological and organizational infrastructure underlying a successful SaaS business. Using an agent-based simulation model, we study the interdependence between cannibalization and scalability in monopolistic and duopolistic software markets.

Keywords: software as a service, cannibalization, scalability, agent-based modelling and simulation

1. INTRODUCTION

1.1. Evolution of Software Delivery and Pricing Models

In the last decade the software industry has witnessed the emergence of the so-called software as a service (SaaS) delivery model, whereby vendors provide web-based, outsourced software applications (SIIA 2001), dispensing customers with most installation, operation and maintenance activities otherwise needed at their premises. Moreover, SaaS solutions are usually coupled with subscription or usage-based pricing models (Lehmann and Buxmann 2009), lowering the initial investment in comparison with packaged software.

As a matter of fact, several beholders of the software industry agree that the adoption of SaaS applications has gained momentum: Information Systems researchers (Benlian, Hess, and Buxmann 2009), IT market analysts (Gartner 2010), and investors, who, in the first quarter of 2011, gave SaaS public companies an average market evaluation 6.5 times their annual revenues. This trend, in turn, urges incumbent vendors, which built their dominant market positions on the on-

premises delivery model, to launch SaaS counterparts to their established software products, in coexistence or as replacements.

We focus on two key challenges inherent in such a strategic response. On the one hand, it is indispensable for the incumbent to understand the dynamics of revenue cannibalization and its consequences on profitability and on the positioning of the firm vis-a-vis the competition. On the other hand, it is necessary to achieve the degree of technological and organizational scalability required to profitably implement a SaaS strategy.

1.2. Cannibalization

Cannibalization is the switching of sales from an existing product to a new one of the same firm (Newman 1967). It is an issue of paramount importance for an incumbent vendor venturing into SaaS, given the intrinsic degree of substitutability between the already established on-premises products and their SaaS siblings. This may indeed put pre-existent revenue streams and market shares at stake.

As a case in point, let us consider how competition is unfolding in the office automation market. Microsoft is the dominant player with 6 billion dollars revenue from that segment in the second quarter of 2011 (as a term of comparison, 5 billion was the revenue from the Windows OS). However, the entry into the market of free online office applications (such as those by Google) has pushed Microsoft to respond with the development of two SaaS alternatives to its well-known Office suite: a free, ad-supported one with limited functionalities and a subscription-based one with enhanced collaboration features. The delicate challenge is for Microsoft to tame the cannibalization effect this move may engender, i.e., to avert a financially harmful drift of on-premises customers to its own SaaS offerings.

Cannibalization may also represent a deliberate, offensive product strategy, pursued to drive growth (McGrath 2001). As a matter of fact, some on-premises vendors have successfully managed the transition to a hybrid or purely SaaS model. Concur Technologies, for instance, paired its on-premises offerings with the Application Service Provider model (predecessor of SaaS) in the late 90s already, and then transitioned to become a purely SaaS player just as this delivery model emerged (Warfield 2007). Analogously, Ariba started

the transition in 2006 and gradually ported all its applications to a SaaS model, now the generator of most of its revenues (Wainwright 2009). Both companies initially went after a SaaS-enabled market expansion aimed towards the mid-segment but eventually – in sharp contrast with the Microsoft scenario – found it profitable to deliberately cannibalize their on-premises customers along the way.

1.3. Scalability

Understanding the financial and competitive consequences of revenue cannibalization and then attempting to avert it or to ride it are not the only concerns facing incumbents. The SaaS delivery model poses a scalability threat as well, both from a technological and from an organizational perspective.

This threat stems from the peculiarities of this newly addressable market. Since SaaS lowers the technological and financial requirements for a software purchase, the market swells in number of potential buyers while the average financial and technological resources available to them decrease. As a matter of fact, since a SaaS offering is hosted and operated by the provider and accessed through the World Wide Web, very simple applications that do not demand supplemental integration and customization virtually appeal to any organization meeting the minimum technical requirement of an available Internet access. From a financial point of view, the SaaS subscription fees dilute over time the investment for the license of a given software functionality. Therefore, small and medium-size companies can, in spite of their usually more limited IT budget and technical personnel, enter application markets once populated by large enterprises only. This is exemplarily shown for the European Union in Table 1. It evidently represents a huge market opportunity to be tapped into by software vendors in terms of number of potential new accounts (enterprises) and users (employees).

Table 1: Key indicators per enterprise size class in the EU-27 area; source: Eurostat (# as of 2010, * as of 2008, † as of 2007)

Size class	Small	Medium	Large
Employees	10-49	50-249	250 or more
Number of enterprises*	1,408,000	228,000	45,000
Persons employed* (million)*	27.9	23.4	45.2
% enterprises that employ IT/ICT specialists†	9%	25%	53%
% enterprises with an ERP system#	17%	41%	64%
% enterprises with a CRM system#	23%	39%	54%

This opportunity for market expansion has its downsides: though larger, the new potential market is more costly to be reached and to be served. Figure 1 compares the trend in total operating expenditures over total revenues for the two leading business application vendors (SAP and Oracle), which have historically kept

it in the 60-70% range, and two SaaS competitors, unable, even after having successfully ridden a steep learning curve, to lower it under the 90% mark (Salesforce) or to reach operational profitability at all (NetSuite).

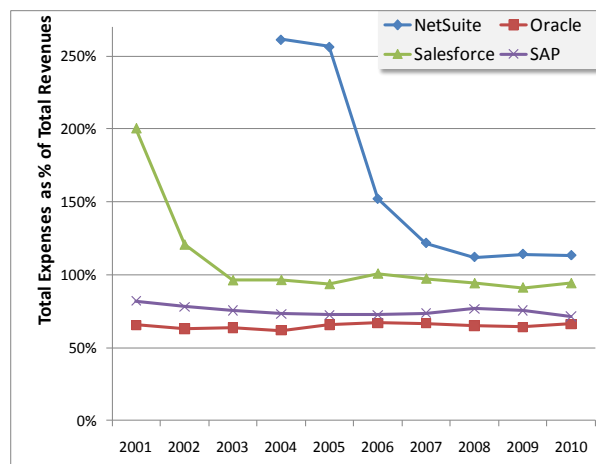


Figure 1: Total Operating Expenses as a Percentage of Total Revenues for selected Software Vendors (source: corporate financial reports)

As a matter of fact, a SaaS software vendor must bear the additional costs of setting up and operating the technological infrastructure needed to deliver the SaaS application. Moreover, beyond those technological repercussions, scalability issues engage the SaaS provider on an organizational level as well, for this new, more fragmented segment of software buyers imposes to think a series of processes anew. For instance marketing and sales, where using dedicated sales team for each account as it is the habit with large enterprises is not possible on a large scale, and other means, such as tele-sales and innovative internet-based funnels, need to be employed.

1.4. The Interplay of Cannibalization and Scalability

Incumbent software vendors introducing SaaS are confronted with a typical new product introduction problem, where the new product may divert current customers from other offerings of the same firm, instead of attracting new buyers or drawing from a competitor's customer base (Kerin, Harvey, and Rothe 1978), further complicated by the trade-off between a more saturated but highly profitable current software market of large enterprises and a fast-growing but less profitable potential SaaS market.

Scalability of the SaaS business is certainly a prerequisite to target market expansion. Without it vendors face the risk of not being able to satisfy demand – i.e., failing to build the appropriate level of capacity to ride growth – or doing so inefficiently – i.e., failing to reach the scale economies that make market expansion a profitable endeavour at all.

Changing perspective, limiting scalability can be a radical lever to avert cannibalization, for it puts an upper bound on the volume of intrafirm switching cus-

tomers. This may expose the incumbent's flank to competition though, whereby customers might switch to a competitor instead.

This work is organized as follows: the reasons that led to the choice of Agent-Based Modelling and Simulation as our research methodology are summarized in the following section (2). In section 3 we describe the model we based our simulation experiments on, which are then detailed and discussed in section 4. Eventually, we conclude in section 5, mentioning the limitations of the present work and possible future developments.

2. RESEARCH METHODOLOGY

2.1. Agent-Based Modelling and Simulation to Study Microeconomics

A market represents an excellent example of Complex Adaptive System (CAS), a collection of adaptive agents (suppliers and customers) concurrently engaged in local interactions (commercial transactions). Local interactions produce higher level conditions (market prices, bandwagon effects, etc.) impacting in turn the way those same interactions will evolve.

Among the paradigms used to investigate such systems we chose Agent-Based Modelling and Simulation (ABMS). In ABMS each interacting component is modelled as an autonomous decision-making agent with attributes and rules defining his behavioural characteristics and how those are to evolve or adapt (North and Macal 2007). This approach lends itself neatly to the exploitation of microeconomic constructs in modelling agents' behaviours and interactions (game theory, for instance, to dictate an agent's strategic response) and is therefore especially suitable to study a CAS populated by economic entities. In fact, the study of economics with ABMS has reached such a respectable status to beget its own specific field of research, called Agent-Based Computational Economics (Tsfatsion 2002).

In this work we use ABMS to investigate at a microeconomic level a stylized business application software market where a multi-product incumbent vendor runs the risk of revenue cannibalization. ABMS suits perfectly the study of this phenomenon since, offering the possibility to observe the behaviours and decisions of individual buyers, it allows a disaggregated analysis of the components of demand. This means identifying exactly which customers switch between software applications of the same vendor (cannibalization), leave for a competitor (competitive draw), or enter the market for the first time (market expansion).

3. MODEL

We model a closed, vertically-differentiated software application market. The market structure is a monopoly with a single vendor selling both an on-premises application and a SaaS one for the first series of experiments, and a duopoly consisting of the same vendor plus a purely SaaS challenger for the second set of experi-

ments. Given space constraints, in this section we describe only the most salient features of the model.

3.1. Software Application

A software application is characterized by the features or benefits it provides (its "quality") to its users, the price to be paid to obtain those benefits (in terms of amount and distribution over time of the fees) and the infrastructure on which it is deployed.

When the application is delivered as SaaS, it will be deployed on an infrastructure operated by the software vendor and priced under subscription terms, with an initial activation charge at the time of purchase and an anticipated, recurrent fee for each period of the simulation in which it is used. When the application is delivered on-premises, the price structure will follow the typical enterprise application pricing model and be once again made up of two components: an initial charge to purchase the licenses and an anticipated, periodical maintenance fee as a percentage of the initial charge.

While the fee structure we employ for the two delivery modes is the same, the proportion between initial and periodical charge differs, with on-premises application weighting more on the front – annual maintenance rates are around 20% of the initial investment for licenses (Buxmann, Diefenbach, and Hess 2011) – and SaaS diluting the expenses more over time.

3.2. Software Application Vendors

The software vendors are price-making suppliers of the same class of business application (e.g., ERP, CRM, etc.) but with different price-quality schedules and delivery models. In the case of a SaaS the vendor must also operate the infrastructure on which the application is deployed. In each simulation period vendors collect the due payments from the customers that adopted one of their applications and bear the costs of the SaaS infrastructure.

3.3. SaaS Infrastructure and Scalability

The infrastructure is made up of a set of technological or organizational resources (e.g., servers, sales representatives) characterized by a certain individual performance. The overall performance is, however, not just the sum of these components, and depends on the level of scalability of the infrastructure. We use Amdahl's law (Amdahl 1967) to account for this issue and formalize the degree of non-scalability of the infrastructure through a so-called contention rate, which exerts a negative impact on the ability to efficiently scale (and, as we will see, to compete) growing exponentially with the scale requirement (Shalom and Perry 2008).

The maximum total capacity K of an infrastructure with N resources of throughput τ each, designed to have a CnR rate of contention is:

$$\kappa = \frac{\tau}{CnR + \frac{1 - CnR}{N}} \quad (1)$$

Equation (1) gives the total capacity a certain infrastructure can attain in a given period. For instance, 20 resources with throughput of 1000 customers per period each, arranged in an architecture designed to have a 20% contention rate, would generate a total capacity of 4167 customers per period. Doubling the resources (i.e., scaling out of 20 additional resources) 4545 customers could be served (an 8% increase). However, the maximum achievable capacity would be bounded to less than 5000 customers per period, no matter how many additional resources are thrown in. Reducing contention would be a much effective lever: decreasing the contention rate of 5% would increase capacity by 25% (to 5195 customers per period).

3.4. Software Application Customers

Customers are current or potential adopters of a software application sold in the market. The decision to adopt an application is made on the basis of the obtained surplus. The surplus for the i -th customer of type θ when adopting an application j is:

$$S_{ij} = \theta_i Q_j - TCO_j(T) + \alpha X_j \quad (2)$$

The first term of equation (2) is the willingness to pay of a customer with marginal valuation of quality θ for an application of quality Q . θ is an input parameter set randomly for each consumer at simulation start (drawn from a uniform distribution with support between 0 and 1). The second term is the present value of the total cost of ownership of the application (detailed below). The third addend is the network externality derived from all consumers that already adopted an application with the same delivery model (i.e., the relevant network X_j is the total number of SaaS customers if j is one of the SaaS applications, or the total number of on-premises customers if j the incumbent's on-premises application).

The total cost of ownership over a horizon of T years is computed for both on-premises and SaaS applications employing the formula for the present value of an annuity:

$$TCO_j(T) = \varphi_{t_0} + \varphi_t + \frac{\varphi_t}{r} (1 - (1 + r)^{-T+1}) \quad (3)$$

where φ_{t_0} is initial charge (activation of the SaaS subscription or on-premises license charge), φ_t the anticipated periodical charge (the subscription fee or the maintenance fee respectively), and r the annual interest rate.

When taking a purchase decision, a consumer first calculates (2) for every available application, then adopts the one with highest non-negative surplus among those with available capacity offered in the market. Although we do not explicitly model switching costs, the consumer that has already adopted an application considers the initial charge a sunk cost and drop φ_{t_0} from

equation (3) when comparing the surplus of her current choice with other alternatives in the market.

The offerings in the market have a certain initial market share each in terms of pre-assigned customers (the incumbent's on-premises one being the largest), but the overall addressable market includes potential customers that will take their first buying decision during the simulation.

3.5. Implementation and Runtime Environment

The whole model was implemented in Java using the ABMS open-source toolkit Repast Symphony 2.0 (North *et al.* 2007).

4. EXPERIMENTS

An experiment consisted of 10 replications of 21 periods of length (the equivalent of three 7-years software application life cycles in the temporal scale we chose) for each model configuration, where a model configuration was given by the contention rate of the incumbent's SaaS infrastructure, specified in a 0%-50% interval with 5% steps. Each experiment was conducted in different growth and competitive scenarios as detailed in the two following sub-sections.

4.1. Experiments in a Monopolistic Market

Our first series of experiments dealt with a monopoly in two scenarios: one of high growth of the SaaS segment, where this has a total size (in terms of number of potential customers) 10 times the on-premises segment, and one of low growth, where it merely matches the on-premises segment's size.

In a monopolistic situation the decisions of the incumbent is linked to the trade-off between revenue cannibalization and market expansion. If the potential market tapped into with SaaS is large enough in size to offset the effect of cannibalizing the high-margin on-premises sales, then the vendor should pursue a high-capacity strategy and, therefore, invest in a scalable infrastructure. Otherwise cannibalization could be averted by limiting the capacity offered in the market with a more conservative strategy. Conversely, a company that has not yet reached the needed level of scalability would unprofitably pursue growth in the SaaS segment and should refrain from it.

Examining the results of these first experiments, it can be seen that, in case of high-growth in the SaaS segment, the monopolist may indeed offset (in terms of sales volume) revenue cannibalization with market expansion by pursuing a high-scalability strategy (Figure 2). On the contrary a low-scalability strategy allows minimizing revenue cannibalization in a low-growth scenario, where no significant market expansion would be possible anyway (Figure 3). Please note that throughout the remainder of this section we calculated total cannibalized revenues in terms of the projected on-premises revenues lost when the customer switch to SaaS (i.e., the discounted stream of maintenance fees, as expressed by eq. 3).

The specific contribution margins of the two software products will dictate the overall effect on the monopolist's profit. Given the higher margins enjoyed in delivering on-premises applications (see introduction), mirrored in our model, a multi-product monopolist in a low-growth scenario would be better off slowing the rate of SaaS adoption among its own customers by limiting the offered capacity (Figure 4). On the contrary, being able to scale to expand into the SaaS segment would be, in case of high growth, the more profitable strategy.

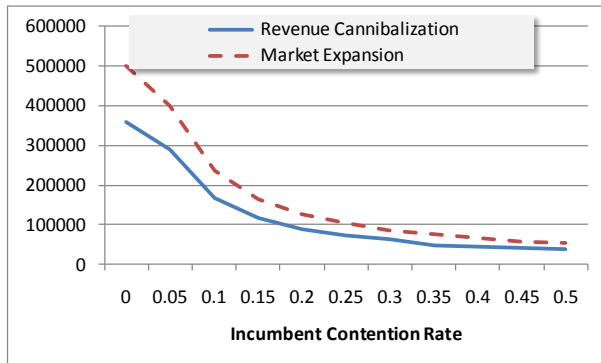


Figure 2: Total Cannibalized On-Premises Revenues and Total SaaS Market Expansion in a Scenario of High Growth (Average for 10 replications)

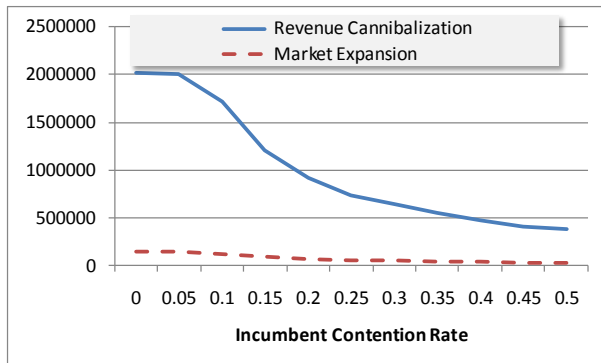


Figure 3: Total Cannibalized On-Premises Revenues and Total SaaS Market Expansion in a Scenario of Low Growth (Average for 10 replications)

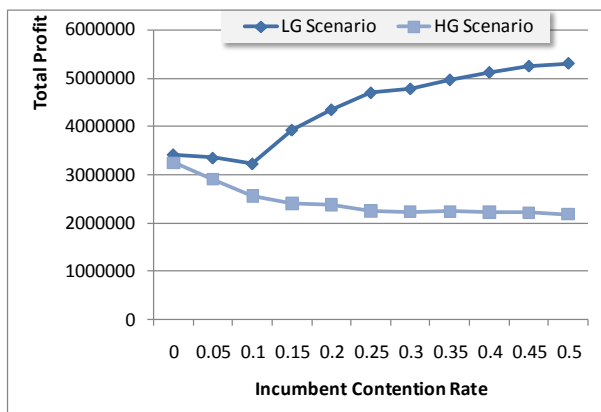


Figure 4: Incumbent's Total Profit in the two Monopolistic Scenarios (HG = High Growth, LG = Low Growth; Average for 10 replications)

4.2. Experiments in a Duopolistic Market

In the presence of a SaaS challenger the incumbent's on-premises customers can switch to either the incumbent's SaaS offering or the competitor's one, adding the risk of competitive draw to the strategic considerations of the incumbent. This risk can be more or less pronounced depending on the scalability of the challenger's SaaS infrastructure. We therefore define four basic scenarios, showed in Table 2.

Table 2: Basic Scenarios for the Duopolistic Market

		Challenger's Scalability	
		Low (CnR = 0.3)	High (CnR = 0.05)
Growth in the SaaS Segment	Low (1X)	Scenario LG1	Scenario LG2
	High (10X)	Scenario HG1	Scenario HG2

In confronting a high-scalable SaaS challenger it always pays for the incumbent to be able to match the competitor's scale, because this allows at least retaining through cannibalization customers that would otherwise be lost (Scenario LG2, Figure 5) if not even offsetting any competitive draw or cannibalization effect by riding growth (Scenario HG2, Figure 6).

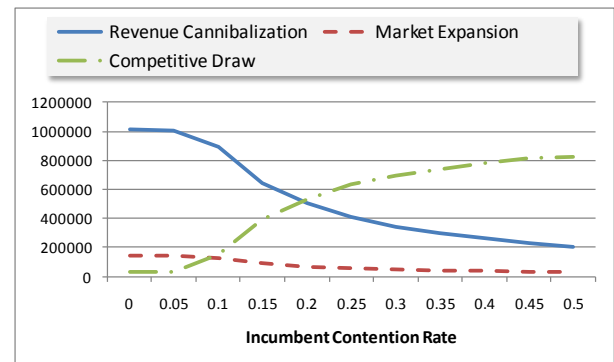


Figure 5: Total Cannibalized On-Premises Revenues, Total SaaS Market Expansion, and Total Competitive Draw of On-Premises Revenues by the SaaS Challenger in Scenario LG2 (Average for 10 replications)

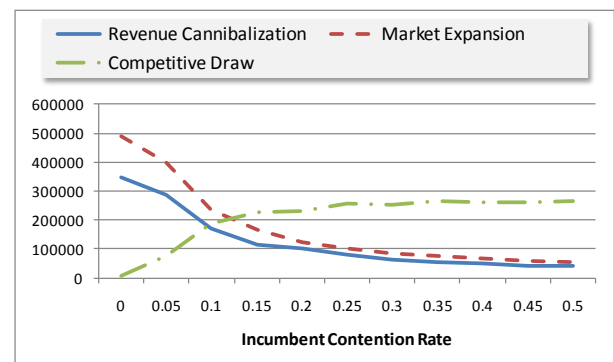


Figure 6: Total Cannibalized On-Premises Revenues, Total SaaS Market Expansion, and Total Competitive Draw of On-Premises Revenues by the SaaS Challenger in Scenario HG2 (Average for 10 replications)

As shown in Figure 7, the incumbent's total profit is generally higher in case of high-growth and negatively correlated with contention, except for the particular case of low growth and presence of a non-scalable challenger (scenario LG1), where the option to limit capacity as a lever to control cannibalization could still be viable. This is due to the lower risk of losing relevant market shares to a poorly-scalable competitor.

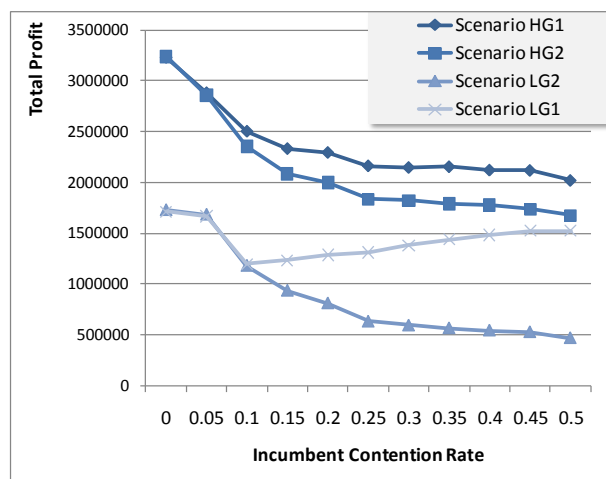


Figure 7: Incumbent's Total Profit in the Identified Market Scenarios (see Table 2; Average for 10 replications)

5. CONCLUSION

This work showed the multi-faceted interdependence of cannibalization and scalability in determining the success of a SaaS strategy pursued by an on-premises vendor, either in a monopolistic position or as an incumbent challenged by a SaaS competitor.

Given the lower margins of a SaaS offering, the monopolist prefers to avoid cannibalization by limiting scale, unless the achievable market expansion proves substantial. In the presence of a SaaS challenger, instead, revenue cannibalization may be for the incumbent a necessary evil whereby customers are retained against the threat of competitive draw. Scalability represents then a key requirement for the incumbent to ride SaaS adoption, cannibalize and possibly expand the market.

These findings were obtained by going after specific strategic interdependences in simplified market scenarios. The modelled market landscape and competitive dynamics should be extended to get a more realistic and comprehensive picture of the trends currently affecting the software industry. Moreover, a thorough validation of the experimental outcomes based on empirical market data ought to be performed.

ACKNOWLEDGMENTS

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) under promotional references 01IA08003A (Project PREMIUM|Services)

REFERENCES

- Amdahl, G., M., 1967. Validity of the single processor approach to achieving large scale computing capabilities. *Proceedings of the April 18-20, 1967, spring joint computer conference* (pp. 483-485).
- Benlian, A., Hess, T., and Buxmann, P., 2009. Drivers of SaaS-adoption—an empirical study of different application types. *Business & Information Systems Engineering*, 1(5), 357–369.
- Buxmann, P., Diefenbach, H., and Hess, T., 2011. *Die Softwareindustrie: Ökonomische Prinzipien, Strategien, Perspektiven*. Springer, Berlin.
- Gartner, 2010. Gartner Survey Indicates More Than 95 Percent of Organizations Expect to Maintain or Grow Their Use of SaaS Through 2010. Available at: <http://www.gartner.com/it/page.jsp?id=1361216> [Accessed June 9, 2011].
- Kerin, R., A., Harvey, M., G. and Rothe, J., T., 1978. Cannibalism and new product development. *Business Horizons*, 5(21), 25–31.
- Lehmann, S., and Buxmann, P., 2009. Pricing strategies of software vendors. *Business & Information Systems Engineering*, 1(6), 452–462.
- McGrath, M., E., 2001. *Product strategy for high technology companies: accelerating your business to web speed*. McGraw-Hill.
- Newman, J., W., 1967. *Marketing management and information: a new case approach*. R. D. Irwin.
- North, M., J., Howe, T., R., Collier, N., T., and Vos, J., R., 2007. A Declarative Model Assembly Infrastructure for Verification and Validation. *Advancing Social Simulation: The First World Congress* (pp. 129-140).
- North, M., J., and Macal, C., M., 2007. *Managing business complexity: discovering strategic solutions with agent-based modeling and simulation*. Oxford University Press, USA.
- Shalom, N., and Perry, G., 2008. Economies of Non-Scale. *Nati Shalom's Blog*. Available at: http://natishalom.typepad.com/nati_shaloms_blog/2008/06/economies-of-no.html [Accessed June 26, 2011].
- SIIA, 2001. *Software as a service: Strategic backgrounder* (Vol. 21). Retrieved from <http://www.siia.net/estore/ssb-01.pdf>.
- Tesfatsion, L., 2002. Agent-based computational economics: growing economies from the bottom up. *Artificial life*, 8(1), 55-82.
- Wainwright, P., 2009. Ariba's Journey to Software as a Service - The Connected Web. *ebiz*. Available at: <http://www.ebizq.net/blogs/connectedweb> [Accessed June 17, 2011].
- Warfield, B., 2007. Interview: Concur's CEO Steve Singh Speaks Out On SaaS/On-Demand. *Smoothspan Blog*. Available at: <http://smoothspan.wordpress.com> [Accessed June 17, 2011].

SECURITY IN SENDING AND STORAGE OF PETRI NETS BY SIGNING AND ENCRPTION

Íñigo León Samaniego^(a), Mercedes Pérez de la Parte^(b), Eduardo Martínez Cámara^(b),
Juan Carlos Sáenz-Díez Muro^(a)

^(a) University of La Rioja. Industrial Engineering Technical School. Department of Electrical Engineering. Logroño, Spain

^(b) University of La Rioja. Industrial Engineering Technical School. Department of Mechanical Engineering. Logroño, Spain

inigo.leon@gmail.com, mercedes.perez@unirioja.es, eduardo.martinezc@unirioja.es, juan-carlos.saenz-diez@unirioja.es

ABSTRACT

The aim of this paper is double. On the one hand, to provide a standard way to hide all or part of a Petri net that could contain sensitive information, such as a company that represents a secret production process through Petri nets (privacy). On the other hand also as standard ensure that Petri net has not been altered (integrity) and that who sends or firm that Petri net is who he say he is (non-repudiation).

To ensure the privacy of an entire Petri net (or a part of it) the best solution is not to prevent access to such information, such as hiding in a safe or behind a firewall, but encrypt that information, even being to view. Today it is easier to open a safe or circumvent a firewall than to break an encryption standard algorithm (which, incidentally, is impossible nowadays).

As for the integrity and non-repudiation, the solution again is not to deliver the Petri net 'in hand' to avoid disruptions and to know who delivers it (since we are in the Internet age). The solution is to digitally sign all or part of a Petri net so that reliably to know who has performed the firm, and be able to detect any unauthorized modification of any of the signed data.

The aim of this paper is to show how to encrypt the selected part of the graph and to sign the Petri net, so that the obtained file compliances with the desired signature and encryption. So, in this final file, all the information (and only that) referred to the shaded part is encrypted and will not be interpretable. In particular, anything will be know about the nodes p1 and p2 or transitions t1 and t3: their constitute a secret process. In addition, this file will contain additional information that will verify the integrity of the file to prevent anyone to modify and information about who has signed this Petri net. The solution we propose is to use PNML representation of Petri nets and XMLEncryption standards for encryption and for signing XMLSignature.

Keywords: Petri nets, Encryption, Digital signature, Privacy, Integrity, Authentication, non-repudiability

1. INTRODUCTION

This paper consits on the application on Petri Nets of some of the latest standard technologies used in computer security. The idea is to provide security and protection of information in data storage and sharing. In particular, we will achieve privacy, authentication, integrity and non-repudiability data. To achieve this, we introduce some concepts such as XML, digital signature, encryption and PNML (Petri Net Marked Language).

Throughout the whole paper standard Technologies are used, but, in order to implement them, in some cases it is necessary to introduce a transformation to the data (without loss of information).

1.1. Privacy

This term is related with the prevention of unauthorized access to information. The solution is not to prevent physically access to such information, eg in a safe hiding or behind a firewall, but to encrypt the information. Nowadays it is easier to open a safe or to circumvent a firewall than to break an encryption standard algorithm (which today is impossible).

1.2. Integrity

The integrity of the data will be obtained if we can avoid or at least detect unauthorized modification of information.

1.3. Authentication

Authentication ensures that people assuring that they say or sign the data, are actually who they say they are. This avoids receiving data from a person posing as another.

1.4. Non-repudiability

It will be obtaided if we can prevent anyone saying that has not sent or modify something done. It should be possible to assure that a preson have done something.

The solution for authentication, integrity and non repudiability fails to deliver the information 'in hand'

to avoid disruptions and to know who gives it, as we are in the era of Internet and technology. The solution is then to digitally sign all or part of the data so that we know who has made the signing of a reliable and be able to detect unauthorized modification of any of the signed data.

1.5. XML

XML is a metalanguage for defining other languages. XML is not really a particular language, but a way of defining languages for different needs. XML is also a standard way to exchange structured information. It is based on distributed hierarchical labels containing data. The XML files are text files. The work is based on this format for implementing information security.

1.6. PNML

Marked Petri Net Language (PNML) is an XML language designed to represent Petri nets. With this language a Petri net can be stored in a text file (XML), without loss of information.

1.7. Digital certificate

A digital certificate is a digital file non-transferable and non-modifiable generated by a trusted third party called Certificate Authority (CA), that associates a public key to a person or entity. For a certificate to perform its tasks need to use a private key that only the owner possesses.

1.8. Digital signature

It is equivalent to the conventional signature. It is an addition to the document you signed and indicates that it agrees with what is said in it. The digital signature provides authentication features, integrity, and non repudiation. Computationally speaking, it is a process that transforms the original message using the private key, and anyone with the signer's public key can verify this.

1.9. Encrypt and Decrypt

Encryption is the process to convert in unreadable some information considered as important. Decoding is the reverse: from the encrypted content becomes legible original content. Keys are used to encrypt and decrypt. In the case that the key to encrypt and decrypt is the same, it is called symmetric encryption. If encryption is made with a key but decryption is made with a different key, it is called asymmetric encryption.

2. APPROACH

The goal of this work is to hide all or part of a Petri net that may contain sensitive information, such as a company that represents a secret production process through Petri nets (privacy). On the other hand, another goal is to ensure with standard resources that a Petri net has not been altered (integrity), and that the sender or firmer of the Petri net is who sais to be (authentication)

and furthermore it may not have been another (non-repudiation).

Let us suppose we have the following Petri net.

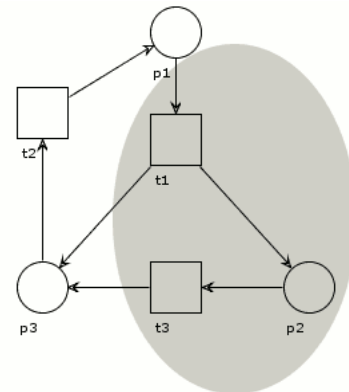


Figure 1: Petri net with a part that want to be hidden

It will be proposed how to encrypt the selected part of the graph, and to sign the Petri net, so that the obtained file meets the desired signature and encryption. So in this final file, all the shaded information (and only that) is encrypted and will not be interpretable. In particular, we will not know anything about the node p2 or transitions t1 and t3: it is a secret process. In addition, this file will contain additional data that will verify the integrity of the file to prevent that anyone modify it, as well as data about who has signed this Petri net.

3. TECHNOLOGIES

3.1. XML Encryption

Encryption is a standard of XML files. It can be used symmetric or asymmetric encryption, but in this case, it is preferable to use symmetric encryption, because it is computationally less demanding.

The idea behind this encryption is to replace the XML elements that want to be encrypted by another piece of XML that contains encrypted information and data about the algorithms used for encryption.

When a file non-XML is encrypted, the only option is to encrypt it completely. When we apply this technology to XML, it permits to define specific fragments of the document that want to be encrypted or even to transform the document before applying encryption.

Whatever the origin of data, the result is always an XML element. Typically, this document has all the information needed to be deciphered. The information that can be found is:

- Encryption algorithm: is the name of a method for encrypting information. It may not be included, being necessary to be know by both the part that encryptes the file and the part that decryptes it.
- Encrypted information: this part must always be present.
- Name of the password used: it is optional. It is used when there is a set of keys, and have to be also

known by both the part that encrypts the file and the part that decrypts it.

- Encrypted password: it is optional. The part that encrypts the document must have a public or a private key. With this key it can encrypt the password used to encrypt the content. The part that decrypts the document must have the other key.

Below is an example of XML Encryption. This is the original document:

```
<?xml version='1.0'?>
<PaymentInfo xmlns='http://example.org/paymentv2'>
  <Name>John Smith</Name>
  <CreditCard Limit='5,000' Currency='USD'>
    <Number>4019 2445 0277
    5567</Number>
    <Issuer>Example Bank</Issuer>
    <Expiration>04/02</Expiration>
  </CreditCard>
</PaymentInfo>
```

Figure 2: Original document with XML Encryption

This is the document after encrypt the credit card (Figure 3). In this example we have only the encrypted information and we have no information about the key or the encryption algorithm.

```
<?xml version='1.0'?>
<PaymentInfo xmlns='http://example.org/paymentv2'>
  <Name>John Smith</Name>
  <CreditCard Limit='5,000' Currency='USD'>
    <Number>
      <EncryptedData
        xmlns='http://www.w3.org/2001/04/xmlenc#'
        Type='http://www.w3.org/2001/04/xmlenc#Content'>
          <CipherData>
            <CipherValue>A23B45C56</CipherValue>
          </CipherData>
        </EncryptedData>
    </Number>
    <Issuer>Example Bank</Issuer>
    <Expiration>04/02</Expiration>
  </CreditCard>
</PaymentInfo>
```

Figure 3: Encrypted document with XML Encryption

3.2. XMLSignature

It is a standard of digital signature of files, not necessarily XML files. However, the final file is always an XML document. It requires digital certificates and public and private keys for its operation. There are three alternatives:

- Envelope: The result is the original XML file to which a signature element is added in the XML file itself.
- Enveloping: The result is an XML file with the signature, and within it, there are the original elements of the original XML file.
- Detached: The result is the original file and, separately, an XML file with the signature of that file.

It really does not matter which one to use. They are simply different ways of organizing the generated signature.

```
<Signature xmlns="http://www.w3.org/2000/09/xmldsig#">
  <SignedInfo xmlns="http://www.w3.org/2000/09/xmldsig#">
    <CanonicalizationMethod
      Algorithm="http://www.w3.org/TR/2001/REC-xml-c14n-20010315"
      xmlns="http://www.w3.org/2000/09/xmldsig#" />
    <SignatureMethod
      Algorithm="http://www.w3.org/2000/09/xmldsig#rsa-sha1"
      xmlns="http://www.w3.org/2000/09/xmldsig#" />
    <Reference URI=""
      xmlns="http://www.w3.org/2000/09/xmldsig#">
      <Transforms
        xmlns="http://www.w3.org/2000/09/xmldsig#">
        <Transform
          Algorithm="http://www.w3.org/2000/09/xmldsig#enveloped-signature"
          xmlns="http://www.w3.org/2000/09/xmldsig#" />
        </Transforms>
      <DigestMethod
        Algorithm="http://www.w3.org/2000/09/xmldsig#sha1"
        xmlns="http://www.w3.org/2000/09/xmldsig#" />
      <DigestValue
        xmlns="http://www.w3.org/2000/09/xmldsig#">
        Olyx+K28+cp7kuUgcnANiTBdUJwg=
      </DigestValue>
    </Reference>
  </SignedInfo>
  <SignatureValue xmlns="http://www.w3.org/2000/09/xmldsig#">
    ZvZrUd7G4mZsDnBavbnZoFUmms5J2OBdkQ+looDLn95ndGydrq6uPQ==
  </SignatureValue>
  <KeyInfo xmlns="http://www.w3.org/2000/09/xmldsig#">
    <X509Data xmlns="http://www.w3.org/2000/09/xmldsig#">
      <X509Certificate
        xmlns="http://www.w3.org/2000/09/xmldsig#">
        IICmDCCAYCEfmm8wCwYHkOZLzjgEAWUAMDIXCzAJBgNVBAYTAKVTRREwDwYDQkEwBdXRi
        pYTEQMA4GA1UEAxMHVXN1YXJpb2ZaEwF0w0OAZjMjUyMTUyVjFvYzA0MjUyMTUyMTUyMTUy
        BgNVBAYTAKVTRREwDwYDQkEwBdXRiBnRwYTEQMA4GA1UEAxMHVXN1YXJpb2ZaEwF0w0OAZj
        MjUyMTUyMTUyMTUyMTUyMTUyMTUyMTUyMTUyMTUyMTUyMTUyMTUyMTUyMTUyMTUyMTUyMTUy
        BgcqhkiOQAQBMIBHwKBggQD9f1OBHxUSKVLFSpw7OTn9hG3UjzVRAADHj+AtIEmaUVQJCJR+1k9
        Jv6v8X1ujD2y5tVbnEBO4AdNGyZmC3a5iQpaSfn+gEaxAwk+7qdf+8Yb+Dx58aophUPBPuD
        9fPHsMCNVQTWHaRMvZ1864Ydcq77IAxmd0UgBxwVVAJdgU18VlwwMspk5gqLrhAwwWbZ1AoGB
        AFPhIXWmz3ey7yPda4v715ik+7+jrqMXTAS9B4JnUjVXjrrUjUjmcQcGgYC0SRZxi+hmkBYT
        t88Jm0zIpuE8FnqLVHyNkOCjrh4rs6Z1kWbjfhw6ITVj8ttiegEK08yk8b6oUJZCJqFP4VrlnwaS
        I2ZegHtYJWQBDv+z0kqA4GEAAKBgDUPDwDZFXmZha74VNmgyFslLM01Wkwi7nbt9UJFJALk71
        iFpozeZMP2u05oYst2nbnkCsIhziuaNjmykz3Vf3+Pml3sQE58SxwJBRUJMEZUTA2006WD3x9N9
        IZ3ybc4WQimB8ekIjyBExSkSTueAzvTA8hN2+Rvgh8MarOzmMAsGByqGSM44BAMFAAMvADAsAhQH
        4+nQZdFwvstyQrqt02h9MJEGlUEYVDvxygkCmrlIA0sQLtaCs0Qo=
      </X509Certificate>
    </X509Data>
    <KeyValue xmlns="http://www.w3.org/2000/09/xmldsig#">
      <DSAKeyValue
        xmlns="http://www.w3.org/2000/09/xmldsig#">
        <P xmlns="http://www.w3.org/2000/09/xmldsig#">
          HTRV8mZgt2uZUkWkn5oBHSjlsJPu6nXrfGG7V+GqkYVDwT7gbTxR7DAJUE1oWkL2df0u
          K2HXkUylgMzndFIAcc=
        </P>
        <Q xmlns="http://www.w3.org/2000/09/xmldsig#">
          I2BQjUjC8ykyrmCouECBYHPU=
        </Q>
        <G xmlns="http://www.w3.org/2000/09/xmldsig#">
          9+GghdabPd7LvkTcNrhXuXmU7v6OuqC+VdMCz0HgmDRWVVeOutRZT+ZxBxCBGLRJRfEj6EwoFh3
          zwkyjMm4TwwEotUJf04K0uHiuzpnWRBqNcJohNWLx+2J6ASQZkZxvqhRklm9ghWUwFbPK
          Z16Ae1UJZAFMO7PSSo=
        </G>
        <Y xmlns="http://www.w3.org/2000/09/xmldsig#">
          N8PDEnkVcytmFrvhU2aDwYUJszTXADXudu31QVMkAuTWWI+mjN5kw/a7RkYh3advGkZWHOKK
          5o0mKTNyNVf4+YvexARLxLHAKFFQwTZRMDDTTPYIE3L2Jn3KJbZCKYH4ogniEHFKTIO54DO
          9MDwc3b5G+ChXExo7OE=
        </Y>
      </DSAKeyValue>
    </KeyValue>
  </KeyInfo>
</Signature>
```

Figure 4: XML obtained after applying XMLSignature

A signature as must contain, accordingly with XML Signature:

- Canonicalization method: Two XML documents are equivalent if they represent the same information. A method of canonicalization transforms an XML document into another equivalent one. All XML documents equivalent, since they are canonicalized using the same method, result in the same XML. It is applied before signing. If a method is not specified, one of them is assigned by default.
- Reference: There may be several references within a single firm. In each reference the part of the document that is signed and the hash algorithm used are indicated. A summary algorithm generates a sequence of bytes of fixed length from contents of arbitrary length. This sequence of bytes is different for each content.
- Information on the key signature can optionally include the data necessary for validation. This part can indicate the public key directly, through a sequence of characters that identifies it or through a URL. Additionally it can also have more information about who has signed it: name, organization, country...

• Transformations: It is possible that what want to be signed is not the complete document, but some information of it. With the changes you can do almost anything, from selecting only certain parts, to change the structure of XML, or to include other XML fragments. If it is not necessary to apply any transformation before signing you can skip this part.

The end result of applying XML Signature is an XML element of the form shown in Figure 4.

4. PROPOSAL

4.1. Encryption

Here are presented 4 equivalent representations of a PN: as graph, code, matrix, and PNML.

Graph:

(see figure 1)

PNML:

<pre><?xml version="1.0" encoding="UTF-8"?> <pnml> <net type="http://www.informatik.hu-berlin.de/top/pntd/ptNetb" id="noID"> <place id="p3"> <name> <text>p3</text> </name> </place> <place id="p2"> <name> <text>p2</text> </name> </place> <place id="p1"> <name> <text>p1</text> </name> </place> <transition id="t3"> <name> <text>t3</text> </name> </transition> <transition id="t2"> <name> <text>t2</text> </name> </transition> <transition id="t1"> <name> <text>t1</text> </name> </transition> <arc id="a9" source="t3" target="p3"> <inscription></pre>	<pre><text>1</text> </inscription> </arc> <arc id="a11" source="p2" target="t3"> <inscription> <text>1</text> </inscription> </arc> <arc id="a1" source="p1" target="t1"> <inscription> <text>1</text> </inscription> </arc> <arc id="a2" source="t1" target="p2"> <inscription> <text>1</text> </inscription> </arc> <arc id="a3" source="t1" target="p3"> <inscription> <text>1</text> </inscription> </arc> <arc id="a4" source="p3" target="t2"> <inscription> <text>1</text> </inscription> </arc> <arc id="a7" source="t2" target="p1"> <inscription> <text>1</text> </inscription> </arc> </net> </pnml></pre>
--	---

Figure 7: Example of a Petri net defined by PNML.

Matrix:

	p1	p2	p3
t1	-	1	1
t2		1	0
t3		0	-1

Figure 6: Example of a Petri net defined by its incidence matrix.

Code:

```
if (p1>0) then
  p1 <- p1 - 1
  p2 <- p2 + 1
  p3 <- p3 + 1
if (p2>0) then
  p2 <- p2 - 1
  p3 <- p3 + 1
if (p3>0) then
  p3 <- p3 - 1
  p1 <- p1 + 1
```

Figure 5: Example of a Petri net defined by code.

The proposed solution is to use the PNML representation of Petri nets, and from it to use XMLEncryption standards for encryption and XMLSignature for the signature.

A little example will be developed to show that it reduces to perform the signature and / or encrypted operations on a Petri subnet of the original network, and that also the matrix associated to the network more appropriated can be selected.

A Petri subnet is a submatrix of the matrix associated to the network. In the matrix, the rows are associated to transitions, and the columns to nodes. This way we can easily show that there exists a single matrix associated to a PN, but if we exchange two rows or two columns of the matrix, the result also defines the same Petri net (or more precisely, it defines an equivalent Petri net).

	p1	p2	p3		p1	p2	p3		p2	p1	p3
t1	-1	1	1	t1	-1	1	1	t1	1	-1	1
t2	1	0	-1	t3	0	-1	1	t3	-1	0	1
t3	0	-1	1	t2	1	0	-1	t2	0	1	-1

Figure 8: Example of equivalent Petri nets

All these would be equivalent representations. It can be shown that indeed an equivalence relation is between matrices M and M'. The equivalence relation is M r M 'if we can get from M to M' by swaping rows and/or columns. To test compliance reflexive relations (reflexive), symmetric (symmetric) and transitive (transitive).

Let T be the set of transformations of a matrix of order nxm (n rows and m columns). Let be Tfij, with i <= n, and j <= n, the transformation that exchanges the row i with row j. Obviously Tfij = Tfji. Similarly, we define Tckl with k <= m and l <= m, as the transformation that exchanges the column k to column l. Similarly, Tckl = Tclk. Let T = {Tabs, s> = 1 | = Tfij Tabs, a = i, b = j, i <= n, j <= n or = Tckl Tabs, a = k, b = l, k <= m, l <= m} where s is a sequence of consecutive natural numbers beginning with 1. Thus Tab1 is the first transformation, Tab2 the second, ..., and Tabn is the nth. Thus we have an ordered set of transformations applied to one mxn matrix in a particular order. It can be shown that the order in which transformations are

applied does not influence the final result, but it is not necessary for our purpose.

Let r be the following relationship we want to study: A matrix M is related to an N , $M r N$, if you can get from M to N with a finite number of transformations. A transformation will be an exchange of two rows or two columns. Let show that this relation is of equivalence:

- Reflexive: $M r M$. Obviously, since you do not need any exchange of rows or columns. In this case $T = \emptyset$.

- Symmetric: if $M r M'$ then $M' r M$. If from M p exchange operations are made from in rows and / or columns to arrive at M' , from M' the same operations are performed in reverse order in order to arrive at M . Thus if $T = \{T_{abs}, s = 1 .. p\}$ and the number of transformations is $p > 0$, then $T = \{T_{ab}(n-s+1), s = 1 .. p\}$ is the set of transformations leading from M' to M .

- Transitive: if $M r M'$ and $M' r M''$ then $M r M''$. Let $T1 = \{T_{abs}, s = 1 .. p\}$ of size p the set of transformations that lead from M to M' and let $T2 = \{T_{abr}, s = 1 .. q\}$ q -sized, the set of transformations that lead from M' to M'' . Then $T = \{T_{abt} \text{ with } t = s \text{ if } t \leq p \text{ and } t = p + r \text{ if } t > p\}$ is a sequence of transformations that lead from M to M'' .

Therefore it is shown that r is an equivalence relation. Thus, we can say that a Petri net corresponds to an equivalence class of the relation r . Thus, we can choose on what representative of the class to make the transformations.

With this in mind, given a matrix representing a Petri net, if we eliminate some row and / or a column, a valid subnet results. This subnet is what we want to process. As a row is associated to a place and a transition to a column, any subset of places and transitions can be selected as a valid subnet.

Following the selection of places and transitions that are to be processed, a matrix having first nodes and transitions th eones that we want to encrypt can be chosen as representative of the Petri net. In our case:

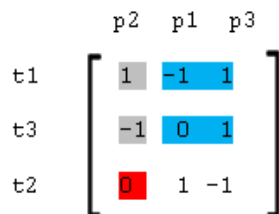


Figure 9: Petri net representing the equivalence class

Interpreting the Figure 9, different parts can be seen:

- The gray part corresponds to the parts that are completely into the process. In this case $p2$, $t1$, $t3$, the arc from $t1$ to $p2$, and the arc from $p2$ to $t3$. It is denoted as hidden subnet.

- The blue arcs indicate arcs (> 0) that part from a hidden transition but become to a visible place or from a visible place to a hidden transition (< 0). It is denoted as hidden transitions subnet.

- The red part corresponds to the arcs starting from a hidden node to a visible transition (< 0) or from a visible transition to a hidden node (> 0). It is denoted as hidden nodes subnet.

- The uncoloured part are nodes, transitions, and arcs that are not hidden. It is denoted as visible subnet.

Thus, a Petri net that wants to be encrypted can be represented by a matrix as follows:

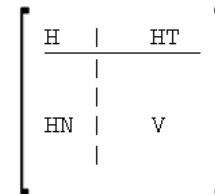


Figure 10: Parts in an ordered encrypted Petri net

Being H the hidden subnet, HT the hidden transitions subnet, HN the hidden nodes subnet, and V the visible subnet.

What will be encrypted, in order to not to give information about the structure, corresponds to the matrices H , HT and HN , which are those affected by any hidden element.

Let us now see how would be the PNML representation associated with this subnet. Within the document PNML three main elements appears:

- place: defines a place with an id and a name (a column of the matrix).

```
<place id="p1">
  <name>
    <text>nodo 1</text>
  </name>
</place>
```

- transition: defines a transition, also with an id and a name (a row of the matrix).

```
<transition id="t2">
  <name>
    <text>transicion 2</text>
  </name>
</transition>
```

- arc: defines an arrow with an ID, from one node to a transition or from a transition to a node, defined by its id (one matrix element different from 0).

```
<arc id="a2" source="t1" target="p2">
  <inscription>
    <text>1</text>
  </inscription>
</arc>
```

Note that no matter the order in which these elements are in the PNML file. Just as there are several matrices that represent the same net, the same goes for files PNML. The order in the file provides no information. It is similar that appear first all places, then all the transitions and finally all the arcs, that all of them

interspersed with each other. Thus, once we have defined what is the subset of nodes and transitions that we want to encrypt, what will be done in the PNML file is to join all these nodes, transitions and arcs that contain as the origin or the end any of these nodes and transitions, and include a new XML element called 'subnet' within the PNML document, with a concrete and unique id. Later it would be indicated what does this id is used for.

Therefore, in the example, the subset would be {p2, t1 and t3}; the new file applying these changes to the original PNML would result this way, grouping these three elements together with the arcs that have one of them as a source or destination; that is, everything associated with the matrices H, HT and HN. The visible subnet, V, remains out of this item.

```
<?xml version="1.0" encoding="UTF-8"?>
<pnml>
  <net type="http://www.informatik.hu-berlin.de/top/pntd/ptNetb" id="noID">
    <subnet id="subnet1">
      <place id="p2">
        <name>
          <text>p2</text>
        </name>
      </place>
      <transition id="t1">
        <name>
          <text>t1</text>
        </name>
      </transition>
      <transition id="t3">
        <name>
          <text>t3</text>
        </name>
      </transition>
      <arc id="a9" source="t3" target="p3">
        <inscription>
          <text>1</text>
        </inscription>
      </arc>
      <arc id="a11" source="p2" target="t3">
        <inscription>
          <text>1</text>
        </inscription>
      </arc>
      <arc id="a1" source="p1" target="t1">
        <inscription>
          <text>1</text>
        </inscription>
      </arc>
      <arc id="a2" source="t1" target="p2">
        <inscription>
          <text>1</text>
        </inscription>
      </arc>
    </subnet>
  </net>
</pnml>
```

Figure 11: PNML of the example, grouping the elements of H, HT y HN.

Notice that this PNML file does not meet the official PNML grammar, but these changes can always be undone to leave the original PNML file (that is, the transformation is reversible). The goal is not that the encrypted and/or signed document meets the grammar, but that, after decrypt and/or check the signature, the original file can be obtained. To obtain the original file it is only necessary to take the elements from the 'subnet' labels.

Note that multiple subnets can be encrypted just considering all of them as a single subnet with the union of the nodes and transitions of them.

Once we got the PNML file in this format, we can apply the encryption via XML Encryption. The final file is shown in Figure 12.

Notice that the content of element 'subnet' no longer exists and has been replaced by an element 'EncryptedData'. The subnet has already been

encrypted. A possible graph representation would be Figure 13, where the subnet is the visible subnet V and the subnets H, HT, and HN are hidden in the black box.

```
<?xml version="1.0" encoding="UTF-8"?>
<pnml>
  <subnet id="subnet1">
    <net type="http://www.informatik.hu-berlin.de/top/pntd/ptNetb" id="noID">
      <xenc:EncryptedData xmlns:xenc="http://www.w3.org/2001/04/xmlenc#"
        Type="http://www.w3.org/2001/04/xmlenc#Element">
        <xenc:EncryptionMethod
          Algorithm="http://www.w3.org/2001/04/xmlenc#aes128-cbc"
          xmlns:xenc="http://www.w3.org/2001/04/xmlenc#" />
        <xenc:CipherData
          xmlns:xenc="http://www.w3.org/2001/04/xmlenc#" />
        <xenc:CipherValue
          xmlns:xenc="http://www.w3.org/2001/04/xmlenc#"
          W1r1njyJlYOM9IA9qCw6GVk2L4pUjQD2GGVoU
          9lVZ0wKqH18y3l3G8Fy4l5K3G8g1e1HRFqe
          7Rt8FiXZgGMeYnQp6oB6ckKp3KFKVqatc9
          rAVzOgC7XAwIoe61HRFqe6RRVzXjNM GY8FY4l5K
          dI8NVPQmUSD7NRtrR6YnQp6oB6ckKp3
          SWr1njyJlYOM9IA9qCw6GVk2L4pUjQD2GGVoU
          9lVZ0wKqH18y3l3G8Fy4l5K3G8g1e2xN4U7x
          7Rt8FiXZgGMeYnQp6oB6ckKp3KFRVzXjN
          AtVzOgC7XAwIoe61HRFqe6RRVzXjNMLU5ZgGMeYn
          lY8NVPQmUSD7NRtrR6YnQp6oB6G8FY8F
          dW1r1njyJlYOM9IA9qCw6GVk2L4pUjQ3l3GY8FY
          9lVZ0wKqH18y3l3G8Fy4l5K3G8g1e2xN4U7x
          b7Rt8FiXZgGMeYnQp6oB6ckKp3KFKG8g1e2
          nAVzOgC7XAwIoe61HRFqe6RRVzXjNMLU5T1HRFqe
          L18NVPQmUSD7NRtrR68=
        </xenc:CipherValue>
        </xenc:CipherData>
      </xenc:EncryptedData>
    </subnet>
  </net>
</pnml>
```

Figure 12: PNML after applying XML Encryption.

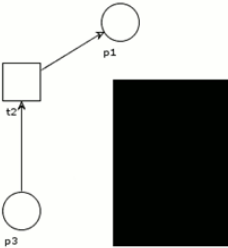


Figure 13: Graphic representation of the Petri net after applying XML Encryption.

Once encrypted information, even the number of nodes, transitions, and arcs that are contained in that black box is unknown. The final matrix associated is represented in figure 14. Black areas are those for which we have no information, and even the size of the matrix is unknown.

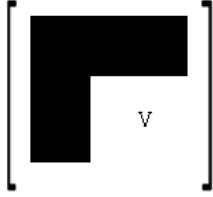


Figure 14: Matrix representation of the Petri net after applying XML Encryption.

Note that in the hidden network no arc comes in or comes out. This is a security decision. However, we could define the arcs that go from inside out or from outside in, hiding the place or the transition of destination in the hidden subnet, replacing the node / transition id of origin or destination inside the hidden net by the own id of the net. Thus, the final file would be as follows:

- Ekelhart A, Fenz S, Goluch G, et al. XML security - A comparative literature review. 2008. JOURNAL OF SYSTEMS AND SOFTWARE Volume: 81 Issue: 10 Pages: 1715-1724
- Meadors K. Secure electronic data interchange over the Internet. 2005 IEEE INTERNET COMPUTING Volume: 9 Issue: 3 Pages: 82-89
- Selkirk A. XML and security. 2001 BT TECHNOLOGY JOURNAL Volume: 19 Issue: 3 Pages: 23-34
- Selkirk A. Using XML security mechanisms. 2001 BT TECHNOLOGY JOURNAL Volume: 19 Issue: 3 Pages: 35-43
- Nordbotten NA. XML and Web Services Security Standards. 2009 IEEE COMMUNICATIONS SURVEYS AND TUTORIALS Volume: 11 Issue: 3 Pages: 4-21 Published: 2009
- Brooke PJ, Paige RF, Power C. Document-centric XML workflows with fragment digital signatures. 2010 SOFTWARE-PRACTICE & EXPERIENCE Volume: 40 Issue: 8 Pages: 655-672

PETRI NET TRANSFORMATION FOR DECISION MAKING: COMPOUND PETRI NETS TO ALTERNATIVES AGGREGATION PETRI NETS.

Juan Ignacio Latorre-Biel^(a), Emilio Jiménez-Macías^(b)

^(a) Public University of Navarre. Department of Mechanical Engineering, Energetics and Materials. Campus of Tudela, Spain

^(b) University of La Rioja. Industrial Engineering Technical School. Department of Electrical Engineering. Logroño, Spain

^(a)juanignacio.latorre@unavarra.es, ^(b)emilio.jimenez@unirioja.es

ABSTRACT

The design and operation of discrete event systems (DES), including their control, require taking decisions in order to guarantee an expected behavior. Usually, this behavior can be characterized by means of performance measurements. To take decisions may require choosing the best solution that optimizes a cost function and complies with certain restrictions, i.e. solving an optimization problem. In the field of DES modeled by Petri nets, it is a classical problem to optimize the initial marking and the sequence of priority assignment to the firing of transitions involved in conflicts (Piera and Music 2011). This problem may be solved by means of simulation and by optimization based on simulation (Piera *et al.*, 2004). The second approach can use a heuristic search to find the best configurations to solve. On the other hand, in the first approach is a human operator who should make this choice and can skip the best solutions to be tested. In the cases, less studied in the literature, of requiring an optimization of the structure of the Petri net, the classical approach is similar to the simulation in the previous problem: several feasible structures are chosen and they are simulated or optimized. If the human operator does not choose the best solutions, the result of the decision taking may be poor. A field of research that has taken the interest of the authors consists of applying a heuristic search to find the best structure for a Petri net (Latorre *et al.* 2009). This kind of optimization problem requires an adequate formalism as the compound Petri nets or the alternatives aggregation Petri nets to perform an efficient solving process (Latorre *et al.* 2011). The transformation from the first formalism to the second one is presented in this paper and illustrated with an example. Its utility arises when it is required to compare the performance of both formalisms for a particular case or when it is easier to model a DES as compound Petri

net but the optimization process is based on the second formalism.

Keywords: decision support system, compound Petri net, alternatives aggregation Petri net, simulation, optimization.

1. INTRODUCTION

Decision making on discrete event systems is a common and difficult task associated to the design and operation of discrete event systems. The complexity in the behaviour of most discrete event systems associated to technological solutions requires the development and use of decision support systems (Jiménez *et al.*, 2005). Even with the help of computers, the analysis of the behaviour of DES under any possible scenario is usually intractable. (Piera *et al.*, 2004). For this reason diverse techniques have been created and applied from state space reduction to the use of metaheuristics to perform efficient searches in the state space. One interesting methodology broadly analysed and with important results that improve the verification, validation and performance measurement of the DES is the net transformation (Berthelot, 1987), (Silva, 1993).

This paper is focussed on the net transformation applied to two formalisms that allow representing an undefined Petri net. The undefined Petri nets constitute an abstraction of the model of an undefined discrete event system with alternative structural configurations among which one should be chosen for the definition of the DES. The two formalisms that will be considered are the compound Petri nets and the alternatives aggregation Petri nets (Latorre *et al.* 2009). Both formalisms may arise from a natural approach to model a DES with alternative structural configurations: the alternative Petri nets. When there are structural similarities between the alternative structures it is likely that the representation of the undefined

Petri net in the form of a compound Petri net or an alternatives aggregation Petri net is more compact than considering the equivalent set of alternative Petri nets, that is to say different models for every structure. A more compact model usually implies that the exploration of the space state of the model of the DES is more efficient than the search performed by means of the set of alternative Petri nets (Latorre *et al.*, 2010b).

On the other hand, the research is active in the field of determining the conditions where the compound Petri net is more efficient than an equivalent alternatives aggregation Petri net and vice versa. For this reason, the comparison between the two formalisms may lead to transformations among them. Furthermore, the compound Petri nets may serve of “formalism bridge” or “origin” towards the process of obtaining an alternatives aggregation Petri net or a disjunctive coloured Petri net. This last formalism can be obtained almost immediately from an alternatives aggregation Petri net and is very useful because of the software developed for the CPN that can be applied for the validation, verification or performance optimization of this formalism.

2. DEFINITIONS

The compound Petri nets can be defined in the following way:

Definition 1. Compound Petri net.

A compound Petri net is a 7-tuple $R^c = \langle P, T, F, w, \mathbf{m}_0, S_\alpha, S_{val\alpha} \rangle$, where

- i) S_α is the set of undefined parameters of R^c .
- ii) $S_{str\alpha} \neq \emptyset$ is the set of undefined structural parameters of R^c , such that $S_{str\alpha} \subseteq S_\alpha$. Notice that S_α is the set of undefined parameters of R^c .
- iii) $S_{val\alpha}$ is the feasible combination of values for the undefined parameters .

□

A compound Petri net can be considered as a parametric Petri net with undefined structural parameters.

The structural parameters refer to the elements of the incidence matrix of a Petri net. If a Petri net has undefined structural parameters it has a structure with certain freedom degrees that should be specified by a decision from the set of feasible combinations of values for them. In summary, a the undefined structural parameters are present in models that correspond with DES with undefined structure, in process of being designed, modified or controlled.

On the other hand, an alternatives aggregation Petri net may be defined as indicated below:

Definition 2. Alternatives aggregation Petri net system.

An alternatives aggregation Petri net system, R^A , is defined as the 8-tuple:

$$R^A = \langle P, T, \text{pre}, \text{post}, \mathbf{m}_0, S_\alpha, S_{val\alpha}, S_A, f_\Lambda \rangle$$

Where,

- P is the set of places.
- T is the set of transitions.
- pre is the pre-incidence matrix, also called input incidence matrix.
- post is the post-incidence matrix, also called output incidence matrix.
- \mathbf{m}_0 is the initial marking that represents the initial vector of state and is usually a function of the choice variables.
- S_α is a set of undefined parameters.
- $S_{val\alpha}$ is the set of feasible combination of values for the undefined parameters in S_α .
- S_A is a set of choice variables such that $S_A \neq \emptyset$ and $|S_A| = n$.
- $f_\Lambda: T \rightarrow \mathbf{f}(a_1, \dots, a_n)$ assigns a function of the choice variables to each transition t such that $\text{type}[f_\Lambda(t)] = \text{Boolean}$.

□

Where a set of choice variables is given by:

$$\text{Let } c_{str} \in C_{str} = \{1, 2, \dots, m_{strq}\} \subseteq \mathbb{N}^*$$

A set of choice variables can be defined as $S_A = \{a_1, a_2, \dots, a_{mstrq} \mid \exists! a_i=1, i \in C_{str} \wedge a_j=0 \forall j \neq i, j \in C_{str}\}$

Furthermore, the dynamic behaviour of an alternatives aggregation Petri net is given by an enabling rule that differs slightly from most of the formalisms based on Petri nets. The firing rule is the one of a generalized Petri net.

Definition 3. Enabled transition.

Given an alternatives aggregation Petri net R^A with an associated set of choice variables $S_A = \{a_1, a_2, \dots, a_n\}$, let us consider the following decision:

$$a_i = 1 \Rightarrow a_i = 0 \forall j \in \mathbb{N}^* \text{ such that } 1 \leq j \leq n \wedge j \neq i$$

A transition $t_j \in T$ in an alternatives aggregation Petri net is said to be enabled if

$$m_i \geq \text{pre}(p_i, t_j) \forall p_i \in {}^o t_j \wedge f_\Lambda(t_j) = 1$$

□

3. TRANSFORMATION ALGORITHM

The transformation from a compound Petri net to an alternatives aggregation Petri net can be performed by the following algorithm:

A compound Petri net is a very compact representation of a discrete event system with an undefined structure. Nevertheless, it contains some undefined structural parameters that require to complement the model with a set of feasible combinations of values for the undefined structural parameters. On the other hand, the alternatives aggregation Petri net, might be more compact for certain systems and this fact may lead to more efficient optimization algorithms. The main reason is that the alternatives aggregation Petri net can profit from similarities in the subnets of the different structures that can be chosen for the original DES and on the other hand it can be constructed in a way that it lacks completely of undefined structural parameters. This last property implies that the model does not require an additional set of feasible combinations for the undefined structural parameters.

For these reasons it is going to be presented an algorithm to perform the transformation from a compound Petri net into an alternatives aggregation Petri net, where both models are equivalent or, what is the same, their graphs of reachable markings are isomorphous.

Let us consider a compound Petri net R^c .

Step 1.

Define

$$\{ S_{valstr\alpha}(R_1), S_{valstr\alpha}(R_2), \dots, S_{valstr\alpha}(R_{n_r}) \},$$

a partition of $S_{valstr\alpha}(R^c)$.

Define a set of choice variables from R^c and

$$\{ S_{valstr\alpha}(R_1), S_{valstr\alpha}(R_2), \dots, S_{valstr\alpha}(R_{n_r}) \}$$

in the way $S_A = \{ a_1, a_2, \dots, a_{n_r} \mid \exists! a_i=1, 1 \leq i \leq n_r, a_j=0 \forall j \neq i \}$, where (by definition)

$\text{card}(\{ S_{valstr\alpha}(R_1), S_{valstr\alpha}(R_2), \dots, S_{valstr\alpha}(R_{n_r}) \}) =$

$$\text{card}(S_A) = n_r.$$

Step 2.

Create a bijection between the elements of the partition

$$\{ S_{valstr\alpha}(R_1), S_{valstr\alpha}(R_2), \dots, S_{valstr\alpha}(R_{n_r}) \}$$

and the elements of S_A (choice variables).

Step 3.

Replicate every transition t_i into a set

$$\{ t_i^1, t_i^2, \dots, t_i^{n_r} \}, \text{ where } n_r = \text{card}(S_A) \text{ and}$$

$i) \text{ pre}(p_j, t_i^q), \text{ post}(t_i^q, p_k) \in S_{valstr\beta}(R_q)$, where $S_{valstr\alpha}(R_q) \cup S_{valstr\beta}(R_q) = S_{valstr\gamma}(R_q)$ and the couple of sets $S_{valstr\alpha}(R_q)$ and $S_{valstr\beta}(R_q)$ stand for the set of feasible combination of values for the undefined structural parameters of R^c and for the defined ones respectively.

$ii) \text{ The choice variable } a_q \text{ is associated to the transition } t_i^q \text{ as a boolean condition to allow its enabling.}$

Step 4.

Transform the resulting AAPN from **step 3** into an equivalent PN by removing the non-connected places and transitions.

Step 5.

Apply a reduction rule to quasi-identical transitions associated to different choice variables, obtaining an equivalent AAPN with a reduced set of transitions.

Step 6.

Apply simplification rules to remove the unnecessary choice variables and functions of choice variables to the transitions of the AAPN. As a result a simpler AAPN than the original basic AAPN is expected. □

4. APPLICATION EXAMPLE OF THE TRANSFORMATION ALGORITHM

In order to illustrate the application of this algorithm, the following examples can be considered.

4.1. Example 1.

In the **figure 1** it can be seen a compound Petri net in both representations, a graphical one and another matrix-based one based on the incidence matrix.

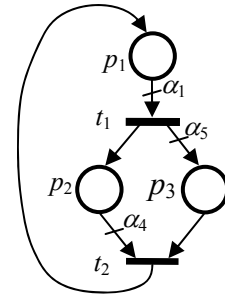


Fig. 1. Compound PN.

The set of structural parameters of the compound Petri net R^c is:

$$S_{str\gamma}(R^c) = \{ \alpha_1, \beta_2, \beta_3, \alpha_4, \alpha_5, \beta_6 \} = \\ S_{str\alpha}(R^c) \cup S_{str\beta}(R^c), \text{ where} \\ S_{str\alpha}(R^c) = \{ \alpha_1, \alpha_4, \alpha_5 \} \text{ and } S_{str\beta}(R^c) = \\ \{ \beta_2, \beta_3, \beta_6 \}$$

On the other hand, the set of feasible values for the structural parameters of R^c can be written as follows:

$$S_{valstr\beta}(R^c) = \{ (1,1,1) \} \\ S_{valstr\alpha}(R^c) = \{ (1,0,1), (0,1,0), (2,0,1), (0,1,2) \}$$

$$\mathbf{W}(R^c) = \begin{pmatrix} t_1 & t_2 \\ -\alpha_1 & \beta_2 \\ \beta_3 & -\alpha_4 \\ \alpha_5 & \beta_6 \end{pmatrix} \begin{matrix} p_1 \\ p_2 \\ p_3 \end{matrix} \\ = \begin{pmatrix} t_1 & t_2 \\ -\alpha_1 & 1 \\ 1 & -\alpha_4 \\ \alpha_5 & 1 \end{pmatrix} \begin{matrix} p_1 \\ p_2 \\ p_3 \end{matrix}$$

Fig. 2. Matrix-based representation of the compound Petri net.

Finally, it is possible to determine the set of feasible values for every undefined structural parameter of R^c .

$$S_{val\alpha_1}(R^c) = \{ 0,1,2 \}$$

$$S_{val\alpha_4}(R^c) = \{ 0,1 \}$$

$$S_{val\alpha_5}(R^c) = \{ 0,1,2 \}$$

The first partition of $S_{valstr\alpha}(R^c)$ has order two and can be represented as follows:

$$\Pi_1(S_{valstr\alpha}(R^c)) = \{ S_{1valstr\alpha}(R^c), S_{2valstr\alpha}(R^c) \} \\ S_{valstr\alpha}(R^c) = S_{1valstr\alpha}(R^c) \cup S_{2valstr\alpha}(R^c) \\ S_{valstr\alpha}(R^c) = \{ (1,0,1), (0,1,0), (2,0,1), (0,1,2) \} \\ S_{1valstr\alpha}(R^c) = \{ (1,0,1), (0,1,0) \} \\ S_{2valstr\alpha}(R^c) = \{ (2,0,1), (0,1,2) \}$$

In order to know the number of undefined structural parameters associated to every subset of the partition, it is necessary to analyse every parameter of $S_{valstr\alpha}(R^c)$.

On one hand, the first subset of the partition will be considered:

$$S_{1val\alpha_1}(R^c) = \{ 0,1 \}$$

$$S_{1val\alpha_4}(R^c) = \{ 0,1 \}$$

$$S_{1val\alpha_5}(R^c) = \{ 0,1 \}$$

As a consequence there will be three undefined structural parameters associated to this subset of the partition, since any of them can take values from a set with more than one element:

$$S_{1str\alpha}(R^c) = \{ \alpha_1^1, \alpha_4^1, \alpha_5^1 \}$$

On the other hand, the second subset of the partition will lead to:

$$S_{2val\alpha_1}(R^c) = \{ 0,2 \}$$

$$S_{2val\alpha_4}(R^c) = \{ 0,1 \}$$

$$S_{2val\alpha_5}(R^c) = \{ 1,2 \}$$

In this case there will be another new three undefined structural parameters associated to this subset of the partition, since any of them can take values from a set with more than one element:

$$S_{2str\alpha}(R^c) = \{ \alpha_1^2, \alpha_4^2, \alpha_5^2 \}$$

As a result, it is possible to see how this partition of $S_{valstr\alpha}(R^c)$, $\Pi_1(S_{valstr\alpha}(R^c))$, from a compound Petri net with three undefined structural parameters leads to a representation with six undefined structural parameters. This representation can be a set of compound alternative PN or an aggregations alternative Petri net. The AAPN that results from the replication of the transitions of the compound PN R^c according to this partition is represented in **figure 3**.

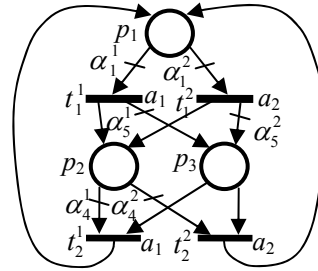


Fig. 3. Graphical representation of an AAPN obtained from a first partition of $S_{valstr\alpha}(R^c)$.

$$\mathbf{W}(R^d) = \begin{pmatrix} t_1^1 & t_1^2 & t_2^1 & t_2^2 \\ -\alpha_1^1 & -\alpha_1^2 & 1 & 1 \\ 1 & 1 & -\alpha_4^1 & -\alpha_4^2 \\ \alpha_5^1 & \alpha_5^2 & 1 & 1 \end{pmatrix} \begin{matrix} p_1 \\ p_2 \\ p_3 \end{matrix} \\ \begin{matrix} a_1 & a_2 & a_1 & a_2 \end{matrix}$$

Fig. 4. Matrix-based representation of an AAPN obtained from a first partition of $S_{valstr\alpha}(R^c)$.

As a conclusion, it is possible to state that despite the fact that the original compound Petri net R^c has only three undefined structural parameters $S_{str\alpha}(R^c) = \{ \alpha_1, \alpha_4, \alpha_5 \}$, the resulting AAPN obtained by a replication of the transitions based on this first partition of

$S_{valstr\alpha}(R^c)$ has six undefined structural parameters:

$$S_{str\alpha}(R^d) = S_{1str\alpha}(R^c) \cup S_{2str\alpha}(R^c) = \{ \alpha_1^1, \alpha_1^2, \alpha_4^1, \alpha_4^2, \alpha_5^1, \alpha_5^2 \}$$

4.2. Example 2.

The second partition of $S_{valstr\alpha}(R^c)$ has order two and can be represented as follows:

$$\begin{aligned} \prod_2(S_{valstr\alpha}(R^c)) &= \{ S_{1valstr\alpha}(R^c), S_{2valstr\alpha}(R^c) \} \\ S_{valstr\alpha}(R^c) &= S_{1valstr\alpha}(R^c) \cup S_{2valstr\alpha}(R^c) \\ S_{valstr\alpha}(R^c) &= \{ (1,0,1), (0,1,0), (2,0,1), (0,1,2) \} \\ S_{1valstr\alpha}(R^c) &= \{ (1,0,1), (2,0,1) \} \\ S_{2valstr\alpha}(R^c) &= \{ (0,1,0), (0,1,2) \} \end{aligned}$$

In order to know the number of undefined structural parameters associated to every subset of the partition, it is necessary to analyse every parameter of $S_{valstr\alpha}(R^c)$.

On one hand, the first subset of the partition will be considered:

$$\begin{aligned} S_{1val\alpha_1}(R^c) &= \{ 1, 2 \} \\ S_{1val\alpha_4}(R^c) &= \{ 0 \} \Rightarrow \text{In this subset of the second partition } \alpha_4 \text{ is no longer an undefined structural parameter but a defined one: } \beta_4^1. \\ S_{1val\alpha_5}(R^c) &= \{ 1 \} \Rightarrow \text{In this subset of the second partition } \alpha_5 \text{ is no longer an undefined structural parameter but a defined one: } \beta_5^1. \end{aligned}$$

As a consequence there will be only one undefined structural parameters associated to this subset of the partition, since only α_1^1 can take values from a set with more than one element:

$$S_{1str\alpha}(R^c) = \{ \alpha_1^1 \}$$

On the other hand, the second subset of the partition will lead to:

$$\begin{aligned} S_{2val\alpha_1}(R^c) &= \{ 0 \} \Rightarrow \text{In this subset of the second partition } \alpha_1 \text{ is no longer an undefined structural parameter but a defined one: } \beta_1^1. \\ S_{2val\alpha_4}(R^c) &= \{ 1 \} \Rightarrow \text{In this subset of the second partition } \alpha_4 \text{ is no longer an undefined structural parameter but a defined one: } \beta_4^1. \\ S_{2val\alpha_5}(R^c) &= \{ 0, 2 \} \end{aligned}$$

This case will provide with another single undefined structural parameters associated to the corresponding subset of this second partition, α_5^2 , since it is the only one that can

take values from a set with more than one element:

$$S_{2str\alpha}(R^c) = \{ \alpha_5^2 \}$$

As a result, it is possible to see how this second partition of $S_{valstr\alpha}(R^c)$, $\prod_2(S_{valstr\alpha}(R^c))$, from a compound Petri net with three undefined structural parameters leads to a representation with only two undefined structural parameters. This representation can be a set of compound alternative PN or an aggregations alternative Petri net. The AAPN that results from the replication of the transitions of the compound PN R^c according to this partition is represented in figure 5.

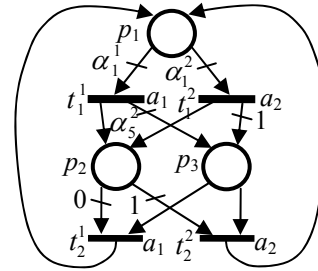


Fig. 5. Graphical representation of an AAPN obtained from a 2nd partition of $S_{valstr\alpha}(R^c)$.

$$\mathbf{W}(R^d) = \begin{pmatrix} t_1^1 & t_1^2 & t_2^1 & t_2^2 \\ -\alpha_1^1 & -\alpha_1^2 & 1 & 1 \\ 1 & 1 & 0 & -1 \\ 1 & \alpha_5^2 & 1 & 1 \\ a_1 & a_2 & a_1 & a_2 \end{pmatrix} \begin{matrix} p_1 \\ p_2 \\ p_3 \end{matrix}$$

Fig. 6. Matrix-based representation of an AAPN obtained from a second partition of $S_{valstr\alpha}(R^c)$.

As a conclusion, it is possible to state that despite the fact that the original compound Petri net R^c has only three undefined structural parameters $S_{str\alpha}(R^c) = \{ \alpha_1, \alpha_4, \alpha_5 \}$, the resulting AAPN obtained by a replication of the transitions based on this second partition of $S_{valstr\alpha}(R^c)$ has only two undefined structural parameters:

$$S_{str\alpha}(R^d) = S_{1str\alpha}(R^c) \cup S_{2str\alpha}(R^c) = \{ \alpha_1^1, \alpha_5^2 \}$$

It is interesting to notice that it depends on the parameters of the transformation algorithm (in this case the chosen partition), that the size of the resulting model is more or less compact.

5. CONCLUSIONS AND FURTHER RESEARCH

In this paper it has been described a transformation algorithm between two formalisms that represent an undefined Petri net. This algorithm develops a link that allow to

obtain an alternatives aggregation Petri net and the subsequent disjunctive coloured Petri net from a compound Petri net or even from another formalism that had been previously transformed in the former. This transformation constitutes an important step in the research of the conditions where one of the two involved formalisms is more efficient in the application of a decision making algorithm related to a discrete event system. The transformation allows as well transforming in a certain case a less promising formalism to a more promising one.

The future research leads to the characterization of the decision problems to forecast the performance of the different formalisms in the exploration of the state space that requires the decision making based on discrete event systems.

References

- Berthelot, G., “Transformations and decompositions of nets” in “Petri Nets: Central Models and Their Properties, Advances in Petri Nets”. Lecture Notes in Computer Science, Brauer, W., Reisig, W., and Rozenberg, G. (eds.), vol. 254-I, pp. 359–376. Springer, 1987.
- Cassandras, Christos G., Lafortune, S., “Introduction to Discrete Event Systems”. Second Edition, Springer, 2008
- David, R., Alla. H., Discrete, Continuous and Hybrid Petri nets, Springer, 2005
- Haddad, S. and Pradat-Peyre, J.F., “New Efficient Petri Nets Reductions for Parallel Programs Verification”. Parallel Processing Letters, pages 101-116, World Scientific Publishing Company, 2006.
- Jiménez, E., Pérez, M., Latorre, J.I., “On deterministic modelling and simulation of manufacturing systems with Petri nets”. Proceedings of European Modelling Simulation Symposium. Marseille, pp. 129-136. Marseille. 2005
- Latorre, J.I., Jiménez, E., Pérez, M., “Macro-Reachability Tree Exploration for D.E.S. Design Optimization,” Proceedings of the 6th EUROSIM Congress on Modelling and Simulation (Eurosims 2007). Ljubljana, Slovenia, September 2007.
- Latorre, J.I., Jiménez, E., Pérez, M., Blanco, J., “The problem of designing discrete event systems. A new methodological approach,” Proceedings of the 21st European Modelling and Simulation Symposium (EMSS 09). Puerto de la Cruz, Spain, vol. 2, pp. 40-46, September 2009.
- Latorre, J.I., Jiménez, E., Pérez, M., “Control of Discrete Event Systems Modelled by Petri Nets” Proceedings of the 7th EUROSIM Congress. Prague. Sept. 2010
- Latorre, J.I., Jiménez, E., Pérez, M., “The alternatives aggregation Petri nets as a formalism to design discrete event systems.” International Journal of Simulation and Process Modeling, Special Issue. 2010
- Latorre, J.I., Jiménez, E., Pérez, M., “Efficient Representations Of Alternative Models Of Discrete Event Systems Based On Petri Nets”. Proceedings of the UKSim 13th International Conference on Computer Modelling and Simulation. Cambridge, United Kingdom, March 2011.
- Peterson, J.L. “Petri Net Theory and the Modelling of Systems”, Prentice Hall, Englewood Cliffs, 1981.
- Petri, Carl A. (1962). “Kommunikation mit Automaten”. Ph. D. Thesis. University of Bonn (German).
- Piera, M.A. y Music, G. “Coloured Petri net scheduling models: Timed state space exploration shortages”, Mathematics and Computers in Simulation. Elsevier. 2011(In press).
- Piera, M.À., Narciso, M., Guasch, A., Riera, D., “Optimization of logistic and manufacturing system through simulation: A colored Petri net-based methodology,” Simulation, vol. 80, number 3, pp 121-129, May 2004
- Silva, M. “Introducing Petri nets”, In Practice of Petri Nets in Manufacturing”, Di Cesare, F., (eds.), pp. 1-62. Ed. Chapman&Hall. 1993

IMPROVEMENTS IN THE OPTIMIZATION OF FLEXIBLE MANUFACTURING CELLS MODELLED WITH DISCRETE EVENT DYNAMICS SYSTEMS: APPLICATION TO A REAL FACTORY PROBLEM

Diego R. Rodríguez^(a), Mercedes Pérez^(b) Juan Manuel Blanco^(b)

^(a)Ikerlan, Paseo J. M. Arizmendiarieta, 2, 20500 Arrasate-Mondragón, Guipuzcoa, Spain

^(b) University of La Rioja. Industrial Engineering Technical School. Department of Mechanical Engineering. Logroño, Spain

^(c) University of La Rioja. Industrial Engineering Technical School. Department of Electrical Engineering. Logroño, Spain

^(a) droduiguez@ikerlan.es, ^(b) mercedes.perez@unirioja.es, ^(c) juan-manuel.blanco@unirioja.es

ABSTRACT

Modeling of Flexible Manufacturing Systems has been one of the main research topics dealt with by researchers in the last decades. The modeling paradigm chosen can be in many cases a key decision that can improve or give an added value to the example modeling task. Here, the modeling tasks are represented by one of the mostly used one in the academia, the Petri Net paradigm, and in particular the Stochastic Petri Net models. These models constructed will be used to optimize the performance measures that could be interesting for the production systems. The production indicators used here are related with the productivity of the systems and its efficiency in the production . We have added here and extra element to the optimization problem that is related with the energy consumption during the productive process. This will add an extra value to the actual scenario, introducing energy efficiency terms and information into the models and into the optimization process. These productivity and energy consumption measures could be included into an optimization process by changing a certain number of parameters into the model. A two phase optimization process has been applied to the real example we have considered and an improvement of this two phase methodology has been applied, comparing these results with the original two phase method to check whether which approach is more appropriate. Previous results obtained for this systems are improved by this new piece of research.

Keywords: Stochastic Petri nets, flexible manufacturing, simulation , performance measures, energy consumption.

1. INTRODUCTION

Flexible Manufacturing Systems and their representation in an adequate model that expresses their behavior the more accurately is a typical topic treated by many researchers. Here, a comparison between two optimization approaches using a novel representation of the energy consumption process associated to the model

is presented. The model representation is done through stochastic Petri nets.

Petri nets have shown their capacities to represent the behaviors that Flexible Manufacturing Systems pose, and specially concurrency and resources representation that are typical features of Manufacturing Systems. Stochastic Petri nets have been used largely to represent systems where an stochastic behavior is associated to tasks. This modeling method has some lacks when dealing with complex models where the state space is clearly untreatable and even simulation can be a great time consuming task. Here we will consider only simulation approaches due to this complexity previously mentioned

Here, a particular optimization problem will be introduced. This particularity lies on the introduction of terms related to the energy consumption associated to the machining processes and the inclusion of these data into the optimization problem and in particular into the goal function. This will make that the energy consumption information will be considered as a new term of the optimization function.

In order to have the best possible results of this optimization process with a contained computational effort a couple of optimization approaches have been considered. The first one is a two phase optimization method where the second phase uses the solution obtained from the first one, while the second approach takes advantage of the information extracted from the solutions visited during the first phase to reduce the complexity of the problem we are considering here.

The rest of the paper is as follows, in section 2 the FMS that will be used along this paper will be explained and all the elements that will be of interest to be represented in our model will be enumerated. Later on, in sections 3 Petri net model will be depicted. In section 4, the optimization will be depicted and the approaches are extensively commented. Finally, the results we are

interested in are represented associated to the two approaches in section 5 where a comparison of the results is shown. Finally some conclusions are presented in section 6.

2. DESCRIPTION OF THE FMS

The Manufacturing system initially considered is able to perform window frames with the following different features:

Feature 1

The first feature to be considered when modeling the system is the type of window where the frame will be included:

- Accessible window,
- tilt and turn window,
- Slide window
- Frames without any other element.

Feature 2

This feature is related with the presence of a crosspiece that goes horizontally from one extreme to the other of the window frame.

- With crosspiece
- Without crosspiece

Feature 3

The number of leafs that compose the window is the next differentiation element.

- One leaf
- Two leafs

It was considered a third leaf in the initial modeling constraints but finally it was considered that the third leaf could be added as a future improvement of the manufacturing system.

Feature 4

The last feature is related with the size of the window that will change the treatment or steps that must be followed in case of considering one size or the other.

- Big size
- Little size

Considering all the features depicted here, there are finally 32 different types of products that the manufacturing cell will be able to produce.

Apart from these types of windows, a set of accessories can be added to the different products. These accessories are:

- Box and Guide to include into this device the blind that can be integrated into the window.
- Drip edge to get all the water that can slide through the window

Initially we will describe the processes signing them in bold letters and describing who the operators that perform every process are:

- Operator 1 performs **the selection of the materials** needed to complete satisfactorily the aluminum profiles depending on the frame to be produced
- Operator 1 also supervises the **cutting of the PVC profiles** in the corresponding machine.
- Operator 1 performs the operation of **introduction of the reinforcement**
- Operator 1 supervises operations perform inside the Numerical Control Machine. These operations are 6 different operations that must be performed. We do not enter into more details about these operations because is not the objective of this thesis.
- Finally operator 1 checks the correctness of the reinforcement screwing operation.
- Operator 2 performs the following operations
 - Selection of Material for the reinforcement cutting
 - Supervises the reinforcements cutting in the corresponding machine
 - Distributes the reinforcements to the corresponding profile
 - Performs an extra operation that is the leaf cutting that corresponds to the inverse leaf for the windows that are composed of two leafs
- Operator 3 performs the following tasks:
 - Once all the operations are completed in the Numerical Control Machine, he distributes the completed pieces in the corresponding carriage.
 - For the pieces where the soldering is not needed this operator will retest the strip/post
- In case the previous parts are frames they should be soldered and cleaned passing to operator 4
- Operator 5 will distribute the pieces after coming from the previous task.
- There is one decision important in the production process at this point that is the presence of a crossbar in the window. Depending on this the pieces will take a way or another.
- Operator 6 is in charge of inserting the crossbar into the window frame.
- After this operation all the frames (independently of having or not crossbar) continue to the next operations jointly
- Now the system will need to know if the window is a two leafs window then operator 6 will fix the inverse leaf
- Once finished this operation all the frames will pass to the ironwork placing. This operation will be accomplished by operator 7
- Operator 8 will continue being in charge of the following:

- In case the window is a frame he will place the locks and the hinges
- In any case he will hang the window in a place where the other operators will take it to perform the following operations.
- After all these processes if the window has a box in its features there will be 7 possible configurations related with the box placing. All these operations related with the box are performed by operators 9 and 9 bis. These seven options are:
 1. Guide Assembly + Box + Drip Edge + Silicone Insert
 2. Guide Assembly + Box + Silicone Insert
 3. Guide Assembly + Drip Edge + Silicone Insert
 4. Box + Drip Edge + Silicone Insert
 5. Guide Assembly + Silicone Insert
 6. Drip Edge + Silicone Insert
 7. Box + Silicone Insert
- Once the box assembly process is completed the next phase will be glaze the window in case is needed. This task is performed to all windows independently of the box presence.
- Operator 10 will continue with:
 - Glazing the window
 - Inserting the reeds into the window
- Operator 11 will:
 - Disassemble the leaf/frame
 - Pack the finished window
- There is an extra operator in the system, operator 12, that will perform ancillary operations helping operator 3 with the distribution of wagons
- Operator 13 selects and distributes the glass supplying them previous to the operation of locks and hinges
- Finally, operator 14 performs the task of cut and distribution of reeds previous to the operation related with them.

Once the operations and the initial operators that are performing the tasks we will consider a table with all the operations and a task numbering that will help us when modeling this example is presented.

The following table presents a description of the different tasks that have to be performed and whom is responsible of performing them.

Task	Description	Performed By
Task1	the selection of the materials	Operator 1
Task2	cutting of the PVC profiles	Operator 1
Task3	Introduction of the reinforcements	Operator 1
Task4	Numerical Control Machine 6 Operations	NCM
Task5	Reinforcements material selection	Operator 11
Task6	Reinforcements Cutting	Operator 11
Task7	Reinforcement distribution	Operator 11
Task8	Screwing reinforcements	Operator 1 and Machining Center
Task9	Leaf cutting	Operator 2
Task10	Inverse Leafs distribution	Operator 2
Task11	Wagon distribution	Operator 3
Task12	Retest the strip/post	Operator 3
Task13	Crossbar distribution	Operator 3
Task14	Soldering and cleaning	Operator 4
Task15	Frame distribution	Operator 5
Task16	Crossbar Mounting	Operator 6
Task17	Locks and hinges fixing	Operator 8
Task18	Window hanging	Operator 8
Task19	Inverse leaf mounting	Operator 6
Task20	Box assembly (with all options)	Operator 9
Task21	Glazing	Operator 10
Task22	Insert the reeds	Operator 10
Task23	Glass selection and distribution	Operator 13
Task24	Reeds cut and distribution	Operator 14
Task25	Disassemble leaf/frame	Operator 11
Task26	Pack finished window	Operator 11

The flow of parts of the systems under study is represented in figure 1.

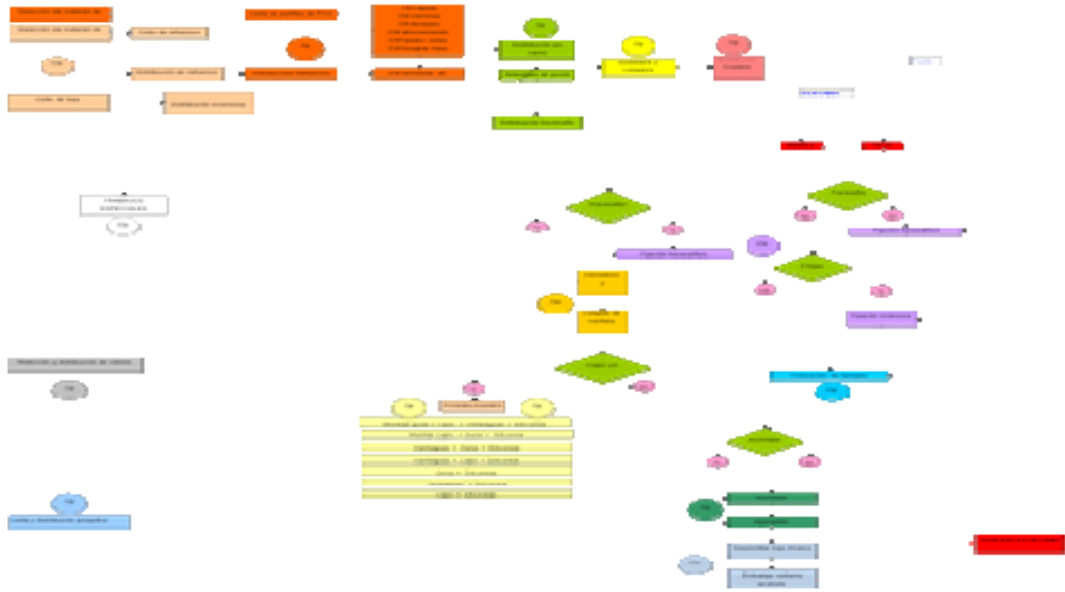


Figure 1. Flow model of the example

3. STOCHASTIC PETRI NET MODEL

Here we present the Petri net it has been modeled using stochastic PNs.

The complete model is represented by the following figure.

1, while T53, T511, T521, T441 and T4111 represent the five operations that the second operator can perform. Finally, the machining tool availability is represented by place Machining_TOOL1 and the operation is shown under transition T421.

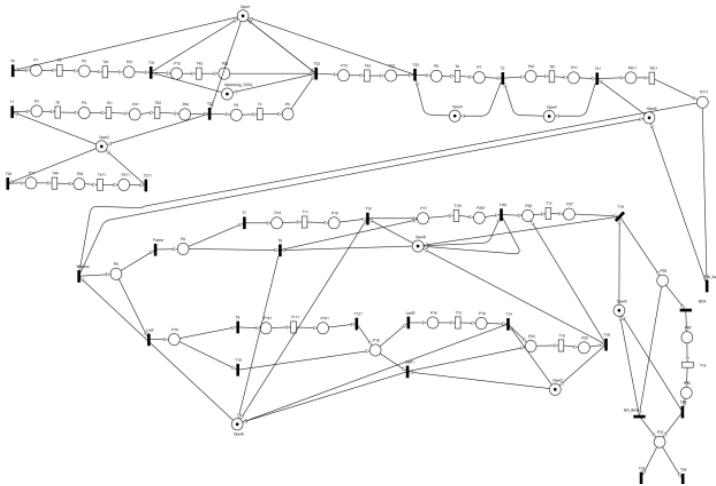


Figure 2. Complete Petri Net model

This complete Petri net model shown before will be more clearly presented in the next figures where it will be divided in substructures that will help understanding the modeling issues.

Figure 3 presents the operations where operators 1, 2 and the numerical control machine are involved. Places Oper1 and Oper2 represent the availability of the operators when marked. Transitions T45, T412, T32 and T431 represent the 4 operations that can be performed or supervised by Operator

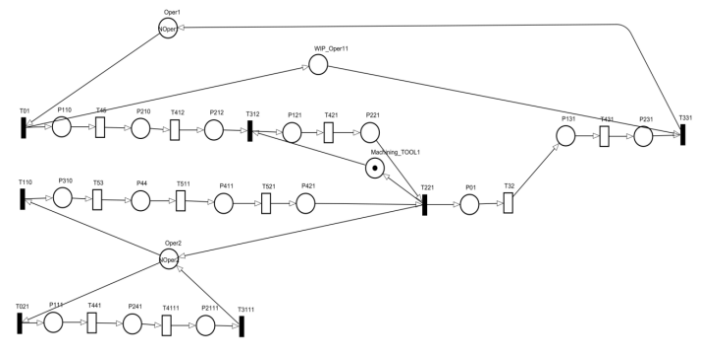


Figure 3. Petri Net model Operator 1 and 2 and NCM from example 2.

Figure 4, represents the operators 3 and 4 and due to their simplicity, because they are only performing an operation we have considered that a simple operator can cover each one of the tasks associated. There is no competition for the operator tasks.

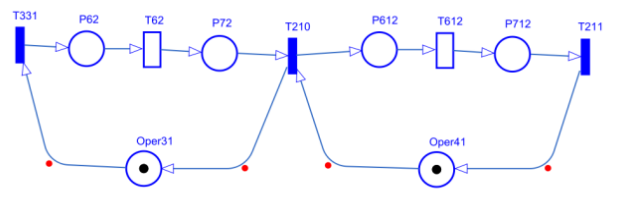


Figure 4 Petri Net model Operator 3 and 4 from example 2.

The next figure (Figure 5) represents the tasks where operators from 5 to 9 are involved. This Petri net model represents most of the decisions that must be taken (depending on the type of final product that the FMS is generating). After operator 5 performs its task (transition T611) then the raw parts will take one way or another depending on the type of final product (window or frame). If it is a window will continue through transition window and then a second decision should be taken depending if what has to be built is a leaf of this window or a frame of it (transitions Leaf or Frame). All these operations will be supervised by operator 6. Then operators 7 and 8 will perform their tasks associated to them (transitions T15, T151 and T17). Finally, operator 9 will perform its operation represented by transition T19 but before that a decision should be taken regarding the presence of a BOX in the window structure represented by immediate transitions BOX and NO_BOX.

The last Petri net submodel is represented in Figure 6, where operators from 10 to 14 are modeled. These operators generally are performing simpler operations than the previous ones and their model representation is simpler also.

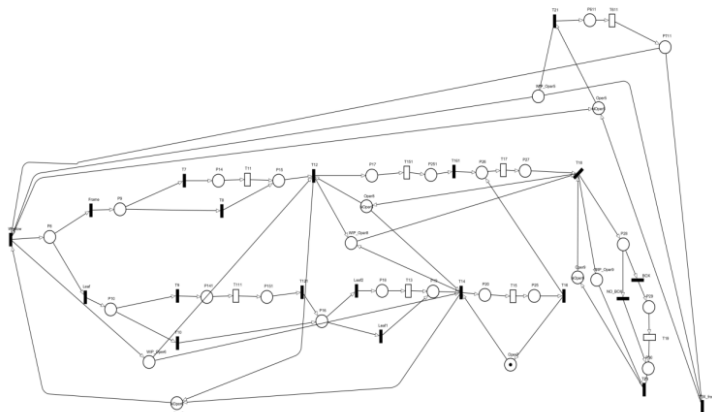


Figure 5. Petri Net model Operators 5 to 9 from example 2.

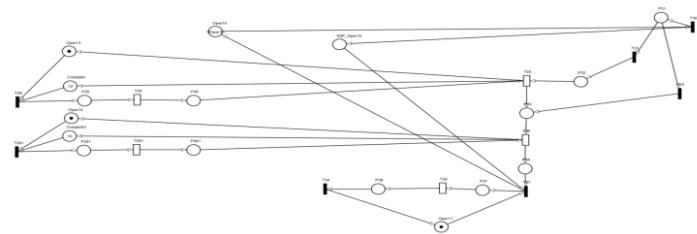


Figure 6. Petri Net model Operators 11 to 14 from example 2.

4. TWO PHASE OPTIMIZATION APPROACHES

The search space corresponding to the optimization problem that it is solved is composed by the following variables:

- Variable \rightarrow **NOper1** is an integer variable that represents the number of operators that will perform the operations initially assigned to operator 2.
- Variable \rightarrow **NOper2** is an integer variable that represents the number of operators that will perform the operations initially assigned to operator 2.
- Variable \rightarrow **NOper5** is an integer variable that represents the number of operators that will perform the operations initially assigned to operator 5.
- Variable \rightarrow **NOper6** is an integer variable that represents the number of operators that will perform the operations initially assigned to operator 6.
- Variable \rightarrow **NOper8** is an integer variable that represents the number of operators that will perform the operations initially assigned to operator 8.
- Variable \rightarrow **NOper9** is an integer variable that represents the number of operators that will perform the operations initially assigned to operator 9.
- Variable \rightarrow **NOper10** is an integer variable that represents the number of operators that will perform the operations initially assigned to operator 10.
- Variable \rightarrow **Mach_Delay** is a real variable that represents the time that in average takes to the Numerical Control Machine to perform the different tasks.
- Variable \rightarrow **PROB_BOX** is a real variable that represents the percentage of windows that has a box inside its structure.
- Variable \rightarrow **PROB_FRAME** is a real variable that represents the percentage of windows that will be a fixed frame window without any leaf (or with a unique one)
- Variable \rightarrow **PROB_WINDOW** is a real variable that represents the percentage of products that will have a window structure instead of a frame one.

The search space considered for this example will be the one shown in the following text box

Parameter definitions:							
#	name	type	minimum	maximum	initial	delta	temp
0	NOper1	INT	1	10	1	0.9	1
1	NOper2	INT	1	10	1	0.9	1
2	NOper5	INT	1	10	1	0.9	1
3	NOper6	INT	1	10	1	0.9	1
4	NOper8	INT	1	10	1	0.9	1
5	NOper9	INT	1	10	1	0.9	1
6	NOper10	INT	1	10	1	0.9	1
7	Mach_Delay	REAL	1	4	1	0.01	1
8	PROB_BOX	REAL	0.05	0.95	0.5	0.01	1
9	PROB_FRAME	REAL	0.05	0.95	0.5	0.01	1
10	PROB_WINDOW	REAL	0.05	0.95	0.5	0.01	1

The optimization function will include the performance measures we are interested in and are mainly related with the throughput of the system and the profit for a 8 hour shift, combined with the costs associated with the presence of more operators in the different positions of the FMC.

The manner how is represented this utilization in a PN model is by the formula represented below that will be explained later on.

MEASURE Profit

$$P_{\{P37>0\}} * 720000 - (P_{\{P34>0\}} + P_{\{P26>0\}} + P_{\{P12>0\}}) * 2880 * 0.3 - 40 * (N_{Oper1} + N_{Oper2} + N_{Oper5} + N_{Oper8} + N_{Oper9} + N_{Oper10}) - 20000 * Mach_Delay;$$

The first expression of this formula ($P_{\{P37>0\}}$) represents the throughput of the whole system given that place P37 is the place positioned just before being performed the last task (done by operator 11). This expression will represent the probability that there is more than zero tokens in place P37, and this is exactly the meaning of the throughput (considering that the maximal number of tokens in place P37 is 1 because there is a P-invariant that contains this place and also places P37 and Oper11). The amount of 720.000 corresponds to the gain that the company is having considering a mean selling price for all the windows produced of 25 Euros per unit produced and considering that there is a shift of 8 hours (28800 seconds).

The next term ($E_{\{P34\}} + E_{\{P26\}} + E_{\{P12\}}) * 30 * 8 * 0.1$) corresponds to the energy consumption term that is associated with the use of the different machines that are involved in the process. In this particular case there are three operation machines that are represented in places P34, P26 and P12. Computing the utilization of these machines during a shift of eight hours and considering the mean cost of the energy equal to 30 kwh and considering a cost of energy equal to 0.1 €/kwh

The next part corresponds to the cost associated with the utilization of the different operators that has been estimated in 40 Euros for each worker and for and 8 hour shift.

Finally, the last part corresponds to the cost associated to the inclusion into the system of a quicker Numerical Machine Center that will increase the price according to the mean operational speed (20000 €/second)

5. RESULTS AND COMPARISON

The results we are interested to compare between the two approaches previously shown are related with productivity measures. It will be considered the number of pieces produced per time unit for each type of product (32 different types can be produced in the FMS). Another performance measure we will consider will be the utilization of the different operators that are present into the system

Another important comparison measure will be how efficient is the convergence process for the two models and the accuracy they can reach. Also the computational time that the computer will be calculating the measures will be another measure of how good the simulation process is with respect to the colored and the stochastic models.

Also given the complexity of the model is important to consider the quality of the solution obtained combined with more qualitative measures more related with the computational effort and the number of iterations done during the optimization process.

In order to present all this information, the following tables are presented. There are 5 experiments or optimization methods we have applied.

Experiments
EXP1: Two Phase Approach: Temp_anneal_scale parameter 100
EXP2: Two Phase Approach: Temp_anneal_scale parameter 50
EXP3: Two Phase Approach: Temp_anneal_scale parameter 20
EXP4: Two phase Approach with Reduction of Search Space in the second Phase
EXP5: Two phase Approach with Temperature Parameter in variables in second Phase

Experiment	Time (Minutes)	Simulations	Profit
EXP1	2765.33	3505	271179.9
EXP2	1090.33	1802	266887.5
EXP3	589.7	707	269647.6
EXP4	7286	2351	245302.2
EXP5	289.12	184	252209.5

Experiment	QUALITY	Timing	Simulations
EXP1	100.00%	100.00%	100.00%
EXP2	98.42%	39.43%	51.41%
EXP3	99.43%	21.32%	20.17%
EXP4	90.46%	50.12%	67.08%
EXP5	93.00%	10.46%	5.25%

6. CONCLUSIONS AND FUTURE RESEARCH

Here we have presented a set of approaches that have been applied to a real Flexible Manufacturing System where a first approach to the introduction of energy consumption information is introduced into the optimization problem adding an extra value to the optimization process and giving another solution to the companies in order to reduce the expenses associated with this energy consumption.

Some possible future research topics will be related with the introduction of more energy related information into the models so that the optimization process will be more concentrated into this topic.

REFERENCES

- Aarts, E., Korst, J. "Simulated Annealing and Boltzmann Machines", Wiley (1989)
- Ajmone Marsan, M., Balbo, G., Conte, G., Do-natelli, S., Francheschinis, G. "Modelling with Generalized Stochastic Petri Nets", Wiley (1995)
- Balbo, G., Silva, M.(ed.), "Performance Models for Discrete Events Systems with Synchronisations:

- Formalism and Analysis Techniques” (Vols. I and II), *MATCH Summer School, Jaca* (1998)
- DiCesare, F., Harhalakis, G., Proth, J.M., Silva, M., Vernadat, F.B. “Practice of Petri Nets in Manufacturing”, *Chapman-Hall* (1993)
- Ingber, L. “Adaptive simulated annealing (ASA): Lessons learned”, *Journal of Control and Cybernetics*, 25 (1), pp. 33–54 (1996)
- Rodriguez, D. “An Optimization Method for Continuous Petri net models: Application to Manufacturing Systems”. European Modeling and Simulation Symposium 2006 (EMSS 2006). Barcelona, October 2006
- M. Silva. “Introducing Petri nets, In Practice of Petri Nets in Manufacturing” 1-62. Ed. Chapman&Hall. 1993
- Zimmermann A., Rodríguez D., and Silva M. Ein effizientes optimierungsverfahren für petri netzmodelle von fertigungssystemen. In Engineering komplexer Automatisierungssysteme EKA01, Braunschweig, Germany, April 2001.
- Zimmermann A., Rodríguez D., and Silva M. A two phase optimization method for petri net models of manufacturing systems. *Journal of Intelligent Manufacturing*, 12(5):421–432, October 2001.
- Zimmermann, A., Freiheit, J., German, R., Hommel, G. “Petri Net Modelling and Performability Evaluation with TimeNET 3.0”, 11th Int. Conf. on Modelling Techniques and Tools for Computer Performance Evaluation, LNCS 1786, pp. 188-202 (2000).

A SIMULATION-BASED CAPACITY PLANNING MODEL: A CASE STUDY IN A CONTRACT FURNISHING SME

Nadia Rego Monteil^(a), David del Rio Vilas^(b), Diego Crespo Pereira^(c), Rosa Rios Prado^(d), Arturo Nieto de Almeida^(e)

^{(a), (b), (c), (d), (e)} Integrated Group for Engineering Research (GII). University of A Coruna, Spain

^(a)nadia.rego@udc.es, ^(b)daviddelrio@udc.es, ^(c)dcrespo@udc.es, ^(d)rrios@udc.es, ^(e)anieto@udc.es

ABSTRACT

The contract furnishing sector –stores, hotels and public facilities– is characterized for working under a MTO philosophy, having worldwide clients, a flexible and highly manual process and a high mix of products. In the presence of these sources of variability, the lack of a proactive planning drives to outsourcing and cost overruns. This paper presents a case study of capacity assessment in a Spanish SME of manufacturing, distribution and assembly of contract furnishing. To do so, a Monte Carlo simulation approach was adopted, with stochastic values for demand, process and product parameters. Based on historical data and expert interviews a spreadsheet-based model was proposed in order to represent the variability. As a result, capacity anticipation under different scenarios was provided: (i) at present conditions, (ii) with an increased demand and, (iii) in case of a change in the type of production orders.

Keywords: Capacity Planning, Monte Carlo Simulation, Demand Forecasting, Contract Sector, SME

1. INTRODUCTION

Contract furnishing sector is made up by companies that offer a complete furnishing service to hotels, stores, offices and public buildings, including design, manufacturing, distribution and final on-site assembly. These companies typically work under a Make to Order (MTO) philosophy, meaning that the manufacturing starts only after a customer's order is received.

The case study is a family-owned Spanish medium enterprise. Its main client is one of the world's largest fashion distributors representing circa 60% of the company's production and sales. Being a SME and having powerful worldwide clients is a complex balance; failing or even delaying an order is not an option. However, projects' planning is usually carried out in a reactive manner making subcontracting necessary in order to meet the tight deadlines.

Uncertainty is a well-known characteristic of make-to-order (MTO) philosophy based companies. It is difficult (if not impossible) to predict the time at which a customer will place an order, the order due date, its quantity and nature and accordingly, the process and material requirements to fulfil it. Generally, orders' uncertainty affects more to SMEs because of their weaker negotiation position (Achanga, 2006). Usually, diversifying is a way to reduce the impact of such condition. For instance, civil engineering companies try to combine projects for public clients with other works for private ones. Also shipyards do the same by offering both ship repairing and maintenance activities and new shipbuilding. However, this product variability and simultaneous production with shared resources make the planning and scheduling problem even more complicated. The available effective capacity to be used during the execution depends on the operational dynamics, which in turn depends on planning decisions. This circularity in planning is more complex in unsynchronised shops where the variety of products follows diverse routings (Albey and Bilge 2011). All these factors - uncertainty in demand, product variability, simultaneous conditions and diverse routings- are present in the normal activity of this case study.

As said before, the short period of time between the effective confirmation of an order and the moment the manufacturing process actually starts does not allow a proactive planning. The impact on the production levels may lead to bottleneck formations, subcontracting, difficulties to meet the due dates and expensive raw material acquisition. In other words, competitiveness relies on efficient production and capacity planning.

Clearly, deterministic processing times and deterministic demand rates based models are inappropriate to consider uncertainty in models. Mula et al. (2006) have carried out a literature review of the uncertainty treatment on production planning models. The use of simulation in production planning has been considered for several authors in the literature. For instance, Albey and Bilge (2011) state that simulation

is a “natural solution” to represent the complex dynamic operational behaviour of unsynchronised shops. Several works propose simulation production planning. Hung and Leachman (1996) collect flow time statistics instead of machine capacities while simulating a production plan. Byrne and Bakir (1999) solve a multi-period multi-product production planning problem by using a hybrid simulation-analytical approach. Kim and Kim (2001) update both machine capacities and flow times by collecting relevant statistics during simulation.

However, Buxey (2005) points out three reasons why business has ignored researcher’s efforts so far. Those are basically related to (i) their huge need of data –impossible, difficult or expensive–, (ii) to the model underlying assumptions which make them inappropriate for the process or (iii) because they are too complex and they are seen as worthless by the managers. Moreover, empirical evidence shows that practitioners using advanced planning methods are on average less satisfied with their plans than those who use simpler and less accurate methods (Jonhson and Mattson 2003). In that sense, Tenhiälä (2011) links appropriate capacity planning techniques with process types, set-up and nature. He also highlights the wide use of non-systematic planning methods and claims that optimization is not always desirable in complex real-world planning situations so more pragmatic research in operations management is needed.

Considering these circumstances, the main goal of the study was obtaining an estimate for the fitted allocation of production resources (both in time and quantity) for the following year under two plausible demand scenarios:

1. The next year presents the same tendency than the previous ones.
2. Changes in the type of orders may occur:
 - a. There is the possibility of establishing an important contract with a chain of resort hotels, increasing the workload around a 20%.
 - b. Their main client may change the type of order, from the complete interior store manufacturing one to the partial refurbishing of actual stores, which would mean smaller orders (-25%)

To do so, we decided to build a spreadsheet-based simulation model aiming at representing how demand implied workload on the different work centres. Variability modelling is obtained from historical data-based probability distributions for demand, process and product parameters. Monte Carlo simulation is then used for the risk assessment of the solutions provided.

2. CASE STUDY

Demand is composed of several components. On the one hand, 60 % of production is dedicated to their main client. Although time deadlines are strict, at least

orders are given with a certain level of anticipation. On the other hand, 40 % of production serves a variety of hotels, stores and business with more flexibility in due dates, but with much more uncertainty both in its amount and nature and in its occurrence. After a failed statistical attempt of characterizing this sort of demand, it was decided not to include it within the scope of the study. Besides, due to the possibility of delay in these type of order deliveries, it has been considered that the company’s capacity estimation is mainly influenced by the rest of the “predictable” orders workload. From now on, demand and orders will refer to the main client’s one.

The normal company’s operating scheme is now briefly described. The whole process starts when the fashion distributor provides the furnishing company with its own interior designs catalogue for all the stores of the year. This catalogue allows a basic preliminary estimate of materials as well as an early and conceptual technical design of the furnishing elements. From that moment on, specific orders may be placed in the form of a store plan, between 10 weeks -normal order- and 5 weeks -urgent order- before the opening date, where all the furnishings should be assembled and ready to be used.

Once the order arrives, the number of pieces of furniture and the amount and type of materials is calculated by the Technical Department. Then, the material supplying and the generation of production orders (in term of parts’ groups that are manufactured together) may start. From that time, the manufacturing process takes place sequentially through four manufacturing sections. Each of them, with different machines, operators and responsible, is now described. Also a fifth section is included, although it does not take place within the production plant.

1. Machining, comprising cutting and machining operations (drills, shapers, CNC machining centre, etc.).
2. Finishing, consisting of sanding, varnishing and automatic and manual painting operations.
3. Cabinet making, where carving operations, assembly and subassembly take place. They are the most qualified and experienced workers.
4. Packing, that can be either automatic or manual packing. All the finished packages are placed on the storage area until the whole order is ready so it can be dispatched.
5. Assembly. Finally, the pieces of furniture and the assembly workers are sent to the final location. Currently, the cabinetmakers are also in charge of these operations.

Some of the reasons for the complexity when applying a planning method are described now:

- There is a high uncertainty on demand, considering that orders are different in quantities and due dates.

- The manufacturing process is defined in the literature as an HMLVS, i.e. High-Mix Low-Volume Simultaneous production (Prasad, 2011). Its main characteristic is that product mix and bottlenecks keep changing frequently over time.
- Distribution and final assembly take place out of the manufacturing plant, increasing the uncertainty due to unpredictable external events. Besides, assembly and cabinet-making compete for the same resources.

3. APPROACHING THE PROBLEM

The followed methodology is shown in Figure 1. From historical demand data (2003-2010) we tried to determine the number and dates of the orders that could be usually expected any year, according to the past events. The same information was used for determining how project usually developed throughout the process operations. From its analysis, probability distributions are obtained in order to represent the following events:

- Monthly probability of a particular number of order deliveries.
- For each order:
 - Sizing. Distribution of the global amount of hours and its allocation among the different main operations.
 - Timing. Distribution of the operations durations and their respective delays.

Probability distributions are the input data of the model. Regarding the system modelling, time and process considerations have to be regarded. On the one hand, a marketing medium term approach (one year) has been proposed as a useful way of linking the managerial issues with the production requirements. Intermediate-range planning is facilitated by aggregating the many products of a company into a single unit of output (Aggregate Production Unit). In our case, each order will be translated into working hours. Also, we have considered as the planning period time the working week, because it is the usual unit in which production activities are referred within the company. On the other hand, the process has been divided in their main operations, that is to say, Machining, Finishing, Cabinetmakers, Packing and Assembly. A backward scheduling has been adopted, starting from the assembly tasks and finishing in the machining.

The resulting workload -the output data- is then analysed both in its quantity (maximum and average) and attending to its distribution along the year. The workload sizing and timing will be the basis of the capacity planning. Decisions regarding capacity planning have different risk implications. For the assessment of capacity planning, three different parameters will be used:

1. Number of operators, related to the direct manufacturing costs.
2. Failure probability, related to the risk of a plan. It is calculated as the number of weeks where the required workload exceeds the available workload (given a certain capacity level).
3. Occupation level. The average occupation level is related with the efficiency of the plan.

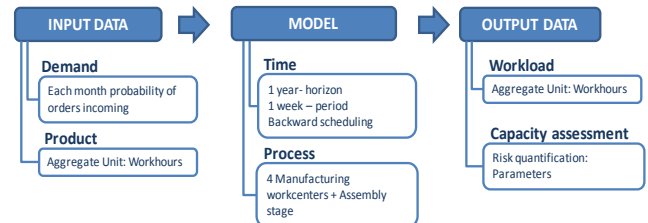


Figure 1: Model Development

4. MODEL DEVELOPMENT

Attending to the eight years record, when characterizing the occurrence of demand in terms of how likely is to have a fixed due date, a set of five month's behaviours within a year was identified. Autocorrelation tests were performed without leading to the identification of any significant pattern. By means of Maximum Likelihood Estimation five Poisson distributions were found to model the orders fulfilment process. When one or more orders are initially expected to be fulfilled in the same month, we have considered that they all will have to be ready for the first week (so assuming a worst case). As a result of the demand estimation, at each simulation N-orders arrive at a certain delivery week (D_w)

The amount of manufacturing and assembly hours (H) is composed of two terms. First, the average value was obtained from historical data and verified with the responsible of the Production Department. Then, a variable noise was modelled. The distribution of total hours in each process sector was quite regular (distribution of orders in work centres is shown in Figure 2). In order to simplify the model it was considered fixed (p_1, p_2, p_3, p_4, p_5).

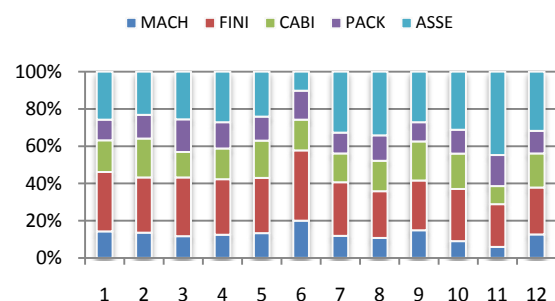


Figure 2: Monthly Distribution for the 2006 Year Orders Attending to the Defined Process Sectors

Given a certain delivery week, (D_w), the process scheduling is established backwards. Accordingly, the

variety on operations lengths was modelled. For example, most times (63% of frequency) an order machining takes place in four weeks. During these four weeks, 32% of the machining hours happen the first week, 41% the second, 15% the third and 13% the fourth. However, 37% of times it takes place in three weeks. This has been done for all the work centres of the process.

In addition, production orders have to go through the whole process in a certain order attending to technological constraints. The sequential progress of the order is characterized by the following parameters:

- x: End of packing – start of assembly delay.
- r1: Start of cabinetmaker work – start of packing delay.
- r2: Start of finishing – start of cabinet maker work delay.
- r3: Start of machining – start of finishing delay

The variety on delays between operations has been modelled in the same way as the variety on lengths. Graphically, the evolution of an production order is depicted in Figure 3.

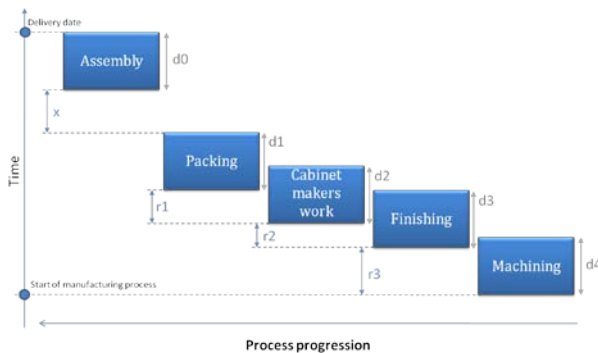


Figure 3. Process Evolution (for the Centres, including the Assembly).

Each order has to progress along the different operations according to the evolution parameters (x, d1, r1, d2, r2, d3, r3 and d4). Those different operations take a fixed rate of the global H, (p1, p2, p3, p4 and p5), so the amount of hours of each process (d_{in}) can be obtained as a vector:

$$d_{in} = HT (p_1, p_2, p_3, p_4, p_5) \quad (1)$$

All the production sequence is included in a 5x D matrix (M), where the dimension depends on the final duration of the manufacturing and assembly process, as it follows:

$$D = d_0 + x + d_1 + r_1 + r_2 + r_3 \quad (2)$$

Each element on M is the fraction of the corresponding production department dedicated hours (in columns) for a particular week (in rows). This way, M shows the evolution of the different operations for a single order.

As a result, the number of hours per operation, week and order (K_i) can be obtained:

$$(K)_i = d_{in} \cdot M \quad (3)$$

The global amount of work, K_t is the composition of each K_i in a common time axis. The workload (Q), a 52x5 matrix, is obtained by adding the same operations each week.

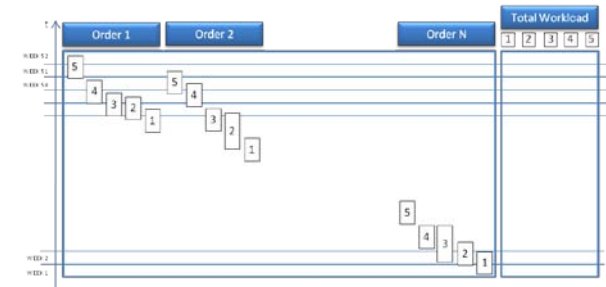


Figure 4: Workload Matrix Composition

In the same execution, there is a variation in terms of amount of hours, operations lengths and delays. Between two successive runs there is variation in terms of different number of incoming orders and their corresponding delivery date. So, we can talk of an inter and intra- year variation.

5. RESULTS

5.1. Workload sizing

The average workload, in hours per week, is the average of the average workload of each operation. The maximum workload is the average of the maximum values. The obtained distributions for average and maximum cabinet makers workload after 1000 simulations are shown in Figure 5. The results for the rest of departments are described on Table 1.

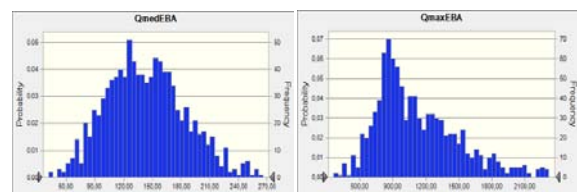


Figure 5. Average (left) and Maximum (right) Distribution of Workload for the Cabinet Department.

Table 1. Average and maximum workload per year

Workload	Mac	Fin	Cab	Pac	Asse
Average (h/week)	120.2	145.4	216.8	110.6	294,52
Max (h/week)	548.7	616.6	1597.0	449.3	3268.9

However, as a result of (i) the variability in number of hours and time evolution for every order, and (ii) the seasonable behaviour of the demand, one single value is not accurate enough to describe the expected workload. Therefore, results will be presented

dividing in quarters of year (Q1, Q2, Q3 and Q4) and a range of more likely values (range of 50% confidence level) are presented together with a boxplot. In Table 2, for instance, machining workload in the first thirteen weeks has been between 643 hours and 1636 hours in 500 of the 1000 simulations.

As it can be noticed, second and third quarters of the year show the greater values of both workload and variability in all the departments except in assembly. More details in the year distribution of the workload will be found in section 5.2.

Table 2. Workload's Departments by Quarter

D.	Range	
M.	Q1 643-1636	
	Q2 977- 2148	
	Q3 1618-3027	
	Q4 493-1340	
F.	Q1 782-1942	
	Q2 1148-2530	
	Q3 2005-3687	
	Q4 645-1611	
C.	Q1 1054-2891	
	Q2 1644-3868	
	Q3 2912-5682	
	Q4 776-2374	
P.	Q1 1054-2891	
	Q2 1644-3868	
	Q3 2912-5682	
	Q4 776-2374	
A.	Q1 2192-4680	
	Q2 1466-3859	
	Q3 4611-8471	
	Q4 1027-2949	

5.2. Workload timing

Yearly workload distribution (along 52 weeks) is now showed for the different departments on Figure 6. On Table 3 all operations workload distribution is described in terms of shape and Coefficient of Variation (CV), and maximum and minimum level of occupations.

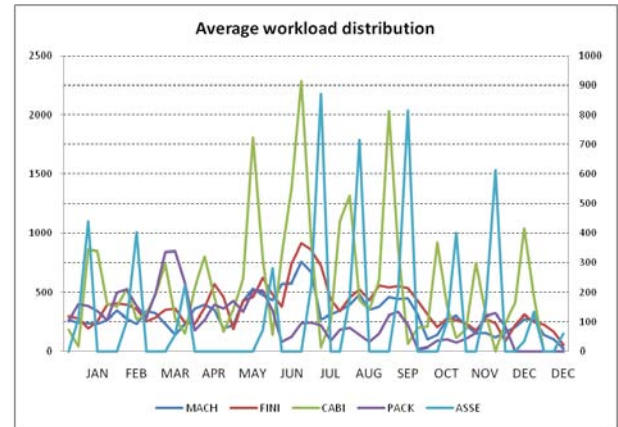


Figure 6: Year Workload Distribution for the different Departments

Table 3: Operation Workload Distribution: CV, Shape, Maximum and Minimum Occupation

Op.	CV	Shape	Maximum Occupation.	Minimum Occupation
Mac	1,20	Smooth	August-September May-June	December
Fin	1,14	Smooth	May-July	December-January and April
Cab	1,64	Concentrated in short periods	One or two weeks in June, July and August	March and December
Pac	1,07	Smooth	End of July, beginning of August	End of October, beginning November
Asse	2,29	Concentrated in very short periods	One week in August and one in September	March-April

The different shapes of the workload are related to the tasks nature. For instance, cabinet makers and assembly respective workload appear in a more concentrated way. Besides, these end stages of the process strongly influence the product quality. According to both factors, these work centres are considered the most critical in the process.

Currently, assembly and cabinetmakers share the same labours. This condition aims at reducing the impact of the concentrated assembly works. However, it also implies that sometimes the end of the process is almost unfilled. As a result the work flow is sometimes interrupted. It has been advised to the managers to separate these work centres and to try to negotiate a resource pool (variable number of working hours that

can be compensated with longer holiday days) with the workers.

The plant total workload can be obtained by adding the different operations (machining, finishing, cabinet makers and packing) each quarter. Results are shown on Figure 7.

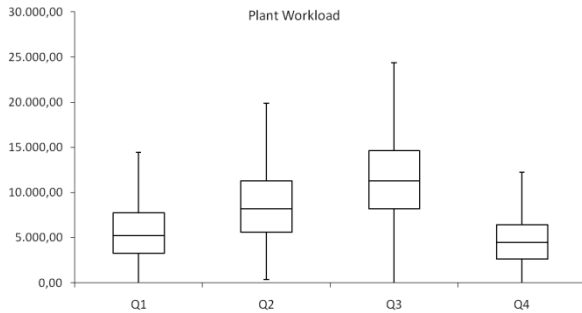


Figure 7: Plant Workload

First and fourth quarter of the year usually present the lowest level of work while the third quarter reaches its top. This information is useful for setting prices and hiring policies.

5.3. Capacity Estimation

According to the workload on each operation, different levels of capacity can be established. In a first approach, a constant annual level of capacity will be studied. Those levels have implications in terms of probability of risk, efficiency and of course, cost (parameters introduced on section 3). Capacity range for each department will be studied between a minimum and maximum level. The adopted criteria are that the minimum/maximum level for each department corresponds with covering the average/maximum workload values (Table 1).

The results are shown in Figure 8 (only for one operation). As it could be expected, the higher the number of operators is, the less the failure probability (lower blue line) and the occupation level are (red line).

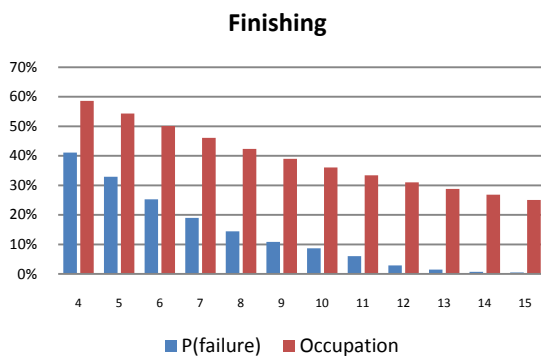


Figure 8: Operators in Machining as a Function of Failure Probability (lower bars) and Occupation Level (upper bars).

These graphs could be used for aiding in the decision process of establishing the capacity level. For example, the managers might decide that a 20% failure probability for the Finishing section is acceptable (one of five weeks the expected amount of work exceeds the Finishing capacity), which would imply that seven labours would be enough.

As it has been stated, each department workload reaches different levels and presents different behaviours. So, different capacity levels would be needed for meeting the same requirements (in terms of failure probability). However, experience shows that it is not necessary the same confidence level in all departments. For instance, a delay in machining would be much less severe than a delay in assembly. Being machining the first department, the excess of workload would probably be transferred to the following week without more consequences, while an excess of workload in assembly would probably lead to an eventual delay in the order. Accordingly, an operation cost indicator for each department can be built as follows:

$$C = n + 52 \cdot \frac{c_F}{c_H} \cdot P(\text{failure}) \quad (1)$$

Being C an indicator of the cost of operating with n workers in a certain department, $\frac{c_F}{c_H}$ the relation between the cost of a failure and the cost of hiring an extra worker. This cost increases with the number of workers and decreases with the failure probability. We could say, for instance, that a single failure in Assembly would cost 3 times more than hiring an extra worker, while a failure in Machining would only cost 0.5 times more (these values have been chosen for illustration purposes, and do not have to be close to reality). When representing these expressions, in Figure 9, the optimum number of workers (7 for machining and 16 for assembly) is obtained.

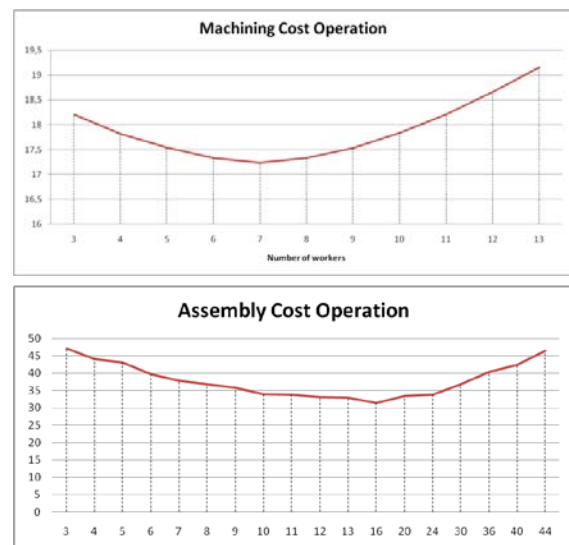


Figure 9. Machining and Assembly Cost Indicator depending on the Number of Workers.

5.4. Increased demand

According to the commercial department's expectations, a new contract with an important international hotel client would imply a 20% increase in the number of orders. As a result, the average workload is a 20 % higher compared to the previous situation. The maximum workload is 12-16 % higher (Figure 10), depending on the work centre. Assembly is the department that shows the higher increase of maximum workload. Referring to the CV, a smoothing effect in every task is observed. However, the lowest decrease in CV corresponds to the Assembly. It appears to be the strictest in its concentrated nature.

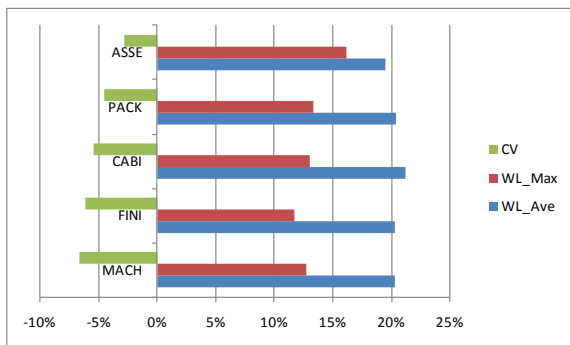


Figure 10: Workload Distribution when Demand increases 20%

5.5. Changes in the “order type”

Before a plausible change in the main client's type of works, from “complete new stores” to refurbishing existing ones, the Sales Department forecasts smaller projects (25% less, in average) but an increase in the number of orders (20%). Were this demand scenario, the workload would change as showed in Figure 11. The average workload decreases around a 10% per work centre. A similar smoothing effect in every task takes place. However, it is remarkable that the maximum workload decreases a 17% for the cabinet centers whilst the assembly only decreases a 14%. In fact, assembly has the lower decrease in maximum workload. It can be concluded that assembly department is more sensitive to changes in the number of orders than it is in the orders' size.

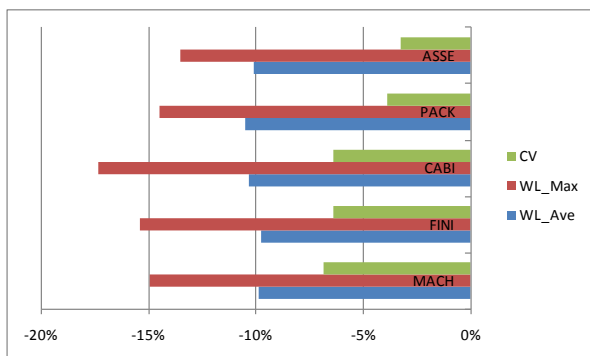


Figure 11: Workload Distribution when the Type of Order Changes.

6. VERIFICATION AND VALIDATION

For verification and validation purposes, the observed sample distribution (historical data) was compared with the modelled distributions in terms of four variables. The first is the number of orders, which accounts for demand level. The second is the total amount of workload hours per year. Yet the number of hour per quarter of year, week or department would have been a more accurate indicator for comparing results, this information was not available within the company's historical data. The third parameter accounts for the time gap between the start of the manufacturing and the final delivery. This value is obtained from the simulation as the addition of each operation estimate represented by different probability distributions. Then, it is compared to the historical time interval between the order incoming and the final delivery. Finally, the number of hours of each order is compared with the actual values from the database. The p-values for the null hypothesis of the averages being the same and the standard deviations being different are shown in Table 4 and Table 5. Model values for the average and standard deviations come from 1000 simulations so they were considered with a negligible error. They were then compared to the available number of observations (n) in the real data.

Table 4. Test T for Differences in Variable Means

Factors	Average Model	Average Data	N	p-value
Orders / year	15,1	15,28	7	0,956
Hours / year	46764,6	47867,1	7	0,903
Weeks / order	11,4	9,6	106	0,001
Hours / order	3094,6	3109,8	106	0,903

Table 5. Test χ for Differences in Variables Standard Deviations

Factors	Std. Dev. Model	Std. Dev. Data	n	p-value
Orders / year	3,8	6,9	7	0,003
Hours / year	12719,9	22976,2	7	0,003
Weeks / order	5,3	5,8	106	0,103
Hours / order	835,6	1275,0	106	0,001

It can be concluded that the estimations of demand per year, workload per year and workload per order averages do not significantly differ from those observed in the historical data (Table 4). On the other hand, significant differences were found in weeks per order average. This might be explained by the aforementioned differences in time durations, but further research should be conducted in order to assess the practical relevance of such a difference.

Significant differences in estimated standard deviations were found for all the tested variables (Table 5). This suggests that the model systematically

underestimates variability levels in workload rates and, consequently, the forecasted failure probabilities. A plausible explanation is given by the way that the annual demand is generated. Although standard autocorrelation tests did not show any autocorrelation patterns in monthly orders, this is not a sufficient proof for independence. Only under actual independence among monthly demands, annual demand variability would be accurately estimated from monthly variability. Available data were not enough to conduct a more profound autocorrelation analysis. This shortcoming of the model can be corrected by adopting a more risk-averse position in the decision making process.

7. CONCLUSIONS

A workload and capacity planning based on historical data in a Spanish SME of manufacturing, distribution and assembly of contract furnishing has been presented. A simulation approach has been adopted in order to represent the high variability associated to their MTO philosophy and job-shop production schema. The spreadsheet-based model within a Monte Carlo simulation approach allows introducing stochastic values for demand, process and product parameters. As a result, workload estimation under different scenarios was provided. Also, by means of a set of three general performance parameters – labour costs, failure probability and occupation level- the assessment of the production resources necessary to cope with the corresponding workload is achieved. When complemented with overall cost information, this planning methodology can be the basis for optimised capacity estimation according to the nature of each department. This work aims at connecting the operational level with strategic considerations by means of a simple but comprehensive and precise tool for decision making.

REFERENCES

- Albey E., Bilge U., 2011, “A hierarchical approach to FMS planning and control with simulation-based capacity anticipation.” *International Journal of Production Research*, Vol. 49, pp. 3319-3342.
- Achanga P., Shehab E., Roy R., Nelder G., 2007, “Critical success factors for lean implementation within SMEs”, *Journal of Manufacturing Technology Management*, Vol. 17, No. 4, pp. 460-471
- Byrne M.D., Bakir M.A., 1999, “Production planning using a hybrid simulation-analytical approach”, *International Journal of Production Economics*, Vol. 59, pp. 305–311
- Bitran G., Yanesse, H., 1984, “Deterministic approximation to stochastic production problems”, *Operations Research*, Vol. 32, pp. 999-1018
- Boiteux O., Forradella R., Palma R., Guiñazo H., 2010, “Modelo matemático par la planificación agregada de la producción de Impsa”, *Iberoamerican Journal of Industrial Engineering*, Vol. 2, No.2, pp. 90-112
- Buxey G., 2005, “Aggregate planning for seasonal demand: reconciling theory with practice”, *International Journal of Operations & Product*, Vol. 25 No.11, pp. 1083-1100
- Geneste L., Grabot B., Letouzey A., 2003, “Scheduling uncertain orders in the customer-subcontractor context”, *European Journal of Operational Research*, Vol. 147, pp. 297-311
- Hung Y.F., Leachman, R.C., 1996, “A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations”, *IEEE Transactions on Semiconductor Manufacturing*, Vol. 9, pp. 257–269.
- Jonsson P., Mattsson, S-A., 2003, “The implications of fit between planning environments and manufacturing planning and control methods” *International Journal of Operation & Production Management*, Vol. 73, pp. 165–173
- Kim B., Kim S., 2001, “Extended model for a hybrid production planning approach” *International Journal of Production Economics*, Vol. 73, pp. 165–173
- Mula J., Poler R., García-Sabater J.P., Lario F.C., 2006, “Models for production planning under uncertainty: A review”, *International Journal of Production Economics*, Vol. 103, pp. 271-285
- Prasad V., 2011, “Production Scheduling for Job Shops”, White Paper of OPTISOL (Optimal Solutions for Production Scheduling), available http://www.optisol.biz/job_shop_scheduling.html, 30/03/2011
- Tenhiälä A., 2010, “Contingency theory of capacity planning: The link between process types and planning methods”, *Journal of Operations Management*, Vol. 29, pp. 65-77
- Thompson S.D., Davis W.J., 1990, “An integrated Approach for Modeling Uncertainty in Aggregate Production Planning”, *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 20, No. 5, pp. 1000-1012

AUTHORS BIOGRAPHY

Nadia Rego Monteil obtained her MSc in Industrial Engineering in 2010. She works as a research engineer at the Integrated Group for Engineering Research (GII) of the University of A Coruna (UDC), where she is also studying for a PhD. Her areas of major interest are in the fields of Ergonomics, Process Optimization and Production Planning.

David del Rio Vilas holds an MSc in Industrial Engineering and has been studying for a PhD since 2007. He is Adjunct Professor of the Department of Economic Analysis and Company Management of the UDC. He has been working in the GII of the UDC as a research engineer since 2007. Since 2010 he works as a

R&D Coordinator for two different privately held companies in the Civil Engineering sector. He is mainly involved in R&D projects development related to industrial and logistical processes optimization.

Diego Crespo Pereira holds an MSc in Industrial Engineering and he is currently studying for a PhD. He is Assistant Professor of the Department of Economic Analysis and Company Management of the UDC. He also works in the GII of the UDC as a research engineer since 2008. He is mainly involved in the development of R&D projects related to industrial and logistical processes optimization. He also has developed projects in the field of human factors affecting manufacturing processes.

Rosa Rios Prado works as a research engineer in the GII of the UDC since 2009. She holds an MSc in Industrial Engineering and now she is studying for a PhD. She has previous professional experience as an Industrial Engineer in an installations engineering company. She is mainly devoted to the development of transportation and logistical models for the assessment of multimodal networks and infrastructures.

Arturo Nieto de Almeida received his PhD in Economics from the UDC in 2010. He has been Associate Professor of the Department of Economic Analysis and Company Management of the UDC during the last 18 years. He also owns a management and technical consulting firm in A Coruna (Spain).

PN AS A TOOL FOR INNOVATION IN INDUSTRY: A REVIEW

Jesús Fernández de Miguel^(a), Julio Blanco Fernández^(b), Mercedes Pérez de la Parte^(b)

^(a) Grupo ECO3G, Logroño, La Rioja, Spain

^(b) University of La Rioja. Industrial Engineering Technical School. Department of Mechanical Engineering. Logroño, Spain

^(a) jesusfernandez@grupoeco3g.com, ^(b) julio.blanco@unirioja.es mercedes.perez@unirioja.es

ABSTRACT

PNs are a basic tool widely used in scientific and technical fields, in areas as diverse as automation, computer science, management, modelling, simulation, optimization, etc. Many important research groups around the world work both in the progress of the PN in themselves (behaviour, properties, analysis, etc.) and its application fields of science, industry, or services. However, such an important tool for industrial development experiences a lack of exploitation in the productive sector, compared with the potential applicability of Petri nets in industry and services management.

In this paper, we review the actual use of the PN in industry, especially in the high Ebro Valley, in Spain, where ECO3G company is dedicated to the management of innovation in the industry for many years and therefore it has information based on experience. The review is also extrapolated to the world, based on information in the scientific and technical literature, and in the mentioned ECO3G experience in innovation, and national and European research projects. It also presents special attention to the use of PN in patents and industrial property as well as the know-how in companies of different industrial sectors. Finally, an analysis of the cause of this lack of exploitation in the productive sector, and a study of ways to improve this current state, are developed.

Keywords: workstation design, work measurement, ergonomics, decision support system

1. INTRODUCTION

There are certain characteristics of the PN that make them especially suitable for modelling discrete systems, among which are:

- Easy representation of concurrent systems (with parallel evolutions and synchronizations)
- The ability to condensation in a simple model of an underlying state space that suffers the state explosion problem that makes it impossible in practice the exhaustive analysis of actual cases by capacity problems of computational effort.
- The duality of graphic/mathematical representation, which allows on one hand to intuitively

and easily model by graphic systems, and secondly to analyze the properties of the model through mathematical and computational techniques

- The richness of existing modelling formalisms in the paradigm of the PN
- The richness of knowledge generated in previous research works on behaviour, properties, and techniques of analysis or simulation of PN.

From the foregoing it would appear that the PN constitute a tool fully integrated in the industry due to the benefits that could be obtained in productive systems, and some service companies, with their use. However its use in industry and services is much lower than expected, given its potential as a source of innovation.

In this paper, we review the actual use of the PN in industry, especially in the high Ebro Valley, in Spain, where ECO3G company is dedicated to the management of innovation in the industry for many years and therefore it has information based on experience. The review is also extrapolated to the world, based on information in the scientific and technical literature, and in the mentioned ECO3G experience in innovation, and national and European research projects.

It also presents special attention to the use of PN in patents and industrial property as well as the know-how in companies of different industrial sectors.

Finally, an analysis of the cause of this lack of exploitation in the productive sector, and a study of ways to improve this current state, are developed. All the analysis is developed in a general way, and also applied to the most important industrial sectors of La Rioja (the Country of the authors in Spain).

2. POTENTIAL APPLICABILITY OF PN IN THE AREA (LA RIOJA)

In this first phase of study we conducted an analysis of applications made in 5 sectors that clearly characterize the industry of La Rioja (Spain) and make it recognizable both nationally and in some cases internationally. So, these sectors have been chosen for its economic and strategic importance within the region, seeking to evaluate in this article their degree of

maturity for a potential introduction of the PN in their productive activity. Thus, as will be founded later, we have chosen the following sectors:

- Wine Production: the most recognizable industrial sector in the region through its designation of Origin Denomination and greater international presence through exports. It is also one of the most proactive R&D areas in recent years.

- Auxiliar Automotive industry: which encompasses all techniques characteristic of the automotive production; in La Rioja they are represented mainly in the rubber industry, metal, and chemical auxiliary. Their presence in the value chain is significant nationally through its connection to some of the major automotive plants in Spain and Europe.

- Footwear Sector: La Rioja, and more specifically its West region (Arnedo), are the basis of the footwear producer industrial district of Spain, and one of the top 5 in Europe, even being a leader in subsectors such as safety shoes. The growth of investment in R&D in this sector is significant.

- Sector snuff: Imperial Tobacco has a plant in La Rioja that has nearly 2000 employees (representing a significant 1.5% of the population of the Community according to the latest data) and is now the largest company in Spain of the group and one of the most prominent in Europe. In recent years, Imperial Tobacco, earlier known as Altadis, has been gathering at its plant in La Rioja many production plants in Spain, gradually increasing investment and providing greater strategic importance to this factory. The special feature of this product makes that it is included in the food industry.

3. OVERVIEW OF THE APPLICATION OF PN

Research in PN is now at the point of greatest activity in the world, although it could stand a greater intensity of research groups in Asia, perhaps encouraged by a strong tradition in the introduction of new production techniques that improve productivity.

However, we determined that since the 90's there were some references focusing on the applicability, especially as a complement to the development of R&D activities in the productive environment. For example, in (Japan-USA, 1992) already pointed to the PN as a particularly useful tool in addition to production techniques such as mechatronics, microrobotics or CAD-CAM techniques. Surprisingly, the research on PN was originally closely tied to certain sectors of technics to which they intended to supplement, especially through a very focused approach to automate certain tasks or at least integrating them. So in 2001, works such as Zhan and Luo (2001) already pointed to the usefulness of PN in highly automated industrial environments, such as the snuff, proposing a PN modeling that would allow assessing the productivity of selected alternative processes, mainly in packaging lines. Jeng et al. (2004) seeking the application of PN in the development of new types of semiconductors in the electronics industry. It is particularly interesting the approach that the work gives to the use of PN as a

resource for the evaluation of alternatives after the research phase of R&D projects, avoiding costly processes of reanalysis and correction of the prototypes, with the consequent consumption of resources. In other sectors, such as footwear, Carpanzano et al. (2004) posed the NP as a complement to the development of a modular production that is controlled by flexible production systems (RMS) integrated directly into the production line. Finally, sectors such as furniture had similar proposals such as Gradisar and Music (2007) where, through tools such as MATLAB, proposed the definition of an algorithm to determine the most appropriate control strategy for a given environment and productive process. Thus, through these representative examples we can see how the transfer of PN counts with relevant background and case studies that allow us to appreciate the competitive advantage that they provide to the industry.

However, in the last years, it can be determined that there is a clear evolution towards a more experimental research and somewhere, more focused on sectors with higher technological capability and maturity in the production side, where the most representative example is the automotive. Influenced by Japanese production techniques arising under the Automobile Industry (JIT-Kanban, 5S, LEAN ...) the PN found in this sector increased responsiveness and a route application much more straightforward than in traditional sectors, as is highlighted in Miao and Xu (2009). References in this field are numerous in recent years (2009-2011), from the most theoretical ones, such as Zhang et al. (2010), which seeks to define the technique of detection and evaluation of the critical points of contradiction as the key to improving the definition of PN based on classic models in engineering solutions, to the most applied, such as Wang (2009), which works in the application of fuzzy PN in the production chain, especially in technical delivery, targeting a logistics guidance that subsequently has been consolidated.

In parallel, there are numerous research groups working in the development of PN powerful techniques of a more global approach, and from which can be implemented oriented tools for various industrial sectors, almost always with a productive approach. Representative examples of this more general research are Han et al. (2009) or Li et al. (2010) and Wang (2009), and even more and Chen Xiao (2010), applying DSM mathematical techniques, or Xu et al. (2010), who posed a transfer of know-how to the field of logistics, which currently represents one of the busiest lines in applied research in PN, especially in seaports, airports, docks or logistics centers of activity.

Finally, in this overview is intended to highlight the link between PN with applied R&D projects in industry, a trend that has manifested itself more strongly nowadays. Bartz (2010) proposed an improvement in the management of the information accumulated in the development of R&D project exemplified in the case of the automobile. Especially interesting is the

contribution of the work in the search for synergies through the application of techniques WfMC (Workflow Manager Coalition). This also allows a more precise analysis of the different possibilities around a project with a more flexible decision-making and effective management emerged from the data acquired during the analysis of the production process. Currently the firm cooperation in R&D in both national and European projects is essential for peak performance in their development. Applied and reference examples that illustrate the effectiveness of PN in research and development is the case of China's research project about optical LAMOST (Modeling of control system based on LAMOST for Petri net workflow, X. Lu) or Costa et al. (2010) which highlight some of the PN modeling tools more useful for decision making in a project, supported by graphical and multilanguage code editors.

4. PETRI NETS AND INDUSTRIAL PROPERTY. PATENTS

A search of the most important patents around the PN has been developed as a basis for this work. It is difficult to find patents around the PN as advances are usually framed mainly in the field of Intellectual Property through publications such as those discussed above. Still, the application of PN in certain production processes means that there are interesting references to cases of successful applied research in PN. The results are essentially codes G06 (computers, calculators and counting) and especially G05 (Control systems or automatic control in general; functional elements, monitoring or testing devices or elements).

The geographic distribution of these patents, as in the case of articles presents a greater intensity in the Asian region, with significant references to leading companies in sectors as disparate MICROSOFT CORP (Constructing Petri Nets from traces for diagnostics, US2008320437 (A1)), SCHNEIDER ELECTRIC Automotion (Method for orchestrating services of a service-oriented orchestration and automation system machine, US2010292810 (A1)), SAMSUNG (Configuring learning petri nets, EP1335320 (A2)) and even EXXONMOBIL (System and method for abnormal event detection of continuous operation in the Industrial Processes, WO2006031635 (A2)) that have patents applied directly to their main lines of work which are closely linked to the most cutting edge in the industry today. Such records allow us to appreciate that macroindustrial level, the PN has not only proven its effectiveness, but the companies that have chosen to integrate them have ended up reaching patents reflect the success and utility of PN as a tool.

Within the references can be found in the attached document include, for direct application to fields of industry Virtual Production Control System and Method and Computer Program Product thereof (US2011040596 (A1) is a reference to recent virtual control system a manufacturing industry, Information processing method for Evaluating and using

biochemical pathway models using clinical data (WO02099569 (A2)) which shows the usefulness of PN in the evaluation methods or fields of biochemistry and method for System abnormal event detection of continuous operation in the Industrial Processes, WO2006031635 (A2) of 2006 already laid the foundations for industrial operation from the use of PN.

In addition, other references as traces from Constructing Petri Nets for diagnostics, US2008320437 (A1), or Method to Improve Unfolding in Petri Nets, US2009172013 (A1) represent the results of experimental research lines and emerging result of the search for a deeper knowledge in the field of methodology and definition of the PN.

5. METODOLOGY

The introduction of PN in R&D projects of La Rioja should be closely linked to a flexible methodology that allows business to appreciate their added value without consuming a large amount of resources. Keep in mind that with few exceptions, such as Imperial Tobacco, we speak of SMEs with a very short experience in terms of R&D and that assess the return on investment in these technologies through the results of billing partners.

Similarly, being traditional sectors, a methodology of integration of PN into the Business activities should be sought more than a methodology of implementation, looking for an easy transition. This will need to define an integration strategy. This strategy, through the University of La Rioja, the Administration or the social agents involved in research should seek to organize the integration initiatives to support more effectively the strategies of both lines of business and technology in the companies. That is, to seek the development of the technology without consuming resources that could make companies perceive a threat to their business activity. Thus, in a first phase of diagnosis, we define the response of each company or sector to some fundamental questions:

- How can PN provide greater business value through the integration of applications or their improvements?
- Where do we start? What is the critical technical area that needs to be revised or which is the one that the company intended for R&D?
- In what order should be integrated processes and applications that the company can carry out easily?
- How to take advantage of the investments?
- Can the company receive aids from the Administration to facilitate the transition?
- How to ensure that investment in integration is maintained over time?
- How to organize the integration effort? Does the company have staff trained to understand or at least apply the methodology defined by the PN?
- How to ensure that they are aligned with the defined strategy? Does the company have ability or experience in cooperation other companies, Universities, or agents of innovation?

- Is it necessary prior training?
- Is the company capable of perceiving the received added value through the use of PN to deal with future phases of the research and the innovation?

Once defined the particular scope of the proposal in each sector through the diagnosis, the development of a PN model is the next step, in companies with capacity to develop it, prioritized by the needs and the benefits of a detailed description of the behaviour and the knowledge of how key initiatives could be undertaken for research proposals to facilitate the evolution of the maturity level of integration required to support the proposed initiatives. At this point we should also assess the investment capacity of the company before moving to the next phases. Not surprisingly, obtaining a model would constitute for 90% of the companies a significant progress in its strategic capacity of production management.

Later, in a more advanced stage of the research, the following stages could be to perform simulations, analysis, implementation, and a final phase of the optimization of the process studied under the PN model in order to obtain the final diagnosis. All these phases depend on the outcome of further analysis of individual sectors, and particularly their degree of maturity and the awareness that the industrial structure of La Rioja and the Government have towards this type of research initiatives.

In summary, the methodology of strategic development for the diagnosis-modelling phase consists of three levels of information acquisition, which are consolidated to yield an overview of the baseline, the state of the art and skills of each company or sector, the gap between the current models and the models obtained by PN, and a plan of initiatives to reduce this gap. This process can be seen in Figure 1.

Data collection and the diagnosis is made through three fundamental points:

- Current business and productive strategy: must be done with site visits to each company or industry in order to define the vision and business systems. Current ways of business, production model, systems and technology, experience in R&D issues and business needs; it is done with sessions of interviews to the various areas involved and their managers

- Architecture or state of applications and technology: based on questionnaires given to the responsible for application and production areas. Formats should be simple, very protected, allowing us to capture data from the production or R&D project for further analysis as accurately as possible.

- Development of diagnosis: With the information gathered, the analysis of the current situation in each company or sector will be developed, where the technical maturity will be evaluated, as well as the sensitivity to the integration of new technology, its perception, or the elements of methodology of integration previously used by the organization that may be useful, as well as the objectives of subsequent phases.

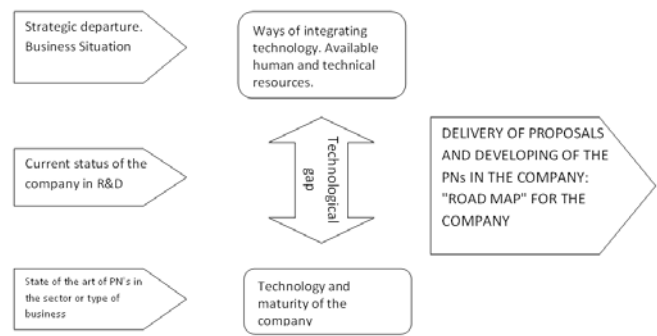


Figure 1: Methodology of analysis of PN implementation in companies of La Rioja

Finally, it would be very interesting to develop a deliverable for Business named "Strategy and foundations of a PN model for production management", which quantifies technical gap, the technical means available and necessary, and a series of proposals reflected in a starting "road map" that the companies should take to achieve the successful implementation of the methodology of pre-PN phase of modeling, simulation and analysis. Additionally a "cash flow" or plan of ROI showing the added value provided globally by the PN process, including potential or repayable financial support of the Administration, Could complement the document.

6. CONCLUSIONS AND MAIN RESULTS.

Accordingly to this preliminar analysis, industry from La Rioja is based primarily on the so-called "traditional" sectors, especially if we value that the production of wine, shoes and furniture, account for over 50% of economic activity in La Rioja. Additionally, these companies have a strategic organizational model and very traditional: the transition from the first generation of many SMEs to the second generation, better trained and adapted to the new methodologies, is being much slower than in other regions, mainly because the good economic figures before the crisis occasioned that the crisis affected La Rioja later than the other regions.

Therefore, in order to deal with projects like the proposed around the PN, the region has a first line of strategic-cultural handicap because R&D has been seen by many of the companies more as a complement (or even something unnecessary) than to normal and necessary activity. This has led many companies to face later incorporation into the new strategic models, despite the existence of financial and human resources provided by the authorities that have facilitated and facilitate nowadays the transition.

Therefore, in general, we can say that the industry of La Rioja lacks mature enough to tackle a research project of optimization of the production based on PN (ranging from modeling to optimization). It is therefore

necessary to address the first phase of awareness the support of the Administration and later to deal modelling projects on companies that presents a sufficient level of maturity through specific programs related to projects or implementations of projects of R&D, as a supplement to for example the implementation of nacional regulations UNE 166000 (Management of R&D and innovation). So a series of conclusions and general recommendations (as well as some more specific ones) are made in this work, for the analyced sectors, in order to promote the integration of PN techniques in the industry of La Rioja.

6.1. General recommendations

The industry of La Rioja must become sensitized in the application of innovative techniques and the tools provided by the PN for the implementation of R&D projects. In order to achieve greater maturity, it would be interesting to carry out some of these initiatives:

- Campaign awareness pre-production enhancement techniques. Disclosure of them, especially the PN through specific workshops or presentations, in situ, in an informative and forthcoming way.

- To adopt systems and formats that fit the business profile of La Rioja industry. If the initial maturity is low, we should perform a simple and close approximation to the strategic management of the companies that show, especially through case studies similar to those described above, the interest of the PN for modelling, simulation and optimization. Approaches would also be interesting in order to perform return on investment that the company should carry out in PN, demonstrating their added value.

- To seek support from national and regional administrations and to work in cooperation with companies and industry organizations for the development of PN. Bringing the triple helix model (Industry, University, Government) to companies in a tangible and easily way. Given that La Rioja has a University and an active Development Agency, it would be interesting to carry out a series of audits on industrial organization and R&D in which the balance of the interest of the inclusion of PN within the Rioja business network could be determined.

- That such cooperation includes, where possible, tools of financial support that at least allow companies to assess their baseline pre-PN modeling with a low cost. This would be particularly interesting to search for knowledge transfer from the incorporation of technologists from the University of La Rioja, because of its proximity and experience to businesses. We also consider that such support should take precedence collaboration between companies, the aforementioned incorporation of technologists, and that the studies should be addressed as a whole (or cluster) in order to optimize resources and achieve that the PN have a

global impact instead of only local (without particular solutions).

- As seen below, support for research in some of the most characteristic sectors should increase, like wine, which has found an interesting niche development that should be addressed as a way of obtaining value for the sector and by extension for La Rioja.

We now proceed to discuss in Table 1 specific proposals for each of the sectors analyzed in. It is noted that there are some priority sectors of application, such as wine or snuff, if what is sought is that research has a high degree of novelty. However, if the focus is on modernizing the production model in La Rioja, the focus should be more towards furniture and footwear as most representative sectors.

REFERENCES

- 1992 Japan - USA Symposium on Flexible Automation
Advances in Materials Manufacturing Science and Technology XIII: Modern Design Theory and Methodology, MEMS and Nanotechnology, Material Science and Technology in Manufacturing, Materials Science Forum, Volume 628 629, 2009, 734p
- Bartz, R. 2010 Contribution to a workflow-based information management in automotive testing and data análisis, Proceedings of the IEEE International Conference on Industrial Technology, 2010, Article number 5472558, Pages 1026-1031
- Carpanzano, E.a , Cataldo, A.a , Tilbury, D.b 2004 Structured design of reconfigurable logic control functions through sequential functional charts, Proceedings of the American Control Conference, Volume 5, 2004, Pages 4467-4471
- Chen, J.a , Zhang, L.-W.b , Luo, J.-Q.b 2009 Study on reconfiguration cost modelling of Reconfigurable Manufacturing System IET Conference Publications Volume 2009, Issue 556 CP, 2009
- Cicarelli, F., Furfaro, A., Nigro, L. 2010 A service-based architecture for dynamically reconfigurable workflows, Journal of Systems and Software, Volume 83, Issue 7, July 2010, Pages 1148-1164
- Cicarelli, F., Furfaro, A., Nigro, L. 2010 Using time stream Petri Nets over a service architecture for workflow modelling and enactment Spring Simulation Multiconference 2010, SpringSim'10, 2010, Article number 131
- Costa, A.a c , Gomes, L.a c , Barros, J.P.b c c , Oliveira, J.a , Reis, T.a 2010 Petri nets tools framework supporting FPGA-based controller implementations , Proceedings - 34th Annual Conference of the IEEE Industrial Electronics Society, IECON 2008, 2008, Article number 4758345, Pages 2477-2482

SECTOR	STRATEGIC MATURITY	MATURITY IN R&D	APPLICABILITY OF PETRI NETS	ADDED VALUE OF PETRI NETS
Wine Sector	Medium-Low. Big companies are adapted to the new environment but there are plenty of small traditional wineries	Medium. There are R&D projects mainly related to big companies. Currently there is a tendency to increase investment in this area, although there are still many companies that do not have wine experts or technicians and R&D departments.	Very high. There are hardly any references. The traditional origin of production processes has hindered the technology transfer.	It is recommended the bid for product differentiation, using the PN to study the impact in production and distribution of new wines based on R&D from new varieties.
Auxiliary Automotive Sector	Medium-High. It is a sector heavily influenced by the groups to which they provide, which are generally part of the same group. Thus many companies already assume management techniques from its parent companies. In any case it is a sector that traditionally has bet, also in La Rioja, by advanced management techniques.	Medium-High. There are R&D projects for years and many companies have their R&D to continuously generate ideas. However a need exists for greater cooperation between companies although there are already ongoing corrective measures.	Medium. In this case the degree of novelty of the application of PN is quite low. Locally, it may be considered a productive innovation, although it PN integration would consist, in the first phase, mainly in integrating solutions or models already implemented successfully.	PN are recommended to be used as support tool for optimization and improvement of productive activity. Also as supplement in the R&D carried out in this sector.
Footwear Sector	Medium-Low. The sector in La Rioja presents a significant gap between large firms with capacity and experience in strategic management and other family companies in a intergenerational transition period.	Medium. In the last 5 years, companies have tried, more or less, to develop, internally or with their suppliers, improvements to compete with the Asian market. In addition, the presence of the Technology Centre in the environment is a basic support.	Medium-High. PN can be considered a new and important support in decision-making in R&D, especially in the case of safety footwear.	In this case it is proposed that PN are used to support the transition from classical to a production model based on the R&D through a gradual adjustment: NP are proposed as a tool for analysis and development of projects.
Snuff Sector	Very high. Imperial Tobacco is a multinational group that applies the most modern production techniques and R&D results both to productive and strategic management.	Very high. It has different areas of R&D and is continually developing new projects to adapt to legislation and consumer, or as a strategy of differentiation.	High. While there are solutions in the industry previously, Imperial Tobacco has both the capacity and interest enough to develop projects based on PN from modeling to optimization. Perhaps this is the most suitable plant for this in La Rioja.	In this case the contribution that can bring Petri nets is much lower than in other sectors less mature. That is why we must seek that PN add value in any area of special interest, highlighting the possible improvement in product distribution and logistics.
Furniture Sector	Medium-Low. In La Rioja some traditional companies coexist with companies applying more productive techniques. However, the most typical profile is a cooperative or a family company with classic strategic-functional structure.	Medium. There exists very interesting R&D projects in course, but generally developed through collaboration with suppliers or on topics of little technological interest. It is facing an evolution in R&D to provide competitive advantage	Medium. There exist previous success cases and companies that have sensitivity towards innovation. However, it is necessary a strategic and productive evolution in the short to medium term to advance from only modeling to simulation and optimization.	The Rioja furniture companies need to vary its traditional production model to proprietary products, characterized by modularity, flexibility of production and innovation. All these values can be provided through PN models.

Table 1: Summary of analysis of PN applicability in the main industrial sectors of La Rioja

- Gradišar, D.a b , Mušič, G.a 2007 Production-process modelling based on production-management data: A Petri-net approach, International Journal of Computer Integrated Manufacturing, Volume 20, Issue 8, December 2007, Pages 794-810
- Han, K.-H.a , Yoo, S.-K.b , Kim, B.c 2009 Integration of UML and Petri Net for the process modeling and analysis in workflow applications Proceedings of the 13th WSEAS International Conference on Computers - Held as part of the 13th WSEAS CSCC Multiconference, 2009, Pages 255-262
- Jeng, M.a , Xie, X.b , Chung, S.-L.c 2004 ERCN* Merged Nets for Modeling Degraded Behavior and Parallel Processes in Semiconductor Manufacturing Systems, IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans. Volume 34, Issue 1, January 2004, Pages 102-112
- Kleyner, A.a , Volovoi, V.b 2010 Application of Petri nets to reliability prediction of occupant safety systems with partial detection and repair Reliability Engineering and System Safety Volume 95, Issue 6, June 2010, Pages 606-613
- Li, X.-P.a c , Zhao, W.b c c , Liu, D.-X.a c c , Yuan, C.-Y.a , Zhang, S.-K.b c c , Wang, L.-F.b c c , 2010, A supply chain modeling technology based on RFID discovery service , Tien Tzu Hsueh Pao/Acta Electronica Sinica Volume 38, Issue 2A, February 2010, Pages 107-116
- Mendes, J.M.a , Restivo, F.a , Leitão, P.b , Colombo, A.W.c 2010 Petri net based engineering and software methodology for service-oriented

- industrial automation IFIP Advances in Information and Communication Technology Volume 314, 2010, Pages 233-240
- Miao, Z., Xu, K.-L. 2009 Research on control policy for lean production systems based on petri net, 2009 International Conference on Information Management, Innovation Management and Industrial Engineering, ICIII 2009 Volume 2, 2009, Article number 5370440, Pages 557-560
- Telmoudi, A.J.a , Nabli, L.b , M'hiri, R.c 2009 Modeling method of robust control laws for manufacturing system to temporal and non temporal constraints through Petri nets, International Review on Computers and Software, Volume 4, Issue 2, March 2009, Pages 266-277
- Wang, H., Dong, T., Zhang, J., Wang, H. 2010 Simulation and optimization of the camshaft production line based on Petri net Advanced Materials Research Volume 139-141, 2010, Pages 1506-1509
- Wang, J. 2009 Automotive supply chain performance influencing path analysis based on fuzzy petri net , 2009 International Conference on Information Management, Innovation Management and Industrial Engineering, ICIII 2009 Volume 1, 2009, Article number 5368162, Pages 359-362
- X.Lu 2010 Modeling of control system for LAMOST based on Petri net workflow, Proceedings of SPIE - The International Society for Optical Engineering, Volume 7738, 2010, Article number 77381I
- Xiao, R., Chen, T. 2010 Research on design structure matrix and its applications in product development and innovation: An overview, International Journal of Computer Applications in Technology Volume 37, Issue 3-4, March 2010, Pages 218-229
- Xu, Y., Zhang, M., Tang, S., 2010 Research on workflow model of cooperation between 4PLs and 3PLs based on Petri-net 3rd International Conference on Information Management, Innovation Management and Industrial Engineering, ICIII 2010; Kunming; 26 November 2010 through 28 November 2010; Category number P4279; Code 83865
- Yu, X.-L., Jiang, J., Xia, B.-Q., Pan, Z.-K. 2010 Petri-net-based analysis method for grid services composition model CICC-ITOE 2010 - 2010 International Conference on Innovative Computing and Communication, 2010 Asia-Pacific Conference on Information Technology and Ocean Engineering 2010, Article number 5439240, Pages 180-184
- Zhan, Y.-D., Luo, Y. 2001 Modeling and simulation research of material handling automatic system based on Petri Net , Xitong Fangzhen Xuebao/Acta Simulata Systematica Sinica, Volume 13, Issue 4, 2001, Pages 501-504
- Zhang, D., Zhang, P., Jiang, P., Tan, R. 2010 Contradictions determination method in product design using Petri net 5th IEEE International Conference on Management of Innovation and Technology, ICMIT2010; Singapore; 2 June 2010 through 5 June 2010; Category number CFP10795-ART; Code 81208
- Zhao, N., Dong, S., Ding, W., Chen, L. 2010 The modularization of the tobacco distribution center simulation ICLEM 2010: Logistics for Sustained Economic Development - Infrastructure, Information, Integration - Proceedings of the 2010 International Conference of Logistics Engineering and Management Volume 387, 2010, Pages 2271-2277

STOCHASTIC OPTIMIZATION OF INDUSTRIAL MAINTENANCE STRATEGIES

Castellanos F.^(a), Wellens A.^(b)

Facultad de Ingeniería – UNAM, México

^(a) Ing.fcastellanos@gmail.com, ^(b) wann@unam.mx

Abstract

The dynamics of business requires industries to produce more and more at the lowest cost, highest quality and a high level of reliability (availability and reliability of their equipment) to meet the stringent technical, economical and legal requirements, and to remain efficiently and competitively in the market.

The aim of this research is the development of a mathematical simulation model for the optimization of maintenance policy in a complex industrial production system, in which maintenance practices and production are integrated into a single procedure, in which control variables are the maintenance policies. This was done using a combination of maintenance management techniques and stochastic processes.

This work combines an efficient methodology for maintenance management, such as RCM, with mathematical modeling, to be able to optimize the availability of a complex system.

To be able to compare different maintenance strategies and the corresponding reliability, availability and economics, the model's output variables are the system and parts availability in a given time.

Two different simulation strategies were used, on one hand being the modeling of the complex system using an Excel database and the interaction of variables through constraints and likelihood behavior, and on the other hand Matlab's Simulink / SimEvents, feeded by equipment characteristics, behavior, the companies' environment, needs and maintenance strategy actions.

Keywords: *reliability, maintenance, stochastic processes, simulation*

1. Introduction

Since even industrial equipment of the same type tends to behave in a different way, maintenance techniques consider equipment as independent items, which mean that maintenance decisions are made for individual equipment. The common practice is that decisions are made regarding the most problematic equipment and afterwards these decisions are generalized to other equipment of the same type. This causes maintenance to be effective, although usually with very high costs and unnecessary in some equipment.

Industrial equipment computers have a feature that generally is not taken into account in the determination of its reliability: the equipment may have failures that do not limit their production capacity, but increase the proportion of defective products, operation or preparation time, etc.

This paper discusses a computer application designed as a decision support tool for selection of maintenance activities taking into account the risks and costs associated with choosing different maintenance strategies. Rather than searching for a solution to a problem: "what maintenance strategy would lead to the best reliability and dependability parameters of system operation", in this approach different maintenance scenarios can be examined in "what-if" studies and their reliability and economic effects can be estimated.

The proposed model represents the reality of the machines and can be successfully implemented as a maintenance management system as for example Reliability Centered Maintenance (RCM). The model makes use of stochastic process theory, specifically semi-Markov chains in continuous time (SMC). The

first difficulty to model the problem is its mixed discrete/continuous condition. On one hand, the equipments degree of damage is considered to be continuous in time; on the other hand the maintenance inspection system which detects the state of the team and take actions for it, is discrete in time. This makes the mathematical system to predict the total system state at time t a hybrid stochastic system (HSP) and thus quite complex. To overcome this problem, simulation will be used to optimize the system. The second difficulty with respect to the modeling of the system is that the behavior of a real maintenance system is usually unknown: due to its nature, accurate data is difficult to obtain as in case of actual failure, inspection and reparations are carried out, so future state of the equipment without maintenance cannot be known. To overcome this limitation, subjective information from maintainers is taken into account in the model data.

1.1 Other research

The simulation of maintenance models generally do not represent a complete the reality of the machines by focusing on very specific aspects of individual machines. In the aging model proposed by Jaroslaw Sugie refers to a maintenance model which does not depend on changes in the maintenance, rather focuses on the deterioration of the machine, so it does not propose changes to improve efficiency the machine, its life or reduce costs.

This system proposes a new approach, the machine as part of a system which responds to the decisions arising from the maintenance management and selection policies. This model shows costs of each policy, use of the machine and economic performance in the system.

2. Modeling System

The deterioration process of the equipment will, to a large extent, depend on the adopted inspection and maintenance policy. However, it is often difficult to determine with reasonable confidence just what the best frequency of equipment inspection is, or what should be inspected. As a result, some

maintenance procedures may cost more money than they should or essential equipment may be unnecessarily taken out of service for prolonged periods of time.

Preventive maintenance policies are either aimed at detecting deterioration of the equipment before it fails or simply are adopted on the a priori assumption that the equipment has deteriorated and requires replacement. In either case, there is a need to select the maintenance frequency and extent so that the desired objectives are achieved. As well, the operators want to maximize benefits (which would results in reduced equipment deterioration and reduced cost of equipment replacement and repairs) and to minimize the costs of the maintenance activities.

2.1 Life cycles

The simulation model will determine the maintenance strategy that maximizes the availability of equipment; the availability is linked to the equipment's aging and corresponding life cycle, which in turn is also linked with the maintenance activities. These life curves represent the relationship between the equipment's technical or financial condition, and time. Since there are many uncertainties in predicting equipment life, the probabilistic analysis of failure rates should be done carefully to construct and evaluate the life curves properly. Figure 1 shows an example of a life curve and its modeled equipment status over time with different maintenance policies. The simulation results are similar to the presented ones.

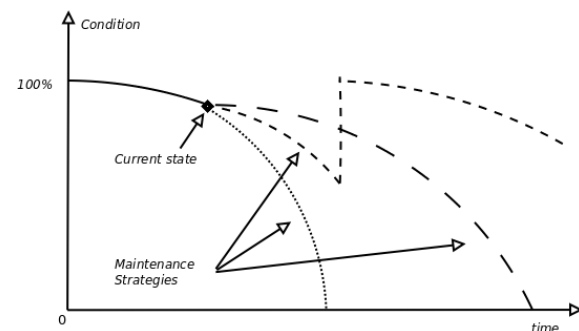


Figure 1: Typical life cycle of industrial equipment.

The three examples of maintenance policies are:

- 1) Stop all maintenance actions.
- 2) Continue current maintenance policies.
- 3) Reduce partially existing maintenance policies after some time.

2.2 Application Scenarios

Optimization of maintenance policy using the proposed tool is applicable to any company that meets the following initial conditions:

- Being a medium sized manufacturing industrial
- Have a maintenance department or maintenance operations
- The machinery must be analyzed and removable/reparable

All variables in the model must be analyzed for each case study, the values that will in many cases are subjective information by the maintenance technicians.

2.3 Model Explanation

The simulation model consists of a machine that has a number of stages during his life. Each stage is represented as $D1$ to $D4$, each has different levels of productivity, efficiency and quality of product produced. For example, $D1$ represents brand new machine with 0 defects and 100% efficiency, $D4$ represents the machine in the last stage of their productive life with low efficiency and many defective products.

The transitions between states are represented by Dn a Markov chain with a continuously variable λ which determines the values of the exponential distribution for the state changes represented as $PDij$. When the team arrives at $D4$ and changes state again becomes F means failure of the equipment.

During his lifetime the machine have a system to prevent this reaches F , this is maintenance. Each determined time inspections are performed (In) to the machine to determine the degree of wear and take decisions about whether to be left like this, send it to light maintain (MnI)

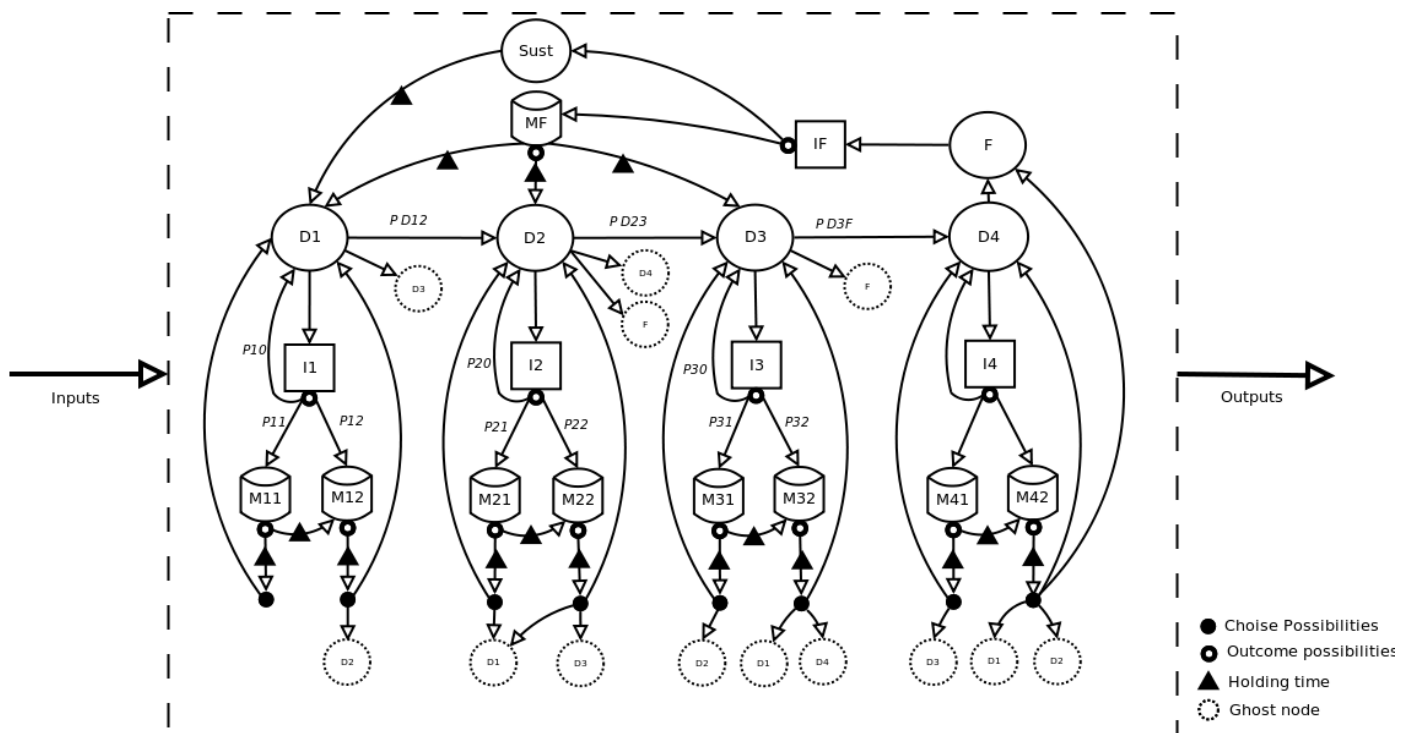


Figure 2. Graphical representation of the proposed Markov model.

or intensive maintenance (*Mn2*). These decisions will be taken by the analysis and experience of the inspector. This value should be studied for each simulated case and will likely *Pn0*, *Pn1* and *Pn2*. The inspection system is represented by a discrete markov chain imbued. The maintenance system is represented by a semi-markov chain.

If the machine was sent to maintenance, it going to be out from production system, this will cause delays also maintenance generates costs and a timeout to return to activities. Waiting times and maintenance costs are normally distributed with μ and σ , which are greater in *Mn2* case.

At the end of the holding time before maintenance, the machine will return to production at some stage *Dn* which will be simulated by a binomial distribution. Depending on the situation the machine and maintenance process, as result may have a minor, higher or same wear state after maintenance.

When the machine reaches state *F* has a special status of the inspection (*IF*) which determines the conditions of the machine and if it can be repaired through an exhaustive maintenance (*MF*) or must be replaced completely (*Sust*). In *MF* has high waiting times and high costs, the machine could return to any of the first

3 states of deterioration, depending on the skill of workers; in *Sust* have a high cost for the acquisition of new machinery. If the system reaches the final state of *Sust*, the simulation stops because it would be the life of a second machine. The graphic model system shows in Figure 2.

The model input variables are divided into 2 types: variables and variables of computer maintenance. Parameters of the equipment are all variables that are intrinsic to the equipment, facilities, the external environment as demand and costs. Maintenance variables are variables that can simulate different maintenance strategies, these variables are: time between inspections, experience levels of inspectors and inspection fidelity.

These variables fed the model and responds with results, also varying levels to design the best strategy. The model results are divided into 3 categories, each of which will have a different priority in each case: financial, time available and maintainability.

The machinery is in a industrial system of production inputs and outputs. This system is designed so that each time *t* (representing 1 day) has outputs useful for decision-making as downtime, number of parts produced, defects,

	Replica	Time available	Garanty cost	Maintenance Cost	Total Cost	Number of maintenance	Time until failure	
1	default maintenance policy	1	81%	\$450.00	\$6,300.00	\$6,750.00	21	1173
		2	85%	\$512.00	\$6,900.00	\$7,412.00	23	1150
		3	83%	\$571.00	\$6,900.00	\$7,471.00	23	1144
2	Inspections each 10 days	1	92%	\$270.00	\$9,600.00	\$9,870.00	32	1250
		2	93%	\$356.00	\$8,700.00	\$9,056.00	29	1310
		3	89%	\$381.00	\$8,400.00	\$8,781.00	28	1330
3	Inspections each 20 days	1	95%	\$255.00	\$8,400.00	\$8,655.00	28	1850
		2	93%	\$456.00	\$9,300.00	\$9,756.00	31	1890
		3	94%	\$421.00	\$9,600.00	\$10,021.00	32	1880
4	Inspections each 40 days	1	92%	\$295.00	\$7,500.00	\$7,795.00	25	1173
		2	89%	\$396.00	\$9,300.00	\$9,696.00	31	1150
		3	72%	\$378.00	\$8,400.00	\$8,778.00	28	1189
5	No inspections	1	100%	\$15,145.22	\$0.00	\$15,145.22	0	986
		2	100%	\$14,416.24	\$0.00	\$14,416.24	0	1009
		3	100%	\$13,825.46	\$0.00	\$13,825.46	0	992

Table 1: Main model results

delays, etc. Those are transformed into costs for the analysis of the financial part of the model. It takes into account each day that the machine is under maintenance or inspection to assess the percentage of the working life of machinery. The model collected how often the machine has been sent to light or intensive maintenance to determine the maintainability of the equipment.

2.4 Setting the model

The proposed model has great flexibility in representing various types of maintenance systems and situations, which may range from corrective maintenance, preventive and situational analysis of the team.

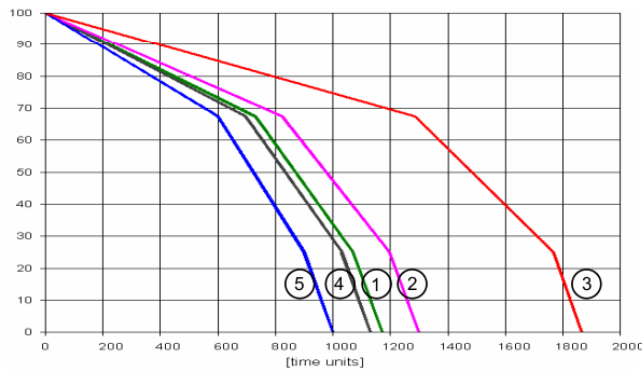


Figure 3: Results of life curve of equipment for default and generated maintenance policy from markov model.

As input variables are the specific qualities of the maintenance system, the main ones: the set time between inspections, the rigor of the inspection, maintenance time and costs and the length of time between the various states of impairment, as well the company variables such as demand, production costs, sales costs, compliance costs and rework costs.

3. Application example

Model was used to simulate production machinery within a company dedicated to selling radiator cooling tubes. The system was tested in 5 different conditions: default policy, inspections of 10, 20 and 40 days, without inspections. There were 3 replicates of each type and the results were

as follows, Figure 3 shows the curves of life of different models to the point of wear.

As noted on table 1, you have different factors to select the most appropriate policy. If you search the availability of the machine would select the alternative 5, if you want to reduce costs is due to select Alternative 1 and to extend the life of machine to the maximum recommended Scenario 3.

4. Conclusions

The optimization of maintenance activities through the implementation of techniques such as simulation, situational analysis and maintenance management techniques play a key role in industries which increase their competitiveness by achieving better results with high quality at lower cost possible.

The paper presents a markov model adaptation method that allows adjustment of the basic model to user-expected changes in maintenance policy. The model has the ability to adjust to working environments commonly found in industries and can be extended to represent more complex systems, confirming the effectiveness of a system or proposing a new one with a tuning of policies.

References

- **Abdel-Hamed, Mohamed S. (2004). Optimal predictive maintenance policies for a deteriorating system: The total discounted cost and the long-run average cost cases. Communications in statistics.**
- **Ascher H.; Feingold H. (1984). Repairable systems Reliability. Modelling Inference, misconceptions and their causes. Marcel- Dekker**
- **Christer, A.H. (1999). Developments in delay time análisis for modelling plant maintenance. Journal of Operations Research Society.**
- **Jaroslav Sugie (2007). Modeling Changes in Maintenance Activities through Fine-Tuning Markov Models of Ageing Equipment. 2nd International Conference on Dependability of Computer Systems.**

BIOGRAPHICAL NOTES

Francisco Castellanos obtained his bachelor degree in Industrial Engineering (Tecnológico de Monterrey, México). Currently is studying a master's degree on System engineering at UNAM.

Ann Wellens is a chemical engineer with postgraduate studies in Industrial Administration (KUL, Belgium) and a master degree in Environmental Engineering (UNAM, Mexico). At the moment she is a full-time lecturer in the Systems Department of the Industrial and Mechanical Engineering Division of the National University of Mexico (UNAM). She has been working in air pollution issues for the last 15 years, dictating courses, collaborating in research projects and participating in conferences related with mathematical modeling of air pollution dispersion and statistics.

DESIGN AND DEVELOPMENT OF DATA ANALYSIS MODULES FOR THE AERMOD AND CALPUFF SIMULATION MODELS

Wellens A.^(a), García G.^(b)

^{(a)(b)}Facultad de Ingeniería, UNAM-MÉXICO

^(a)wann@unam.mx, ^(b)gamarzaid@gmail.com

ABSTRACT

Mathematical models for atmospheric dispersion are being used in a wide variety of industrial applications. Even simplified models have improved their formulation incorporating up-to-date knowledge regarding micrometeorology and dispersion, and can be used to estimate air pollution concentrations around, for example, industrial facilities. Two dispersion models based on Gaussian modeling of the dispersion plume, are of special interest in the simulation of small and medium scale dispersion: AERMOD and CALPUFF are recommended by the US EPA to determine air pollution dispersion.

Both models are freely distributed by EPA, although independent developers offer graphical interfaces (for example ISC-AERMOD, View, Breeze, CALPUFF View) to be able to integrate in a friendly way the topographic, land use and meteorological data, and to represent the results graphically. Although these graphical interfaces are quite complete, specific research projects may need some data manipulation not provided by these interfaces.

This paper proposes some external modules to AERMOD and CALPUFF that extract and prepare in a specific way the output data for certain research needs, such as comparison of the model data with DOAS measurements, etc.

Keywords: consequence analysis, dispersion modeling, CALPUFF, AERMOD.

1. INTRODUCCIÓN

Mathematical models are used extensively in a variety of applications related to the study of air pollution. Examples are, among others, emission modeling, pollutant dispersion modeling, determination of the minimum chimney height, safety audit studies or the modeling of accident consequences. There exists a big variety of models that differ in application type, generated model output, spatial scale, temporary resolution, complexity, method of solution, reference system and required resources.

Pollutants enter the environment in diverse ways. The dispersion of industrial chimney pollutants depend on many correlated factors, as for example:

- The physical and chemical nature of the effluent.
- The meteorological characteristics in the environment.
- The location of the chimney with respect to possible obstructions for the free movement of air.
- The nature of the area located downwind the chimney.

With very few exceptions, the basic approach of the current regulative platform of the EPA for air pollutant modeling in the surroundings of an industrial source has been maintained fundamentally without changes from the beginning of the air programs, approximately 30 years ago. Most used models have been Gaussian ones; they give quick results, but their development is based on quite severe assumptions. Nevertheless, in the last years significant scientific advances have been reached: these have been incorporated in the ISCST3 model (Industrial Source Complex – Short Term Model) to design advanced Gaussian models able to evaluate pollutant transport at long distances and in complex topographical and meteorological conditions. Two of these advanced models are CALPUFF and AERMOD.

1.1 The Gaussian model

The Gaussian model is a particular solution of the general equation for pollutant concentration transport.

Figure 1 illustrates the problem to be studied, including the used coordinate system where the origin is located at the base of the chimney.

In a simple Gaussian model the concentration c of a compound located at the coordinate point (x, y, z) can be described by:

$$c(x, y, z) = \frac{Q}{u 2\pi \sigma_y \sigma_z} \exp \left[-\frac{1}{2} \frac{y^2}{\sigma_y^2} + \frac{(z - H_e)^2}{\sigma_z^2} \right]$$

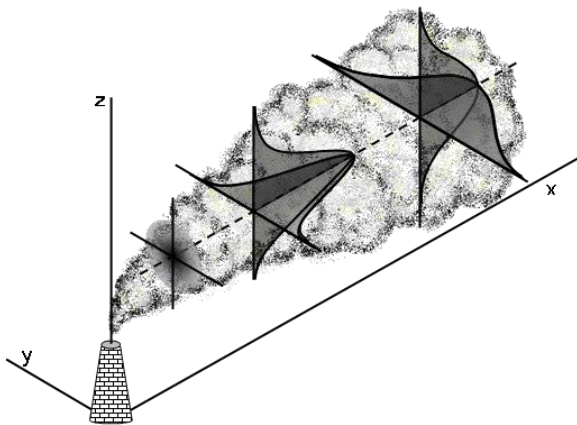


Figure 1. The Gaussian dispersion model.

1.2 The AERMOD model

Of all available models, few are widely accepted. The US Environmental Protection Agency (EPA) proposes and recommends AERMOD to model the dispersion of pollutants of fixed sources.

AERMOD, developed by the US-EPA and the American Meteorological Society, was designed to support the EPA's regulatory modeling programs (EPA 2010). AERMOD is a regulatory steady-state plume modeling system and includes a wide range of options for modeling air quality impacts of pollution sources, making it a popular choice among the modeling community for a variety of applications (Lakes AERMOD VIEW USER GUIDE). Together with the AERMOD code, the US-EPA provides three complementary components: AERMAP (AERMOD Terrain Preprocessor) AERMET (AERMOD Meteorological Preprocessor) and AERSURFACE, a tool that produces surface characteristics data.

As AERMOD includes recent scientific knowledge with respect to the understanding of the planetary boundary layer, it has a more realistic approach in treating plume interaction with the earth's surface than older Gaussian air dispersion models like ISCST3. Since December 2006, AERMOD replaced ISCST3 as the standard regulatory model.

AERMOD, as any other mathematical dispersion model, provides only an estimate of the atmospheric concentration of environmental pollutants, and its results depend on the quality of the corresponding input data, and the methodology used for its determination.

1.3 The CALPUFF model

CALPUFF is an advanced non-steady-state meteorological and air quality modeling system, adopted by the U.S. Environmental Protection Agency in its Guideline on Air Quality Models as the preferred model for assessing long range transport of pollutants and their impacts on Federal Class I areas and on a case-by-case basis for certain near-field applications involving complex meteorological conditions. The

modeling system consists of three main components and a set of preprocessing and postprocessing programs. The main components of the modeling system are CALMET (a diagnostic 3-dimensional meteorological model), CALPUFF (an air quality dispersion model), and CALPOST (a postprocessing package). Each of these programs has a graphical user interface (GUI). In addition to these components, there are numerous other processors that may be used to prepare geophysical (land use and terrain) data in many standard formats, meteorological data (surface, upper air, precipitation, and buoy data), and interfaces to other models such as the Penn State/NCAR Mesoscale Model (MM5), the National Centers for Environmental Prediction (NCEP) Eta/NAM and RUC models, the Weather Research and Forecasting (WRF) model and the RAMS model.

This model has been evaluated and improved by institutions and/or groups like the Interagency Workgroup on Air Quality Modeling (IWAQM-US), the EPA and other north-American and foreign organizations, and at present it is one of the most used models due to the EPA support.

Both AERMOD and CALPUFF are models recommended by the United States Environmental Protection Agency and have been used in a variety of applications and countries. In México, they have been used for example by the Universidad Nacional Autónoma de México in air pollution studies around petroleum refineries or electricity generation facilities (Ruiz Suárez *et al.*, 2010; Jazcilevich *et al.* 2009; Grutter *et al.* 2008). However, the nature of these research projects requires transformation of model results provided by AERMOD or CALPUFF to be able to compare them with real-time measurements.

2. PROPOSED METHODOLOGY

Both AERMOD and CALLPUFF are written in Fortran, which is a not a very flexible programming language. For the proposed external modules, new programming technologies will be used, independently of the original Fortran code of the specific model. In this stage, the development platform .NET is proposed, as it is a powerful system, which provides quick and reliable results.

The methodology used in the.NET platform will help to improve significantly the tasks and the interaction of the modules to be developed with the original AERMOD and CALPUFF code. In .NET a three layer architecture is used, in which every part of the programming is organized in the most efficient way in order to access the information rapidly and efficiently. The programming time is short due to the large number of tools provided in .NET and the versatility of its use.

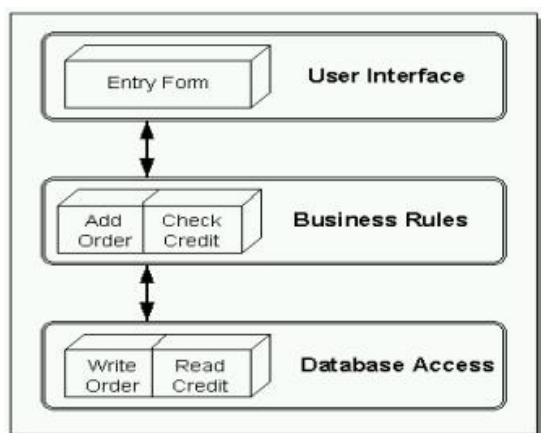


Figure 4. Architecture of the three .NET layers.

Initially, two modules will be developed: the first integrates concentration values at different vertical levels to a sole integrated concentration (in ppm m) to be compared with DOAS integrated concentration data. The other one integrates monthly or seasonal data in an annual concentration, as – due to the size of the meteorological input data – the yearly average cannot be defined in a sole simulation run. In the future, a third module is planned, determining the region where predefined ambient standards are violated when AERMOD or CALPUFF surface concentrations are known.

Once the requirements of each module are defined, the graphical interfaces are developed; the trial version will be evaluated to correct programming errors or to obtain information for future implementations.

3. CASE STUDY

3.1 Study area

The infrastructure of the Mexican Federal Commission of Electricity (CFE) includes 154 energy generation facilities. The number of thermoelectric generation plants distributed in Mexico is 79, of which the Petacalco thermoelectric facility is one of the most important. The Plutarco Elías Calles facility in Petacalco has a capacity of 2100 MW, in six production units. The electric power produced is transported through fifteen transmission lines between 115 and 400 kV.

The plant uses coal as a primary fuel to produce high pressure steam (between 120 and 170 kg/cm²) and high temperature (of the order of 520°C), to move the electrical generator connected to the rotor of the steam turbine.

The electricity generation plant is an important pollution source, emitting among other SO₂, NO_x, particulate matter and CO.



Figura 3. Petacalco thermoelectricity facility.

3.2 Objectives

The general target is to compare the Gaussian model of dispersion of atmospheric pollutants AERMOD and CALPUFF with the spectroscopic skill DOAS in the industrial Petacalco complex, the Warrior's State by means of the creation of external modules to these models that realize the above mentioned comparison and help to work in a practical way the information.

The specific targets of the same sound the following ones:

1. To use the skill DOAS to estimate by implication and wind below the entire emission of SO₂ as well as its spatial distribution about the industrial Petacalco complex.
2. To shape the dispersion of this pollutant gas with AERMOD and CALPUFF using the topography and available meteorological information.
3. To evaluate the exits of the models AERMOD and CALPUFF using the sets of meteorological data, with the results of the measurements of the DOAS.
4. To use the modules designed to work with the information of exit of both measurements.

3.3 DOAS

The Optical Spectroscopy of Distinguishing Absorption (DOAS, for its initials in English) is a method to determine the gas concentrations in the ambience by means of the analysis of the light, principally in the spectral status of the ultraviolet and visible one. The light that travels across the ambience is partially absorbed by the gases along the covered trajectory.

Analysis in the vertical one

For this project, the radiation dispersed by the blue sky is collected by means of a telescope and it analyzed espectroscópicamente to obtain column concentrations

in units of ppm*m. The concentration in column of a gas represents the concentration integrated along an indefinite trajectory.

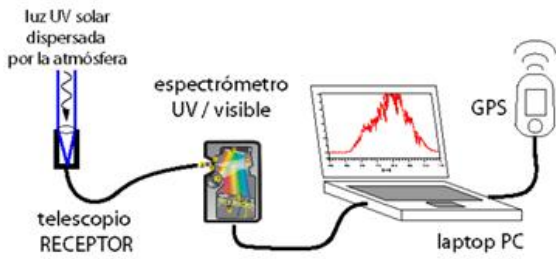


Figure 7. DOAS function

The remarks with the DOAS, on having been measured from a vehicle realizing passages below the pen, give the possibility of estimating the gas flows making use of the speeds of spread of the pen.

The column concentration measures itself to this skill while it passes below a pen or cloud of pollutant gases. The target of this strategy is to obtain the measurement of a "slice" of the pen that disperses perpendicularly over the measurement trajectory.

The facilities chimneys emit, among other pollutants, SO₂ from 100m high chimneys. Besides this facility, four minor SO₂ pollution sources, with lower emission heights (20 to 53 m), are found in the southwest area of the modeling region (Figures 8 and 9). To be able to evaluate CALPUFF model results, Differential Optical Absorption Spectroscopy (DOAS) was used to obtain experimental information on the column concentrations of SO₂ in concentrations of ppm*m. The DOAS technique collects scattered radiation by the blue sky with a telescope and by analyzing the absorbed radiation spectroscopically, integrated SO₂ concentration in the vertical column can be obtained.

Observations were performed with the passive DOAS technique assembled on a van traveling around the industrial complex, below the emitted gas plumes. The traversals downwind were used to evaluate CALPUFF performance. Figure 8 shows a specific DOAS transversal for May 12, 2009 around the electricity generation facility, obtained between 17:03 and 17:19). However, as CALPUFF provides point concentrations, and not vertically integrated concentrations, an external Fortran module was written to integrate CALPUFF point concentrations in different layers in a column concentration comparable with DOAS results. To assure an appropriate horizontal resolution, only 5 vertical layers of different height could be considered in a first approach, as CALPUFF's number of receptors is limited. At the moment, the programming is being extended to include more vertical layers and refine the results.

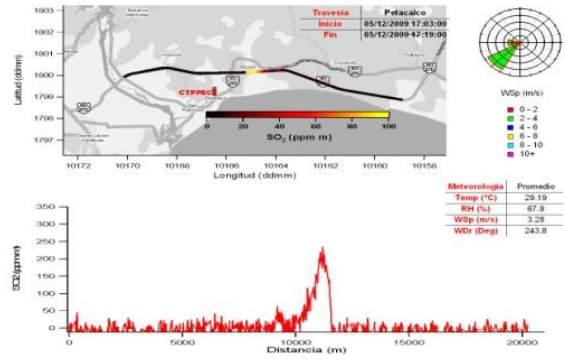


Figure 8. DOAS transversal to be compared with CALPUFF simulation: May 12, 2009.



(a) 2 m



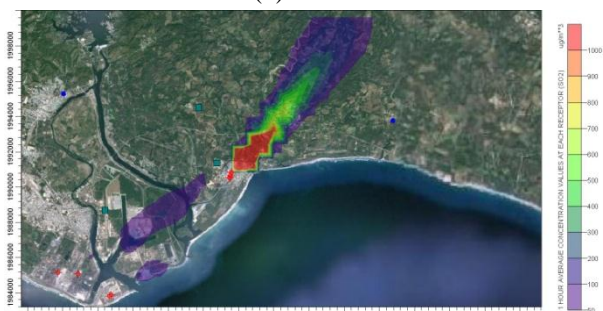
(b) 12 m



(c) 60 m



(d) 200 m



(e) 500 m

Figure 9. Point concentrations at different heights.

CALPUFF, May 12, 2009 (17:00 a 18:00).

As can be seen in Figure 9, the resulting concentration is quite different in different vertical layers: at lower height (below 100 m), the smaller sources in the southwest region of the domain generate higher SO_2 concentrations, while at higher heights (see for example at 500 m), the electricity generation facility is becoming more and more important.

The resulting vertically integrated concentration of SO_2 is quite different from the default surface concentration (see Figure 9(a)) given by CALPUFF or in its case AERMOD, as concentrations in different vertical layers differ a lot. Comparison of CALPUFF/AERMOD results with DOAS was considerably better when comparing the vertically integrated concentration instead of the point concentration at the surface (Figure 10).

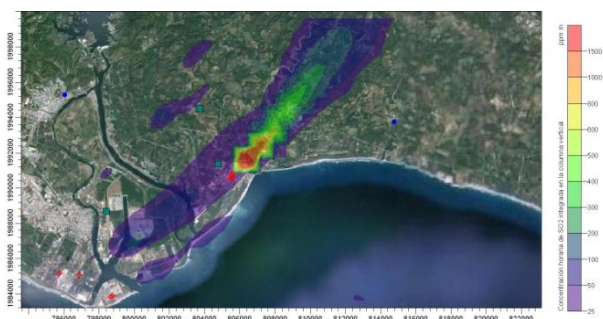


Figure 10. Integrated SO_2 concentration in the vertical column (concentrations in ppm m). CALPUFF, May 12, 2009 (17:00 a 18:00)

As could be observed after analysis of modeling results for different simulation dates and hours, the final

integrated concentration, and thus the quality of the comparison between CALPUFF and the DOAS measurements, depended strongly of the number and value of the chosen heights to obtain the integrated concentration. The extern module for vertical integration of CALPUFF model results is being adapted at the moment to be able to take into account easily receptors at more than 5 heights and to change in an easy way the chosen heights to integrate, as both variables depend on the specific information of the sources in the modeling domain.

4. CONCLUSIONS

Extern modules for CALPUFF and AERMOD were developed to adapt standard output concentrations to specific research needs. Fortran models were written for vertical integration of point concentration in different layers, for integration of simulation data for different trimesters into an annual result, among others. These Fortran models were integrated to the CALPUFF and AERMOD simulation models to offer more flexibility in the results. These extern modules are being adapted at the moment to be able to make them more flexible to research needs and specific case studies.

REFERENCES

- Ruiz Suárez, L.G., Grutter de la Mora, M., Rosas Pérez I., Torres Jardón R., García Reynoso, A., Granada Macías, L.M., Torres Jaramillo, J.A., Wellens Purnal, A., Padilla Gordón, H., Belmont Dávila, R., García García, A., Rebulloza, R., Basaldud, R. (2010), Diagnóstico ambiental de la zona de influencia de la CTPPEC durante la construcción – puesta en servicio de la Unidad 7. Subproyecto aire. Informe final, presentado por el CCA-UNAM a CFE.
- Jazcilevich, A., Siebe, C., Wellens A., Rosas, I. (2009), *Impacto ambiental y en la salud de los habitantes de las actividades mineras en el Distrito Molango, Hidalgo. Subproyecto: estudios ambientales*. Informe final de la segunda etapa del proyecto, proyecto 100662 del CCA-UNAM, presentado al International Development and Research Center of Canada (IDRC).
- Grutter de la Mora M., García Reynoso A., Torres Jardón R., Limón Sánchez T., Wellens Purnal A., Basaldud R., García Escalante J. (2008), *Emisiones a la Atmósfera y Calidad del Aire en la Central Termoeléctrica Plutarco Elias Calles*. Primer informe parcial. Proyecto CFE/PUMA-UNAM. Mayo de 2008.
- EPA (2010) Support Center for Regulatory Air Models, Technology Transfer Network. <http://www.epa.gov/scram001/>. Fecha de consulta: enero 2010.

BIOGRAPHICAL NOTES

Gamar Castillo G. obtained his bachelor degree in Industrial Engineering (UNAM, México).

Ann Wellens is a chemical engineer with postgraduate studies in Industrial Administration (KUL, Belgium) and a master degree in Environmental Engineering (UNAM, Mexico). At the moment she is a full-time lecturer in the Systems Department of the Industrial and Mechanical Engineering Division of the National University of Mexico (UNAM). She has been working in air pollution issues for the last 15 years, dictating courses, collaborating in research projects and participating in conferences related with mathematical modeling of air pollution dispersion and statistics.

DEVELOPMENT OF A SIMULATION TOOL FOR CONSEQUENCE ANALYSIS IN INDUSTRIAL INSTALATIONS

Pérez V. ^(a), García G. ^(b), Ávila M.G. ^(c), Castellanos F. ^(d), Wellens A. ^(e)

Facultad de Ingeniería, UNAM-MEXICO

^(a)victorphz@comunidad.unam.com.mx, ^(b)gamarzaid@hotmail.com, ^(c)gavila@hotmail.com,
^(d)ing.fcastellanos@gmail.com, ^(e)wann@unam.mx

ABSTRACT

At present, a variety of models exist to estimate the magnitude of the consequences of different types of accidents, most of them having an empirical basis. These models aren't for exclusive use in industrial facilities and can also be used for consequence analysis in for example chemical or petroleum industries or in environmental impact studies. Although a wide variety of mathematical expressions are available for events as unconfined or confined explosions, liquid spills or gas leaks, releases of hazardous materials, BLEVES or fires, few simulation packages exist that can assist a non-specialist in decision making related to consequence analysis. In general, existing simulation packages are not freely available for undergraduate students due to their costs, accessibility or required knowledge.

This paper describes the development of a didactic tool for the simulation of accident consequences, for academic and free use in the Departamento de Ingeniería Mecánica e Industrial of the Universidad Nacional Autónoma de México. The software is being developed using Visual Basic, as this is a widespread developing platform, easily accessible and quick to assimilate among the students of the different disciplinary areas of the engineering faculty.

Keywords: Consequence Analysis, Dispersion Modeling, CALPUFF, AERMOD.

1. INTRODUCTION

Currently, due to the large amount of consumer goods that are required to meet the needs of human beings, is process-intensive industrials as well as more raw materials in most cases are classified high risk by flammability and instability of these (usually fuel).

No wonder that major recorded accidents occur:

- Chemical industry
- Oil industry
- Transportation of materials
- Accidents industrials within plants (mainly boilers and leakage of materials).

Predicting this type of accident is difficult especially if there is not a vitacor maintenance or there is no process for responding to incidents of any kind.

For this reason, it is important to keep in mind that while an incident is not present, does not mean it can not happen.

To estimate consequences of the eventualities of this nature has been developed simulation models are empirically based and that the occurrence and some kind of accident will always occur in an unpredictable manner so that the models allow an estimate of the magnitude of the consequences that could trigger an unwanted event before.

Based on historically recorded events such methods have been tested and types of incidents that are triggering the methods of these have been classified into the following groups:

- BLEVE
- Leaks
- Fires
- UVCE
- Explosions
- Detonations
- Implosions

In this work has been developed a computational tool (software) that grouped the different types of incidents for which has developed a simulation model in order to provide a tool to determine the main variables considered to estimate the magnitude of the consequences that may occur in the event of the event (accident).

In these models, use is made of physical parameters to estimate the conditions that affect the severity of the event (usually environmental conditions) as well as being specific to each type of incident.

Later succinctly addresses each of these events and shows the implementation within the computational tool presented here.

2. PROPOSED METHODOLOGY

The implementation of the models related to the analysis of consequences in case of accidents industrials described here, has developed a computational tool (software) in order to make a contribution to the academic faculty for teachers to students in different disciplinary fields in the Faculty of Engineering, UNAM, although initially considered as a support to those related to Industrial Engineering.

In this first version of the tool has been considered as a development platform to Visual Basic as the language of simplicity, versatility, syntax and semantics digestible and available to students without further delay.

It is considering the software is enriched in the near future in order that it can be a powerful tool strong and competitive open source and freely distributed through information technologies, including Web.

Overall, the development of the tool and its expected evolution is intended to have a robust tool for analysis of impact and implications for coping with the lack of resources needed for the acquisition of commercial tools such as the CALPUFF or AERMOD (although their focus is oriented towards the analysis of dispersion).

3. DEVELOPMENT

3.1. BLEVE

The BLEVEs (Boiling liquid expanding vapor explosion) is a type of mechanical explosion which corresponds to a special case of catastrophic explosion of a pressure vessel in which it occurs a sudden escape to the atmosphere of a large mass of fluid (gas or liquefied pressure) overheated.

The main feature is that the explosive expansion of the entire mass of liquid evaporates suddenly, increasing its volume 200 times.

The variables that most interest for this type of incident are:

- Critical temperature and pressure (thermodynamics).
- Fragments of materials.
- Number of fragments.
- Initial speed.
- Average speed.
- Overpressure.
- Speed of fragments in function away.
- Maximum distance range of the fragments.
- Thermal radiation.
- Diameter of the fireball.

- Initial diameter at ground level Hemisphere.
- Height of the fireball.
- Duration of the fireball.
- Thermal radiation received.

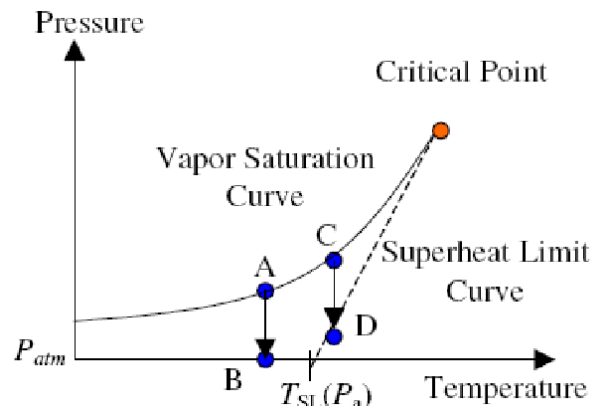


Figure 1. : Reid's 'Superheat Limit Temperature' theory for BLEVE formation.

3.2. FIRES

Due to the nature of the different materials used in industry and the environmental conditions that combine at some point, the fires have originated in material spills and from this state branches into evaporator, scattering clouds flammable pool fire, flash fire type, dispersions with concentration and toxic clouds depending on the case (see Figure 3).

To analyze the effects that are of interest to such incidents, the main variables of interest are related to thermal phenomena: thermal radiation, atmospheric transmissivity (τ), burning rate, flame height and diameter of the fire.

When facilities are closed and contained, is formulated in a manner equivalent to the radius of the rectangle of a puddle.

In the case of fire in the form represented in the figure, comparable to a liquid fire poured into a bucket, pond or swimming pool angular.

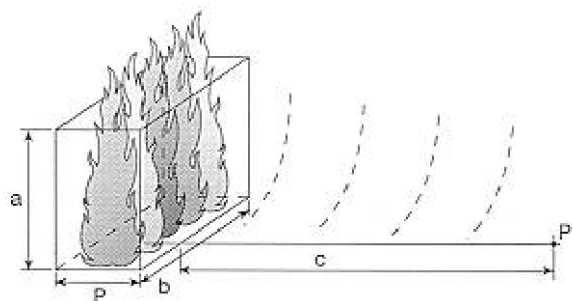


Figure 2. Features pool-type fire.

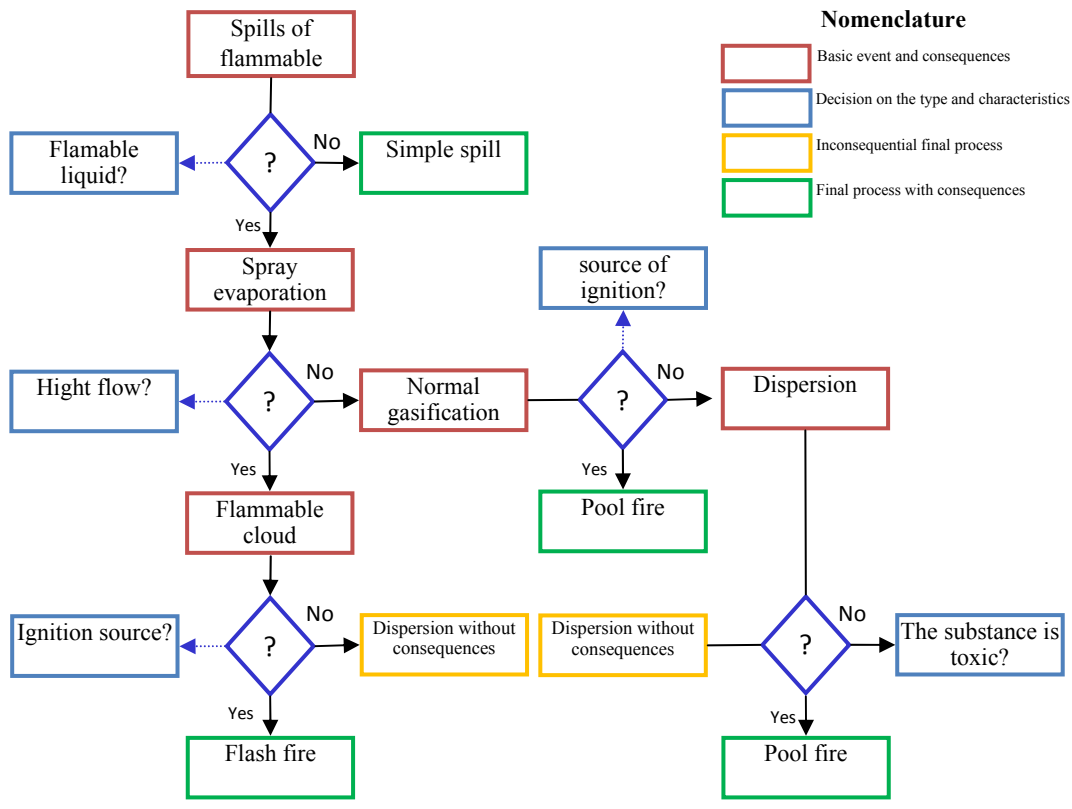


Figure 3. Classification of consequences

3.3. VAPOR CLOUD EXPLOSION NOT CONFINED (UVCE)

Such events can be defined as: Explosion explosive cloud of flammable gas which is in a large space, which pressure wave reaches a maximum pressure of about 1 bar in the ignition.

Unconfined explosions occur outdoors and are usually caused by a quick release of a flammable fluid with a moderate dispersion to form a large flammable cloud of air and hydrocarbon.

The parameter generally defined and measured is the pressure generated by the pressure wave as they propagate undisturbed through the air. Figure 4 shows graphically the value of the pressure versus time.

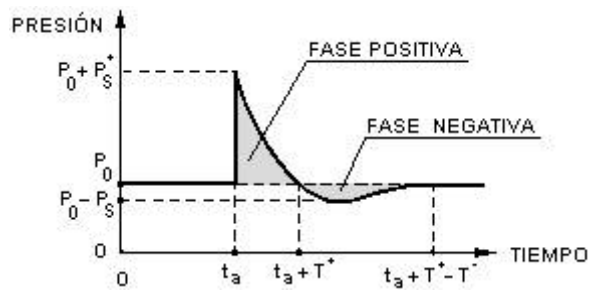


Figure 4. Variation of the pressure wave of an explosion.

In general, the vapor cloud explosions are not confined explosions and rarely have become detonations.

In the case where an explosion is not reached, there would be a quick way to fire blaze that could be defined as a fire called progressive diffusion or premixed flame at low speeds without causing pressure wave. Its most important effect would be the thermal radiation.

The parameter that most interested in this type of event corresponds to the pressure as shown in the figure 4, the pressure can be determined is maximum incident overpressure, maximum lateral pressure, dynamic pressure and overpressure reflected.

TNT Assessment Model

It is based on the hypothesis of equivalence in blast effects from a given mass of inflammable material and a TNT.

In the explosion of a vapor cloud the shape of the initial wave of the explosion is different than an explosion of TNT, but from a distance both can be considered equal to that shown in Figure 4. The model establishes the following relationship:

$$W = \frac{\eta M - E_v}{E_c TNT}$$

W = TNT equivalent mass (kg)
 M = Mass of flammable substance released (kg)
 η = Performance (effectiveness) of the explosion empirical (0.01 to 0.10).
 E_c = Lower heat of combustion of flammable gas or vapor ($\frac{kJ}{kg}$).
 E_{cTNT} = Heat of combustion (detonation) of TNT ($4437 \text{ A } 4765 \frac{kJ}{kg}$)

Once you calculate the TNT equivalent mass is used in Figure 5 for the most important parameters in terms of climbing distance Z.

$$Z = \frac{R}{\sqrt[3]{W}}$$

R = Real distance in meters (m)

W = Equivalent mass in kg.

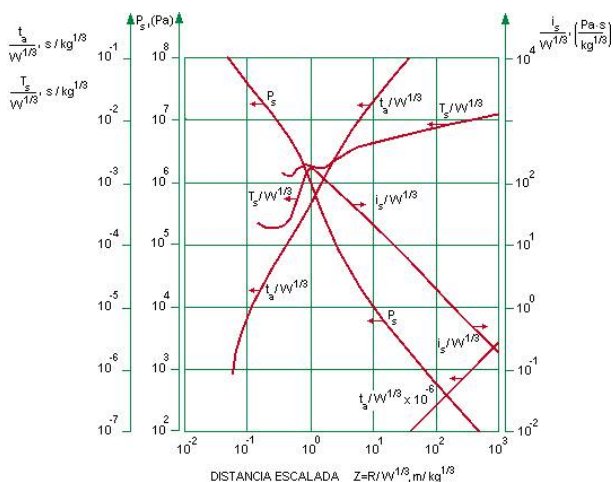


Figure 5. Parameters of the explosion based on the distance climbing.

P_s = Maximum incident pressure in Pascals (Pa)
 t_s = Specific impulse in Pascal · Second (Pa · s)
 t_a = Arrival time of the shock wave in seconds (s)
 T_s = Duration of the overpressure positive phase of the shock wave in seconds (s)

4. RESULTS

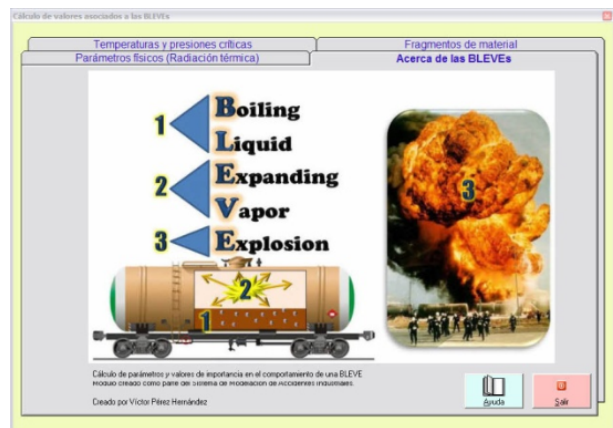


Figura 6. BLEVE Module

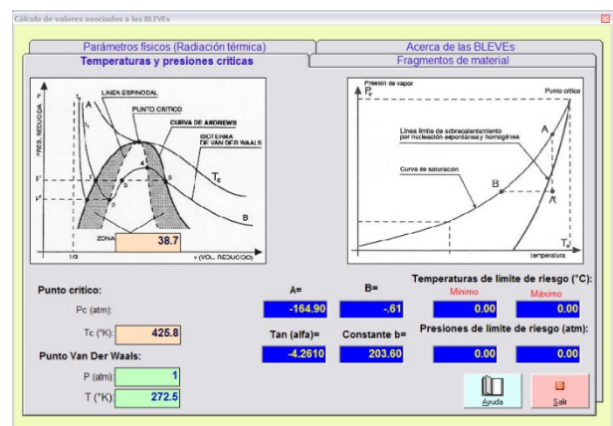


Figura 7. BLEVE - Determination of F and T.

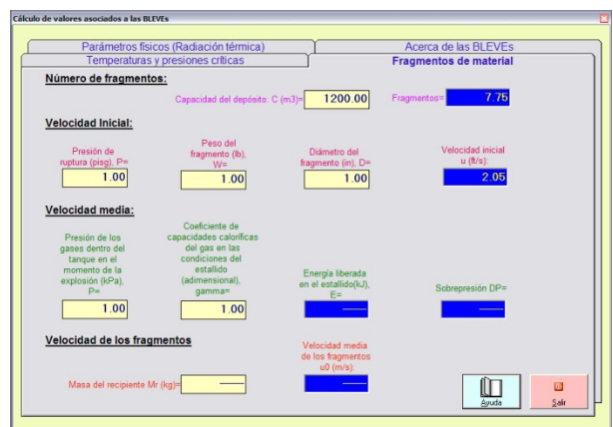


Figura 8. BLEVE - Determination of fragments



Figura 9. Liquid discharges

- MINISTERIO DE TRABAJO Y ASUNTOS SOCIALES, ESPAÑA (2003). NTP 326: Radiación térmica en incendios de líquidos y gases.
- MINISTERIO DE TRABAJO Y ASUNTOS SOCIALES, ESPAÑA (2003). NTP 362: Fugas en recipientes y conducciones: emisión en fase líquida.
- MINISTERIO DE TRABAJO Y ASUNTOS SOCIALES, ESPAÑA (2003). NTP 599: Evaluación del riesgo de incendio criterios.
- S.Q. Ingeniería LTD. Respuesta a las consultas de la SEREMI de Salud Región Metropolitana. Ministerio de Salud, Colombia.
- Ribera Balboa, Rubén Darío (2002). Metodología para la evaluación del riesgo en el transporte terrestre de materiales y residuos peligrosos. CENAPRED (Centro Nacional de Prevención de Desastres, Secretaría de Gobernación), México.
- Wellens Purnal, Ann (2011). BLEVE: Boiling liquid expanding vapor explosión, Evaluación de la radiación térmica.

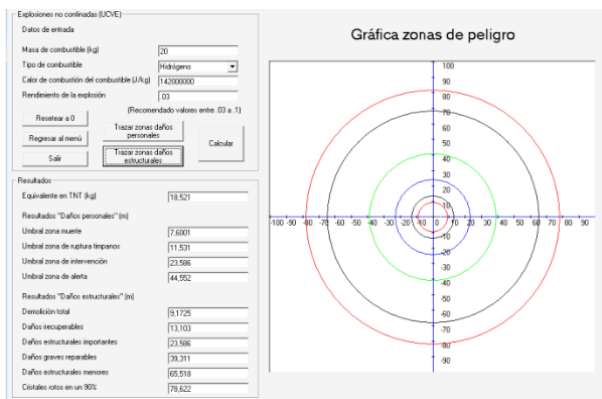


Figura 10. TNT equivalent explosion model

BIOGRAPHICAL NOTES

Ann Wellens is a chemical engineer with postgraduate studies in Industrial Administration (KUL, Belgium) and a master degree in Environmental Engineering (UNAM, Mexico). At the moment she is a full-time lecturer in the Systems Department of the Industrial and Mechanical Engineering Division of the National University of Mexico (UNAM). She has been working in air pollution issues for the last 15 years, dictating courses, collaborating in research projects and participating in conferences related with mathematical modeling of air pollution dispersion and statistics.

Francisco Castellanos D. obtained his bachelor degree in Industrial Engineering (ITESM Campus Central de Veracruz, México).

Gamar Castillo G. obtained his bachelor degree in Computation Engineering (UNAM, México).

María Guadalupe Ávila G. obtained his bachelor degree in Computation Engineering (UNAM, México).

Víctor Pérez H. obtained his bachelor degree in Computation Engineering (UNAM, México).

5. CONCLUSIONS

A simulation model was developed in Visual Basic; it is able to determine the adverse consequences of dangerous physical phenomena in industrial accidents. At this stage, the tool determines typical values related with the occurrence of BLEVES, fireballs, fires, UVCE explosions and spills. Future work includes integration the different modules in an expert system, able to define which type of accident can occur depending on existing ambient conditions, migrate the system to a more powerful developing platform and publish the system in web.

REFERENCES

- LESS FP (1995). Loss prevention in the process industries. Vols. 1, 2 y 3. Segunda edición. Ed. Butterworth-Heinemann.
- MINISTERIO DE TRABAJO Y ASUNTOS SOCIALES, ESPAÑA (2003). NTP 293: Explosiones BLEVE (I): evaluación de la radiación térmica.
- MINISTERIO DE TRABAJO Y ASUNTOS SOCIALES, ESPAÑA (2003). NTP 321: Explosiones de nubes de vapor no confinadas: evaluación de la sobrepresión.

USING MARKOV CHAIN AND GRAPH THEORY CONCEPTS TO ANALYZE BEHAVIOR IN COMPLEX DISTRIBUTED SYSTEMS

Christopher Dabrowski^(a) and Fern Hunt^(b)

U.S. National Institute of Standards and Technology

^(a)cdabrowski@nist.gov, ^(b)fhunt@nist.gov

ABSTRACT

We describe how a Discrete Time Markov chain simulation and graph theory concepts can be used together to efficiently analyze behavior of complex distributed systems. Specifically, the paper shows how minimal s-t cut set analysis can be used to identify state transitions in a directed graph of a time-inhomogeneous Markov chain, which when suitably perturbed, lead to performance degradations in the system being modeled. These state transitions can be then be related to failure scenarios in which system performance declines catastrophically in the target system being modeled. Using a large-scale simulation of the grid system, we provide examples of the use of this approach to identify failure scenarios. Preliminary experiments are reported that show this approach can be applied to problems of significant size. The approach described here combines techniques whose use together to analyze dynamic system behavior has not previously been reported.

Keywords: time-inhomogeneous Discrete Time Markov chain; distributed system; minimal s-t cut set.

1. INTRODUCTION

In large-scale, dynamic distributed systems, such as computing grids, the interactions of many independent components can lead to emergent system-wide behaviors with unforeseen, often detrimental, outcomes (Mills and Dabrowski 2008). To ensure availability and reliability of computing services in such environments, new techniques will be needed to rapidly assess trends and predict changes in system behavior caused by such factors as shifts in workload, modifications to system configurations, policy changes, or failures.

In earlier work (Dabrowski and Hunt 2009), we described a succinct Discrete Time Markov chain (DTMC) representation for analyzing the behavior of a grid computing system in order to identify potential failure scenarios in which system-wide performance collapses. In this representation, the stochastic characteristics of Markov chains were used to summarize the evolving state of a system, in which dozens of users and grid service providers interacted to process over 1000 grid computing tasks over simulated time durations (Mills and Dabrowski 2008). To capture change in system behavior over time, the DTMC

representation was made *time inhomogeneous*—also referred to as *piecewise homogeneous* (Rosenberg, Solan, and Vielle 2004)—in which a set of transition probability matrices (TPMs) was used to model successive time periods. The time-inhomogeneous TPM set could be perturbed by systematically changing the values of related state transition probabilities to examine alternative system execution paths. State transitions were deemed *critical state transitions* if they could be perturbed to cause system performance to decline drastically. Once identified, these critical transitions could then be related to events such as faults, policy changes, and workload shifts, in order to describe failure scenarios in a target system being modeled. The perturbed TPM set could be used to simulate the rate at which performance declines in response to such events and to establish thresholds, beyond which increased incidence of failure caused performance collapse. This initial approach, however, required exhaustive search of the TPM set in order to find failure scenarios.

To overcome this limitation, this paper extends (Dabrowski and Hunt 2009), to analyze the DTMC as a directed graph and to use minimal s-t cut set analysis (Tsukiyama, Shirakawa, Ozaki, and Ariyoshi 1980) to identify critical state transitions, which if perturbed, reveal potential performance collapses in the target system being modeled. We show that the use of minimal s-t cut set analysis reduces the computation needed to find critical transitions, and can thus be applied to more complex problems in comparison to the exhaustive search methods used in our initial approach (Dabrowski and Hunt 2009). Further, we show that minimal s-t cut analysis can also find combinations of critical transitions that represent more complicated failure scenarios, which our initial approach also could not do. In experiments, our new approach is applied to analyze grid system simulations with different durations and workloads. We assess the use of minimal s-t cut set analysis to predict failure scenarios, using a detailed, large-scale grid simulation as a proxy for a real-world system. The application of minimal cut set analysis to a grid system parallels our use of this method to analyze dynamic behavior in cloud computing systems, which we report in (Dabrowski and Hunt 2011). The use of this method in two different domains is essential to investigating the generality of the approach.

This paper also considers whether minimal s-t cut set analysis can be applied to large problems. Since larger directed graph problems can potentially contain many minimal s-t cut sets, we investigate the use of a cut set identification algorithm that can be bounded to run within reasonable time limits. We evaluate the effectiveness of this algorithm in identifying critical transitions in DTMCs with up to 50 states and 160 state transitions. To our knowledge, the use of minimal s-t cut set analysis to guide perturbation of a time-inhomogeneous DTMC has not previously been studied as an approach for analyzing dynamic behavior in complex distributed systems.

This paper is organized as follows. Section 2 reviews related work on using Markov chains to analyze dynamic systems. Section 3 summarizes the DTMC for the grid system example used here. Section 4 defines minimal s-t cut sets and describes their use in finding critical state transitions. Section 5 presents an algorithm for computing minimal s-t cut sets in large DTMC graphs and analyzes its performance. Section 6 discusses future work and concludes.

2. RELATED WORK

The method discussed in this paper is distinguishable from the well-known use of DTMCs to provide quantitative measures of system performance and reliability, which we review in (Dabrowski and Hunt 2009). Instead of measuring system reliability, we use DTMCs to examine alternative execution paths in dynamic systems in order to identify failure scenarios.

Both perturbation analysis and graph theory have previously been applied to DTMCs, but for different purposes than we intend. Perturbation analysis of DTMCs has been the topic of theoretical (Schweitzer 1968; Hassin and Haviv 1992) and computational study (Meyer 1989; Stewart and Dekker, 1994). Other researchers have used system performance gradients that are based on key decision parameters to perturb Markov models (Ho and Li 1988; Suri 1989; Cao and Zhang 2008). While gradient-based approaches demonstrated potential in modeling performance change, some issues involving computation of gradients required further research to fully resolve (Cao and Zhang 2008). Also, gradient-based approaches appear to be geared for system optimization, rather than for examining alternative execution paths to identify situations in which performance degrades.

Graph-theoretic methods have also been used previously to study dynamic behavior in Markov chain models. For example, graph decomposition has been used to calculate stationary probability distribution vectors of Markov chains (Benzi and Tuma 2002; Gambin, Kryzanowski and Pokarski 2008; as well as to measure how perturbation affects stationary distributions (Solan and Vielle 2003). Minimal cut set analysis has been used on topology graphs of avionics system components to identify the shortest sequence of component failures (Tang and Dugan 2004). However, these applications of graph theory have been targeted

for analysis of specific subsystems in their respective domains, rather than using minimal s-t cut set analysis to identify global failure scenarios in the manner we envision. Finally, there are maximum-flow algorithms (Ford and Fulkerson 1962; Goldberg and Tarjan 1988), well-known graph-theoretic methods that find s-t cut sets on the basis of flow levels. These algorithms could potentially be used to identify critical state transitions. However, because these algorithms use flows, they are distinguishable from Markov chain approaches and so best merit separate investigation.

3. THE DISCRETE TIME MARKOV CHAIN

The DTMC model of a grid system was developed by observing a large-scale grid computing simulation (Mills and Dabrowski 2008). This section overviews the DTMC model, with full details in (Dabrowski and Hunt 2009). The DTMC model of the grid system simulates the progress of over 1000 computing tasks from the time they are submitted by a user to the grid for execution to the time they either complete or fail. Figure 1 shows a state diagram of this system, which describes the lifecycle of a single task. This model has 7 states: an *Initial* state, where a task remains prior to submission; a *Discovering* state, during which service discovery directories are accessed to locate grid service providers who are able to execute the task; a *Negotiating* state during which a Service Level Agreement (SLA) for task execution is negotiated with a provider; a *Waiting* state in which tasks reside that are temporarily unsuccessful in discovery or negotiation; and a *Monitoring* phase in which a provider who has entered into an SLA executes a task by a deadline. Transitions between states, shown in Figure 1 by arrows, represent actions taken by the grid system to process a task. All tasks eventually enter either the *Tasks Completed* or *Tasks Failed* state, which are the *absorbing states* of the Markov chain, because once entered, a task cannot leave. A Markov chain with these characteristics is called an *absorbing chain*.

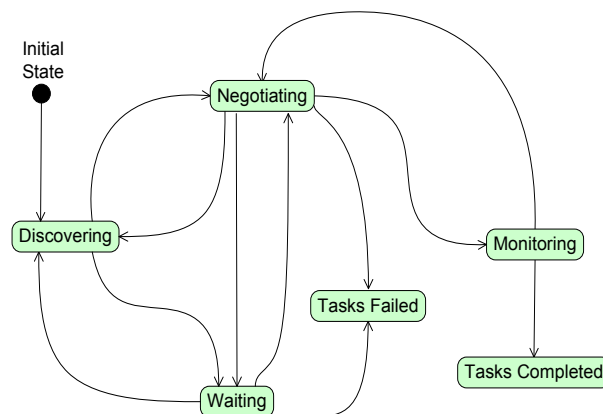


Figure 1: State model of grid computing system.

The large-scale grid simulation was observed over extended durations to accumulate frequencies for the state transitions shown in Figure 1, compute transition probabilities, and form TPMs. We computed transition probabilities by determining where state transitions

occur in the large-scale model code and inserting counters at those places. State transition probabilities were derived as follows. Given states $s_i, s_j, i, j = 1 \dots n$ where $n=7$, p_{ij} , is the probability of transitioning from state i to state j , written as $s_i \rightarrow s_j$. This probability is estimated by calculating the frequency of $s_i \rightarrow s_j$, or f_{ij} , and dividing by the sum of the frequencies of s_i to all other states s_k , as shown in equation (1)

$$P_{ij} = \frac{f_{ij}}{\sum_{k=1}^n f_{ik}} \quad (1)$$

Here i and j may be equal, to allow for self transitions, which are counted if the task process remained in a state longer than a discrete time step, chosen to be 85 s. The resulting TPM is a 7×7 stochastic matrix, where rows stand for the state the transition originates from, or the *from state*, i , and columns represent states the transition goes to, or the *to state*, j . Figure 2 shows an example of such a TPM. Each element in this TPM contains a p_{ij} , where i and j are the *from* and *to* states, respectively. As in any stochastic TPM, the transition values of all row elements must sum to 1.0.

	Initial	Wait	Disc	Ngt	Mon	Comp	Failed
Initial	0.9697	0	0.0303	0	0	0	0
Wait	0	0.8363	0.0673	0.0918	0	0	0.0046
Disc	0	0.0355	0.6714	0.2931	0	0	0
Ngt	0	0.4974	0.0182	0.2882	0.1961	0	0.0001
Mon	0	0	0	0.0003	0.9917	0.0080	0
Comp	0	0	0	0	0	1.0	0
Failed	0	0	0	0	0	0	1.0

(a)

	Initial	Wait	Disc	Ngt	Mon	Comp	Failed
Initial	0.9997	0	0.0003	0	0	0	0
Wait	0	0.7612	0.0460	0.1911	0	0	0.0017
Disc	0	0.0686	0.6084	0.3230	0	0	0
Ngt	0	0.2401	0.0062	0.2378	0.4801	0	0.0358
Mon	0	0	0	0.0007	0.9902	0.009	0
Comp	0	0	0	0	0	1.0	0
Failed	0	0	0	0	0	0	1.0

(b)

Figure 2: (a, b) Summary TPMs for the grid system over (a) 8- and (b) 640-hour durations. The summary TPMs are weighted averages of their component time period TPMs, in which the weight of each time period TPM is determined by the relative number of transitions in the time period.

Separate observations were made to create two cases: (a) one in which the system executes for 8 hours with varying workload; and (b) a 640-hour execution that reaches near steady state. To create time-inhomogeneous representations for the two cases, the total duration of each was divided into equal periods of 7200 s and a TPM was computed for each period. Figure 2 shows the summary TPMs for the two cases,

which are weighted averages of their respective time period TPMs. The weights are based on the relative number of transitions in each period.

3.1. Simulating System Behavior

To simulate system behavior over time, a well-known DTMC method was employed, which we refer to as *Markov simulation*. For further description, see (Hunt, Morrison, and Dabrowski 2011; Dabrowski and Hunt 2009). In Markov simulation, multiplication of time period TPMs is used to advance the system state in discrete time steps of a fixed duration, h . Here, $h = 85$ s. Since a time period covers a duration of $d_{period} = 7200$ s, each time-period TPM is made to represent $S = d_{period}/h$, or 85, steps. Thus, an 8-hour Markov simulation, with a 2-hour period for residual clean-up, covers 5 time periods, consisting of a total of 425 time steps. Correspondingly, a 640-hour Markov simulation with a clean-up period covers 321 time periods with over 27, 000 time steps.

In Markov simulation, the state of the system can be summarized at any time step in an n -element state vector v , where n is the number of states in the related Markov chain. Each of the n elements in v represents the proportion of tasks in one of the n states of the DTMC. For the Markov chain of the grid system, the $n=7$ elements are ordered so as to correspond to the states in Figure 1. Thus, the first element in v contains the proportion of tasks in the *Initial* state, the second contains the proportion of tasks in the *Waiting* state, and so forth. In Markov simulation, the vector v represents the system state at different time steps, such that the vector v_m represents the system state at time step m . To evolve the system state by one discrete time step, the vector v_m is multiplied by the TPM, Q^{tp} , for the applicable time period tp to produce a new system state v_{m+1} , as shown in equation (2):

$$(Q^{tp})^T * v_m = v_{m+1}, \text{ where } tp = \text{integral value } (m/S) + 1 \quad (2)$$

where T indicates a matrix transpose. Starting with v_1 , which represents a system state with a value of 1.0 for the *Initial* state (see Figure 1) and 0 for all other states (i.e., all tasks begin in the *Initial* state), equation (2) is repeated for 425 time steps to evolve the system state over 8 simulated hours. This results in a system state vector, v_{425} at the end of the simulated 8 hours. To simulate 640 hours, equation (2) is repeated for 27, 000 time steps to produce a state vector, v_{27000} , at the end of 640 hours. In both cases, repeated application of equation (2) causes the proportion of tasks to be distributed over the 6 states other than the *Initial* state (i.e., all states have transitioned out of *Initial*). In an absorbing chain, tasks eventually transition into one of the absorbing states, i.e., the *Tasks Completed* or *Failed* states, where they remain permanently. Thus, a measure of the performance of a system is the proportion of tasks that enter the *Tasks Completed* state, because this absorbing state represents tasks that have succeeded. On the other hand, a performance collapse may be

simulated either when a large proportion of tasks enter the *Task Failed* state, or when they are otherwise prevented from entering *Tasks Completed*. Figure 3 shows that *Markov simulation* of the grid system closely approximates the performance of a large-scale simulation in both the 8- and 640-hour cases in terms of proportion of tasks that enter the *Tasks Completed* state.

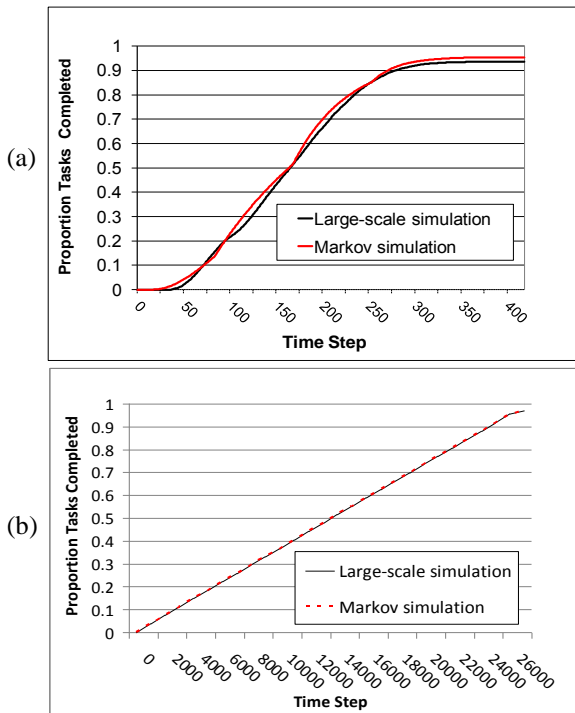


Figure 3: Performance of Markov chain and large-scale simulations as measured by *Tasks Completed* over: (a) 8 hours (5 time-period TPMs – 421 time steps with an extra cleanup period), and (b) 640 hours (321 time-period TPMs – 27000 time steps with cleanup). A time period represents 7200s and a time step represents 85 s.

3.2. Perturbing Critical State Transitions

To identify critical state transitions, we described a perturbation algorithm in (Dabrowski and Hunt 2009), which systematically raises and lowers all feasible combinations of non-zero state transition probabilities in individual rows of a TPM for a DTMC. The different combinations of changed values are then evaluated by Markov simulation in order to explore potential alternative system executions. This algorithm can be applied to exhaustively perturb all feasible combinations of state transition probabilities in all rows of a TPM. The algorithm outputs a set of individual critical state transitions, which when perturbed to extreme values, cause system performance to degrade drastically. We must omit the full description to the perturbation algorithm due to lack of space. In (Dabrowski and Hunt 2009), we showed that exhaustive application of this algorithm could replicate (with good agreement) scenarios in which performance drastically degraded in the large-scale grid simulation.

Figure 4 provides an example of a critical state transition, *Negotiating* \rightarrow *Monitoring*, identified by the perturbation algorithm. The figure shows the impact of a set of related perturbations, in which lowering the probability of transition to 0 for *Negotiating* \rightarrow *Monitoring* causes the proportion of *Tasks Completed* to fall to 0 in the Markov simulation (blue curves). The perturbation of this transition models a failure scenario in which negotiations for SLAs fail, due to events such as system-wide viruses which gradually affect all providers; hence, tasks cannot progress to the *Monitoring* state. Figure 4 also shows the result altering the target large-scale grid simulation (red curve), to randomly fail negotiations with systematically increased incidence. The figure shows that both the Markov and large-scale simulation curves exhibit low thresholds for the rate of successful negotiation, below which there is a sharp drop in *Tasks Completed*. In the Markov simulation this threshold is below 0.05, while in the large-scale simulation, the threshold is slightly higher at 0.15. However, both curves are sufficiently similar, so that the perturbation algorithm could be used to forecast that increased incidence of failed negotiation will eventually lead to a system performance collapse. In (Dabrowski and Hunt 2009), we provide the complete results of applying the perturbation algorithm. Though its computational cost prohibits use on large problems, the perturbation algorithm provides a baseline for assessing the use of minimal s-t cut set analysis.

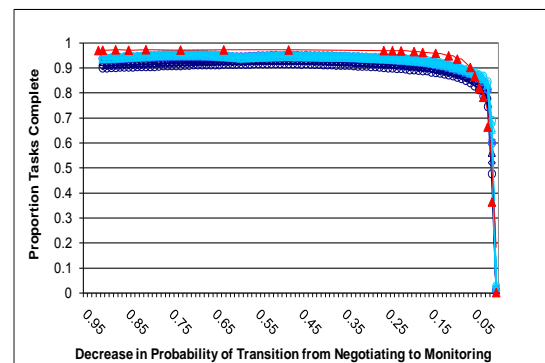


Figure 4: Perturbation of *Negotiating* State to reduce the probability of transition from *Negotiating* \rightarrow *Monitoring* while raising the probability of transition from *Negotiating* \rightarrow *Waiting* in the 640-hour case. The blue curve shows the proportion of *Tasks Completed* estimated by the perturbation algorithm. Large-scale simulation results are denoted by red triangles.

4. MINIMAL S-T CUT SET ANALYSIS IN A MARKOV CHAIN MODEL

This section describes how identifying minimal s-t cut sets on paths between an *Initial* state and a desired absorbing state can be used to identify critical state transitions in a DTMC, which if perturbed, lead to system performance degradations. In contrast to the perturbation algorithm, which can identify only single state transitions that are critical, minimal s-t cut set analysis identifies combinations of critical state

transitions, an important benefit for analysis of more complex problems. In Section 5, we describe an algorithm for finding minimal s-t cut sets and show its effectiveness for large problems. Our approach does not use flow levels to identify minimality, but instead uses cardinality and other factors discussed below.

4.1. Definitions

In graph theory, a graph $G(V, E)$ consists of a set of vertices V connected by edges from the set E . A directed graph is a graph in which edges can be traversed in only one direction. A Markov chain is a directed graph, in which vertices correspond to states and directed edges correspond to state transitions. A directed path through this graph is a sequence of state transitions from one state to another. In this problem, the directed paths of most interest are non-cyclic paths that lead from the *Initial* state to one of the two absorbing states: *Tasks Completed* or *Tasks Failed*. This paper considers only paths to *Tasks Completed*. Figure 5 shows two such paths.

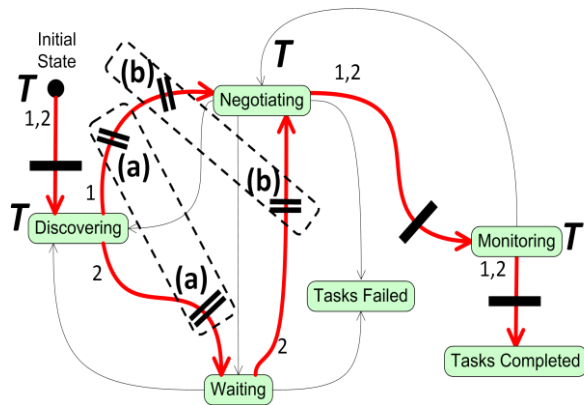


Figure 5: There are 2 directed paths (in red) from the *Initial* state to *Tasks Completed*, labeled 1 and 2. Three single-transition s-t cuts (minimal s-t cut sets consisting of one state transition) are marked by single bars. Two multiple transition s-t cuts (s-t cut sets with two transitions each) are marked by double bars: (a) *Discovering* \rightarrow *Negotiating* and *Discovering* \rightarrow *Waiting*; and (b) *Discovering* \rightarrow *Negotiating* and *Waiting* \rightarrow *Negotiating*. Trap states are denoted by *T*.

A set of one or more edges, which if removed, disconnects all paths between two vertices s and t is referred to as an *s-t cut set* (Tsukiyama, Shirakawa, Ozaki, and Ariyoshi 1980). An s-t cut set is a *minimal s-t cut set* if removal of any edge from the cut set reconnects s and t . By finding minimal s-t cut sets consisting of state transitions that disconnect the *Initial* and *Tasks Completed* states, it is possible to know where reducing the related transition probabilities to 0 prevent tasks from completing. In this paper, minimal s-t cut sets with a single member will be referred to as *single-transition s-t cuts*, while those with more than one member are *multiple-transition s-t cuts*. State transitions that are members of a minimal s-t cut set are critical state transitions as defined above.

4.2. Identifying Minimal s-t Cut Sets in the Grid Markov Chain Model

In Figure 5, there are 3 single-transition s-t cuts: *Initial* \rightarrow *Discovering*, *Negotiating* \rightarrow *Monitoring*, and *Monitoring* \rightarrow *Tasks Completed*. Figure 4 shows that reducing the probability of transition for *Negotiating* \rightarrow *Monitoring* to 0 using Markov simulation causes the proportion of tasks reaching *Tasks Completed* to drop to 0. The same result occurs when the other two single-transition s-t cuts, *Initial* \rightarrow *Discovering* and *Monitoring* \rightarrow *Tasks Completed*, are similarly perturbed (see Dabrowski and Hunt 2009). Use of the exhaustive perturbation algorithm confirmed that the 3 single-transition s-t cuts identify state transitions, which if reduced to 0, cause the proportion of tasks reaching the *Tasks Completed* state to fall to 0 (see Section 5). These 3 single-transition s-t cuts are critical state transitions that clearly relate to resource allocation and task execution functions. Figure 5 also shows two multiple-transition s-t cuts, labeled (a) and (b), which disconnect all paths between the *Initial* from the *Tasks Completed* state. Both multiple-transition cuts consist of two transitions. In a multiple-transition s-t cut, lowering transition probabilities to 0 of all transitions in the cut set reduces the proportion of *Tasks Completed* to 0. Multiple-transition s-t cuts identify situations where a combination of state transitions is critical and together describe circumstances that degrade system performance. We return to multiple-transition s-t cuts in Section 5.

4.3. Identifying Trap States

The previous discussion considered only state transitions between different states. However, in a DTMC, a state may also transition to itself in the next discrete time step and remain in the same state. In this paper, this is referred to as a *self-transition*. If a self-transition probability is near 1, the task may stay in the state for a long time. Such a state effectively becomes a *trap state*. Figure 6 shows an example of how a *trap state* affects performance, when the self-transition probability of the *Discovering* state is raised to 1. As the self-transition probability approaches 1, tasks are increasingly stalled in *Discovering*, so that they cannot proceed to other states and complete by their deadlines. The evolution of *Discovering* into a trap state may correspond to a real-world failure scenario in which service discovery is impaired by directory failures, so that information about existing grid services cannot be retrieved. As the incidence of directory failures increases, the length of time to complete service discovery for all tasks also increases, until finally no task can progress beyond the discovery stage to begin negotiation. Figure 6 also shows how the large-scale simulation behaves when the equivalent failure is introduced. In the latter, the failure is modeled by systematically increasing the frequency of directory access failure. As in the example discussed in Section 3.2, the perturbation shown in Figure 6 predicts how this failure scenario impacts the large-scale simulation.

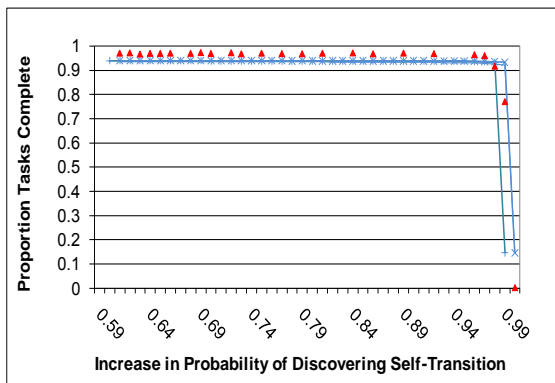


Figure 6: Perturbation of *Discovering* to increase the self-transition probability to 1, while decreasing the transition probability from *Discovering* to other states to 0 in 640-hour case. The blue curve shows the proportion of tasks completed as estimated by the perturbation algorithm. Values from the large-scale simulation are shown by red triangles.

A trap state is distinguishable from a permanent absorbing state, because the latter always has a self-transition probability of 1, while in the former, the transition probability varies. The concept of an s-t cut set can be extended to include vertices whose removal cuts all paths from s to t . A minimal set of such elements (edges and vertices) is a *minimal s-t separating set* (Hayakawa, Tsukiyama and Ariyoshi 1999) a topic we leave for future work.

5. PERFORMANCE OF MINIMAL S-T CUT SET ANALYSIS

This section shows that minimal s-t cut set analysis is an effective and efficient means of finding critical state transitions and trap states. The section also shows the potential applicability of this method to large problems. Section 5.1 first describes a well-known algorithm, known as the *node contraction algorithm*, which we adapted to find minimal s-t cut sets in a directed graph. The node contraction algorithm is considered probabilistic (Karger and Stein 1996), because it can find solutions with a probability which can be increased to a high level by repeatedly executing the algorithm. Though it is not guaranteed to find all minimal s-t cut sets, the computational cost of node contraction can be bounded, making it potentially applicable to large problems where use of exhaustive methods would be infeasible. The node contraction algorithm also finds critical transitions that are part of multiple transition s-t cuts, which the perturbation algorithm cannot find.

Section 5.2 shows effectiveness and efficiency of node contraction for the grid system problem. Here, the node contraction algorithm is able to find all individual critical transitions and trap states that were found using the perturbation algorithm, but at far less cost. Section 5.3 then reports experiments on the use of node contraction for finding critical state transitions in large Markov chain problems. To do this, the performance of the node contraction algorithm is tested by comparing it to the performance of an algorithm (Provan and Shier

1996) that, unlike node contraction, enumerates *all* minimal s-t cut sets in a directed graph, and thus finds all critical transitions. While, like other algorithms of this type, the time complexity of (Provan and Shier 1996) prohibits practical use on many large problems, the algorithm provides a good baseline for testing the effectiveness of node contraction. (The complexity of the minimal s-t cut set enumeration algorithm described in (Provan and Shier 1996) is $O(|E|)$ for each s-t cut set that exists, where $|E|$ is the number of edges in the graph). To examine the potential for scalability of minimal s-t cut set analysis, we wish to know what proportion of minimal s-t cut sets (and thus critical transitions) can be found by node contraction in large problems and the related computational cost.

5.1. Overview of the Node Contraction Algorithm

In this section, we summarize our implementation of the node contraction algorithm, with pseudo-code in (Dabrowski, Hunt and Morrison 2010). Though the time complexity of node contraction algorithms for directed graphs has not been studied, efficient versions of this algorithm for undirected graphs find a minimum cut with a complexity of $O(|V|^2)$, where $|V|$ is the number of vertices in the graph (Karger and Stein 1996). While this cost is significant, the algorithm can be used on large problems by controlling the number of executions, as will be discussed in Section 5.3.

The node contraction algorithm operates by randomly choosing two vertices connected by an edge and replacing these vertices with a single, new vertex. The new vertex assumes the edges by which the two replaced vertices were connected to the remainder of the graph (i.e., the edges of replaced vertices become the edges of the new vertex) and takes up the edges that connected the two replaced vertices. The result of each contraction is recorded. The process of randomly selecting pairs of vertices repeats until only two large, *mega-vertices* remain. The directed edges between the two remaining mega-vertices c_1 and c_2 , and the directed edges between vertices $\langle v_1, v_2 \rangle$, $v_1 \neq v_2$, in which v_1 was replaced by c_1 and v_2 was replaced by c_2 , constitute a minimal s-t cut set of the graph. The node contraction algorithm was modified for an absorbing Markov chain problem to prevent the two vertices representing the *Initial* state, s , and the *Tasks Completed* absorbing state, t , from being contracted into the same vertex. This ensures that the *Initial* state, s , and *Tasks Completed* state, t , will not both end up in either c_1 or c_2 . In this way, the edges between the two remaining mega-vertices, c_1 and c_2 , together with the vertices each has absorbed, yield a minimal s-t cut set of state transitions, which if removed, disconnect the *Initial* and absorbing state (*Tasks Completed*).

Since the algorithm randomly selects two connected vertices to combine, repeated applications produce different cut sets. The more the algorithm is repeated, the greater the chances that a large proportion, if not all, of the minimal s-t cut sets of interest will be obtained. Hence, the operation of the algorithm is

considered to be probabilistic. Because the number of repetitions can be controlled, computation cost can be bounded. Further, cut sets can identify potential trap states, which exist when all transitions in the cut set originate in one state. Perturbation then need be applied only to the transitions in the s-t cut set, in order to generate curves of tasks completed, such as appear in Figures 4 and 6, and to identify performance thresholds.

5.2. Comparing the Perturbation Algorithm with Minimal s-t Cut Set Analysis

Table 1 compares the result of applying the perturbation algorithm described in Section 3 with the result of minimal s-t cut set analysis using node contraction, when both are used to identify individual critical state transitions and trap states. The perturbation algorithm was applied to the 5 rows representing non-absorbing states (labeled a-e) in the time period TPMs for the 8- and 640-hour cases. The combinations of row elements representing the transition probability being decreased and increased appear in the two leftmost columns. For each such combination of transitions, the next two columns show the proportion of *Tasks Completed* for the 8- and 640-hour cases as the transition probability being decreased falls to 0. The rightmost column indicates if the state transition being reduced corresponds to a single-transition s-t cut (see Figure 5).

Table 1 shows that all combinations where perturbation causes a decline in the proportion in *Tasks Completed* to 0 correspond to single- transition s-t cuts.

There are 7 such combinations, and all correspond to single-transition s-t cuts that are verified by large-scale simulation. In no case, did node contraction find an s-t cut that did not correspond to such a drastic reduction. For instance, in Table 1(c), rows 10–12, when the probability of transition from the *Negotiating* state to *Monitoring* (i.e., *Negotiating* → *Monitoring*) is reduced to 0, the proportion of *Tasks Completed* falls to 0. This is shown in Figure 4. Figure 5 shows that *Negotiating* → *Monitoring* is a single-transition s-t cut.

Note that in Table 1(d), row 3, reducing the probability of *Monitoring* self-transition while raising the probability of *Monitoring* → *Negotiating* also caused a severe decline in the proportion of *Tasks Completed*. This happens because the probability of transition for *Monitoring* → *Tasks Completed* is very low (see Figure 2), and so the probability of *Monitoring* self-transition must be very high to ensure tasks remain in the *Monitoring* state long enough to eventually transition to *Tasks Completed*. Thus, reducing the probability of *Monitoring* self-transition to 0 while raising the probability of *Monitoring* → *Negotiating* prevents tasks from reaching *Tasks Completed*—and acts like a single-transition s-t cut on *Monitoring* → *Tasks Completed*. However, because the transition probability of *Monitoring* → *Tasks Completed* is not lowered to 0 by this perturbation, some tasks are able to complete. Table 1 also contains 3 rows that show only partial reductions (Table 1 (a), rows 5 and 6, and Table 1 (b), row 5). These correspond to the state transitions

Table 1: Correspondence of results of applying the perturbation algorithm to the TPMs for the 8- and 640-hour cases with single-transition s-t cuts found by the node contraction algorithm. The perturbations represented by individual rows correspond to single-transition s-t cuts in Figure 5 as follows: Table (c) rows 10–12 to *Negotiating* → *Monitoring*; Table (d) rows 3, 5, and 6 to *Monitoring* → *Tasks Completed*; and Table (e) row 1 to *Initial* → *Discovering*. Note also the explanation in the text for Table (d) row 3. Perturbations verified by large-scale simulation are bolded and shaded.

(a) row = Discovering						(b) row = Waiting					
	Element reduced→0	Element raised	Proportion of <i>Tasks Complete</i>		s-t cut exists		Element reduced→0	Element raised	Proportion of <i>Tasks Complete</i>		s-t cut exists
			8-hour	640-hour					8-hour	640-hour	
1	<i>Waiting</i>	<i>Discovering</i>	0.957	0.935	No	1	<i>Waiting</i>	<i>Discovering</i>	0.974	0.937	No
2	<i>Waiting</i>	<i>Negotiating</i>	0.959	0.935	No	2	<i>Waiting</i>	<i>Negotiating</i>	0.981	0.939	No
3	<i>Discovering</i>	<i>Waiting</i>	0.939	0.935	No	3	<i>Discovering</i>	<i>Waiting</i>	0.937	0.934	No
4	<i>Discovering</i>	<i>Negotiating</i>	0.963	0.935	No	4	<i>Discovering</i>	<i>Negotiating</i>	0.963	0.936	No
5	<i>Negotiating</i>	<i>Waiting</i>	0.894	0.933	No	5	<i>Negotiating</i>	<i>Waiting</i>	0.818	0.843	No
6	<i>Negotiating</i>	<i>Discovering</i>	0.651	0.932	No	6	<i>Negotiating</i>	<i>Discovering</i>	0.939	0.932	No
(c) row = Negotiating						(d) row = Monitoring					
1	<i>Waiting</i>	<i>Discovering</i>	0.974	0.937	No	1	<i>Negotiating</i>	<i>Monitoring</i>	0.982	0.937	No
2	<i>Waiting</i>	<i>Negotiating</i>	0.985	0.938	No	2	<i>Negotiating</i>	<i>Tasks Comp</i>	0.982	0.938	No
3	<i>Waiting</i>	<i>Monitoring</i>	1.000	0.939	No	3	<i>Monitoring</i>	<i>Negotiating</i>	0.028	0.186	Yes
4	<i>Discovering</i>	<i>Waiting</i>	0.954	0.935	No	4	<i>Monitoring</i>	<i>Tasks Comp</i>	0.980	0.949	No
5	<i>Discovering</i>	<i>Negotiating</i>	0.957	0.935	No	5	<i>Tasks Comp</i>	<i>Negotiating</i>	0.001	0.006	Yes
6	<i>Discovering</i>	<i>Monitoring</i>	0.967	0.936	No	6	<i>Tasks Comp</i>	<i>Monitoring</i>	0.002	0.016	Yes
7	<i>Negotiating</i>	<i>Waiting</i>	0.923	0.931	No	(e) row = Initial					
8	<i>Negotiating</i>	<i>Discovering</i>	0.941	0.933	No	1	<i>Discovering</i>	<i>Initial</i>	0	0	Yes
9	<i>Negotiating</i>	<i>Monitoring</i>	0.988	0.938	No	2	<i>Initial</i>	<i>Discovering</i>	0.970	0.988	No
10	<i>Monitoring</i>	<i>Waiting</i>	0.000	0.000	Yes						
11	<i>Monitoring</i>	<i>Discovering</i>	0.000	0.000	Yes						
12	<i>Monitoring</i>	<i>Negotiating</i>	0.000	0.000	Yes						

In the two multiple transition s-t cuts in Figure 5, which were identified by node contraction, but could not be found by the perturbation algorithm of Section 3.

The perturbation algorithm was also applied to raise the self-transition probability of the 5 non-absorbing states in the grid model to 1. This perturbation caused the proportion of *Tasks Completed* to decline to 0 when applied to 4 of these states: *Initial*, *Discovering*, *Negotiating*, and *Monitoring* states. All 4 are trap states found through node contraction. The fifth state *Waiting*, is not a trap state; but is part of a state transition that is a member of both multiple-transition s-t cuts in Figure 5. Hence, if the self-transition probability of *Waiting* is raised toward 1, there is only a partial reduction in proportion of tasks completed.

Executing the exhaustive perturbation algorithm on non-absorbing rows of the grid model took 56 minutes in the 8-hour case and 4.5 hours in the 640-hour case. In comparison, node contraction needed less than 0.01 s to find all minimal s-t cut sets and trap states. In the 8-hour case, generating Markov simulation curves to reduce the proportion of *Tasks Completed* to 0 for all minimal s-t cut sets and trap states required 244 s, or 7 % of the 56 minutes needed by the perturbation algorithm. For the 640-hour case, these computations took 230 s, or 1.4 % of the 4.5 hours needed by the perturbation algorithm. Thus, minimal s-t cut set analysis needed two orders of magnitude less time than exhaustive application of the perturbation algorithm. All experiments were executed on a workstation with dual quad-core, 3.16GHz processors and 32 GB memory.

5.3. Using Node Contraction to Find Minimal s-t Cut Sets in Large Problems

This section reports the results of experiments on the use of the node contraction algorithm to find critical transitions in large, complex Markov chain models with many multiple transition minimal s-t cuts. These experiments compare the results of using the contraction algorithm to the results produced by the enumeration algorithm of (Provan and Shier 1996) which is guaranteed to find all minimal s-t cut sets and the critical transitions in these cut sets. Here, the criticality of transitions is estimated using measures we define for these experiments. The results show that, with some exceptions, the node contraction algorithm found a large proportion of the most critical cut sets in two orders of magnitude less time than did exhaustive enumeration. While further experiments are needed, these preliminary investigations suggest that minimal s-t cut set analysis can effectively identify critical transitions in large, complex Markov chain graphs as might be encountered in real-world problems.

5.3.1. Experimental Design

To perform these experiments, four Markov chain models were selected, each consisting of 40 or 50 states, from (Boyarsky 1988; Stewart 2004; Jensen and Jessup 1986) and single time-period TPMs were generated using (Hunt 1994). All four problems were originally

ergodic chains, which were suitably modified to be absorbing chains. Though the matrices were sparse, these problems were large and complex, with a very sizable number of minimal s-t cut sets between the *Initial* and absorbing states ($> 4 \times 10^8$ for the largest; see Table 2.) In contrast to the grid system model, minimal s-t cut sets for these problems consisted of multiple state transitions, which could correspond to combinations of circumstances that impact system performance. In (Dabrowski, Hunt and Morrison 2010), the full description of all four Markov chain problems is provided, which we omit here due to lack of space.

To provide a baseline measure for the number of minimal s-t cut sets in these Markov chain graphs, we implemented the minimal s-t cut set enumeration algorithm of (Provan and Shier 1996), which lists all cut sets. To determine which minimal s-t cut sets were most critical, we selected ranking criteria based on the idea that the most critical cut sets will have a small number of state transitions. We chose this basis, because fewer transitions represent smaller combinations of circumstances that are more likely to occur and thus more likely to impact a system. (Note: this intuition is supported in the case of undirected graphs by the finding (Karger 2001) that small cut sets are more likely to disconnect undirected graphs, if edges of the graph that fail independently with a known probability. Also in (Dabrowski and Hunt 2011), we use this ranking criterion to analyze a DTMC for a cloud computing system.) We used this basis to choose 3 ranking criteria. The first criterion, Sort A, ranks minimal s-t cut sets by the fewest number of edges as the primary sorting criterion and lowest total transition probability of edges as the secondary criterion. The second, Sort B, uses only the lowest total transition probability of edges in the cut set as a sorting criterion (which also tends to rank cut sets with few transitions higher). Hence, Sorts A and B are likely to identify minimal s-t cut sets in which small perturbations to the fewest number of state transitions are likely to produce the largest changes. The third ranking criterion, Sort C, uses the least number of edges as a primary sorting criterion and the highest total transition probability of edges as a secondary criterion. Sort C identifies cut sets consisting of state transitions that are more likely to be taken and therefore, if perturbed, more likely to affect system behavior.

5.3.2. Experimental Results

We applied the node contraction algorithm and the enumeration algorithm of (Provan and Shier 1996) to the four TPMs, ranked the minimal cut sets produced by each using the ranking criteria described above, and compared the results to determine the proportion of most highly ranked cut sets that the node contraction algorithm could find. With the exception of Matrix 1, Table 2 shows that, with 100,000 repetitions, node contraction generated 91.4 % of the top 100 ranked cut sets that were generated by the enumeration algorithm for all four TPMs under all three sorting criteria. The contraction algorithm produced these results in 1.3 % of

the time needed by the enumeration algorithm. This amounts to a two-order of magnitude improvement in time. For instance, for Matrices 2 and 3, the algorithm was able to find almost all top 100 minimal s-t cut sets in a relatively small fraction of the number of hours required by the enumeration algorithm. For Matrix 4, the node contraction algorithm could find all the top 100 under sort criteria B and C in about 15 minutes (as opposed to 156.1 hours through enumeration). Here, node contraction was successful despite the fact that Matrix 4 has over 4×10^8 minimal s-t cut sets.

Table 2: Comparison of minimal s-t cut sets generated by the enumeration algorithm of (Provan and Shier 1996) and by the node contraction algorithm. At 10,000 repetitions, node contraction generated 77.2 % (variance 555.2) of the top 100 ranked cut sets in 0.14 % of the time for Sorts A–C. At 100,000 repetitions, node contraction generated 91.4 % (variance 432.0) of the cut sets found by enumeration in 1.3 % of the time.

Number	Order	Minimal s-t cut set enumeration		Proportion (in %) of 100 top-ranked minimal s-t cut sets ranked by criteria A, B that were found by the node contraction algorithm							
		Number of cut sets	Time (in hours)	After 10,000 repetitions				After 100,000 repetitions			
				Time	Sort A	Sort B	Sort C	Time	Sort A	Sort B	Sort C
1	50	530,432	332 s	640 s	80	100	96	---	---	---	---
2	50	28,230,288	21.6	171 s	93	98	65	1710 s	99	100	99
3	50	27,242,634	36.0	218 s	67	100	100	2288 s	88	100	100
4	40	422,060,801	193.6	106 s	30	80	62	1051 s	37	100	100

However, in Matrix 4, the algorithm found only 37 of 100 high ranked minimal s-t cut sets under Sort A. Also, for Matrix 1, Table 2 shows that the node contraction algorithm had to run longer than the enumeration algorithm, before it began to produce a large percentage of highly-ranked cut sets. This difference in performance may be attributable in part to topological characteristics such as vertices (states) with large numbers of edges (state transitions), which increases vertex interconnectivity and impedes the contraction process. This exception suggests that in some cases where TPMs are small, it may be more efficient to enumerate cut sets than to generate them probabilistically. Despite these exceptions, the data shows that the node contraction algorithm can be used to find a high proportion of minimal s-t cut sets representing combinations of critical state transitions in larger Markov chains within reasonable time limits.

6. CONCLUSIONS AND FUTURE WORK

This paper has described an approach for using minimal s-t cut set analysis to guide perturbation of a time-inhomogeneous DTMC in order to understand the potential for failure in grid computing systems. The approach combines multiple techniques in a way not previously reported. In this approach, minimal s-t cut sets are computed for paths from the *Initial* to selected absorbing states in the directed graph of a DTMC.

These cut sets can be used to identify critical state transitions, which if perturbed, reveal areas for potential performance degradation. The perturbation of critical state transitions in turn provides a basis to identify potential failure scenarios that could occur in the target system being modeled. By perturbing critical state transitions incrementally, it is possible to quantitatively measure performance degradation and to predict how the target system being modeled is likely to respond to increased incidence of failure. As we have shown, the stochastic character of the Markov chain representation of the system state enables modeling of systems having large numbers of tasks, while time inhomogeneity allows modeling of system evolution over time. Using a large-scale grid system as a proxy for a real-world system, we used the approach described here to identify failure scenarios in systems that process hundreds of tasks over different durations. Our results showed that minimal cut set analysis could be used to identify (in two-orders of magnitude less time) all of the failure scenarios found using exhaustive search techniques. In addition, this method also discovered failure scenarios that involved multiple state transitions, which the exhaustive search algorithm could not. To find critical state transitions in larger Markov chains, the paper has presented a probabilistic algorithm for minimal s-t cut set analysis. Experimental results in Section 5.3.2 show the potential of this algorithm for efficiently analyzing large, complex problems and finding related critical transitions that represent complicated circumstances.

To further evaluate the utility of minimal s-t cut set analysis, it will be necessary to carry out experiments in which DTMC representations are constructed for different problem domains, such as reported in (Dabrowski and Hunt 2011). The use of this approach will have to be further evaluated on larger problems as we have begun to do in this paper. As part of this effort, it will also be necessary to investigate other methods for finding minimal s-t cut sets in large, complex directed graphs. For instance, there are alternative approaches to node contraction, such as (Curet, DeVinney and Gaston 2000) which could be examined. Another possible method involves use of maximum-flow algorithms (Ford and Fulkerson 1962; Goldberg and Tarjan 1988) to find s-t cut sets that identify critical transitions. These algorithms find s-t cut sets on the basis of maximum flow and minimum capacity. Potentially, maximum-flow algorithms could be adapted to find cut sets and rank them on the basis of their nearness to maximum flow and minimum capacity, rather than the criteria described here. To enable such rankings, the work of (Curet, DeVinney and Gaston 2000; Balcioglu and Wood 2003) could be used. In addition, we are also investigating the use of methods that are not based on graph theory concepts to analyze dynamic behavior in complex systems. In (Hunt, Morrison and Dabrowski 2011), we describe the use of spectral methods for eigendecomposition to identify critical state transitions, and in (Dabrowski, Hunt and Morrison 2010) we employ this technique as a complementary method to

minimal s-t cut set analysis. Beyond this, it is our hope that this paper will provide useful ideas to other researchers studying dynamic behavior in complex systems, and that ultimately, the work will lead to the development of effective tools for this purpose.

REFERENCES

- Balcioglu, A. and Wood, K., 2003. Enumerating Near-Min s-t Cuts. In: D. Woodruff, ed., *Network Interdiction and Stochastic Integer Programming*. Kluwer Academic Publishers, 21–49.
- Benzi, M. and Tuma, M., 2002. A parallel solver for large-scale Markov chains. *Applied Numerical Mathematics*, 41, 135–153.
- Boyarky, A., 1988. A matrix method for estimating the Liapunov exponent of one-dimensional systems. *Journal of Statistical Physics*, 50 (1-2), 213–229.
- Cao, X. and Zhang, J., 2008. Event-Based Optimization of Markov Systems. *IEEE Transactions on Automatic Control*, 53 (4), 1076–1082.
- Curet, N., DeVinney, J. and Gaston, M., 2000. An Efficient Network Flow Code for Finding all Minimum Cost s-t Cutsets. *Computers and Operations Research*, 29, 205–219.
- Dabrowski, C. and Hunt, F., 2009. Using Markov Chain Analysis to Study Dynamic Behavior in Large-Scale Grid Systems. *Proceedings of the Seventh Australasian Symposium on Grid Computing and e-Research—Volume 99*, pp. 29–40. January 21, Wellington (New Zealand).
- Dabrowski, C., Hunt, F. and Morrison, K., 2010. *Improving the Efficiency of Markov Chain Analysis of Complex Distributed Systems*. National Institute of Standards and Technology, Interagency Report 7744.
- Dabrowski, C. and Hunt, F., 2011. Identifying Failure Scenarios in Complex Systems by Perturbing Markov Chain Models. *Proceedings of the 2011 Pressure Vessels and Piping Division Conference*. July 17–21, Baltimore (Maryland, USA). In press.
- Ford, L. and Fulkerson, D., 1962. *Flows in Networks*. Princeton: Princeton University Press.
- Gambin, A., Kryzanowski, P. and Pokarowski, P., 2008. Aggregation Algorithms for Perturbed Markov Chains with Applications to Network Modeling. *SIAM Journal of Scientific Computation*, 31 (1), 45–77.
- Goldberg, A. and Tarjan, R., 1988. A New Approach to the Maximum-Flow Problem. *Journal of the ACM*, 35 (4), 921–940.
- Hassin, R. and Haviv, M., 1992. Mean Passage Times and Nearly Uncoupled Markov Chains. *SIAM Journal of Discrete Mathematics*, 5 (3), 386–397.
- Hayakawa, J., Tsukiyama, S. and Ariyoshi, H., 1999. Generation of Minimal Separating Sets of Graphs. *IEICE Transaction Fundamentals*, E82-A (5), 775–783.
- Ho, Y. and Li, S., 1988. Extensions of infinitesimal perturbation analysis. *IEEE Transactions on Automation Control*, AC-33 (5), 427–438.
- Hunt, F., 1994. A Monte Carlo Approach To The Approximation of Invariant Measures. *Random and Computational Dynamics*, 2 (1), 111–112.
- Hunt, F., Morrison, K. and Dabrowski, C., 2011. Spectral Based Methods That Streamline the Search for Failure Scenarios in Large-Scale Distributed Systems. *Nineteenth IASTAD International Conference on Modeling and Simulation*. June 22–24, Crete (Greece). In press.
- Jensen, R. and Jessup, E., 1986. Statistical Properties of the Circle Map. *Journal of Statistical Physics*, 43 (1–2), 369–389.
- Karger, D. and Stein, C., 1996. A New Approach to the Minimum Cut Problem. *Journal of the ACM*, 43, 601–640.
- Karger, D., 2001. A Randomized Fully Polynomial Time Approximation Scheme for the All-Terminal Network Reliability Problem. *SIAM Review*, 43 (3), 499–522.
- Mills, K. and Dabrowski, C., 2008. Can Economics-based Resource Allocation Prove Effective in a Computation Marketplace? *Journal of Grid Computing*, 6 (3), 291–311.
- Meyer, C., 1989. Stochastic Complementations, Uncoupling Markov Chains, and the Theory of Nearly Reducible Systems. *SIAM Review*, 31 (2), 240–272.
- Provan, J. and Shier, D., 1996. A Paradigm for Listing (s,t)-cuts in Graphs. *Algorithmica*, 15, 351–372.
- Rosenberg, D., Solan, E. and Vielle, N., 2004. Approximating a Sequence of Observations by a Simple Process. *The Annals of Statistics*, 32 (6), 2742–2775.
- Schweitzer, P., 1968. Perturbation Theory and Finite Markov Chains. *Journal of Applied Probability*, 5 (2), 401–413.
- Solan, E. and Vielle, N., 2003. Perturbed Markov Chains. *Journal of Applied Probability*, 40, 107–122.
- Stewart, G., 2004. *MVMRWK: Markov Chain Transition Probability Matrix*. National Institute of Standards and Technology. Available from: <http://math.nist.gov/MatrixMarket/data/NEP/mvmrwk/rw136.html>. [Accessed 27 June 2011]
- Stewart, W. and Dekker, M., 1994. *Numerical Solution of Markov Chains*. Princeton: Princeton University Press.
- Suri, R., 1989. Perturbation Analysis: The State of the Art and Research Issues Explained via the GI/G/1 Queue. *Proceedings of the IEEE*, 77 (1), 114–138.
- Tang, Z. and Dugan, J., 2004. Minimal cut set/sequence generation for dynamic fault trees. *Proceedings of the 2004 Annual Symposium on Reliability and Maintainability*, pp. 207–213. January 26–29, Los Alamitos (California USA).
- Tsukiyama, S., Shirakawa, I., Ozaki, H. and Ariyoshi, H., 1980. An Algorithm to Enumerate All Cut Sets of a Graph in Linear Time per Cutset. *Journal of the ACM*, 27 (4), 619–632.

FORMAL FRAMEWORK FOR THE DEVS-DRIVEN MODELING LANGUAGE

Ufuoma Bright Ighoroje ^(a), Oumar Maïga ^(b), Mamadou Kaba Traoré ^(c)

^(a)African University of Science and Technology, Abuja, Nigeria

^(b)University of Bamako, Mali

^(c)Blaise Pascal University, Clermont-Ferrand 2, France

^(a)uighoroje@aust.edu.ng, ^(b)maigabababa78@yahoo.fr, ^(c)traore@isima.fr

ABSTRACT

The DEVS-Driven Modeling Language (DDML) is a graphical modeling language that is based on Discrete Event System Specification (DEVS). Models built with DDML are highly expressive and communicable and validation of model properties can be done by simulating these models following the DEVS simulator protocol. We can take advantage of the usefulness of formal methods and apply symbolic manipulation and reasoning to deduce properties of models that cannot be derived from simulation. Since DDML focuses on three levels of abstraction in the hierarchy of system specification, we propose to do formal reasoning at each level of abstraction by applying a semantic mapping function to formal methods that can capture the properties of the model at each level. We do this because we can gain more insight about a model by observing different perspectives. This formal framework for DDML is the focus of this paper.

Keywords: DEVS, Formal Methods, DDML

1. INTRODUCTION

Modeling is a way of thinking about systems by developing abstract models, usually at multiple levels of abstractions with each level revealing a perspective of the system that cannot be observed at other levels. Simulation and testing have been employed to verify and validate the properties of abstract models against the system under study by exploring some of the possible behaviors and scenarios. On the other hand, formal analysis provides a very attractive and increasingly appealing compliment to simulation by conducting exhaustive exploration of all possible behaviors through symbolic reasoning exercises. The advantage here is that we can combine formal analysis and simulation to derive properties of models and compare them with system properties. We can also take benefit of the advances in formal analysis and the existing tools that are available to ensure that the models built accurately represent the desired properties of the system. Since DEVS (Zeigler, Praehofer, and Kim 2000) is a “universal” and powerful simulation modeling formalism, integrating methods for formal analysis with DEVS would bring the desired results.

We say DEVS is universal because other formalisms have been proven to have an equivalent (or approximate) DEVS representation. DEVS supports full range of dynamic system representation. A Differential Equation System Specification (DESS) can have an

approximate Discrete Time System Specification (DTSS) by discretization (selection of a sufficiently small constant time interval). A DTSS model, in turn has an equivalent DEVS representation. Quantization of events in a DESS system can result in an approximate DEVS model. As such DEVS approach can be used to model discrete systems and provide approximate representations of continuous and hybrid systems.

DEVS also promotes separation of concerns by separating the model, simulator, and experimental frame. Although DEVS is powerful, it is a semi-formal specification. It is up to the modeler to express the system structure, behavior, and traces in ways that are most appealing. This “freedom” has led to the development of several DEVS based modeling formalisms. To fill these voids, we propose the DDML (DEVS-Driven Modeling Language). DDML combines DEVS, visual modeling, and formal analysis. In addition, DDML provides an approach to unify the two variants of DEVS (Classic DEVS and Parallel DEVS).

This paper is structured as follows. In section 2, we review some related works. In section 3, we present the concrete syntax of DDML. In section 4, we present the formal framework for DDML and show the different levels of abstractions in DDML specification and the properties that can be derived.

2. RELATED WORKS

Related works have addressed formal analysis of DEVS models. These proposals range from formal model-checking of sub-classes of DEVS, or transformation of DEVS into formal methods for verification purposes, generation of traces from DEVS models for testing

Saadawi and Wainer (2010) proposed a subclass of classic DEVS by mapping the time advance to a rational number which they call Rational Time-Advance (RTA) DEVS thus imposing restrictions on the elapsed time to transform RTA-DEVS to Timed Automata (TA) to enable reachability algorithms to be implemented in UPPAL. Earlier, (Saadawi and Wainer 2009) had proposed a technique for verification of DEVS models based on Model-checking. The technique is to specify graphically DEVS models using E-CD + + and transforming these models into TA in UPPAAL.

Hong, Song, Kim, and Park (1997) proposed the Real-Time DEVS formalism (RT-DEVS) which introduces a time advance function that maps each state to a range with maximum and minimum time values. Further work on verifying RT-DEVS has been done by

using timed automata and UPPAAL with transformation from RT-DEVS to UPPAAL. (Hong and Kim 2005) proposed a method of verification of DEVS models in the environment DEVSIM++. The approach is to specify the model in DEVS (operational formalism) and use the temporal logic (TL) formalism assertions to specify the properties and time constraints of the system. They use a projection technique (external TLA) to reduce the state space

Ernesto (2008) worked on the development of an alternative theoretical foundation for DEVS by defining DEVS models as labeled transition systems (LTS). With this, we can use existing tools for LTS to reason with DEVS and compare DEVS with other formalisms defined as LTS.

Hwang and Zeigler (2009) defined a class of DEVS, called finite and deterministic DEVS (FD-DEVS) by introducing assumptions that enable us to define a reachability graph.

Weisel, Petty and Mielke (2005) discussed a formal theory for semantic composability that examines simulation composability using formal definitions and reasoning.

Cristia (2007) proposed a transformation method of DEVS models in TLA+. The main conclusion here is that DEVS models describing discrete event systems can be easily translated into TLA+ specifications. This would be beneficial for DEVS since it lays the basis for a formal semantics of this powerful modeling language.

Hernandez and Giambiasi (2005) showed that verification of general DEVS models through reachability analysis is not decidable. They based their deduction on building a DEVS simulation Turing machine. They argue that reachability analysis maybe possible only for restricted classes of DEVS. This result however was based on introducing state variables into DEVS formalism with infinite number of values.

These works mentioned above have focused on only one aspect in the Zeigler's hierarchy of system specification by transforming DEVS into a formal method that can capture the properties of the model at that level. Then formal reasoning can be done with the formal structure. We argue that more insights about a model can be derived by studying other levels. We realize that one formal method might not be suitable to fully capture all aspects of a system. Hence, we propose a framework that would provide logical semantics for reasoning at different levels of abstraction. Furthermore, we make use of different formal methods at different levels, depending on the expressive power of the method chosen.

This framework that we propose is based on a graphical modeling formalism – the DEVS-Driven Modeling Language (DDML). DDML is inspired by DEVS that combines graphical modeling and formal analysis. It adopts the DEVS simulation protocol in its operational semantics. It is also a unifying framework for the two variants of DEVS (CDEVS and PDEVS).

3. THE DEVS DRIVEN MODELING LANGUAGE (DDML)

Fig. 1 is the concrete syntax of DDML showing its elements and the relationships between elements.

A DDML model interacts with its environment via IOFrames (input and output events) which are realized with input and output ports (represented graphically as arrows). There are two types of models – atomic (describes a system that cannot be decomposed into sub-systems) and coupled (composed of sub-models with couplings between the models). Couplings are partitioned into External Input Coupling (EIC), Internal Coupling (IC), and External Output Coupling (EOC). The select flags within the coupled model are used to resolve synchronization issues between child-models. This corresponds to the select function in CDEVS. Some situations can lead to a voting/Condorcet's paradox. Several flags can be added to indicate paradoxes. Atomic model is drawn as a box with input and output arrows. A coupled model is drawn in a similar shape containing its child-models, their couplings (with lines as shown) and select flags.

State variables are used to partition states in an atomic model. A DDML state is defined by a configuration on a set of finite state variables. Each state has properties based on this configuration. The Initial state is used to define all the state variables and to define the subroutines that are used in other states. An atomic model usually starts from an initial state, and then by external (in response to an external input event), internal (automatically at end of a lifespan), confluent (when there's a conflict in transition), or conditional transitions. It changes its state to passive (lifespan is infinite time), finite (lifespan between 0 and $+\infty$) or transient (lifespan is 0) states. When in a state, the system might undergo some activities. States are shown in boxes with 4 compartments for name, properties, activities and time advance. The transitions are drawn with arrows with the exception of the conditional transition drawn in a diamond box. Note the labels on the transitions. For internal (λ is an output function and assignments refer to the reconfiguration of state variables before entering the next state), external (input is the trigger input event that causes the transition). Condition shows the criteria for particular transitions. The sub-diagram “illustrating transitions” shows the graphical origin of the transitions in DDML notation.

Note the following constraints not shown in the diagram. For EIC, source = self, target \neq self; IC: source \neq self, target \neq self; EOC: source \neq self, target = self. For Transient state: $t_a = 0$; Passive state: $t_a = +\infty$; Finite state: $0 < t_a < +\infty$. “Illustrating Transitions” shows external transition must appear before the edge of the box (top or bottom) and directed towards the lateral sides. Internal transition must originate at the right edge of the box and confluent transition originates from the top right corner and is directed towards the back.

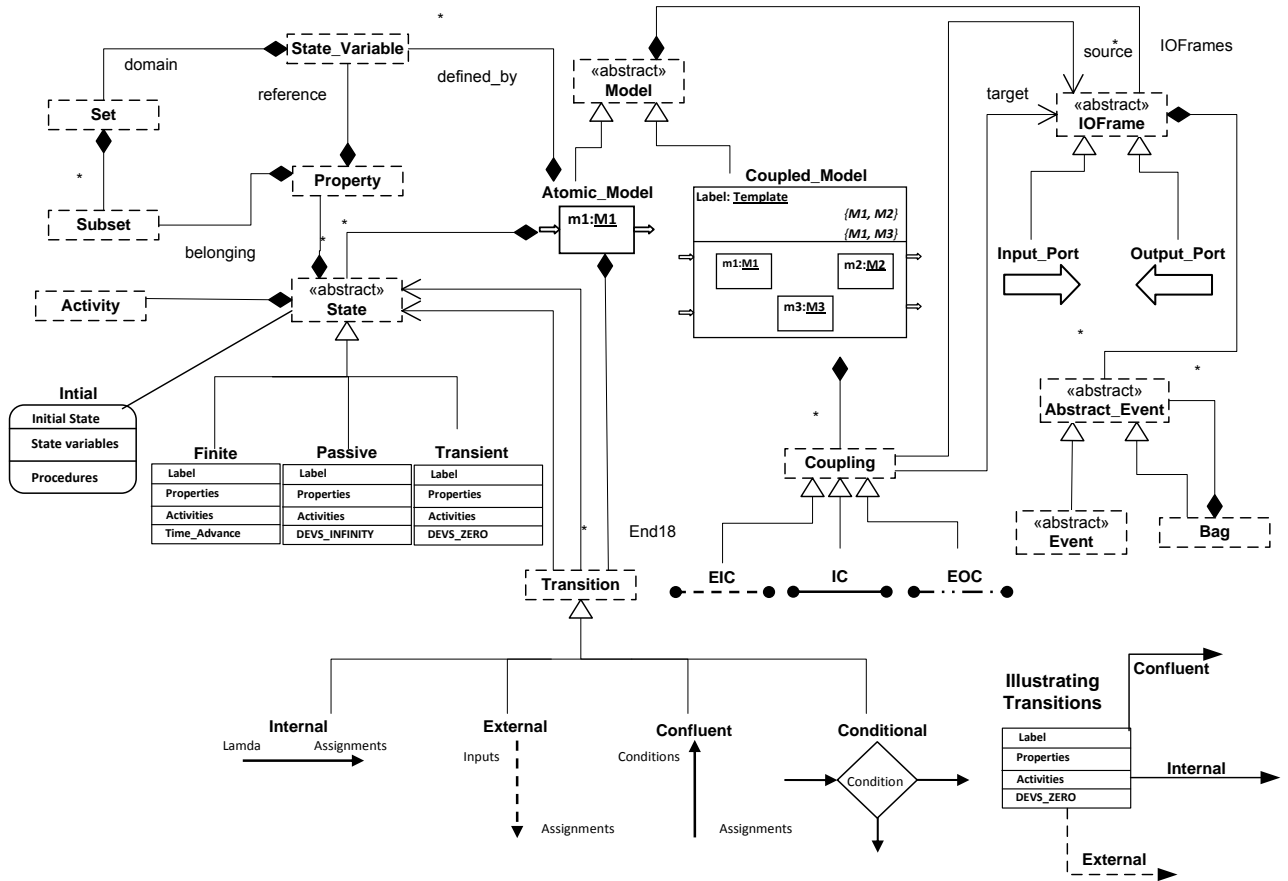


Figure 1: DDML Concrete Syntax showing notations for Coupled Model, Atomic Model, Input and Output Ports, Couplings (EIC, IC, and EOC), States (Initial, Finite, Passive, and Transient), and Transitions (External, Internal, Confluent, and Conditional). Others that do not have graphical notations are components of other graphical elements.

4. FORMAL FRAMEWORK FOR DDML

The DDML paradigm focuses on three levels of abstraction of the hierarchy of system specification (Zeigler, Praehofer, and Kim 2000):

- *Coupled Network (CN)* concerned with structural properties and functional couplings.
- *Input Output System (IOS)* concerned with system dynamics characterized by states and state transitions.
- *Input Output Relation Observation (IORO)* concerned with traces and trajectories of the system.

The formal semantics of DDML shall be expressed at these levels by using different formal methods to represent the corresponding properties that can be get. This is done so that we can derive different insights about the model. We can also take advantage of existing tools for formal analysis to derive properties of the model. Fig. 2 summarizes the formal framework for DDML.

At the CN level, a semantic mapping function maps DDML processes onto concurrent processes defined in CSP (communicating sequential processes) (Hoare 1985). At the IOS level, a semantic function maps the state transition diagram onto an LTS (labeled transition

system). At the IORO level, the system traces (which are expressed as footprints of the state transition system) are mapped onto CTL (computational tree logic). Due to space limitations, we shall show only the semantic mapping at the IOS level and give a taste of the properties that can be derived at other levels.

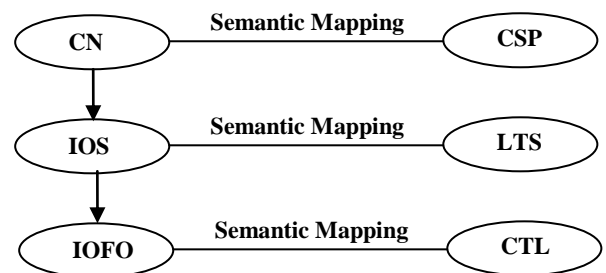


Figure 2: Formal framework for DDML

4.1. DDML at the IOS Level

The IOS level of DDML can be mapped to a labeled transition system (LTS). An LTS is a tuple $(S, \Lambda, \rightarrow)$ where S is a set (of states), Λ is a set (of labels) and $\rightarrow \subseteq S \times \Lambda \times S$ is a ternary relation (of labeled transitions). If $p, q \in S$ and $\alpha \in \Lambda$, then $(p, \alpha, q) \in \rightarrow$ is written as

$p \xrightarrow{\alpha} q$. This represents a transition from state p to state q with label α . Labels can represent different things (typically input expected or triggering actions).

An atomic DDML model is a tuple of the form:

$B = \langle X_B, Y_B, S_B, S_{0B}, \psi, C(V_B), T_{int}, T_{ext}, C(e), OP, \varphi, Act \rangle$
Where

$X_B = \{(p, X_p) | p \in IPorts \text{ and } \text{dom}(p) = X_p\}$

Such that $\#X_B < +\infty$ is the set of input ports;

$Y_B = \{(q, Y_q) | q \in OPorts \text{ and } \text{dom}(q) = Y_q\}$

Such that $\#Y_B < +\infty$ is the set of output ports;

$V_B = \{(v, S_v) | v \in StateVar \text{ and } \text{dom}(v) = S_v\}$

$\#V_B < +\infty$ is the set of state variables;

S_B is a finite set of state clusters;

$S_{0B} \in S_B$ is the initial state of B ;

OP is the set of operations defined on the state variables;

Given that 'ops' is an operation and 'a' is a state, the notation $a.ops$ indicates that 'a' is operated by 'ops' to change the state of 'a'.

$C(V_B)$ is a set of constraints on state variables. A constraint is of the form

$c := v \in D | v \in D \wedge v' \in D' | v \in D \vee v' \in D'$

Where $v = (v_1, \dots, v_n)$ and $D \subseteq \text{dom}(v_1) \times \dots \times \text{dom}(v_n)$,

$v' = (v'_1, \dots, v'_m)$ and $D' \subseteq \text{dom}(v'_1) \times \dots \times \text{dom}(v'_m)$.

$\psi: S_B \rightarrow \mathbb{P}C(V_B)$ is a mapping between each element of S_B and a finite set of conditions on the variables in V_B .

Given $s \in S_B$,

We denote $\bar{s} = \{s' \in \prod_{v \in StateVar} \text{dom}(S_v) | s' \models \psi(s)\}$

The function ψ is defined by the following:

$\psi(S_{0B})$ is verified by the initial state of the system

$\bigcup_{s \in S_B} \bar{s} = \prod_{v \in StateVar} \text{dom}(S_v)$ and $\forall s, s' \in S_B, \bar{s} \cap \bar{s}' = \emptyset$ for $s \neq s'$

$\bar{s}' = \emptyset$ for $s \neq s'$

Act is the set of activities;

$\varphi: S_B \rightarrow \mathbb{P}Act$ is the mapping from states to the set of activities ;

$T_{int} \subseteq (S_B \times \mathbb{R}^+) \times C(V_B) \times Y \times \mathbb{P}OP \times (S_B \times \mathbb{R}^+)$ is the set of internal transitions satisfying:

The internal transition $((s, d), c, l, ops, (s', d')) \in T_{int}$

will be denoted by $(s, d) \xrightarrow{\langle c, l, ops \rangle} {}_{iB} (s', d')$ or

$(s, d) \xrightarrow{\langle c, l, ops \rangle} (s', d')$ to avoid confusion.

$C(e)$ is the set of conditions α of the elapsed time e of the for

$\alpha := e \sim t | e - t \sim t' | e + t \sim t' | e * t \sim t' | \alpha_1 \wedge \alpha_2 | \alpha_1 \vee \alpha_2$

Where $\sim \in \{=, \leq, <, \geq, >\}$, e, t, t' are positive real numbers and α_1 et α_2 are conditions (denoted by $\models \alpha$ if e satisfies the condition α);

$T_{ext} \subseteq (S_B \times \mathbb{R}^+) \times C(V_B) \times X \times C(e) \times \mathbb{P}OP \times (S_B \times \mathbb{R}^+)$ is the set of external transition.

The internal transition $((s, d), c, x, c(e), ops, (s', d')) \in T_{ext}$

will be noted $(s, d) \xrightarrow{\langle x, c(e), ops \rangle} {}_{eB} (s', d')$

or $(s, d) \xrightarrow{\langle x, c(e), ops \rangle} (s', d')$. In particular, if there is no branching condition, the transition

$((s, d), \emptyset, x, c(e), ops, (s', d')) \in T_{ext}$ will be denoted $(s, d) \xrightarrow{\langle x, c(e), ops \rangle} {}_{eB} (s', d')$.

At a given instance of time, the system is in a cluster state $s \in S_B$ with a life span d that is to say the variables satisfy $\psi(s)$. If the lifespan of the current atomic state $s_1 \in \bar{s}$ elapses before an external event occurs then the model output l is sent just before transiting to another cluster state s' such that

$((s, d), l, ops, (s', d')) \in T_{int}$ (internal transition,

$v.ops \models \psi(s')$). When an external event x occurs before the end of the lifespan of the state, the model transits to the cluster state s' such that

$((s, d), x, c(e), ops, (s', d'))$ (external transition,

$v.ops \models \psi(s')$ with a life time of d' .

Given an atomic DDML model

$B = \langle X_B, Y_B, S_B, S_{0B}, \psi, C(V_B), T_{int}, T_{ext}, C(e), OP, \varphi, Act \rangle$

It can be shown that B is equivalent to LTS

$L(A) = \langle S_L, init, \Sigma, D, T \rangle$

Where,

$S_L = Q = \{((s, e), d) | s \in S_B \text{ and } 0 \leq e \leq d\}$;

$init = ((S_{0B}, 0), d)$ is the initial state ;

$\Sigma = (\bigcup_{p \in IPorts} X_p) \cup (\bigcup_{q \in Oports} Y_q) \cup (\mathbb{R}^+ \cup \{0, +\infty\})$ is the set of events (alphabet) ;

$D = \{((s, e), d) \xrightarrow{y} ((s', 0), d') | \exists ops \in \mathbb{P}OP, e = d \text{ and } ((s, d), y, ops, (s', d')) \in T_{int}\} \cup$

$\{((s, e), d) \xrightarrow{x} ((s', 0), d') | \exists ops \in \mathbb{P}OP, 0 \leq e < d \text{ and } ((s, d), x, c(e), ops, (s', d')) \in T_{ext}\}$

$T = \{((s, e), d) \xrightarrow{t} ((s, e'), d) | e' = e + t < d\}$

To illustrate the how DDML can express properties at this level graphically, we shall consider a DDML IOS model of a traffic light. The functional diagram of the TrafficLight is shown in Fig. 3.

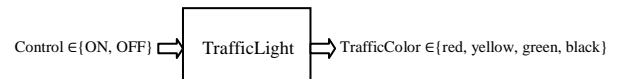


Figure 3: Simplified Traffic Light Atomic Model

The output evens in the TrafficLight system are the displays of colors (red, yellow, green, or black). This controls the flow of cars in a road network. Hence, we have an output port (trafficColors). The TrafficLight can be controlled by a Control (with ON and OFF as possible inputs).

The DDML IOS model of the TrafficLight is shown in Fig. 4. The state of the system is determined by the value of the color attribute. Hence we have the states: STOP, READY_TO_GO, READY_TO_STOP, GO, and OFF. The value of the color attribute is the indicated in the state diagrams. OFF is a passive state (time advance is INFINITY, and it does not undergo any internal transition).

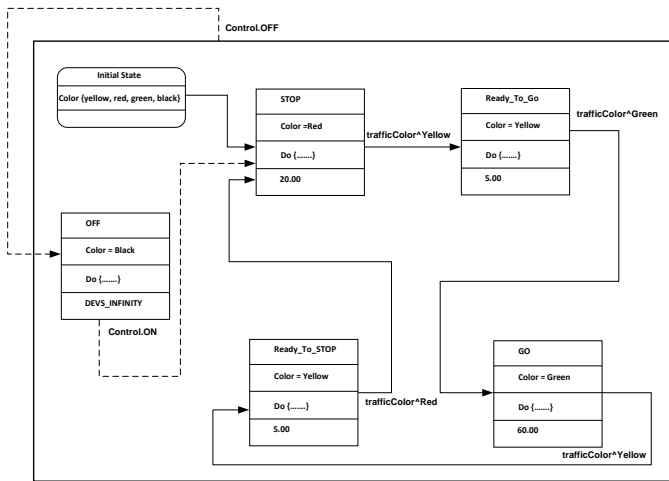


Figure 4: DDML IOS model of TrafficLight

From OFF, when the system receives an ON signal through its Control port (*Control.ON*), it undergoes an external transition to STOP state. The system remains in STOP for 20 seconds before an internal transition to READY_TO_GO. It outputs Yellow through the trafficColor port (*trafficColor^Yellow*). From any state,

if Control.OFF occurs, the system transitions to OFF state (external transition).

We can do formal analysis on this system by transforming it into a labeled transition system and leveraging formal analysis tools like LTSA (Labeled Transition System Analyzer) (Magee and Kramer 2006). LTSA is a model checking tool that uses algorithms to check for desirable and undesirable properties for all possible sequences of events and actions and to see that it conforms to specification. LTSA uses FSP (Finite State Processes) as a concise way to represent an LTS and the properties of the system can be animated and visualized. Fig. 5 is a snapshot of LTSA with our TrafficLight model. Labels are triggers for transition. We can trace all the possible sequences of events and perform checks for safety, deadlocks, and completeness.

For example, we can ask questions like “*how would the system react when it is in READY_TO_GO state (State 2) and it receives Control.ON?*” Analysis shows that the model does not take care of this possibility. This would make us to refine the model to ensure that it is complete.

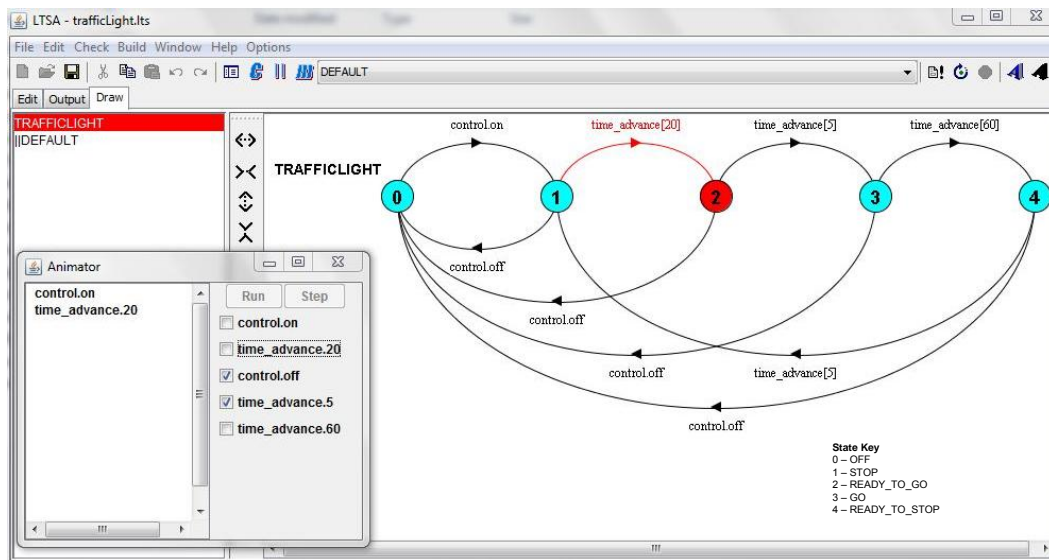


Figure 5: Analysis of TrafficLight in LTSA

4.2. DDML at the CN Level

The focus at the CN level is the structure and functional couplings of the system. To illustrate, we shall consider a model of a road-network (RN). Fig. 6 shows a schematic road network. A more advanced system would contain the road network, traffic lights, and authorization and synchronization channels. The CN model is shown in fig. 7.

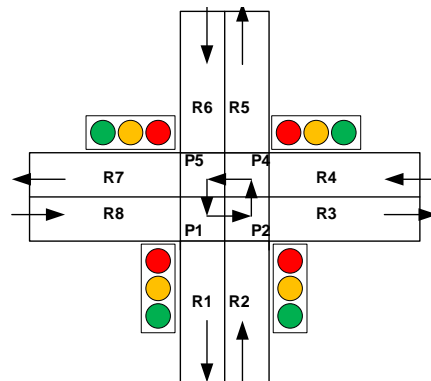


Figure 6: Road Network (RN)

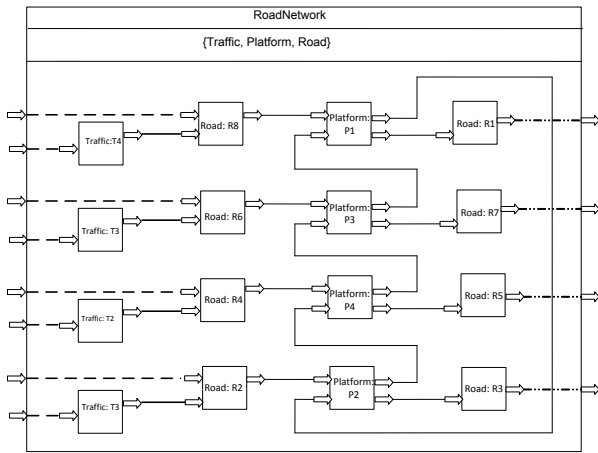


Figure 7: DDML CN model of the RoadNetwork

The coupled model (RN) consists of atomic models: TrafficLights (T1-4), Roads (R1-8) and Platforms (P1-4). The TrafficLights are for flow control of cars. A platform is used for interchange of cars within the road network. There are 4 EIC couplings between RN and R8, R6, R4, and R2. These correspond to the events whereby cars enter the road network. These cars are interchanged between roads and platforms (as shown by the ICs). EOCs (between RN and R1, R7, R5, and R3) include events whereby cars leave the road network. The select flag indicates the priorities when components are imminent.

The formal semantics of DDML CN models can be given in terms of CSP or other process algebras. We can use the FDR (Failures-Divergence Refinement) model checker to check the models for desirable and undesirable properties.

4.3. DDML at the IORO Level

At the IORO level, we can derive properties about the trajectories of the system. These trajectories are footprints of the DDML IOS and they can be mapped to CTL. We can get several footprints depending on the starting state and the sequence of activities that occur. Fig. 8 shows the footprint of the traffic from when it is in an OFF state. It receives an ON signal from its Control port and transitions to the Stop state where it remains for 20 seconds and displays Yellow signal before moving to the READY_TO_GO state where it remains for 5 seconds, and so on.

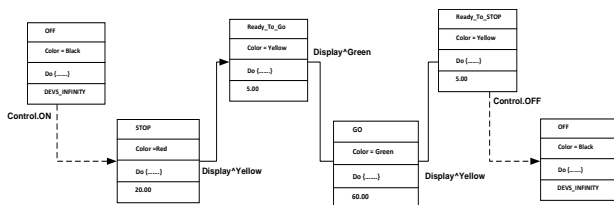


Figure 8: DDML IORO Model of Traffic Light

As explained earlier, this DDML IORO model can be easily mapped to CTL for formal analysis.

5. CONCLUSION

We presented a framework for formal reasoning of DDML models at three levels of abstraction using different formal methods that allow us to draw different insights about the model. We presented the formal semantics of DDML IOS by semantic mapping to LTS. By leveraging model-checking tools like LTSA we could do some analysis to check for desirable and undesirable properties, safety, progress, and safety properties. Future work would investigate how this can be done at the CN and IORO levels.

ACKNOWLEDGMENTS

The work in this paper is partly funded by grants from RAMSES (*Réseau Africain pour la Mutualisation et le Soutien des pôles d'Excellence Scientifique*).

REFERENCES

- Cristia, M. 2007. A TLA+ encoding of DEVS models. *International Modeling and Simulation Multiconference*, Buenos Aires (Argentina), pp. 17–22.
- Ernesto, P. 2008. *Modeling and simulation of dynamic structure discrete-event systems*. Thesis (Ph.D). McGill University.
- Hernandez, A., and Giambiasi, N. 2005. State Reachability for DEVS Models. *Proceedings of Argentine Symposium on Software Engineering*.
- Hoare, C.A.R. 1985. *Communicating Sequential Processes*. Prentice Hall International Series in Computer Science. Prentice Hall.
- Hong, J., Song, H., Kim, T., and Park, K. 1997. A Real-time discrete-event system specification formalism for seamless real-time software development. *Discrete Event Systems: Theory and Applications*, vol. 7, pp. 355–375.
- Hong, K.J., and Kim, T. G. 2005. Timed I/O Test Sequences for Discrete Event Model Verification. AIS 2004, LNAI 3397, pp. 275–284.
- Hwang, M.H., and Zeigler, B. P. 2009. Reachability Graph of Finite and Deterministic DEVS Networks. *IEEE Transactions on Automation Science And Engineering*, 6 (3).
- Magee J., Kramer J. 2006. “*Concurrency: State Models and Java Programs*”. 2nd Edition.
- Saadawi, H., Wainer, G. 2010. From DEVS to RTA-DEVS. *IEEE/ACM 14th International Symposium on Distributed Simulation and Real Time Applications*, 2010 pp.207-210.
- Saadawi, H., Wainer, G. 2009. Verification of Real-Time DEVS Models, *Proceedings of SpringSim Multi Simulation Conference*, San Diego, CA March 2009.
- Weisel, E.W., Petty, M.D., Mielke, R.R. 2005. A Comparison DEVS Semantic Composability Theory. *Proceedings of the Spring 2005 Simulation Interoperability Workshop*
- Zeigler, B., Praehofer, H., Kim, T. 2000. *Theory of Modeling and Simulation*. 2nd Edition. Academic Press.

A METHODOLOGY FOR THE DEVS SIMULATION GRAPH CONSTRUCTION

Adedoyin Adegoke^(a), Ibrahima Amadou^(b), Hamidou Togo^(b), Mamadou K. Traoré^(c)

^(a) African University of Science and Technology, Abuja (Nigeria)

^(b) Université de Bamako (Mali)

^(c) LIMOS, CNRS UMR 6158, Université Blaise Pascal, Clermont-Ferrand 2 (France)

^(a)aadegoke@aust.edu.ng, ^(b)temena2004@yahoo.fr, ^(c)traore@isima.fr

ABSTRACT

Various DEVS (Discrete Event Systems Specification) implementations exist but differ in the approaches considered for use. These approaches include how events are processed, the simulation architecture in use, the existing procedures (set of rules/algorithm), the organizational architecture (of the simulator) and so on. This work attempts to formalize a generic approach to Parallel and Distributed Simulation (PADS) with DEVS as well as embody these approaches by providing formal definitions that can be standardized for use during DEVS implementation. Therefore, we propose a DEVS Simulation Graph which gives basic information about the necessary elements that are useful for the analysis and construction of a DEVS simulator. The major aim is to help identify these elements before implementation and ease the development of a DEVS simulator.

Keywords: PADS, DEVS, Simulation Graph, Simulator

1. INTRODUCTION

DEVS (Discrete Event System Specification) can be called a universal simulation Turing machine as it offers a platform for the modeling and simulation of sophisticated systems in a variety of domains. It unifies various formalisms and provides a general description for the construction and execution of a model from an original system. Due to the separation of concerns in DEVS, the modeler needs to focus only on the models being created avoiding the details about how the simulator was built.

Parallel and Distributed Simulation (PADS) (Fujimoto 1990) has been a widely researched area in recent years. Its prominence offers increase in execution speed, reduction in execution time, execution of larger simulation models and increased processor fault tolerance to a possible failure. In addition, it provides a solution to the scientific need to federate existing and naturally dispersed simulation codes.

Although PADS is a matured field of study, its adaptability to existing modeling and simulation formalisms is an arduous task. PADS with DEVS implementation strategies differ from one another. Based on this heterogeneous factor the intrinsic

elements used in developing DEVS simulators are not formally defined for these strategies. Hence, this work will attempt to identify and capture the elements commonly used in these strategies as well as propose a more generic approach and formal framework which is deemed necessary. These fundamental elements are in terms of the simulator's tree structure, the number of execution streams and the number of computing resources.

The rest of the paper is organized as follows: Section 2 presents a review of some existing works in the area of PADS with DEVS. Section 3 presents the foundations of DEVS simulation i.e. the Simulation Tree, the concept of the Simulation Graph and the fundamental elements. Section 4 presents a methodology and basic operations which are useful for the construction of the Simulation Graph. In section 5, we present a case study and a look at Simulation Graph approaches in existing works. Finally, we conclude in Section 6.

2. PADS WITH DEVS – IMPLEMENTATIONS

We take a look at some implementations which have attempted combining PADS with DEVS.

Himmelspach, Ewald, Leye, and Uhrmacher (2007) proposed a Parallel Sequential Simulator which was implemented to ease the distribution of DEVS models on several physical processors consequently introducing the need for partitioning and load balancing. Also, it proposes performing Sequential and/or Parallel execution.

Parallel variant of the CD++ (Wainer 2009) tool was designed to execute DEVS models on parallel memory architectures i.e. with the idea of distributing the simulating entities on different physical processors.

DEVS-Ada/TW (Christensen and Zeigler 1990) is the first attempt to combine DEVS and Time Warp mechanism for Optimistic Distributed simulation. The hierarchical DEVS model is partitioned at the highest level of the hierarchy and as a consequence, the flexibility of partitioning models is restricted.

The DOHS (Distributed Optimistic Hierarchical Simulation) scheme proposed by Kim, Seong, Kim and Park (1996) is a method of distributed simulation for hierarchical and modular DEVS models. It uses the

Time Warp mechanism for global synchronization. Each node of the simulation tree structure is revised to adapt to a simulation parallel/distributed environment.

There are also other variants that take the structure of non-hierarchical DEVS model to help reduce the cost of exchanged messages in the simulator. This is the case of optimistic simulation in P-CD++ (Qi and Wainer 2007) an optimistic version of the CD++ tool which was developed for the simulation of DEVS and Cell-DEVS models.

Contrary to optimistic approaches, few parallel DEVS simulators belong to the conservative class. In (Zeigler, Praehofer and Kim 2000), a distributed simulation framework (Conservative Parallel DEVS Simulator) is described for non-hierarchical DEVS models using conservative synchronization. In addition, the performance of a conservative approach depends strictly on a good look-ahead.

3. FROM DEVS SIMULATION TREE TO DEVS SIMULATION GRAPH

We interpret the building of a Parallel and Distributed Simulation (PADS) with DEVS as a move from the original Simulation Tree (ST) to a Simulation Graph (SG).

3.1. DEVS Sequential Simulation Tree

DEVS formalism (Zeigler, Praehofer and Kim 2000) specifies system behavior as well as system structure. System behavior in DEVS is described as input and output events as well as states while system structure is built from the composition of atomic or coupled models. A coupled model is composed of several atomic or coupled models and atomic model is a basic component that cannot be decomposed any further.

A DEVS model is built according to specification i.e. Classic DEVS or Parallel DEVS. CDEVS (Classic DEVS System Specification) was introduced in 1976 by Zeigler (Zeigler 1976) to simulate and execute models sequentially on single processor machine. As a solution to the rigidity in CDEVS, the appropriate execution of simultaneous events has led to the concept of process and to PDEVS (Parallel DEVS System Specification) (Chow and Zeigler 1994).

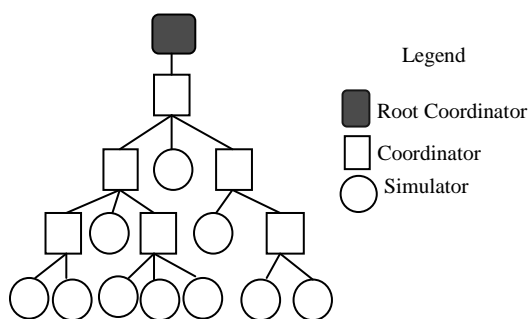


Figure 1: DEVS Simulation Tree

DEVS model execution is driven through the simulation tree which serves as vital element in construction of a DEVS simulator. The tree consists of

simulating nodes which are used for executing DEVS models. These nodes are Coordinators, Simulators and Root Coordinator which are organized in a hierarchy that mimics the hierarchical structure of a DEVS model.

A DEVS atomic model is executed by assigning a simulator to it and to a DEVS coupled model a coordinator is assigned. Root Coordinator is a special coordinator that drives the global aspects of the simulation on a tree; it initializes and ends the simulation (when a termination condition is detected).

3.2. PADS with DEVS Simulation Graph

Due to the increasing number of complex model systems it is necessary to improve efficiencies and performances of DEVS simulators. A typical PADS with DEVS implementation will result in the Simulation Graph (SG). A SG is a representation of the relationship between a DEVS simulator's fundamental elements which are simulation tree, process and processor.

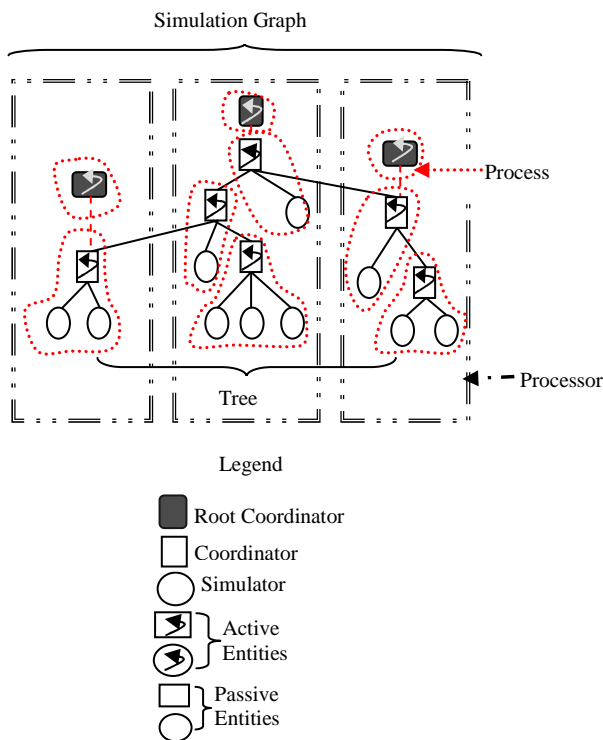


Figure 2: DEVS Simulation Graph

The structure of the DEVS simulator tree can be altered to improve performance of a simulator either by reducing (also known as flattening) the number of nodes on the tree as seen in (Jafer and Wainer 2009) or increasing the number of nodes (with specialized Coordinators and Simulators) as seen in (Troccoli and Wainer 2003).

Also, this tree can be split to form sub-trees based on the analysis of the model's structure. A simulator with single tree structure is designed with the use of a central scheduler called the Root Coordinator while in a multiple tree structure simulator each of the sub-trees has its own central scheduler/Root Coordinator and

different simulation clocks. This is the preferable solution in distributed simulation.

We define a process as a stream of execution. It contains two types of entities during execution; they are active and passive entities. An active entity is an entity that is currently active in an execution stream (e.g. Java threads, ADA Tasks, etc.). While a passive entity is part of an execution stream but not actively involved until it is activated e.g. function calling in Object Oriented Paradigm. We consider that a process would have at most one active entity. If a process has more than one active entity, those entities are then regarded as being autonomous sub-processes. Also, there can be more than one passive entity in a process.

A processor is a computing resource that allows the execution of a program (a process, an entire tree, any other executable code) on itself.

4. METHODOLOGY FOR BUILDING THE SIMULATION GRAPH

The state chart provides an overview of the method and the trajectories describe the set of all possible paths that can be taken during the construction of the Simulation Graph (SG).

The SG construction is driven based on the analysis of the initial Simulation Tree, the available number of Processes and Processors. An overview of this methodological approach is given by the following state chart. It is worth noting that the methodology iterates on each state until some user-defined satisfaction criteria are reached (optimal splitting, optimal clustering, optimal mapping and optimal transformation).

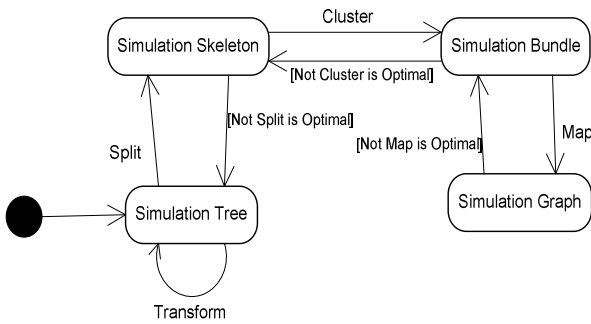


Figure 3: Simulation Graph Methodology

The process of Transformation, Splitting, Clustering and Mapping continues until it is certain that a good performance or speed will be gained during simulation from the new Simulation Graph.

4.1. Simulation Skeleton and Bundle

Informally presented, the Simulation Skeleton is the structure of the simulation protocol that can fit the PADS scheme. The Simulation Bundle is a collection of cluster of nodes. Examples are shown with Figures 4 and 5.

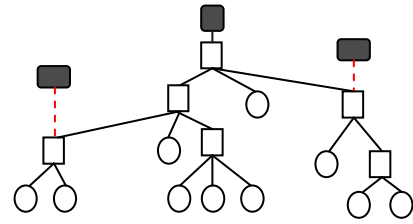


Figure 4: Simulation Skeleton

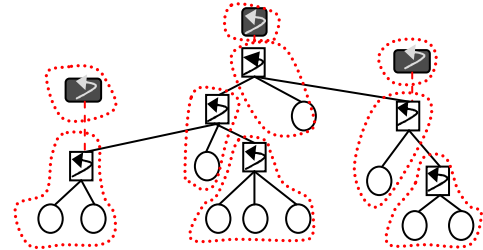


Figure 5: Simulation Bundle

4.2. Simulation Structures Formalized

Definition 1: A Simulation Tree T, can be defined as

$$T = \langle R, N, f \rangle$$

With:

- $R \in N$
- $f: N \rightarrow \wp(N)$ where $\wp(N)$ is Power Set of N
- $f^{-1}(R) = \emptyset$
- $f^{-1}(J) \neq \emptyset, \forall J \in N - \{R\}$
- Cardinal $f(R) = 1$

Where:

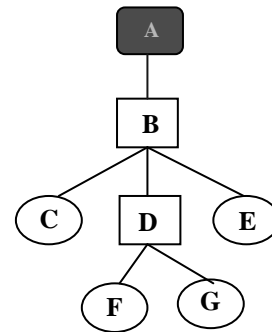
R: the Root Coordinator of the tree

N: the set of nodes of the tree

f: a function that maps a child node to its parent

A node is a "parent" of another node (its child) if it is one step higher in the hierarchy.

For example:



Tree T will be defined as

- $R = \{A\}$
- $N = \{A, B, C, D, E, F, G\}$
- $f(A) = \{B\}$
- $f(B) = \{C, D, E\}$
- $f(D) = \{F, G\}$
- $f(C) = f(E) = f(F) = f(G) = \emptyset$

Definition 2: Simulation Tree T can also be defined as

$$T = \langle R, N, F \rangle$$

With:

- $R \in N$
- $F \subset N \times (N - \{R\})$
- $(a, b) \in F \Leftrightarrow b \in f(a)$

Using Definition 2 for the example above, R and N will be defined as same while F will be

$$F = \{(A, B), (B, C), (B, D), (B, E), (D, F), (D, G)\}$$

Definition 3: A Simulation Skeleton is defined by

$$S = \langle \{R_i\}, N, f \rangle$$

With

- $R_i \in N \forall i$
- $f: N \rightarrow \wp(N)$ where $\wp(N)$ is Power Set of N
- $f^{-1}(R_i) = \emptyset \forall i$
- $f^{-1}(J) \neq \emptyset, \forall J \in N - \cup\{R_i\}$

Definition 4: A Simulation Bundle is formally defined as

$$B = \langle \{R_i\}, N, f, Ps, Cluster \rangle$$

With

- $\langle \{R_i\}, N, f \rangle$ as a skeleton
- Ps is the set of Processes
- $Cluster: N \rightarrow Ps$

Definition 5: A Simulation Graph is defined by

$$G = \langle \{R_i\}, N, f, Ps, Pr, Cluster, Map \rangle$$

With

- $\langle \{R_i\}, N, f, Ps, Cluster \rangle$ is a Simulation Bundle
- Pr is a set of Processors
- $Map: Ps \rightarrow Pr$

4.3. Basic Operations Formalized

In this section we show that moving from a Simulation Tree (ST) to a Simulation Graph (SG) can be decomposed into basic operations and later we will show the methodology that drives this process.

4.3.1. Split

It is a function used for creating a partition of simulating entities/nodes from a simulation tree.

Definition 6: Split: $\tau \rightarrow \Sigma$

With

$$T = \langle R, N, f \rangle$$

$$Split(T) = \langle \{R_i\}, N', f' \rangle$$

Based on the following conditions:

- $R \in \cup\{R_i\}$
- $N' = N \cup \{R_i\}$
- $f'_{/N} = f$

Where τ is the set of all possible trees and Σ is the set of all possible skeletons.

4.3.2. Cluster

This function takes the available number of nodes and groups them into Processes.

Definition 7: Cluster: $N \rightarrow Ps$

Where

Ps is the set of Processes.

Based on the conditions that

- $Cluster^{-1}(p)$ is Connex $\forall p \in Ps$
- $\forall p_i, p_j \in Ps, p_i \neq p_j,$
 $Cluster(p_i) \cap Cluster(p_j) = \emptyset,$

4.3.3. Map

This function takes the set of available Processes and plots them onto the set of available Processors.

Definition 8: Map: $Ps \rightarrow Pr$

Where

- Ps is the set of Processes
- Pr is the set of Processors

Based on the following condition

$$\forall p_i, p_j \in Ps, p_i \neq p_j, Map(p_i) \cap Map(p_j) = \emptyset$$

4.3.4. Transform

Transform is a function used for altering the Simulation Tree structure either by expansion or reduction. This altering is done on the number of available nodes (not including the Root Coordinators) on the Tree and their relationships.

Definition 9: $Transf[Na, Nr, Fa, Fr]: \tau \rightarrow \tau$

$$Transf[Na, Nr, Fa, Fr](\langle R, N, F \rangle) = \langle R, N', F' \rangle$$

With

- $N' = N \cup Nr - Na$
- $F' = F \cup Fr - Fa$

Where

- Na is the set of nodes to be added to N
- Nr is the set of nodes to be removed from N
- Fa is the set of relationships to be added to F
- Fr is the set of relationships to be removed from F

Based on the following conditions:

- $Na \cap N = \emptyset$
- $Nr \subset N - \{R\}$
- $Fa \subset (N \times Na) \cup Na^2 \cup (Na \times N - \{R\})$
- $Fr \subseteq F$

5. APPLICATION

5.1. Case Study

In this study we describe a possible path in the application of the Simulation Graph construction methodology. This application starts with the original DEVS tree. At the Transform stage the tree structure is modified by reduction after which the tree is Split to create a partition of nodes also increasing the number of Root Coordinators on the entire tree. These partitions of nodes are grouped into available number of Processes thereby creating a Simulation Bundle (Figure 6c). At

the final stage (Map), a Simulation Graph was created by mapping the each Process on the Simulation Bundle to an available number of Processors. See Figure 6.

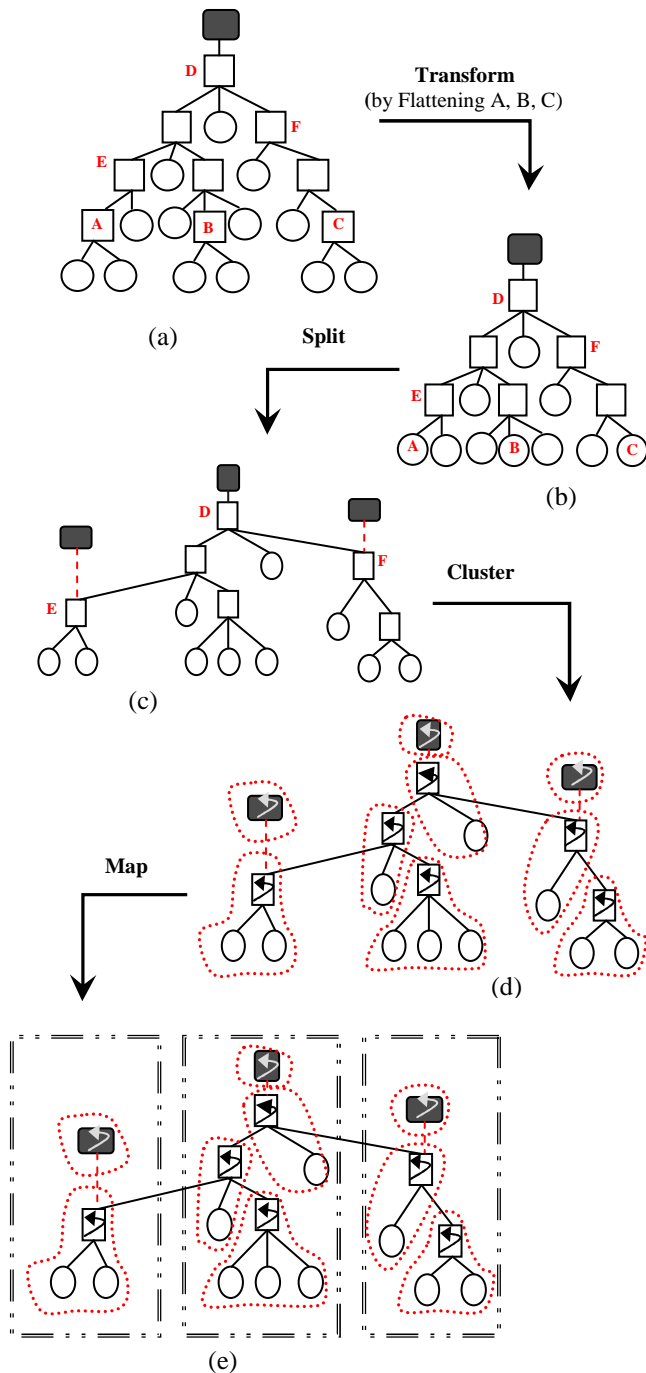


Figure 6: Application of Simulation Graph Methodology

5.2. Works Revisited

In a more general overview, most DEVS implementation decisions have been observed to be based on the fundamental elements i.e. simulation tree, process and processor. We present some of them here and their Simulation Graph.

To explain the strategies in the literature in a more formal way we suggest the Tree-Process-Processor notation. It consists in defining the number of elements

for each aspect of PADS. We use N for “many elements”. For example, a 1-1-1 scheme is a Simulation Graph with 1 tree, 1 process and 1 processor while N-N-1 is a Simulation Graph with many trees, many processes and 1 processor.

PythonDEVS (Bolduc and Vangheluwe 2002) is a 1-1-1 strategy. It uses the CDEVS formalism in specifying models and as a consequence it performs sequential simulation.

The Abstract Threaded Simulator of the James II (Himmelspace and Uhrmacher 2006) package uses a 1-N-1 strategy with its processes created using Java threads. Depending on the memory size of the processor and the model size, the cost of creating threads gets expensive as the number of models increases. This is a critical factor to be considered when using many processes.

In (Troccoli and Wainer 2003), the Parallel CD++ Simulator is a 1-N-N strategy. Also, new simulation nodes were used on the tree thereby expanding it. This methodology better suits distributed simulation but increases the cost of communication between the nodes. Also, Parallel Sequential Simulator by Himmelspace, Ewald, Leye and Uhrmacher (2007) uses this 1-N-N strategy.

The Conservative CD++ (Jafer and Wainer 2010) is an N-N-N strategy. In order to reduce communication costs between the nodes the simulator was flattened. The flattening involves a reduction in the number of nodes on the simulation tree. Some other works that uses the N-N-N strategy include DEVS-Ada/TW (Christensen and Zeigler 1990), DOHS scheme by Kim, Seong, Kim and Park (1996) and Optimistic Parallel CD++ (Qi and Wainer 2007).

Also, we observed that having an implementation which involves the use of N-1-1 strategy or N-1-N strategy is not realistic. The reason for this is execution of each tree is asynchronous and can be done simultaneously using many processes instead.

6. CONCLUSION

This paper is part of a more general research direction in that we investigate various approaches commonly used in PADS with DEVS to build simulators. We presented the Simulation Graph concept to help ease the process of building DEVS simulators and provide a common platform for different implementation strategies. This was achieved through the identification of the fundamental elements used in DEVS simulators and the relationship between them. Thus, with this we were able to provide definitions for a formal framework as opposed to the traditional intuitive way of constructing a DEVS simulator. Also, we presented a possible path in the application of the Simulation Graph and revisited works by identifying their Simulation Graph approach. Further works include automating this process by implementing the methodology in SimStudio package (Traoré 2008).

ACKNOWLEDGMENTS

The work in this paper is partly funded by grants from RAMSES (*Réseau Africain pour la Mutualisation et le Soutien des poles d'Excellence Scientifique*).

REFERENCES

- Bolduc, J., Vangheluwe, H., 2002. *A Modeling and Simulation Package for Classic Hierarchical DEVS*. Technical Report, McGill University, School of Computer Science.
- Chow, A. C., Zeigler, B. P., 1994. Revised DEVS: A Parallel, Hierarchical, Modular Modeling Formalism. *Proceedings of the Winter Simulation Conference*
- Christensen, E.R., Zeigler, B. P., 1990. Distributed Discrete Event Simulation: Combining DEVS and Time Warp. *In Proceedings of the SCS Eastern Multiconference on AI and Simulation Theory and Applications*
- Fujimoto, R. M., 1990. Parallel Discrete Event Simulation. *Communications of the ACM*, 33(10), 30-53.
- Himmelspach, J., Ewald, R., Leye, S., Uhrmacher, A., 2007. Parallel and Distributed Simulation of Parallel DEVS Models. *Proceedings of the 2007 Spring Simulation Multiconference*.
- Himmelspach, J., Uhrmacher, A., 2006, Sequential Processing of PDEVS Models. *Proceedings of the 3rd EMSS*, 239-244.
- Jafer, S., Wainer, G. A., 2009. Flattened Conservative Parallel Simulator for DEVS and CELL-DEVS. *Proceedings of CSE (1)*, 443-448.
- Jafer, S., Wainer, G. A., 2010. Conservative DEVS – A Novel Protocol for Parallel Conservative Simulation of DEVS and Cell-DEVS Models. *Proceedings of 2010 Spring Simulation Conference (SpringSim10), DEVS symposium 168-175*.
- Kim, K. H., Seong, Y. R., Kim, T. G., Park, K. H., 1996. Distributed Simulation of Hierarchical DEVS Models: Hierarchical Scheduling Locally and Time Warp Globally. *TRANSACTIONS of the SCS International 13, no. 3, 135-154*.
- Qi, L., Wainer, G. A., 2007. Parallel Environment for DEVS and Cell-DEVS Models. *SIMULATION* 83(6), 449-471.
- Traoré, M. K., 2008. SimStudio. A Next Generation Modeling and Simulation Framework, *SIMUTools'08, ISBN 978-963-9799-20-21*. March 3-7, Marseille, France
- Troccoli, A., Wainer, G., 2003. Implementing Parallel Cell-DEVS. *Proceedings of the 36th Annual Symposium on Simulation*, 273-277. March 30-April 02.
- Wainer, G. A., 2009. *Discrete-Event Modeling and Simulation: A practitioner's Approach*. New York: CRC Press.
- Zeigler, B. P., 1976. *Theory of Modeling and Simulation*. New York; Wiley-Interscience.

Zeigler, B. P., Praehofer, H., Kim, T. G., 2000. *Theory of Modeling and Simulation. Integrating Discrete Event and Continuous Complex Dynamic Systems*. London; Academic Press.

EFFICIENT EXPLORATION OF COLOURED PETRI NET BASED SCHEDULING PROBLEM SOLUTIONS

Gašper Mušič^(a)

^(a)University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia

^(a)gasper.music@fe.uni-lj.si

ABSTRACT

The paper deals with simulation-optimization of schedules that are modelled by simple Coloured Petri nets (CPNs). CPN modelling is combined by predefined transition sequence conflict resolution strategy to enable generation of neighbouring solutions that are always feasible. This way standard local search optimization algorithms can be effectively applied to CPN models. Modelling approach and neighbourhood construction procedure are explained in detail. Some preliminary results of tests on standard job shop benchmark problems are provided.

Keywords: Petri nets, simulation, optimization, scheduling, local search

1. INTRODUCTION

Among modelling formalisms suitable for description of systems with highly parallel and cooperating activities, Petri nets are perhaps the most widely used one. With Petri nets, production systems' specific properties, such as conflicts, deadlocks, limited buffer sizes, and finite resource constraints can be easily represented in the model (Tuncel and Bayhan 2007).

Optimization of planning and scheduling problems has been investigated within the production control community for several years. This is one of the fields where information technology has an immediate and considerable impact on the efficiency and quality of production control and related manufacturing processes. The simplicity of model building, the possibility of realistic problem formulation as well as the ability of capturing functional, temporal and resource constraints within a single formalism motivated the investigation of Petri net based optimization of planning and scheduling problems.

In our previous work a simulation based optimization approach applying Petri nets was intensively studied, as well as other, more classical approaches, such as dispatching rules and reachability tree based heuristic search (Gradišar and Mušič 2007, Löscher, Mušič and Breiteneker 2007, Mušič, Löscher and Breiteneker 2008, Mušič 2008).

In the reported works, Petri nets based scheduling methods were compared and a certain level of experience was gained about the behaviour of the methods in relation

to different scales of problems. Among others, reachability tree based heuristic search methods (Lee and DiCesare 1994, Yu, Reyes, Cang and Lloyd 2003, Mujica, Piera and Narciso 2010) were of particular interest, since the rich structural analysis framework of Petri nets seemed a promising way to derive a suitable heuristic function that would significantly improve the efficiency of the search and obtained results.

The obtained results meet the expectations for small or moderate size problems. Unfortunately, the results for complex problems, such as standard job-shop benchmark problems (Taillard 1993), are not satisfactorily, even with the advanced heuristic functions.

Local search methods, on the other hand, are widely used in operational research (OR) community (Blazewicz, Domschke and Pesch 1996, Vaessens, Aarts and Lenstra 1996). Their performance is not significantly decreased for larger problems. The optimality of obtained solutions is not guaranteed (Pinedo 2008) but they can be obtained in computational time that is significantly shorter compared to other methods. In particular, the Tabu search algorithms (Dell'Amico and Trubian 1993, Taillard 1994, Nowicki and Smutnicki 1996) are claimed to represent the state-of-the-art by a comfortable margin over the closest competition (Watson, Whitley and Howe 2005).

This motivated the investigations on combination of Petri net modelling approach and local search methods (Löscher, Mušič and Breiteneker 2007). In particular, a combination of efficient generation of feasible neighbouring solutions from Coloured Petri net representation of the problem and local search is tested in this paper.

The paper shows how a Coloured Petri net model of a scheduling problem can be used in conjunction with state-of-the-art local search algorithms provided a special type of parameterized conflict resolution strategy and neighbouring solution generation procedure are adopted. Parameters in a form of sequence vectors are adjoined to shared resources in the system. The makespan of a feasible schedule is calculated through CPN model simulation, which is supervised by the sequence vectors. Constrained permutations on these vectors are used to generate neighboring schedule solutions that are always feasible, which improves the effectiveness of CPN based exploration of solutions compared to previous works.

2. JOB-SHOP SCHEDULING AND LOCAL SEARCH

A job-shop scheduling problem is a well known problem in the field of operations research. It is defined as the determination of the order in which a set of jobs (tasks) $\{J_i | i = 1, \dots, n\}$ is to be processed through a set of machines (resources) $\{M_k | k = 1, \dots, m\}$.

Job i is specified by a set of operations $\{o_j | j = 1, \dots, m_j\}$ representing the processing requirements on various machines. Processing times are assigned to individual operations. Job shop problem assumes that all jobs have to be processed on all machines while the operations processing order (routing) is fixed but not identical for individual jobs. The objective of a job shop scheduling problem is most often to determine a job sequence schedule for every machine, which will minimize the total processing time, i.e., makespan.

This objective can be reached with various strategies including heuristic based dispatching rules, simulation-optimization and local search methods.

Local search is an iterative procedure which moves from one solution in the search space S to another as long as necessary. In order to systematically search through S , the possible moves from a solution s to the next solution should be restricted in some way. To describe such restrictions a neighbourhood structure $N : S \rightarrow 2^S$ is introduced on S . For each $s \in S$, $N(s)$ describes the subset of solutions which can be reached in one step by moving from s . The set $N(s)$ is called the neighbourhood of s . Usually it is not possible to calculate the neighbourhood structure $N(s)$ beforehand because S has an exponential size. To overcome this difficulty, a set AM of allowed modifications $F : S \rightarrow S$ is introduced. For a given solution s , the neighbourhood of s can be defined by $N(s) = \{F(s) | F \in AM\}$.

A general local search method may be described as follows. Each iteration starts with a solution $s \in S$ and then a solution $s' \in N(s)$ or a modification $F \in AM$ which provides $s' = F(s)$ is chosen. Based on the values of the objective function $f : S \rightarrow \mathbb{R}$, $f(s)$ and $f(s')$, the new solution is adopted or discarded. The next iteration starts with either the old or the new solution. Different methods of choice of solution for the next iteration lead to different local search techniques, e.g. simulated annealing, tabu search, and genetic algorithms (Brucker 2001).

Local search algorithms are simple to implement and fast in execution, but they have the main disadvantage that they can terminate in the first local minimum, which might give an objective function that deviates substantially from the global minimum. The useful algorithms should be able to leave the local minimum by sometimes accepting transitions leading to an increase in the objective function. Simulated Annealing is an example of such an approach where cost-increasing transitions are accepted with a non-zero probability which decreases gradually as the algorithm continues its execution (van Laarhoven, Aarts and Lenstra 1992).

The previously described scheduling methods can be used in combination with various modelling formalisms. As mentioned in the introduction, a Coloured Petri net framework will be used here.

3. COLOURED PETRI NETS

Coloured Petri nets (CPNs) used for modelling of scheduling problems in this paper are defined as follows. Note that the definition is different from (Jensen 1997) in the sense that it does not allow for transition guards. Instead it closely follows one of the representations used in (Basile, Carbone and Chiacchio 2007) with an important difference: a different interpretation of transition delays is used, which is closer to that of (Jensen 1997).

A $CPN = (\mathcal{N}, M_0)$ is a Coloured Petri net system, where: $\mathcal{N} = (P, T, Pre, Post, Cl, Co)$ is a Coloured Petri net structure:

- $P = \{p_1, p_2, \dots, p_k\}$, $k > 0$ is a finite set of places.
- $T = \{t_1, t_2, \dots, t_l\}$, $l > 0$ is a finite set of transitions (with $P \cup T \neq \emptyset$ and $P \cap T = \emptyset$).
- Cl is a set of colours.
- $Co : P \cup T \rightarrow Cl$ is a colour function defining place marking colours and transition occurrence colours. $\forall p \in P, Co(p) = \{a_{p,1}, a_{p,2}, \dots, a_{p,u_p}\} \subseteq Cl$ is the set of u_p possible colours of tokens in p , and $\forall t \in T, Co(t) = \{b_{t,1}, b_{t,2}, \dots, b_{t,v_t}\} \subseteq Cl$ is the set of v_t possible occurrence colours of t .
- $Pre(p, t) : Co(t) \rightarrow Co(p)_{MS}$ is an element of the pre-incidence function and is a mapping from the set of occurrence colours of t to a multiset over the set of colours of p , $\forall p \in P, \forall t \in T$. It can be represented by a matrix whose generic element $Pre(p, t)(i, j)$ is equal to the weight of the arc from p w.r.t colour $a_{p,i}$ to t w.r.t colour $b_{t,j}$. When there is no arc with respect to the given pair of nodes and colours, the element is 0.
- $Post(p, t) : Co(t) \rightarrow Co(p)_{MS}$ is an element of the post-incidence function, which defines weights of arcs from transitions to places with respect to colours.

$M(p) : Co(p) \rightarrow \mathbb{N}$ is the marking of place $p \in P$ and defines the number of tokens of a specified colour in the place for each possible token colour in p . Place marking can be represented as a multiset $M(p) \in Co(p)_{MS}$ and the net marking M can be represented as a $k \times 1$ vector of multisets $M(p)$. M_0 is the initial marking of a Coloured Petri net.

3.1. Timed models

As described in (Bowden 2000), there are three basic ways of representing time in Petri nets: firing durations (FD), holding durations (HD) and enabling durations (ED). The FD principle says that when a transition becomes enabled it removes the tokens from input places immediately but does not create output tokens until the firing duration has elapsed. When using HD principle, a firing has no duration but a created token is considered unavailable for the time assigned to transition that created the token. The unavailable token can not enable a transition and therefore causes a delay in the subsequent transition firings. With ED principle, the firing of the transitions has no duration while the time delays are represented by forcing transitions that are enabled to stay so for a specified period of time before they can fire.

The ED concept is more general than HD. Furthermore, in (Lakos and Petrucci 2007) an even more general concept is used, which assigns delays to individual arcs, either inputs or outputs of a transition. This way both ED and HD concepts are covered, and the enabling delay may even depend on the source of transition triggering while holding delay may differ among different activities started by the same transition.

When modelling several performance optimization problems, e.g. scheduling problems, such a general framework is not needed. It is natural to use HD when modelling most scheduling processes as transitions represent starting of operations, and generally once an operation starts it does not stop to allow another operation to start in between. HD principle is also used in timed version of CPNs defined by Jensen, although the unavailability of the tokens is only defined implicitly through the corresponding time stamps. While CPNs allow the assignment of delays both to transition and to output arcs, we further simplify this by allowing time delay inscriptions to transitions only. This is sufficient for the type of examples investigated here, and can be generalized if necessary.

To include a time attribute of the marking tokens, which implicitly defines their availability and unavailability, the notation of (Jensen 1997) will be adopted. Colours are adjoined to token number by 'c notation and coloured tokens are accompanied with a timestamp, which is written next to the token number and colour and separated from the colour by @. E.g., two c-coloured tokens with time stamp 10 are denoted $2 \cdot c@10$. A collection of tokens with different colours and/or time stamps is defined as a multiset, and written as a sum (union) of sets of timestamped coloured tokens. E.g., two c-coloured tokens with time stamp 10 and three d-coloured tokens with timestamp 12 are written as $2 \cdot c@10 + 3 \cdot d@12$. The timestamp of a token defines the time from which the token is available.

Time stamps are elements of a time set TS , which is defined as a set of numeric values. In many software implementations the time values are integer, i.e. $TS = \mathbb{N}$, but will be here admitted to take any positive real value including 0, i.e. $TS = \mathbb{R}_0^+$. Timed markings are represented as collections of time stamps and are multisets over TS : TS_{MS} . By using HD principle the formal representation of a Coloured Timed Petri net is defined as follows.

$CTPN = (\mathcal{N}, M_0)$ is a Coloured Timed Petri net system, where:

- $\mathcal{N} = (P, T, Pre, Post, Cl, Co, f)$ is a Coloured Time Petri net structure with $(P, T, Pre, Post, Cl, Co)$ as defined above.
- $f : Co(t) \rightarrow TS$ is the time function that assigns a non-negative deterministic time delay to every occurrence colour of transition $t \in T$.
- $M(p) : Co(p) \rightarrow TS_{MS}$ is the timed marking, M_0 is the initial marking of a timed Petri net.

3.2. Firing rule

Functions Pre and $Post$ define the weights of directed arcs, which are represented by arc inscriptions in the matrix form. In the case when the all the weights in the matrix

are 0, the arc is omitted. Let $\bullet t_b \subseteq P \times Cl$ denote the set of places and colours which are inputs to occurrence colour $b \in Co(t)$ of transition $t \in T$, i.e., there exists an arc from every $(p, a) \in \bullet t$ to t with respect to colours $a \in Co(p)$ and $b \in Co(t)$.

To determine the availability and unavailability of tokens, two functions on the set of markings are defined. The set of markings is denoted by \mathbb{M} . Given a marking and model time, $m : P \times \mathbb{M} \times TS \rightarrow Co(p)_{MS}$ defines the number of available coloured tokens, and $n : P \times \mathbb{M} \times TS \rightarrow Co(p)_{MS}$ the number of unavailable coloured tokens for each place of a TPN at a given model time $\tau_k \in TS$.

Two timed markings can be added (denoted $+_\tau$) in a similar way as multisets, i.e. by making a union of the corresponding multisets. The definition of subtraction is somewhat more problematic. To start with, a comparison operator is defined. Let M_1 and M_2 be markings of a place $p \in P$. By definition, $M_1 \geq_\tau M_2$ iff $m(p, M_1, \tau_k) \geq m(p, M_2, \tau_k), \forall \tau_k \in TS, \forall a \in Co(p)$.

Similarly, the subtraction is defined by the number of available tokens, and the subtrahend should not contain any unavailable tokens. Let M_1, M_2 and M_3 be markings of a place $p \in P$, $M_1 \geq_\tau M_2$, and $m(p, M_1, \tau_k), m(p, M_2, \tau_k)$, and $m(p, M_3, \tau_k)$, be the corresponding numbers of available tokens at time τ_k , and $n(p, M_2, \tau_k) = 0$. The difference $M_3 = M_1 -_\tau M_2$ is then defined as any $M_3 \in \mathbb{M}$ having $m(p, M_3, \tau_k) = m(p, M_1, \tau_k) - m(p, M_2, \tau_k)$.

Using the above definitions, the firing rule of a CTPN can be defined. Given a marked $CTPN = (\mathcal{N}, M)$, a transition t is time enabled at time τ_k w.r.t occurrence colour $b \in Co(t)$, denoted $M[t_b]_{\tau_k}$ iff $m(p, M, \tau_k) \geq Pre(p, t)(b), \forall p \in \bullet t$. An enabled occurrence transition can fire, and as a result removes tokens from input places and creates tokens in output places. If transition t fires w.r.t occurrence colour b , then the new marking is given by $M'(p) = M(p) -_\tau Pre(p, t)(b)@_{\tau_k} +_\tau Post(p, t)(b)@(\tau_k + f(t, b)), \forall p \in P$. Here the subtraction operation is implemented in such a way that in case of several choices, the token with the oldest timestamp is always removed first. If marking M_2 is reached from M_1 by firing t_b at time τ_k , this is denoted by $M_1[t_b]_{\tau_k} M_2$. The set of markings of TPN \mathcal{N} reachable from M is denoted by $R(\mathcal{N}, M)$.

4. COLOURED PETRI NET MODELLING OF SCHEDULING PROBLEMS

An important concept in PNs is that of conflict. Two transition firings are in conflict if either one of them can occur, but not both of them. Conflict occurs between transitions that are enabled by the same marking, where the firing of one transition disables the other transition.

The conflicts and the related conflict resolution strategy play a central role when modelling scheduling problems. This may be illustrated by a simple example, shown in Figure 1. The example involves two machines $M = \{M_1, M_2\}$, which should process two jobs $J = \{J_1, J_2\}$, and where $J_1 = \{o_1(M_1) \prec o_2(M_2)\}$ and $J_2 = \{o_3(M_1)\}$. Job J_1 therefore consist of two opera-

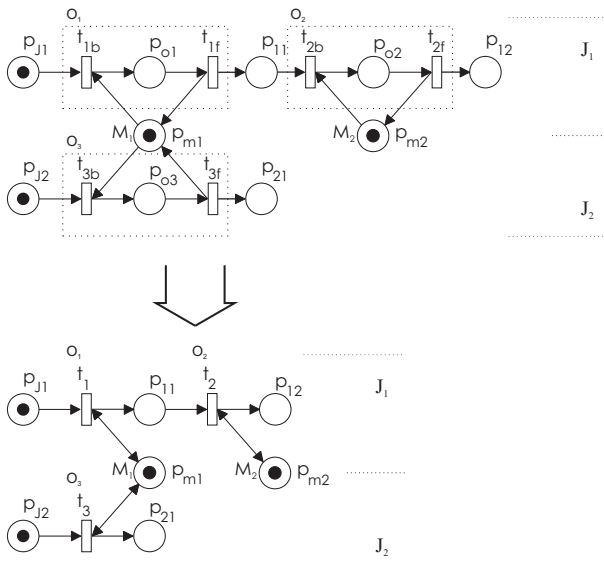


Figure 1: A PN model of a simple scheduling problem

tions, the first one using machine M_1 and the second one machine M_2 , while job J_2 involves a single operation using machine M_1 . Obviously, the two jobs compete for machine M_1 . This is modelled as a conflict between transitions starting corresponding operations.

Place p_{m1} is a resource place. It models the machine M_1 and is linked to t_{1b} and t_{3b} , which start two distinct operations. Clearly, the conflict between t_{1b} and t_{3b} models a decision, whether machine M_1 should be allocated to job J_1 or J_2 first.

Similarly, other decisions are modelled as conflicts linked to resource places. The solution of the scheduling problem therefore maps to a conflict resolution in the given Petri net model.

The transitions that model finishing of operation (t_{if} in Figure 1) are not relevant for scheduling and can be removed. Same holds for intermediate buffer places since the holding duration interpretation of transition delays guarantees that a subsequent transition can not fire before the precedent transition delay expires. The model can be therefore simplified as shown in the lower part of Figure 1. The occupation of a shared resource M_i during the evolution of the system is marked by a presence of unavailable token in the corresponding place p_{mi} .

With the introduction of token and occurrence colours, the resource sharing as described above can be represented in even much more compact model. Several jobs that go through a similar operation sequence can be folded together and represented by a single place/transition sequence with different token colours. The transition occurrence colours enable to distinguish different jobs both in terms of operation durations as well as in terms of their dependence on shared resources. The model from Figure 1 therefore maps to the model in Figure 2. The two jobs are represented by two token colours while a third colour is added to model resource availability. Both remaining transitions appear with two occurrence colours to model different durations where the absence of the second operation in

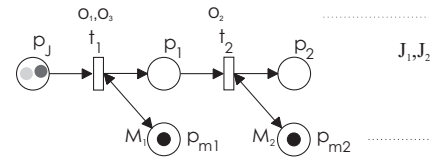


Figure 2: A CTPN model of a simple scheduling problem

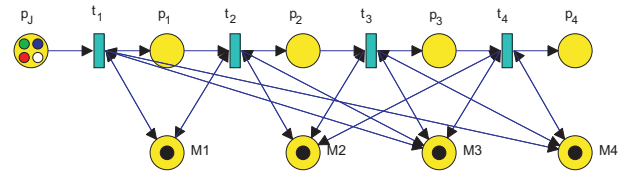


Figure 3: A simple job shop problem

Table 1: Operation durations for a simple job shop problem

Operation \ Job	J_1	J_2	J_3	J_4
o_1	54	9	38	95
o_2	34	15	19	34
o_3	61	89	28	7
o_4	2	70	87	29

Table 2: machine requirements for a simple job shop problem

Operation \ Job	J_1	J_2	J_3	J_4
o_1	3	4	1	1
o_2	1	1	2	3
o_3	4	2	3	2
o_4	2	3	4	4

job J_2 is simply modelled by setting the duration to zero.

A more elaborated example is shown in Figure 3. The model is based on a test example from Taillard (1993). It consists of four jobs and four machines. Every job includes four operations. Operation durations are shown in Table 1 and resource requirements in Table 2. Note that arc weights are not shown in the figure, they will be shown in the sequel. Nevertheless, only the arcs with at least one nonzero weight for any occurrence colour are shown.

Further compaction can be achieved by folding the operation places. Job sequences, operation durations and resource requirements are coded by different sets of colours and corresponding transition guards and expressions (Mujica, Piera and Narciso 2010). Since the transition guards and expressions are not supported by the type of CTPNs used in this paper, this representation can not be used here. The proposed representation is therefore not the most compact one but has the advantage of a very efficient coding in a general mathematical analysis software, e.g. Matlab.

In Matlab, the flow matrices of a CPN can be represented as cell matrices of size $|P| \times |T|$, where each element is a cell containing weight matrix of size $|Co(p)| \times |Co(t)|$. E.g. for example in Figure 3 the corresponding

pre-incidence matrix is

$$Pre = \begin{bmatrix} I_4 & 0 & \cdots & 0 \\ 0 & I_4 & 0 & 0 \\ 0 & 0 & I_4 & 0 \\ 0 & \cdots & 0 & I_4 \\ 0 & \cdots & \cdots & 0 \\ R_{11} & R_{12} & R_{13} & R_{14} \\ R_{21} & R_{22} & R_{23} & R_{24} \\ R_{31} & R_{32} & R_{33} & R_{34} \\ R_{41} & R_{42} & R_{43} & R_{44} \end{bmatrix} \quad (1)$$

where I_4 stands for 4×4 identity matrix and zeros should be interpreted as 4×4 zero matrices. R_{ij} define i -th resource requirements of j -th operation within jobs:

$$\begin{aligned} R_{11} &= \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix} & R_{12} &= \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix} \\ R_{21} &= \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix} & R_{22} &= \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \\ R_{31} &= \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} & R_{32} &= \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix} \\ R_{41} &= \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} & R_{42} &= \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned} \quad (2)$$

$$\begin{aligned} R_{13} &= \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix} & R_{14} &= \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix} \\ R_{23} &= \begin{bmatrix} 0 & 1 & 0 & 1 \end{bmatrix} & R_{24} &= \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \\ R_{33} &= \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} & R_{34} &= \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} \\ R_{43} &= \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} & R_{44} &= \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix} \end{aligned}$$

Furthermore, the cell matrix can be any time converted to an incidence matrix of the corresponding unfolded P/T net and reverse, the P/T net can be folded back. The only information necessary consists of the place and transition colour sets of individual nodes in the CPN.

For example, the Matlab command

```
>> PTPre=cell2mat(Pre)
```

unfolds the pre-incidence matrix from the previous example into a pre-incidence matrix of an equivalent P/T Petri net and the command

```
>> Pre=mat2cell(PTPre, ncolP, ncolT)
```

reproduces back the original cell matrix, provided that vectors $ncolP$ and $ncolT$ contain information about numbers of token and occurrence colours for all $p \in P$ and $t \in T$. E.g., for the above example $ncolP = [4 \ 4 \ 4 \ 4 \ 4 \ 1 \ 1 \ 1 \ 1]^T$ and $ncolT = [4 \ 4 \ 4 \ 4]^T$.

This way the CPN framework can be used to efficiently encode various scheduling problems into a compact representation. Later the CPN representation can be analyzed directly, or can be translated into an equivalent P/T Petri net, which enables the application of standard PN analysis methods as well as PN based scheduling techniques.

4.1. Derivation of optimal or sub-optimal schedules

A derived Coloured Petri net model can be simulated by an appropriate simulation algorithm. During the simulation, the occurring conflicts are resolved 'on the fly', e.g. by randomly choosing a transition in conflict that should fire. Instead, **heuristic dispatching rules** (Haupt 1989), such as Shortest Processing Time (SPT) or Longest Processing

Time (LPT), can be introduced when solving the conflicting situations. By introducing different heuristic dispatching rules (priority rules) decisions can be made easily. In this way, only one path from the reachability graph is calculated, which means that the algorithm does not require a lot of computational effort. The schedule of process operations can be determined by observing the marking evolution of the net. Depending on the given scheduling problem a convenient rule should be chosen. Usually, different rules are needed to improve different predefined production objectives (makespan, throughput, production rates, and other temporal quantities).

A more extensive exploration of the reachability tree is possible by **PN-based heuristic search method** proposed by Lee and DiCesare (1994). It is based on generating parts of the Petri net reachability tree, where the branches are weighted by the time of the corresponding operations. Sum of the weights on the path from the initial to a terminal node gives a required processing time by the chosen transition firing sequence. Such a sequence corresponds to a schedule, and by evaluating a number of sequences a (sub)optimal schedule can be determined.

Recent reports in scheduling literature show an increased interest in the use of **meta-heuristics**, such as genetic algorithms (GA), simulated annealing (SA), and tabu search (TS). Meta-heuristics have also been combined with Petri net modelling framework to solve complex scheduling problems (Tuncel and Bayhan 2007). With such an approach, the modelling power of Petri nets can be employed, and relatively good solutions of scheduling problems can be found with a reasonable computational effort. Compared to reachability tree based search methods, meta-heuristics require less memory.

5. COLOURED PETRI NET SIMULATION BASED EXPLORATION OF THE SOLUTION SPACE

In our previous work (Löscher, Mušič and Breitenacker 2007, Mušič, Löscher and Breitenacker 2008) different ways of solution space exploration were studied. Extensive testing of the reachability tree search based approaches has been performed. The approach is very general, as it can be applied to any kind of scheduling problem that can be represented as a Petri net. Unfortunately, the approach does not perform very well on the standard job shop benchmarks (Mušič 2008). This motivated the exploration of alternative approaches, including local search based techniques.

In (Löscher, Mušič and Breitenacker 2007) the approach is presented, which extends the Petri net representation by sequences and priorities. Priorities are used as a way of parametrizing the conflict resolution strategy. For this purpose a priority ranking is assigned to transitions. If there is a conflict between a pair of transitions the transition with higher priority will fire.

Another way of parametrization is to select disjoint groups of transitions and map them to sequences. A firing list is defined by ordering transitions within the group. During the model evolution a set of sequence counters is maintained and all transitions belonging to sequences are disabled except of transitions corresponding to the current

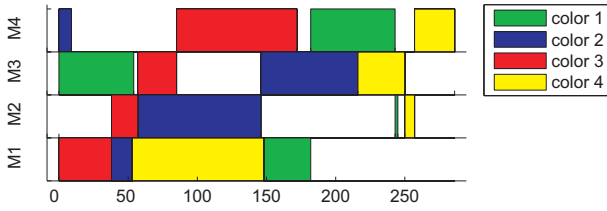


Figure 4: A possible solution of the given job-shop problem

state of the sequence counters. After firing such a transition the corresponding sequence counter is incremented.

This way the transition firing sequence can be parameterized. If the model represents a scheduling problem, the sequence obtained by a simulation run of the Petri net model from the prescribed initial to the prescribed final state is a possible solution to the problem, i.e. it represents a feasible schedule.

E.g., the model from Figure 3 can be simulated by applying SPT rule (Haupt 1989) as a default conflict resolution mechanism. The resulting sequence represents a possible schedule, shown in Figure 4.

The same schedule can be obtained by fixing the sequential order of transitions in conflicts related to shared resources in the system. E.g. in the above example the shared resources are machines M1 to M4. Related sets of transitions are:

$$\begin{aligned} S_{M1} &= \{t_{1,c3}, t_{1,c4}, t_{2,c1}, t_{2,c2}\} \\ S_{M2} &= \{t_{2,c3}, t_{3,c2}, t_{3,c4}, t_{4,c1}\} \\ S_{M3} &= \{t_{1,c1}, t_{2,c4}, t_{3,c3}, t_{4,c2}\} \\ S_{M4} &= \{t_{1,c2}, t_{3,c1}, t_{4,c3}, t_{4,c4}\} \end{aligned} \quad (3)$$

where $t_{i,cj}$ denotes cj occurrence colour of t_i and colour cj corresponds to job J_j .

If these sets are mapped to four independent sequences, and a set of index vectors

$$V = \{V_1, V_2, V_3, V_4\}$$

is adjoined, where V_i is a corresponding permutation of integer values $i, 1 \leq i \leq 4$:

$$\begin{aligned} V_1 &= \{1, 4, 2, 3\} \\ V_2 &= \{1, 2, 4, 3\} \\ V_3 &= \{1, 3, 4, 2\} \\ V_4 &= \{1, 3, 2, 4\} \end{aligned} \quad (4)$$

a supervised simulation run, which forces the prescribed sequential order of conflicting transitions, results in the same schedule as above.

The sequence supervised simulation is implemented by a simple modification of the regular CTPN simulation algorithm. After the enabled transitions are determined in each simulation step, the compliance of the set of enabled transitions to the state of the sequence counters is checked. Transitions that take part in defined sequences but are not pointed to by a counter are disabled.

The exploration of the solution space and the related search for the optimal schedule can then be driven by modifications of sequence index vectors. Such a modification

leads to a neighbourhood solution of a given solution and the related modification is usually defined through a neighbourhood function.

In the work of (Löscher, Mušič and Breiteneker 2007) several neighbourhood functions as well as different local search strategies were implemented in the PetriSimM toolbox for Matlab and some results are shown in (Löscher, Mušič and Breiteneker 2007, Mušič, Löscher and Breiteneker 2008).

5.1. Generation of feasible neighbourhood solutions from a CTPN model

The problem in the previously described approach is that by perturbing sequence index vectors the resulting transition firing sequence may easily become infeasible, which results in a deadlock during simulation. The search procedures implemented in PetriSimM were designed so that such an infeasible solution is ignored and a new perturbation is tried instead. While this works for many problems, in some cases the number of feasible sequences is rather low and such an algorithm can easily be trapped in an almost isolated point in the solution space.

The job shop scheduling approaches reported in the OR literature started to address the issue of efficient neighbourhood generation quite a while ago. With the wide acceptance of the Tabu search algorithm as the most promising methods for schedule optimisation the design of efficient neighborhood generation operator become the central issue and several such operators have been proposed (Blazewicz, Domschke and Pesch 1996, Jain, Ranganwamy and Meeran 2000, Watson, Whitley and Howe 2005).

The question is how to link these operators and related effective schedule optimization algorithms with Coloured Petri net representation of scheduling problems. As mentioned above the Petri net scheduling methods have advantages in unified representation of different aspect of underlying manufacturing process in a well defined framework. Unfortunately, the related optimization methods are not as effective as some methods developed in the OR field. The link of two research areas could be helpful in bridging the gap between highly effective algorithms developed for solving academic scheduling benchmarks and complex real-life examples where even the development of a formal model can be difficult (Gradišar and Mušič 2007).

A possible way of such a link is the establishment of a correspondence of a critical path and the sequence index vectors described previously.

In a given schedule the critical path CP is the path between the starting and finishing time composed of consequent operations with no time gaps:

$$CP = \{O_i : \rho_i = \rho_{i-1} + \tau_{i-1}, i = 2 \dots n\} \quad (5)$$

where O_i are operations composing the path, ρ_i is the release (starting) time of operation O_i , and τ_i is the duration of O_i .

The operations O_i on the path are critical operations. Critical operations do not have to belong to the same machine (resource) but they are linked by starting/ending times.

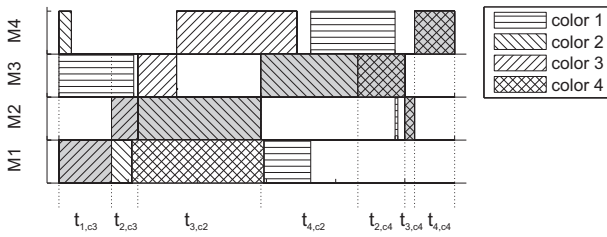


Figure 5: A critical path within a schedule and critical transitions

Critical path can be decomposed in a number of blocks. A block is the longest sequence of adjacent critical operations that occupy the same resource.

The length of the path equals the sum of durations of critical operations and defines the makespan C_{max} :

$$C_{max} = \sum_{O_i \in CP} \tau_i \quad (6)$$

Figure 5 shows a redrawn gantt chart from Figure 4 with indication of the critical path and the sequence of critical operations. The shown critical path consists of 5 blocks.

Critical operations in Figure 5 are denoted by transition labels that trigger the start of a critical operation when fired. A transition that triggers a critical operation will be called a critical transition.

The scheduling literature describes several neighborhoods based on manipulations (moves) of critical operations (Blazewicz, Domschke and Pesch 1996). One of the classical neighborhoods is obtained by moves that reverse the processing order of an adjacent pair of critical operations belonging to the same block (van Laarhoven, Aarts and Lenstra 1992). Other neighbourhoods further restrict the number of possible moves on the critical path, e.g. (Nowicki and Smutnicki 1996).

Clearly every critical transition participates in one of the conflicts related to shared resources, e.g. sets (3) for the given case. If these transitions are linked to predefined firing sequences parameterized by index vectors V_i (4), a move operator corresponds to a permutation of an index vector.

For example, in the schedule shown in Figure 5 a move can be chosen, which swaps the two operations in the third block on the critical path. This corresponds to the swap of transitions $t_{4,c2}$ and $t_{2,c4}$ in the sequence S_{M3} , which is implemented by the exchange of third and fourth element within V_3 index vector:

$$move(V_3) : \{1, 3, 4, 2\} \mapsto \{1, 3, 2, 4\}$$

A new schedule obtained by simulation with modified V_3 is shown in Figure 6.

When the move is limited to swap of a pair of the adjacent operations in a block on the critical path this narrows down the set of allowed permutations. The most important feature of such a narrowed set of permutation on the index vector is that every permutation from this set will result in a feasible firing sequence, i.e. a feasible schedule. Therefore no deadlock solutions can be generated, which

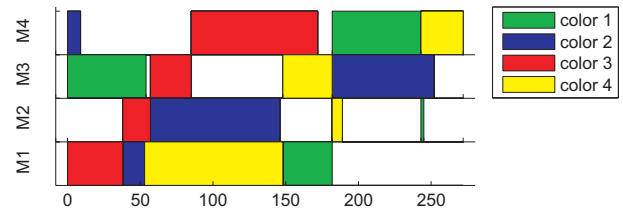


Figure 6: An optimized solution of the given job-shop problem

are often encountered when unrestricted permutations on the index vectors are used.

Based on this observation a set of neighbourhood functions can be defined which limit the permutations of the index vectors in a way that will produce feasible firing sequences only. Several widely used move operators can be implemented. Such a neighbourhood function permits the optimisation of schedules represented as PN or CTPN models by a wide variety of local search optimization techniques.

It is also important to note that such a neighbourhood function is comparable to exploring the reachability tree in an event driven manner. It is possible that certain feasible firing sequence imposes one or more intervals of idle time between transitions, i.e. some transitions are enabled but can not fire due to sequence restrictions. This is different from the exploration in a time driven manner when a transition has to be fired whenever at least one transition is enabled. The difference is important in cases when the optimal solution can be missed unless some idle time is included in the schedule as shown in (Piera and Mušič 2011).

The described neighbourhood generation procedure was coded in Matlab and used in combination with a simple Simulated annealing (SA) search algorithm. Comparison of the minimum makespan for the above job shop problem calculated by the proposed algorithm and some other standard algorithms is shown in Table 3. SA-SPT denotes the combined algorithm with the Simulated Annealing and the SPT rule (Mušič 2009), RT-search stands for a reachability tree based heuristic search (Lee and DiCesare 1994, Yu, Reyes, Cang and Lloyd 2003), and SA-CPN-N1 denotes Simulated Annealing and a CPN-based neighbourhood function as proposed in this paper.

Clearly, the reachability tree based search and SA-CPN-N1 outperform other algorithms with regard to the result. It must be noted, however, that the computational effort in the case of reachability tree based search is much higher.

Table 3: Calculated makespan for a simple job shop problem

Algorithm	Makespan
SPT	286
LPT	341
RT-search	272
SA-SPT	286
SA-CPN-N1	272

Table 4: Calculated makespan for a set of 15 jobs/15 machines problems

Algorithm	Makespan				
	ta01	ta02	ta03	ta04	ta05
SPT*	1462	1429	1452	1668	1618
LPT*	1701	1674	1655	1751	1828
RT-search	1592	1465	1637	1590	1568
SA-SPT	1359	1358	1352	1362	1352
SA-CPN-N1	1299	1326	1357	1353	1344
optimum	1231	1244	1218	1175	1224

* min out of 100 runs

Results in Table 3 were obtained by implementation of simple neighbourhood based on N1 move operator of (van Laarhoven et al. 1992) which was also used in (Taillard 1993) - the notation N1 is taken from (Blazewicz, Domschke and Pesch 1996). Only a single critical path was considered. Other neighbourhoods can be easily implemented and these as well as some other extensions of neighbourhood generation algorithm are currently being tested.

The computational complexity drawback of reachability tree based search is much more obvious with complex problems. Table 4 shows the preliminary results of a set of standard benchmark problems with 15 jobs and 15 machines (Taillard 1993). For reference also the optimal values are listed (source: <http://mistic.heig-vd.ch/taillard/>). The reachability tree based search has to be limited to predefined maximum tree size in order to complete in a reasonable time.

In contrast to that the SA-SPT and proposed SA-CPN-N1 algorithms are able to improve the initial SPT solutions with a moderate effort. A prototype implementation of tabu search algorithm (TS-CPN-N1) has also been tested and the obtained results are comparable to the SA based search. It is expected that the tests with other neighbourhood operators would further improve the obtained results, which is one of the tasks for the future work.

6. CONCLUSIONS

The presented results indicate that the proposed combination of CPN models, sequence based conflict resolution and local search performs relatively well with a moderate computational effort. The approach may be interesting for practice, in particular because of the ability to use various existing PN or CPN models of different problems. In general, any scheduling problem can be optimized that can be represented by a timed Petri net in such a way that relations among jobs and shared resources are fixed and a shared resource always participates in the given job's operation sequence, regardless of the determined schedule.

The Petri net scheduling methods have advantages in unified representation of different aspect of underlying manufacturing process in a well defined framework. The investigations show, however, that the related optimization methods are not as effective as some methods developed in the Operations Research field. The link of two research areas could be helpful in bridging the gap between highly effective algorithms developed for solving

academic scheduling benchmarks and complex real-life examples where even the development of a formal model can be difficult.

ACKNOWLEDGMENTS

The presented work has been partially performed within Competence Centre for Advanced Control Technologies, an operation co-financed by the European Union, European Regional Development Fund (ERDF) and Republic of Slovenia, Ministry of Higher Education, Science and Technology.

REFERENCES

- Basile, F., Carbone, C. and Chiacchio, P., 2007. Simulation and analysis of discrete-event control systems based on Petri nets using PNetLab, *Control Engineering Practice*, 15, 241–259.
- Blazewicz, J., Domschke, W. and Pesch, E., 1996. The job shop scheduling problem: Conventional and new solution techniques, *European Journal of Operational Research*, 93, 1–33.
- Bowden, F. D. J., 2000. A brief survey and synthesis of the roles of time in petri nets, *Mathematical & Computer Modelling*, 31, 55–68.
- Brucker, P., 2001. *Scheduling Algorithms*, Springer-Verlag Berlin Heidelberg.
- Dell'Amico, M. and Trubian, M., 1993. Applying tabu search to the job-shop scheduling problem, *Ann. Oper. Res.*, 41, 231–252.
- Gradišar, D. and Mušič, G., 2007. Production-process modelling based on production-management data: a Petri-net approach, *International Journal of Computer Integrated Manufacturing*, 20 (8), 794–810.
- Haupt, R., 1989. A survey of priority rule-based scheduling, *OR Spectrum*, 11 (1), 3–16.
- Jain, A., Rangaswamy, B. and Meeran, S., 2000. New and "stronger" job-shop neighborhoods: A focus on the method of nowicki and smutnicki(1996), *Journal of Heuristics*, 6 (4), 457–480.
- Jensen, K., 1997. *Coloured Petri Nets: Basic Concepts, Analysis Methods and Practical Use*, Vol. 1, 2 edn, Springer-Verlag, Berlin.
- Lakos, C. and Petrucci, L., 2007. Modular state space exploration for timed Petri nets, *International Journal on Software Tools for Technology Transfer*, 9, 393–411.
- Lee, D. Y. and DiCesare, F., 1994. Scheduling flexible manufacturing systems using Petri nets and heuristic search, *IEEE Transactions on robotics and automation*, 10 (2), 123–132.
- Löscher, T., Mušič, G. and Breiteneker, F., 2007. Optimisation of scheduling problems based on timed petri nets, *Proc. EUROSIM 2007*, Vol. II, Ljubljana, Slovenia.
- Mujica, M., Piera, M. A. and Narciso, M., 2010. Revisiting state space exploration of timed coloured petri net models to optimize manufacturing system's performance, *Simulation Modelling Practice and Theory*, 18, 1225–1241.

- Mušič, G., 2008. Timed Petri net simulation and related scheduling methods: a brief comparison, *The 20th European Modeling & Simulation Symposium*, Campora S. Giovanni (Amantea, CS), Italy, pp. 380–385.
- Mušič, G., 2009. Petri net base scheduling approach combining dispatching rules and local search, *21th European Modeling & Simulation Symposium*, Vol. 2, Puerto de La Cruz, Tenerife, Spain, pp. 27–32.
- Mušič, G., Löscher, T. and Breiteneker, F., 2008. Simulation based scheduling applying Petri nets with sequences and priorities, *UKSIM 10th International Conference on Computer Modelling and Simulation*, Cambridge, UK, pp. 455–460.
- Nowicki, E. and Smutnicki, C., 1996. A fast taboo search algorithm for the job shop problem, *Management Science*, 42 (6), 797–813.
- Piera, M. A. and Mušič, G., 2011. Coloured Petri net scheduling models: Timed state space exploration shortages, *Math. Comput. Simul.*, p. in press.
- Pinedo, M. L., 2008. *Scheduling: Theory, Algorithms, and Systems*, 3rd edn, Springer Publishing Company, Incorporated.
- Taillard, E., 1993. Benchmarks for basic scheduling problems, *European Journal of Operational Research*, 64, 278–285.
- Taillard, E. D., 1994. Parallel taboo search techniques for the job shop scheduling problem, *Inform Journal on Computing*, 6, 108–117.
- Tuncel, G. and Bayhan, G. M., 2007. Applications of Petri nets in production scheduling: a review, *International Journal of Advanced Manufacturing Technology*, 34, 762–773.
- Vaessens, R. J. M., Aarts, E. and Lenstra, J., 1996. Job shop scheduling by local search, *INFORMS Journal on Computing*, 8, 302–317.
- van Laarhoven, P., Aarts, E. and Lenstra, J., 1992. Job shop scheduling by simulated annealing, *Operations Research*, 40, 113–125.
- Watson, J. P., Whitley, L. D. and Howe, A. E., 2005. Linking search space structure, run-time dynamics, and problem difficulty: A step toward demystifying tabu search, *Journal of Artificial Intelligence Research*, 24, 221–261.
- Yu, H., Reyes, A., Cang, S. and Lloyd, S., 2003. Combined Petri net modelling and AI based heuristic hybrid search for flexible manufacturing systems-part II: Heuristic hybrid search, *Computers and Industrial Engineering*, 44 (4), 545–566.

AUTHOR BIOGRAPHY

GAŠPER MUŠIČ received B.Sc., M.Sc. and Ph.D. degrees in electrical engineering from the University of Ljubljana, Slovenia in 1992, 1995, and 1998, respectively. He is Associate Professor at the Faculty of Electrical Engineering, University of Ljubljana. His research interests are in discrete event and hybrid dynamical systems, supervisory control, planning, scheduling, and industrial informatics. His Web page can be found at <http://msc.fe.uni-lj.si/Staff.asp>.

PLANT CAPACITY ANALYSIS IN A DAIRY COMPANY, APPLYING MONTECARLO SIMULATION

Joselito Medina-Marin^(a), Gilberto Perez-Lechuga^(b), Juan Carlos Seck-Tuoh-Mora^(c), Norberto Hernandez-Romero^(d), Isaias Simon-Marmolejo^(e)

^(a,b,c,d)Advanced Research Centre in Industrial Engineering,
Autonomous University of Hidalgo State, Pachuca, Hidalgo, México
^(e)Superior School of Cd. Sahagún, Autonomous University of Hidalgo State,
Cd. Sahagún, Hidalgo, México

^(a)jmedina@uaeh.edu.mx, ^(b)glechuga2004@hotmail.com, ^(c)juanseck@gmail.com,
^(d)nhromero@uaeh.edu.mx, ^(e)isaias_simn@hotmail.com

ABSTRACT

In this paper, results of plant capacity analysis made to a dairy company are reported. This enterprise is one of the best-positioned companies in production and distribution of milk products in México. The enterprise has only one production plant and seven distribution centers. Because of the enterprise does not cover the client demand, the Planning Department planned to buy more equipment, in order to increase the production rate. This analysis was performed to determine the plant capacity, in order to know the quantity of additional equipment that production plant will need. An annual increasing rate was considered in the study, which was calculated with data of two years ago. In this work, the Monte Carlo method was applied to carry out the simulation of the production plant processes, and ProModel software was used to implement the simulation model.

Keywords: production plant, plant capacity analysis, Monte Carlo method, computer simulation

1. INTRODUCTION

Plant capacity means the maximum quantity that can be produced by time unit in the plant with the existing equipment. (Fare, Grosskopf, and Kokkelenberg 1989)

The knowledge about the production plant capacity is very important because it defines the competitive limits of the enterprise, i.e. the plant capacity sets the response rate, the costs structure, the composition of the personnel, and the general strategy for inventory. If the plant capacity is inadequate to satisfy the market demand, a company could lose its clients.

On the other hand, if the plant capacity is excessive, probably the company will reduce the prices of its products to stimulate the demand, underutilize personnel, keep overstocked, and produce other products, less profitable, in order to be in operation.

1.1. Company description

The plant capacity analysis reported in this paper was developed in a dairy company, which for confidentiality reasons the name of the company is omitted.

This company is one of the highest milk distributors in all the Mexican territory, it processes almost 3 million of milk liters every day, and more than 900 million per year. Furthermore, the company produces more than 100 milk products, which generates setup changes.

The company has a raw milk harvesting system, which is collected from ranches of associated ranchers, according to a morning schedule. The raw milk is transported in tankers from the ranches to the production plant, where the raw milk is pasteurized and ultra pasteurized.

The production plant has silos for the harvested milk, silos for the processed milk (pasteurized and ultra pasteurized), and equipment for bottling and packing.

Furthermore, according to client demand, ultra pasteurized milk is added with flavors, which needs silos for the bottling process, where the milk is mixed with the flavor required.

In addition to fluid milk products, the production plant processes products made from milk, such as yogurt, cream, butter, and cheeses.

1.2. Plant capacity concepts

The aim of this work is focused in the plant capacity analysis; hence, following paragraphs presents some plant capacity concepts. (Blackstone 1989)

1. Design Capacity (DC): Is the maximum possible production rate in a process, given the current designs of the product, mixing, operation policies, human resources, plant installations, and equipment.
2. Effective Capacity (EC): Is the maximum production rate that can be obtained in a reasonable way, taking into account preventive maintenance times, setup changes, and production system limitations.

3. Real Capacity (RC): Is the effective production rate achieved in the process. Normally, it is a time function and it changes constantly. RC is affected by the equipment wear, wastes and reworks, limited machinery assembly, employee's absenteeism, inadequate production master planning, and other similar factors that contribute to decrease the real capacity rates.

As a relationship among these concepts, it can be seen that $DC > EC > RC$.

Moreover, according to given concepts, some indicators can be obtained, such indicators are the utilization factor and the efficiency:

$$Utilization\ factor = \frac{Real\ Capacity}{Design\ Capacity} = \frac{RC}{DC} \quad (1)$$

$$Efficiency = \frac{Real\ Capacity}{Effective\ Capacity} = \frac{RC}{EC} \quad (2)$$

1.3. Monte Carlo method basis

Monte Carlo method is a generic form to call to mathematical procedures whose common feature is the use of random numbers and probability distributions, such as normal, exponential, uniform, beta, among others. It uses random variables defined in a finite dimensional space and the expectation value is calculated to find the approximated solution of a problem. (Kalos M.H., et. al., 2008)

Monte Carlo method is widely used, because it can be applied to solve stochastic problems, or those that can be set out in a stochastic way.

2. DESCRIPTION OF PRODUCTION PLANT PROCESSES

The production plant is divided in three main areas: raw milk reception area, fluid milk processing area, and milk derivative processing area.

2.1. Raw milk reception area

This area receives raw milk from dairy farms that belongs to shareholders of the company. Raw milk is transported in tankers with capacity of 25,000 liters. Before the raw milk is received, it is analyzed by the Quality Control Department in order to determine the quality of the product.

There are six reception lines, where each line pumps milk to 40,000 liters per hour, toward one of the four reception silos, with capacity of 150,000 liters each one. One of them stores milk used to produce milk derivatives. In the pumping process, the milk passes through deareators, filters, and coolers. The capacity of these equipments is 40,000 liters per hour.

Moreover, there is one reception line for cream, which pumps cream to 15,000 liters per hour. The cream is pumped into two tanks, with a capacity of 40,000 liters each one.

2.2. Fluid milk processing area

The processing area has two silos with a capacity of 100,000 liters each one, ten silos with a capacity of 150,000 liters each one and one silo with a capacity of 30,000 liters. The last one is only used to process cream.

Before the milk is pumped to process silos, it passes for a clarification process. There are four clarifiers, two of them have a capacity of 25,000 liters per hour, and the other two have a capacity of 30,000 liters per hour.

Fluid milk processing area is divided in two production lines: pasteurized milk line, and ultra pasteurized milk line.

2.2.1. Pasteurized milk lines

In order to produce pasteurized milk, these production lines take fluid milk from process silos; the milk is pumped to a homogenizer, and then, the milk is sent to two pasteurizers, each one with a tank with a capacity of 18,000 liters.

After that, pasteurized milk is bottled by three bottle filling machines with a capacity of 18,000 liters per hour. Sometimes, another machine with a capacity of 9,000 liters is used in this process; this machine is shared with other production lines.

2.2.2. Ultra pasteurized milk lines

The production of ultra pasteurized milk uses seven lines. Five lines are used to process milk in presentation of one liter, and the other two are used to process in presentation of 250 ml. The use of the production line depends on the type of product demanded. These lines produce whole milk, lacto free milk, light milk, cholesterol free milk, and flavored milk, among others.

Each line has between two or four bottle filling machines, with their respective capacities. Table 1 shows the capacities of ultra pasteurized milk lines.

Table 1: Ultra Pasteurized Lines Capacities.

Line	Cap (*)	Product presentation	Bottle filling machine	Cap (*)
I	24	1 liter	b ₁	12
			b ₂	12
II	24	1 liter	b ₃	6
			b ₄	6
			b ₅	6
			b ₆	6
III	24	1 liter	b ₇	6
			b ₈	6
			b ₉	6
			b ₁₀	6
IV	16	1 liter	b ₁₁	6
			b ₁₂	6
V	30	1 liter	b ₁₃	12
			b ₁₄	12
			b ₁₅	6

Table 1: Ultra Pasteurized Lines Capacities. (Cont.)

Line	Cap (*)	Product presentation	Bottle filling machine	Cap (*)
VI	7	250 ml	b ₁₆	1.5
			b ₁₇	1.5
			b ₁₈	1.875
			b ₁₉	1.5
VII	10	250 ml	b ₂₀	5
			b ₂₁	5

*: in thousands of liters per hour.

2.3. Milk derivative processing area

This area is divided in two subareas: cream area and yogurt area.

2.3.1. Cream area

Cream stored in the two cream reception tanks is pumped toward to three tanks used for a standardization process; these tanks have a capacity of 9,000, 9,000, and 14,000 liters per hour, respectively. Then, the cream is pumped to pasteurization and homogenization process.

There are three pasteurizers, with a capacity of 7,000, 4,000, and 7,000 liters per hour, respectively. Every pasteurizer is connected to its homogenizer, where homogenizer speed is synchronized with the pasteurizer speed.

After that, the cream is pumped into four tanks, with a capacity of 5,000 liters each one. The first tank is used to bottle cream manually in a presentation of four liters.

The second, third, and fourth tank, are alternatively used to bottle cream in three bottled lines. Table 2 shows the characteristics of these lines.

Table 2: Cream Bottled Lines

Line	Source tank	Capacity (*)	Presentation
c ₁	2, 3, 4	10	450 ml 900 ml
c ₂	2, 3, 4	10	450 ml
c ₃	2, 3, 4	10	200 ml

*: in thousands of liters per hour.

2.3.2. Yogurt area

Yogurt process is divided according to product presentation. There are four lines in this area: Plant I, Plant II, Plant III, and Plant III-A.

Raw milk stored in one of the four reception silos is used to process milk derivatives. Milk is pumped from this silo to three standardization tanks, then, milk is pumped to one of the four plants depending on the product presentation:

Plant I: Processes drinking yogurt and fruit yogurt.

Plant II: Processes yogurt with cereal, creamy yogurt, and whipped yogurt.

Plant III: Processes whipped yogurt, drinking yogurt and fruit yogurt.

Plant III-A: Processes a type of yogurt mixed with fruits. Moreover, this plant is shared with pasteurized milk line.

3. PROBABILITY DISTRIBUTIONS

In order to take into account the seasonality of the system, the data provide by the company were analyzed for every month. The company had a set of data from three years.

The modeling of real system started from the arrivals of tankers to reception area. Inter arrival times were analyzed, and the probability distributions are shown in table 3.

Table 3: Probability Distributions for Inter Arrival times of tankers.

Month	Probability distribution (in minutes)
January	E(18.4805)
February	E(18.46)
March	E(18.4804)
April	E(19.39)
May	E(18.7)
June	E(18.15)
July	E(18.49)
August	E(18.49)
September	E(18.4)
October	E(18.2)
November	E(18.52)
December	E(19.476)

The next variable to analyze was the milk contents of tankers, although the capacity of tankers is 25,000 liters, its real content is different. Table 4 shows the obtained results.

Table 4: Probability Distributions for Milk Contents of Tankers.

Month	Probability distribution (in thousands of liters)
January	Normal(22.7500, 3.0000)
February	Normal(22.2500, 2.6800)
March	Normal(22.2500, 2.5000)
April	Normal(22.5710, 2.5000)
May	Normal(22.1140, 2.5000)
June	Normal(22.6630, 2.7000)
July	Normal(22.3120, 2.3000)
August	Normal(22.9000, 2.2500)
September	Normal(22.7500, 2.8000)
October	Normal(22.3500, 2.7990)
November	Normal(22.4190, 2.9250)
December	Normal(22.7500, 2.4000)

Production Master Planner provided us data about weekly production for every product. We obtained the corresponding probability distribution for every product in every month of the year. Fitted distributions for some products in January month are shown in table 5.

Table 5: Probability distributions for product depending on client demand.

Product	Probability distribution (in thousands of liters)
Whole milk 1 liter	LogNormal(229537.93, 69776.62)
Light milk 1 liter	Weibull(10.119, 36666.36)
Strawberry flavored milk 1 liter	Normal(63342.21, 16008)
Drinking yogurt 250 ml strawberry-coconut	Triangular(25967.76, 40661.22, 221153)
Drinking yogurt 250 ml pineapple-coconut	LogNormal(130621.49, 47559.78)
Ultra pasteurized milk with fruits 250 ml strawberry	Triangular(0.65235, 0.65235, 35401.05)
Ultra pasteurized milk with fruits 250 ml mango	Triangular(0.00276, 0.00276, 24286.53)
Yogurt with cereal 150 ml strawberry-nut	Uniform(1775.45, 3791.46)
Yogurt with cereal 150 ml peach-nut	Triangular(0.02711, 0.02717, 11662.38278)
Cream 200 ml	Weibull(32.022, 1467150.23)
Cream 450 ml	Weibull(7.0071, 858032.40)
Yogurt 150 g strawberry	Weibull(9.970, 201340)
Yogurt 150 g peach	Normal(831686, 182164.92)

Figure 1 shows the graphic for the probabilistic distribution fitted for the data corresponding to whole milk in presentation of 1 liter.

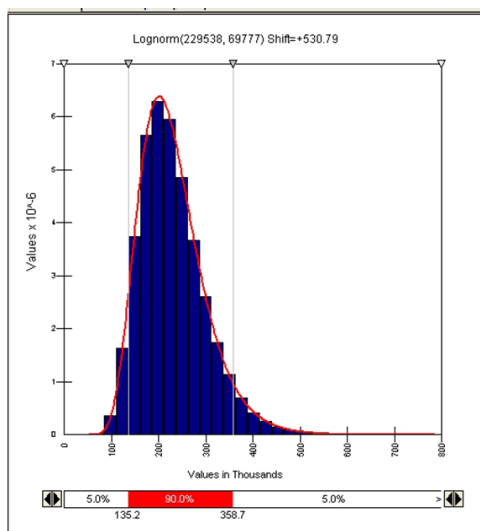


Figure 1: Probabilistic distribution fitted for whole milk of 1 liter.

With the statistical analysis, we found an annual growth rate of 6.16%.

4. COMPUTER SIMULATION

The simulation software used to carry out the simulation model was ProModel.

ProModel Simulation Software provides tools to model and simulate manufacturing process. It has a graphical interface, where the modeler can define entities, locations, processing, arrivals, resources, etc. (Harrel, Ghosh, and Bowden 2003; García, García, and Cárdenas 2006).

The items listed below must be identified from the real system:

Locations: are places used as servers, where entities are processed.

Entities: are dynamic objects that are served by locations.

Processing: model policies are defined in this part.

Resources: used to transport entities among locations, like forklifts, workers, etc.

Arrivals: used to define inter arrival time, it could be a constant value or a probabilistic distribution.

Attributes: used to add values to entities.

Variables: used to save data computed during simulation execution.

Subroutines: used to define procedures in order to improve the software functionality.

In the following subsections the simulation model is described.

4.1. Construction of the simulation model

In this phase of the project, all the elements of the real system that are involved in the product processing were identified, since raw milk arrivals until the bottling filling machines.

Because of fluid milk and cream flows are continuous variables, we considered one entity of milk or cream as 1000 liters of milk or cream, respectively.

4.1.1. Raw milk reception area

Arrival times for tanker were shown in table 3, and the tanker contents in table 4.

Tankers were defined as resources, and the reception silos, valves, and pipes were defined as locations. See figure 2.

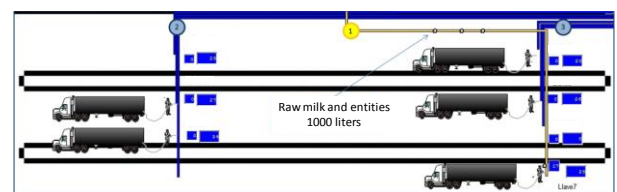


Figure 2: Tankers downloading raw milk.

Raw milk is pumped to reception silos, numbered as 29, 30, 31 and 40. The flow of the raw milk is shown in figure 3; it follows the description given in subsection 2.1.

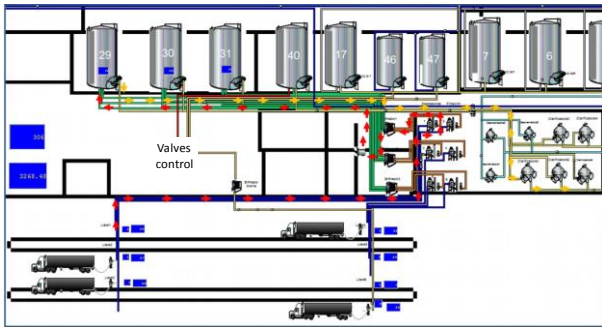


Figure 3: Raw milk pumped toward reception silos (29, 30, 31, and 40).

4.1.2. Fluid milk processing area

Silos, clarifiers, pumps, pipes, homogenizers, pasteurizers, ultra pasteurizers, and bottle filling machines were defined as locations, with their respective processing time.

Figure 4 shows a part of the pasteurized milk area.

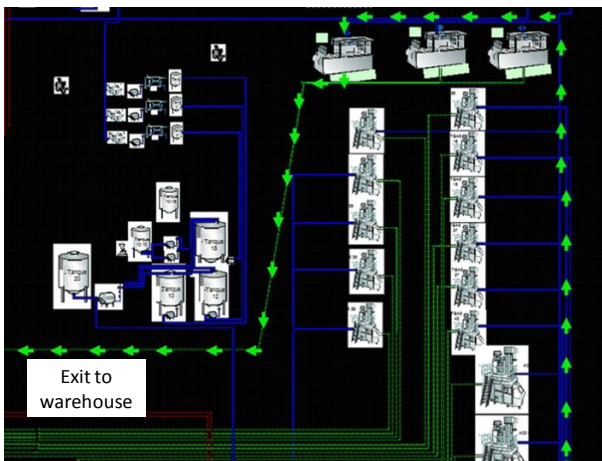


Figure 4: Pasteurized milk process.

4.1.3. Milk derivative processing area

Equipment installed in Cream area and Yogurt area were defined as locations, taking into account their respective processing time.

Figure 5 shows the cream homogenization process, and the four storage tanks.

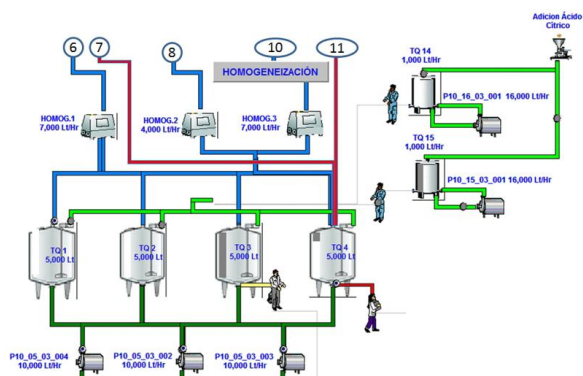


Figure 4: Cream homogenization process.

5. RESULTS AND CONCLUSIONS

The simulation scenarios were executed 100 times and the gathered statistics are summarized in the following tables.

Table 6 shows comparative data about the reception area. In order to get the Efficiency Capacity, time for clean the silos and maintenance time are considered. The table includes Design Capacity (*DC*), Effective Capacity (*EC*), Real Capacity (*RC*), Utilization factor (*UF*), and Efficiency (*Ef*). *DC*, *EC*, and *RC* are denoted in thousands of liters.

Table 6: Raw Milk Reception Statistics

	<i>DC</i>	<i>EC</i>	<i>RC</i>	<i>UF</i>	<i>Ef</i>
Valve 1	6,720	5,586	1,803	26.84%	32.28%
Valve 2	6,720	5,586	2,195	32.68%	39.30%
Valve 3	6,720	5,586	2,777	41.34%	49.72%
Valve 4	6,720	5,586	2,027	30.18%	36.30%
Valve 5	6,720	5,586	1,874	27.90%	33.56%
Valve 6	6,720	5,586	1,384	20.60%	24.78%
Silo 29	150	150	56.68	37.79%	37.79%
Silo 30	150	150	57.85	38.57%	38.57%
Silo 31	150	150	63.48	42.32%	42.32%
Silo 40	150	150	73.44	48.96%	48.96%

Data gathered from fluid milk pasteurization area are shown in table 7.

Table 7: Fluid Milk Processing Statistics

	<i>DC</i>	<i>EC</i>	<i>RC</i>	<i>UF</i>	<i>Ef</i>
Pasteurized milk	7,560	6,331	2,545	33.68%	40.21%
Ultra pasteurized milk	21,903	18,343	11,048	50.44%	60.23%

Finally, data obtained from milk derivative processing area are shown in table 8.

Table 8: Milk Derivative Processing Statistics

	<i>DC</i>	<i>EC</i>	<i>RC</i>	<i>UF</i>	<i>Ef</i>
Cream	1,860	1,626	973	57.95%	59.86%
Yogurt	6,762	6,191	2,481	36.70%	40.08%

Tables 6, 7, and 8 denote the behavior of the real system, and the capacity that the production plant

provides. Nevertheless, it can be seen that the installed equipment is not used at all.

The main obstacle to improve the usage of production lines is the harvesting system, because it is only performed in the morning. Veterinarians of the company declare that it is possible to set a new raw milk harvesting system, where the tankers go to dairy farms twice a day.

On the other hand, a good strategy in the elaboration of the Production Master Planning can reduce time wasted in setup changes, and, in consequence, to improve the efficiency of production lines.

REFERENCES

- Fare, R., Grosskopf, S., Kokkelenberg, E.C., 1989. Measuring plant capacity, utilization and technical change: a nonparametric approach. *International economic review*, 30 (3), 655–666.
- Blackstone, W.H.Jr., 1989. *Capacity Management*., Cincinnati, OH: South-Western
- Banks, J., Carson, J.S., Nelson, B.L., and Nicol, D.M., 2005. *Discrete-Event System Simulation*. USA: Prentice-Hall.
- Kalos, M.H., Whitlock P.A., 2008. *Monte Carlo Methods*. : Wiley-VCH.
- Harrel, C.R., Ghosh, B.K., Bowden, R.O., 2003. *Simulation using ProModel*. : McGraw-Hill Science/Engineering/Math.
- García, D.E., García, R.H., Cárdenas, B.L.E., 2006. *Análisis de sistemas con ProModel*. México:Prentice-Hall.
- Zeigler, B., Kim, T.G., and Praehofer, H., 2000. *Theory of Modeling and Simulation*. New York: Academic Press, New York.

AUTHORS BIOGRAPHY

Joselito Medina-Marin. He received the M.S. and Ph.D. degrees in electrical engineering from the Research and Advanced Studies Center of the National Polytechnic Institute at Mexico, in 2002 and 2005, respectively. Currently, he is a Professor of the Advanced Research in Industrial Engineering Center at the Autonomous University of Hidalgo State at Pachuca, Hidalgo, México. His current research interests include Petri net theory and its applications, active databases, simulation, and programming languages.

Gilberto Perez-Lechuga. He holds a Master Degree in Science, in Operations Research by the National Polytechnic Institute at the Mexico City. Later, he obtained the Engineering Doctor Degree in Operations Research (Stochastic Optimization) by the National Autonomous University of Mexico. His current research is directed to the modeling and optimization of complex

dynamic systems and their applications to stochastic manufacturing systems using emergent techniques.

Juan Carlos Seck-Tuoh-Mora. He received the M.S. and Ph.D. degrees in electrical engineering (option: Computing) from the Research and Advanced Studies Center of the National Polytechnic Institute at Mexico, in 1999 and 2002, respectively. Currently, he is a Professor of the Advanced Research in Industrial Engineering Center at the Autonomous University of Hidalgo State at Pachuca, Hidalgo, México. His current research interests include cellular automata theory and its applications, evolutionary computing and simulation.

Norberto Hernandez-Romero. He received the M.S. degree from Department of Electrical Engineering, Laguna Technological Institute at México, in 2001 and the Ph. D. from Autonomous University of Hidalgo State at México in 2009. Currently, he is a professor of the Advanced Research in Industrial Engineering Center at the Autonomous University of Hidalgo State at Pachuca, Hidalgo, México. His current research interests include system identification, feedback control design, genetic algorithms, fuzzy logic, neural network and its applications.

Isaias Simon-Marmolejo. He received the M.S. degree in Science in Industrial Engineering, graduated from the Autonomous University of Hidalgo State during the period 2007 to 2009. Currently, he works as a research professor in the School of Ciudad Sahagun Autonomous University of Hidalgo State in Tepeapulco, Hidalgo, Mexico and professor at the Technologic Institute of Pachuca in Pachuca of Soto, Hidalgo, Mexico. Years of experience and collaborative relationships and lines of research include Operations Research, Statistical Analysis of Discrete Event Simulation, Logistics and Systems Engineering.

GPGPU PROGRAMMING AND CELLULAR AUTOMATA: IMPLEMENTATION OF THE SCIARA LAVA FLOW SIMULATION CODE

Giuseppe Filippone^(a), William Spataro^(b), Giuseppe Spingola^(c), Donato D'Ambrosio^(d),
Rocco Rongo^(e), Giovanni Perna^(f), Salvatore Di Gregorio^(g)

^(a) ^(b) ^(c) ^(d) ^(f) ^(g) Department of Mathematics and HPCC, University of Calabria, Italy

^(e) Department of Earth Sciences and HPCC, University of Calabria, Italy

^(c) filippone@mat.unical.it, ^(a) spataro@unical.it, ^(b) g.spingola@gmail.com,

^(d) d.dambrosio@unical.it, ^(e) rongo@unical.it, ^(f) gioper86@gmail.com, ^(g) dig@unical.it,

ABSTRACT

This paper presents an efficient implementation of a well-known computational model for simulating lava flows on Graphical Processing Units (GPU) using the Compute Unified Device Architecture (CUDA) interface developed by NVIDIA. GPUs are specifically designated for efficiently processing graphic datasets. However, recently, they are also being exploited for achieving exceptional computational results even for applications not directly connected with the Computer Graphics field. We here show an implementation of the SCIARA Cellular Automata model for simulating lava flows on graphic processors using CUDA. Carried out experiments show that significant performance improvements are achieved, over a factor of 100, depending on the problem size, adopted device and type of performed memory optimization, confirming how graphics hardware can represent a valid solution for the implementation for Cellular Automata models.

Keywords: Cellular Automata, Lava flows simulation, GPGPU programming, CUDA.

1. INTRODUCTION

High Performance Computing (HPC) (Grama et al. 2003) adopts numerical simulations as an instrument for solving complex equation systems which rule the dynamics of complex systems as, for instance, a lava flow or a forest fire. In recent years, Parallel Computing has undergone a significant revolution with the introduction of GPGPU technology (General-Purpose computing on Graphics Processing Units), a technique that uses the graphics card processor (the GPU – Graphics Processing Unit) for purposes other than graphics. Currently, GPUs outperform CPUs on floating point performance and memory bandwidth, both by a factor of roughly 100. As a confirmation of the increasing trend in the power of GPUs, leading companies such as Intel have already integrated GPUs into their latest products to better exploit the capabilities of their devices, such as in some releases of the Core i5 and Core i7 processing units. Although the extreme processing power of graphic processors may be used for general purpose computations, a GPU may not be suitable for every computational problem: only a parallel program that results suitable and optimized for GPU architectures can fully take advantage of the

power of these devices. In fact, the performance of a GPGPU program that does not sufficiently exploit a GPU's capabilities can often be worse than that of a simple sequential one running on a CPU, such as when data transfer from main memory to video memory results crucial. Nevertheless, GPU applications to the important field of Computational Fluid Dynamics (CFD) are increasing both for quantity and quality among the Scientific Community (e.g., Tolke and Krafczyk 2008, Zuo and Chen 2010).

Among the different methodologies used for modelling geological processes, such as numerical analysis, high order difference approximations and finite differences, Cellular Automata (CA) (von Neumann 1966) has proven to be particularly suitable when the behaviour of the system to be modelled can be described in terms of local interactions. Originally introduced by von Neumann in the 1950s to study self-reproduction issues, CA are discrete computational models widely utilized for modeling and simulating complex systems. Well known examples are the Lattice Gas Automata and Lattice Boltzmann models (Succi 2004), which are particularly suitable for modelling fluid dynamics at a microscopic level of description. However, many complex phenomena (e.g. landslides or lava flows) are difficult to be modeled at such scale, as they generally evolve on large areas, thus needing a macroscopic level of description. Moreover, since they may also be difficult to be modelled through standard approaches, such as differential equations Macroscopic Cellular Automata (MCA) (Di Gregorio and Serra 1999) can represent a valid alternative. Several successful attempts have been carried out regarding solutions for parallelizing MCA simulation models (e.g., D'Ambrosio and Spataro 2007). In this research context, the CAMELot virtual laboratory and the libAuToti scientific library represent valid solutions for implementing and automatically parallelizing MCA models on distributed memory machines while, for shared memory architectures, some effective OpenMP parallelizations have been implemented for CA-like models, such as for fire spread simulations, Lattice Boltzmann models or lava flow modeling (Oliverio et al. 2011). However, few examples of GPGPU applications for CA-like models do exist (Tolke 2008) and to our knowledge, none regarding the MCA approach. This paper presents a implementation of a

well-known, reliable and efficient MCA model widely adopted for lava flow risk assessment, namely the SCIARA model (Rongo et al. 2008), in GPGPU environments. Tests performed on two types of GPU hardware, a Geforce GT 330M graphic card and a Tesla C1060 computing processor, have shown the validity of this kind of approach.

In the following sections, after a brief description of the basic version of the SCIARA MCA model for lava flows, a quick overview of GPGPU paradigm together with the CUDA framework is presented. Subsequently, the specific model implementation and performance analysis referred to benchmark simulations of a real event and different CA spaces are reported, while conclusions and possible outlooks are shown at the end of the paper.

2. CELLULAR AUTOMATA AND THE SCIARA MODEL FOR LAVA FLOW SIMULATION

As previously stated, CA are dynamical systems, discrete in space and time. They can be thought as a regular n -dimensional lattice of sites or, equivalently, as an n -dimensional space (called cellular space) partitioned in cells of uniform size (e.g. square or hexagonal for $n=2$), each one embedding an identical finite automaton. The cell state changes by means of the finite automaton transition function, which defines local rules of evolution for the system, and is applied to each cell of the CA space at discrete time steps. The states of neighbouring cells (which usually includes the central cell) constitute the cell input. The CA initial configuration is defined by the finite automata states at time $t=0$. The global behaviour of the system emerges, step by step, as a consequence of the simultaneous application of the transition function to each cell of the cellular space.

When dealing with the modelling of spatial extended dynamical systems, MCA can represent a valid choice especially if their dynamics can be described in terms of local interaction at macroscopic level. Well known examples of successful applications of MCA include the simulation of lava (Crisci et al. 2004) and debris flows (Di Gregorio et al. 1999), forest fires (Trunfio 2004), agent based social processes (Di Gregorio et al. 2001) and highway traffic (Di Gregorio et al. 2008), besides many others.

By extending the classic definition of Homogeneous CA, MCA facilitate the definition of several aspects considered relevant for the correct simulation of the complex systems to be modelled. In particular, MCA provide the possibility to “decompose” the CA cell state in “substates” and to allow the definition of “global parameters”. Moreover, the dynamics of MCA models (especially those developed for the simulation of complex macroscopic physical systems such as debris or lava flows) is often “guided” by the “Minimisation Algorithm of the Differences” (cf. Di Gregorio and Serra 1999), which translates in algorithmic terms the general principle for which natural systems leads towards a situation of equilibrium.

Refer to Di Gregorio and Serra (1999) for a complete description of the algorithm, besides theorems and applications.

2.1. The MCA lava flow model SCIARA

SCIARA is a family of bi-dimensional MCA lava flow models, successfully applied to the simulation of many real cases, such as the 2001 Mt. Etna (Italy) Nicolosi lava flow (Crisci et al. 2004), the 1991 Valle del Bove (Italy) lava event (Barca et al. 1994) which occurred on the same volcano and employed for risk mitigation (D’Ambrosio et al. 2006). In this work, the basic version of SCIARA (Barca et al. 1993) was considered and its application to the 2001 Nicolosi event and to benchmark grids shown.

SCIARA considers the surface over which the phenomenon evolves as subdivided in square cells of uniform size. Each cell changes its state by means of the transition function, which takes as input the state of the cells belonging to the von Neumann neighbourhood. It is formally defined as

$$SCIARA = \langle R, X, Q, P, \sigma \rangle$$

where:

- R is the set of points, with integer coordinates, which defines the 2-dimensional cellular space over which the phenomenon evolves. The generic cell in R is individuated by means of a couple of integer coordinates (i, j) , where $0 \leq i < i_{max}$ and $0 \leq j < j_{max}$.
- $X = \{(0,0), (0, -1), (1, 0), (-1, 0), (0, 1)\}$ is the so called von Neumann neighbourhood relation, a geometrical pattern which identifies the cells influencing the state transition of the central cell.
- Q is the set of cell states; it is subdivided in the following substates:
 - Q_z is the set of values representing the topographic altitude (m);
 - Q_h is the set of values representing the lava thickness (m);
 - Q_T is the set of values representing the lava temperature (K°);
 - Q_o^5 are the sets of values representing the lava outflows from the central cell to the neighbouring ones (m).

The Cartesian product of the substates defines the overall set of state Q :

$$Q = Q_z \times Q_h \times Q_T \times Q_o^5$$

- P is set of global parameters ruling the CA dynamics:
 - $P_T = \{T_{vent}, T_{sol}, T_{int}\}$, the subset of parameters ruling lava viscosity, which specify the temperature of lava at the vents, at solidification and the “intermediate” temperature (needed for computing lava adherence), respectively;

- $\tilde{Pa} = \{a_{vent}, a_{sol}, a_{int}\}$, the subset of parameters which specify the values of adherence of lava at the vents, at solidification and at the “intermediate” temperature, respectively;
 - p_c , the cooling parameter, ruling the temperature drop due to irradiation;
 - p_r , the relaxation rate parameter, which affects the size of outflows.
- $\sigma : Q^5 \rightarrow Q$ is the deterministic cell transition function. It is composed by four “elementary processes”, briefly described in the following:
 - *Outflows computation* (σ_1). It determines the outflows from the central cell to the neighbouring ones by applying the minimisation algorithm of the differences; note that the amount of lava which cannot leave the cell, due to the effect of viscosity, is previously computed in terms of adherence. Parameters involved in this elementary process are: P_T and P_a .
 - *Lava thickness computation* (σ_2). It determines the value of lava thickness by considering the mass exchange among the cells. No parameters are involved in this elementary process.
 - *Temperature computation* (σ_3). It determines the lava temperature by considering the temperatures of incoming flows and the effect of thermal energy loss due to surface irradiation. The only parameter involved in this elementary process is p_c .
 - *Solidification* (σ_4). It determines the lava solidification when temperature drops below a given threshold, defined by the parameter T_{sol} .

3. GPU AND GPGPU PROGRAMMING

As alternative to standard parallel architecture, the term GPGPU (General-Purpose computing on Graphics Processing Units) refers to the use of the card processor (the GPU) as a parallel device for purposes other than graphic elaboration. In recent years, mainly due to the stimulus given by the increasingly demanding performance of gaming and graphics applications in general, graphic cards have undergone a huge technological evolution, giving rise to highly parallel devices, characterized by a multithreaded and multicore architecture and with very fast and large memories. A GPU can be seen as a computing device that is capable of executing an elevated number of independent threads in parallel. In general, a GPU consists in a number (e.g., 16) of SIMD (Single Instruction, Multiple Data) multiprocessors with a limited number of floating-point processors that access a common shared-memory within the multiprocessor. To better understand the enormous potential of GPUs, some comparisons with the CPU are noticeable: a medium-performance GPU (e.g. the NVIDIA Geforce GT200 family) is able to perform nearly 1000 GFLOPS (Giga Floating Point Operations

per Second), while an Intel Core i7 has barely 52 GFLOPS. In addition, the most interesting aspect still is the elevated parallelism that a GPU permits. For instance, the NVIDIA GeForce 8800 GTX has 16 multiprocessors each with 8 processors for a total of 128 basic cores, while a standard multi-core CPU has few, though highly-functional, cores. Another motivation of GPUs increasing utilization as parallel architecture regards costs. Until a few years ago, in order to have the corresponding computing power of a medium range GPU of today (which costs approximately a few hundred Euros), it was necessary to spend tens of thousands of Euros. Thus, GPGPU has not only led to a drastic reduction of computation time, but also to significant cost savings. Summarizing, it is not misleading to affirm that the computational power of GPUs has exceeded that of PC-based CPUs by more than one order of magnitude while being available for a comparable price. In the last years, NVIDIA has launched a new product line called Tesla, which is specifically designed for High Performance Computing.

Supported on Windows and Linux Operating systems, NVIDIA CUDA technology (NVIDIA CUDA 2011a) permits software development of applications by adopting the standard C language, libraries and drivers. In CUDA, threads can access different memory locations during execution. Each thread has its own private memory, each block has a (limited) shared memory that is visible to all threads in the same block and finally all threads have access to global memory. The CUDA programming model provides three key abstractions: the hierarchy with which the threads are organized, the memory organization and the functions that are executed in parallel, called kernels. These abstractions allow the programmer to partition the problem into many sub-problems that can be handled and resolved individually.

3.1. CUDA Threads and Kernels

A GPU can be seen as a computing device that is capable of executing an elevated number of independent threads in parallel. In addition, it can be thought as an additional coprocessor of the main CPU (called in the CUDA context Host). In a typical GPU application, data-parallel like portions of the main application are carried out on the device by calling a function (called kernel) that is executed by many threads. Host and device have their own separate DRAM memories, and data is usually copied from one DRAM to the other by means of optimized API calls.

CUDA threads can cooperate together by sharing a common fast shared-memory (usually 16KB), eventually synchronizing in some points of the kernel, within a so-called thread-block, where each thread is identified by its thread ID. In order to better exploit the GPU, a thread block usually contains from 64 up to 512 threads, defined as three-dimensional array of type `dim3` (containing three integers defining each dimension). A thread can be referred within a block by means of the built-in global variable `threadIdx`.

While the number of threads within a block is limited, it is possible to launch kernels with a larger total number of threads by batching together blocks of threads, by means of a grid of blocks, usually defined as a two-dimensional array, also of type `dim3` (with the third component set to 1). In this case, however, thread cooperation is reduced since threads that belong to different blocks do not share the same memory and thus cannot synchronize and communicate with each other. As for threads, a built-in global variable, `blockIdx`, can be used for accessing the block index within the grid. Currently, the maximum number of blocks is 65535 in each dimension. Threads in a block are synchronized by calling the `syncthreads()` function: once all threads have reached this point, execution resumes normally. As previously reported, one of the fundamental concepts in CUDA is the *kernel*. This is nothing but a C function, which once invoked is performed in parallel by all threads that the programmer has defined. To define a kernel, the programmer uses the `__global__` qualifier before the definition of the function. This function can be executed only by the device and can be only called by the host. To define the dimension of the grid and blocks on which the kernel will be launched on, the user must specify an expression of the form `<<< Grid_Size, Block_Size >>>`, placed between the kernel name and argument list.

What follows is a classic pattern of a CUDA application:

- Allocation and initialization of data structures in RAM memory;
- Allocation of data structures in the device and transfer of data from RAM to the memory of the device;
- Definition of the block and thread grids;
- Performing one or more kernel;
- Transferring of data from the device memory to Host memory.

Eventually, it must be pointed out that a typical CUDA application has parts that are normally performed in a serial fashion, and other parts that are performed in parallel.

3.2. Memory hierarchy

In CUDA, threads can access different memory locations during execution. Each thread has its own *private memory*, each block has a (limited) *shared memory* that is visible to all threads in the same block, and finally all threads have access to *global memory*. In addition to these memory types, two other read-only, fast on-chip memory types can be defined: *texture memory* and *constant memory*.

As expected, memory usage is crucial for the performance. For example, the shared memory is much faster than the global memory and the use of one rather than the other can dramatically increase or decrease performance. By adopting variable type qualifiers, the programmer can define variables that reside in the

global memory space of the device (with `__device__`) or variables that reside in the shared memory space (with `__shared__`) that are thus accessible only from threads within a block. Typical latency for accessing global memory variables is 200-300 clock cycles, compared with only 2-3 clock cycles for shared memory locations. For this reason, to improve performances variable accesses should be carried out in the shared memory rather than global memory, wherever possible. However, each variable or data structure allocated in shared memory must first be initialized in the global memory, and afterwards transferred in the shared one (NVIDIA CUDA 2011b). This means that to copy data in the shared memory, global memory access must be first performed. So, the more his type of data is accessed, the more convenient is to use this type of memory: so, for few accesses it is evident that shared memory is not convenient to use. As a consequence, a preliminary analysis of data access of the considered algorithm should be performed in order to evaluate the tradeoff, and thus, convenience of using shared memory.

4. IMPLEMENTATION OF THE SCIARA MODEL AND EXPERIMENT RESULTS

As previously stated, Cellular Automata models, such as SCIARA, can be straightforwardly implemented on parallel computers due to their underlying parallel nature. In fact, since Cellular Automata methods require only next neighbor interaction, they are very suitable and can be efficiently implemented even on GPUs. In literature, to our knowledge, no examples of Macroscopic Cellular Automata modeling with GPUs are found, while some interesting CA-like implementations, such as Lattice Boltzmann kernels, are more frequent (e.g., Tolke 2008; Kuznik et al. 2010).

In this work, two different implementations of the SCIARA lava flow computational model were carried out, a first straightforward version which uses only global memory for the entire CA space partitioning and a second, but more performing one, which adopts (also) shared memory for CA space substate allocation. What follows is an excerpt of the core of the general CUDA algorithm (cf. Section 2.1):

```
// CA loop
for(int step=0; step< Nstep; step++) {

    // add lava at craters
    crater <<<1, num_craters>>>(Aread,Awrite);
    // s1
    calc_flows<<<dimGrid,dimBlock>>>(Aread,Awrite);
    // s2
    calc_width <<<dimGrid, dimBlock>>>(Aread,Awrite);
    // s3
    calc_temp<<<dimGrid, dimBlock>>>(Aread,Awrite);
    // s4
    calc_quote <<<dimGrid, dimBlock>>>(Aread,Awrite);

    // swap matrixes
    copy<<<dimGrid,dimBlock>>>(Awrite,N,Aread,Substat_N);
}
cudaMemcpy(A, Aread, size, cudaMemcpyDeviceToHost);
// copy data to Host
}
```

In the time loop four basic kernels, `calc_flows`, `calc_width`, `calc_temp` and `calc_quote` are launched corresponding to the four elementary processes of SCIARA, σ_1 , σ_2 , σ_3 and σ_4 , respectively, as described in Spingola et al. (2008). The `crater()` kernel refers to the crater cell(s), which is obviously invoked on a smaller grid than the previous ones. The model was implemented by adopting a system of double matrixes for the CA space representation, one (`Aread`) for reading cell neighbor substates and a second (`Awrite`) for writing the new substate value. This choice has proven to be efficient, since it allows to separate the substates reading phase from the update phase, after the application of the transition function, thus ensuring data integrity and consistency in a given step of the simulation. After applying the transition function to all the cell space, the main matrix must be updated, replacing values with the corresponding support matrix ones (swap matrixes phase). In this implementation, a CA step is simulated by more logical substeps where, after crater cells are updated (by means of the `crater`), lava outflows are calculated according to the σ_1 elementary process. When all outflows are computed, and therefore all outflow substates are consistent, the actual distribution takes place, producing the new value of the quantity of lava in each cell of the CA. Subsequently, each cell reads from a neighbour cell the associated outflow substate corresponding to the quantity of inflowing lava (σ_2 elementary process). In this phase, the σ_3 and σ_4 elementary processes are applied to the new quantity of lava of the cell. At the end of the CA loop, data is copied back to the Host memory by the `cudaMemcpy` function.

Regarding the specific implementation, the first thing to decide on is what thread mapping should be adopted to better exploit the fine-grain parallelism of the CUDA architecture. For example, one might consider using a thread for each row or each column, as occurs in a typical data-parallel implementation (e.g., Oliverio et al. 2011). However, when working in CUDA with arrays, the most widely adopted technique is to match each cell of the array with a thread (e.g., Tolke 2008). The number of threads per block should be a multiple of 32 threads, because this provides optimal computing efficiency (NVIDIA CUDA 2011b) and thus we have chosen to build blocks of size 32×16 , corresponding to the maximum value (512) of number of threads permitted for each block. What follows is an excerpt for defining the grid of blocks that was considered for SCIARA:

```
#define BLOCK_SIZE_X 32
#define BLOCK_SIZE_Y 16
...
int dimX; // CA x dimension
int dimY; // CA y dimension
dim3 dimBlock(BLOCK_SIZE_X, BLOCK_SIZE_Y);

int n_blocks_x = dimX/dimBlock.x;
int n_blocks_y = dimY/dimBlock.y;

dim3 dimGrid(n_blocks_x, n_blocks_y);
```

```
...
// invoke kernel functions
kernel<<<dimGrid, dimBlock>>>(...);
...
```

Once that the grid of blocks (and threads) were defined in this simple manner, kernels are managed so that each cell (i, j) of SCIARA is associated to each thread (i, j). This is simply done, for each invoked kernel (i.e., `calc_flows`, `calc_width`, `calc_temp` and `calc_quote`), by associating each row and column of the CA with the corresponding thread as in this simple scheme:

```
__global__ void kernel(...) {
    int col = blockIdx.x * blockDim.x +
threadIdx.x;
    int row = blockIdx.y * blockDim.y +
threadIdx.y;

    // memory allocation (shared, global, etc)
    ...
    /** transition function for cell[row][col] **
    ...
}
```

5. TESTS AND PERFORMANCE RESULTS

Two CUDA graphic devices were adopted for experiments: a NVIDIA high-end Tesla C1060 and a Geforce GT 330M graphic card. In particular, the Tesla computing processor has 240 processor cores, 4 GB global memory and high-bandwidth communication between CPU and GPU, whereas the less performing graphic card has 48 cores and 512 MB global memory. The sequential SCIARA reference version was implemented on a 2.66 GHz Intel Core i7 based desktop computer. The sequential CPU version is identical to the version that was developed for the GPUs, that is, no optimizations were adopted in the former version. In practice, at every step, the CA space array is scrolled and the transition function applied to each cell of the CA where lava is present.

Many tests have been performed regarding both performance and verification of the accuracy of the results. Regarding performance tests, the best implementation has regarded a version which adopts a hybrid (shared/global) memory allocation.

As known, access to a location in shared memory of each multiprocessor has a much lower latency than that carried out on the global device memory. On the other hand, an access to a shared-memory location necessary needs a first access to global memory (cf. Section 3.2). For this reason, an accurate analysis was carried out in determining how much memory access each thread does for each CA substate matrix. This investigation gave rise to a “hybrid” memory access pattern, where shared memory allocation was adopted for those kernels accessing CA matrixes more than two times. For illustrative purposes, Figure 1 shows how

shared memory is used in the context of our GPU implementation.

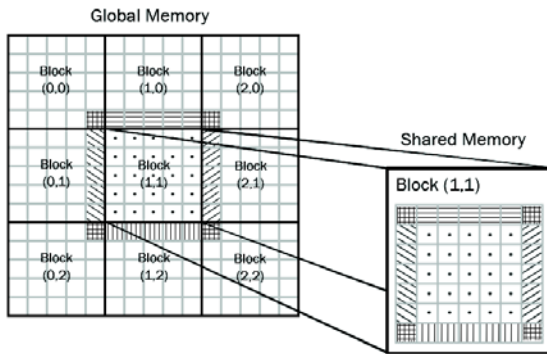


Figure 1: Memory mapping of the CA space allocated in global memory with a portion of shared memory. Shaded areas represent portions of neighbouring block areas which need to be swapped at each CA step to ensure data consistency.

A first test regarded the simulation of well-known and documented real lava flow event, the Mt. Etna Nicolosi event (Crisci et al. 2004) occurred in July, 2001. Table 1 (first row) reports the first results of tests carried out for this experiment, where the CA space is a 819×382 two-dimensional grid. The simulation was carried out for 15000 steps, considering one crater for lava flow emission. In order to further stress the efficiency of the GPU version, further benchmarks experiments were performed by considering four different hypothetical CA spaces, namely 512^2 , 1024^2 , 2048^2 and 4096^2 grids, with cells representing inclined planes, with many craters located over the grid (cf. Table 1 - from second row).

Table 1: Execution times of experiments (in seconds) carried out for evaluating the performance the GPU version of the SCIARA MCA lava-flow model on the considered hardware. The 819×382 matrix refers to the 2001 Mt. Etna event. Other grid dimensions refer to inclined planes. N/A (Not Available) data are due to device lack of memory capacity.

Performance results (in seconds)			
CA dim / Device	Intel i7 (sequential)	Geforce	Tesla
819×382	741	46	11.8
512^2	677	31.4	5.6
1024^2	2716	99.1	21.4
2048^2	11480	344.5	81.1
4096^2	47410	N/A	307

Timings reported for the considered GPU devices indicate their full suitability for parallelizing CA models. Even if applied to a simple MCA model, performance results show the incredible computational power of the considered GPUs in terms of execution

time reduction, significantly outperforming the CPU implementation up to $150\times$ for large grid sizes. Other tests were also performed on a completely global memory version. In this case results, here omitted for brevity, have shown how the use of shared memory can improve performances up to 50%, with respect to the total global memory version.

Eventually, to test if single-precision data can be considered sufficient for SCIARA simulations, tests were carried out on the 2001 lava flow event (15000 CA steps) and compared results produced by the GPU version with those produced by the CPU (sequential) version with single precision data (i.e., float type variables), and those produced still by the same GPU version against a double precision CPU implementation (i.e., double type variables). In each case, comparison results were satisfactory, since the areal extensions of simulations resulted the same, except for few errors of approximation in a limited number of cells. In particular, comparing the GPU version with the CPU single-precision version approximation differences at the third significant digit were only for 4% of cells, while differences were less for remaining cells. Differences were even minor compared to the previous case by considering the single precision GPU version and a CPU version which adopts double-precision variables.

6. CONCLUSIONS

This paper reports the implementation of a Macroscopic Cellular Automata model using GPU architectures. As shown, the CUDA technology, in combination with the an efficient memory management, can produce a very efficient version of the SCIARA lava flow simulator. Although results are indeed already satisfactory, future developments can regard further improvements for both increasing performances and implementing more advanced MCA models.

The results obtained in this work are to be considered positive and extremely encouraging. As confirmed by the increasing number of applications in the field of scientific computing in general, GPGPU programming represents a valid alternative to traditional microprocessors in high-performance computer systems of the future.

ACKNOWLEDGMENTS

This work was partially funded by the European Com-mission - European Social Fund (ESF) and by the Regione Calabria.

REFERENCES

- Barca, D., Crisci, G.M., Di Gregorio, S., Nicoletta, F., 1993. Cellular automata methods for modelling lava flow: simulation of the 1986-1987 eruption, Mount Etna, Sicily. In: Kilburn, C.R.J., Luongo, G. (Eds.), *Active lavas: monitoring and modelling*. UCL Press, London, 12, 291-309.
- Barca, D., G.M. Crisci, Di Gregorio, S., Nicoletta, F. 1994. Cellular Automata for simulating lava

- Flows: A method and examples of the Etnean eruptions. *Transport Theory and Statistical Physics*, 23, 195-232.
- Crisci, G.M., Di Gregorio, S., Rongo, R., Spataro, W., 2004. The simulation model SCIARA: the 1991 and 2001 at Mount Etna. *Journal of Vulcanology and Geothermal Research*, 132, 253-267.
- D'Ambrosio, D., Rongo, R., Spataro, W., Avolio, M.V., Lupiano, V., 2006. Lava Invasion Susceptibility Hazard Mapping Through Cellular Automata. In: S. El Yacoubi, B. Chopard, and S. Bandini (Eds.), *ACRI 2006, Lecture Notes in Computer Science*, 4173, Springer-Verlag, Berlin Heidelberg, 452-461.
- D'Ambrosio, D., Spataro, W., 2007. Parallel evolutionary modelling of geological processes. *Parallel Computing* 33 (3), 186-212.
- Di Gregorio, S., Mele, F., Minei, G., 2001. Automi Cellulari Cognitivi. Simulazione di evacuazione, Proceedings "Input 2001", Seconda Conferenza Nazionale Informatica Pianificazione Urbana e Territoriale, (in Italian) Democrazia e Tecnologia.
- Di Gregorio, S., Rongo, R., Siciliano, C., Sorriso-Valvo, M., Spataro, W., 1999. Mount Ontake landslide simulation by the cellular automata model SCIDDICA-3. *Physics and Chemistry of the Earth, Part A*, 24, 97-100.
- Di Gregorio, S., Serra, R., 1999. An empirical method for modelling and simulating some complex macroscopic phenomena by cellular automata. *Fut. Gener. Comp. Syst.*, 16, 259-271.
- Di Gregorio, S., Umeton, R., Biccocchi, A., Evangelisti, A., Gonzalez, M., 2008. Highway Traffic Model Based on Cellular Automata: Preliminary Simulation Results with Congestion Pricing Considerations. *Proceedings of 20th European Modeling & Simulation Symposium (EMSS)*, pp. 665-674. September 17-19, Campora S.G., CS, Italy.
- Grama, A., Karypis, G., Kumar, V., Gupta, A., 2003. *An Introduction to Parallel Computing: Design and Analysis of Algorithms*, Second Edition. USA: Addison Wesley.
- Kuznik, F., Obrecht, C., Rusaouen, G., Roux, J.J., 2010. LBM based flow simulation using GPU computing processor. *Computers and Mathematics with Applications*, 59, 2380-2392.
- NVIDIA CUDA C Programming Guide, 2011a. Available from: <http://developer.download.nvidia.com/compute/cuda> [accessed 27 June 2011]
- NVIDIA CUDA C Best Practices Guide, 2011b. Available from: <http://developer.download.nvidia.com/compute/cuda> [accessed 27 June 2011]
- Oliverio, M., Spataro, W., D'Ambrosio, D., Rongo, R., Spingola, G., Trunfio, G.A., 2011. OpenMP parallelization of the SCIARA Cellular Automata lava flow model: performance analysis on shared-memory computers. *Proceedings of International Conference on Computational Science, ICCS 2011*, *Procedia Computer Science*, 4, pp. 271-280.
- Rongo, R., Spataro, W., D'Ambrosio, D., Avolio, M.V., Trunfio, G.A., Di Gregorio, S., 2008. Lava flow hazard evaluation through cellular automata and genetic algorithms: an application to Mt Etna volcano. *Fundamenta Informaticae*, 87, 247-268.
- Spingola, G., D'Ambrosio, D., Spataro, W., Rongo, R., Zito, G., 2008. Modeling Complex Natural Phenomena with the libAuToti Cellular Automata Library: An example of application to Lava Flows Simulation. *Proceedings of International Conference on Parallel and Distributed Processing Techniques and Applications*, pp. 44-50. July 14-17, 2008, Las Vegas, Nevada, USA
- Succi, S., 2004. *The Lattice Boltzmann Equation for Fluid Dynamics and Beyond*, UK: Oxford University Press.
- Tolke, J., Krafczyk, M., 2008. TeraFLOP computing on a desktop PC with GPUs for 3D CFD. *Int. Journ. of Comput. Fluid Dynamics*, 22 (7), 443-456.
- Tolke, J., 2008. Implementation of a lattice Boltzmann kernel using the compute unified device architecture developed by NVIDIA. *Comput. Vis. Sci.*, 13 1, 29-39.
- Trunfio, G.A., 2004. Predicting Wildfire Spreading Through a Hexagonal Cellular Automata Model. In: P.M.A. Sloot, B. Chopard and A.G.Hoekstra (Eds.), *ACRI 2004, LNCS 3305*, Springer, Berlin, 2004, 725-734.
- von Neumann, J. (Edited and completed by A. Burks), 1966. *Theory of self-reproducing automata*. USA: University of Illinois Press.
- Zuo, W., Chen, Q., 2010. Fast and informative flow simulations in a building by using fast fluid dynamics model on graphics processing unit. *Build. Envir.*, 45, 3, 747-757.

NEIGHBORHOOD CONCEPT FOR MODELING AN ADAPTIVE ROUTING IN WIRELESS SENSOR NETWORK

Jan Nikodem^(a), Maciej Nikodem^(a), Ryszard Klempous^(a), Marek Woda^(a), Zenon Chaczko^(b)

^(a)Wrocław University of Technology, Poland

^(b)University of Technology Sydney, NSW, Australia

email: {jan.nikodem, maciej.nikodem, ryszard.klempous, marek.woda}@pwr.wroc.pl,
zenon.chaczko@uts.edu.au

ABSTRACT

This paper summarizes the research work of the Wireless Sensor Networks group at the Wrocław University of Technology. The group came up with an innovative technique that uses a number of relations to define activities and manage the network communication. The suggested solution, uses the concept of neighborhood as the primary spatial entity in order to demonstrate how the application of relations could support concurrence of both global and local perspectives. Additionally, a case study is provided to show the merits of the new adaptive routing technique for both semi-static and dynamic environments.

1. INTRODUCTION

In this paper we present recent results of our work on the formalization of processes description in a distributed systems. We focus on routing in Wireless Sensor Network (WSN) because it is really distributed, moreover, routing as a type of activity, especially associated with information, is a basic processes investigated in WSN. Due to the spatial distribution of nodes and constrained resources, the operation of WSN sensors is focused on local activities and mutual communication.

Consequently, our work concentrates on developing the formal methods and techniques necessary to model and evaluate situations in the network, decision processes and implement intelligent behavior while following the general outlines of the network activity. These requirements will entail the investigation of various methods in which intelligent systems could evaluate, interact and self-organize, both individually and cooperatively with other spatial explorers or while interacting with the environment. The Wireless Sensor Network poses stringent communication efficiency requirements in order to sustain its functionality.

The dependability of such a network depends on the energy efficiency, node failures as well as the quality of the radio channel in a vicinity. Therefore, when working on routing algorithms we look for various adaptive solutions that are suitable for WSNs

2. RELATED WORKS

There is a large number of research publications that consider communication activity in WSN's that is related mainly to clustering and routing problems. A number of authors have

proposed sensors' self-configuring [2], self-management [3], [13], [14], adaptive clustering [2], [9], [16] or the concept of adjustable autonomy [6] to efficiently manage data packets in a network. On the other hand there are papers which discuss bio-inspired ideas and tend to isolate some aspects of the natural world for computer emulation. Authors [5] have shown that the communication topology of some biological, social and technological networks is neither completely regular nor completely random but stays somehow in between these two extreme cases. It is worth to mention papers [3], [14], [16] devoted to self-organizing protocols using both random and deterministic elements.

Design challenges in building WSN structure can be described using different mapping functions. Consequently, in WSN literature several various models were proposed [5], [2], [9], [16]. These attempts are based on the representation of the network as the constellation of nodes connected with each other. In their research work, Cohn at al. [5] concentrated on using the regions as the primary spatial entities rather than the traditional mathematical and dimensionless points representing nodes. The authors proposed the concept of vague and crisp regions as a qualitative spatial representation and have argued that such a modification would allow for simplification and the use of standard mathematical topology tools. This approach is a basis for such popular concepts of segmentation in multi-hop networks as: region building and clusterization. The researchers who discussed these issues, have proposed various different methods to determine such structures and pointed out benefits and drawbacks of these approaches.

3. BASIC CONCEPT

The approach presented in this paper is distinctly different from those mentioned in literature. We propose a relational attempt, based on set and relation theories. We consider three basic relations: subordination, tolerance and collision [6, 8]. These relations correspond to network activity, therefore the fourth relation (neighborhood), corresponding to WSN structure, is involved.

3.1. Traditional network partitioning

Network partitioning is a well known idea [7] to solve large and complex problems by segmenting them into smaller and possibly simpler tasks. The most crucial element of such an attempt is to decide how to make a segmentation in

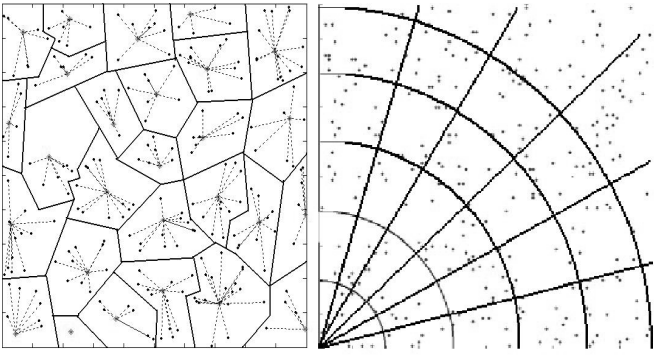


Figure 1. Two methods of WSN partitioning: clusters (left) and regions (right)

order to get a subproblems that can be solved efficiently and, what is more important, can be useful for finding a solution of the original problem. A commonly accepted idea for segmentation of the WSN structure is based on different mapping functions. Hence, in WSN literature [5], [9], [11], [14] several various models were proposed. The most popular concepts of network segmentation for multi-hop networks are regions building, clusterization and neighborhood. Let us come closer to these issues and begin from $Map(X, Y)$ expression that can be defined as a collection of mappings of set X onto set Y (surjection). Further, $Sub(X)$ is defined as a family of all X subsets and segment S as a mapping

$$S \in \{Map(Nodes, Sub(Nodes))\} \quad (1)$$

where

$$S(k)|_{k \in Nodes} := \{y \in Nodes \mid y R_S k\} \quad (2)$$

and k is a segment's main node (segment head).

Based on a different segment relation R_S further dyadic relations are defined. For example, we can build segments which are both mutually exclusive and collectively exhaustive with respect to the set of all network nodes – $Nodes$. Formally, such segments constitute indexed family of sets $\mathcal{S} = \{S_i \mid i \in I\}$ for which following properties are fulfilled:

$$(\forall i \in I)(S_i \neq \emptyset) \wedge \bigcup S_i = Nodes, \quad (3)$$

$$(\forall i, j \in I \mid i \neq j)(S_i \cap S_j = \emptyset). \quad (4)$$

Conditions (3), (4) imply that:

$$(\forall y \in Nodes)(\exists! i \in I \mid y \in S_i) \quad (5)$$

where $\exists!$ means "exists exactly one". We recall that conditions (3), (4) determine a partition of a set $Nodes$. Consequently, \mathcal{S} is a partition of set $Nodes$ and the total number of possible partitions of an n -element set $Nodes$ is expresses by the Bell number B_n . These numbers satisfy a well known recursion formula and soar in the number of network nodes. In clustering and building regions approaches the emphasis is on determining multiple, potentially useful partitions and letting the algorithm decide. However, this lead to a larger consumption of network resources like energy or channel throughput.

In general, research on efficient routing in wireless sensor networks has proceeded along two main approaches: cluster-

In clustering algorithms (R_C) network partition results in family of subsets called clusters. Clustering algorithm decides whether particular node becomes the cluster head or a regular one. As a consequence a specific type of subsets is created Fig.1(left). Considering the pros and cons of clasterization three are of utmost significance:

- it allows to build hierarchical structures with cluster heads and regular nodes,
- it reduces mesh communication, places restrictions on regular nodes activity within cluster,
- it increases efficiency of multi-hop communication since only cluster heads are responsible for message routing.

Nonetheless, clustering results in the existence of exactly one transmission path between any node and the base station Fig.1(left) which is considered as a drawback.

The second commonly accepted approach to network partition is based on region (R_R) concept [15]. That solution is based on obvious and regular network segmentation as presented on Fig.1(right). A regions building approach is derived from both the technological limitation of radio communication and multi-path retransmission. First, based on radio link range, the network is partitioned into *coronas* determined by concentric circles centered in base station. Next, a *pie* structure is determined using a number of angular wedges centered at the base station. Each pie relates to path routing area. Finally, a regular structure consisted of *regions* is created.

From a global (network) perspective the advantages of traditional network partitioning are evident and clearly seen on Fig.1. There is an obvious trade-off; both partitioning methods i.e. sector building and clustering simplify and clarify the global view, but suffer from the reduced possibility of choice in each node. In some specific situations such segmentation of the network into regions or clusters can be very beneficial, because they simplify communication, but this is not the case in general.

3.2. Network segmentation based on neighborhood

In real WSNs, sector building and clustering issues are a mixed blessing: building regions is not practically effective and clustering is not simple at all. Because of that we suggest the use of neighborhoods as the primary spatial entities and show how freak results can be obtained from surprisingly few primitives.

Now we attempt to extend network segmentation approach to neighborhood. Based on (1), (2) and substituting $\mathcal{R}_S = \mathcal{R}_N$ we define the neighborhood \mathcal{N} as follows:

$$\mathcal{N} \in Map(Nodes, Sub(Nodes)). \quad (6)$$

Thus, $\mathcal{N}(k)$ denotes the neighborhood of node k while $\mathcal{N}(S)$ is the neighborhood of set of nodes S defined as:

$$\mathcal{N}(k)|_{k \in Nodes} := \{y \in Nodes \mid y \mathcal{R}_N k\}, \quad (7)$$

$$\mathcal{N}(S)|_{S \subset Nodes} := \{y \in Nodes \mid (\exists k \in S)(y \mathcal{R}_N k)\}, \quad (8)$$

where $y \mathcal{R}_N k$ means that nodes y and k are in relation \mathcal{R}_N .

There are a number of reasons for eschewing a cluster-based approach to adaptive routing in WSN and indeed simply using neighborhood abstraction supported by the standard tools of mathematical topology.

Firstly, clustering is some kind of simplification, that facilitates computation and restricts the set of possible solutions at

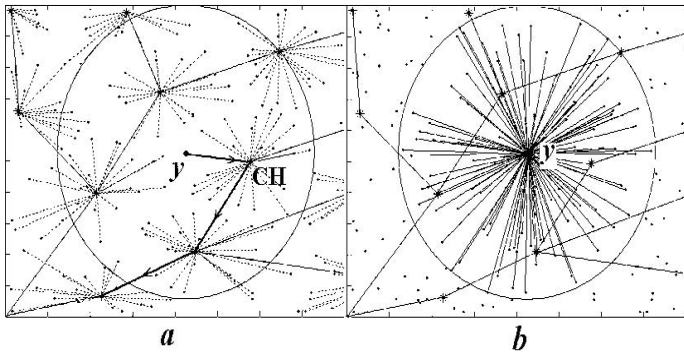


Figure 2. Different perspectives of communication: a). within cluster, b). within neighborhood

the same time. Concerning the cluster (Fig.2a), there is only one way to send data from node y towards base station while neighborhood (Fig.2b) provides evidently more possibilities.

Next, the neighborhood gives a natural way to represent a kind of collective cooperation within some vicinity, that is germane to routing activity. Finally, the neighborhood abstraction is determined by native (mostly technical) constraints – e.g. radio communication range. However, such an approach generalizes the concept of a neighborhood far beyond its intuitive meaning, it also turns out that it is possible to extend neighborhood abstraction to sets (e.g. (8)).

Considering native neighborhoods within WSN network one can define an indexed family of sets $\mathcal{N} = \{N_i \mid i \in I\}$, where I denotes the set of indicies and N_i has the following properties:

$$(\forall i \in I)(N_i \neq \emptyset) \wedge \bigcup N_i = Nodes, \quad (9)$$

$$(\exists \sim i, j \in I \mid i \neq j)(N_i \cap N_j \neq \emptyset). \quad (10)$$

The last property can be rewritten for a single node y as

$$(\forall y \in Nodes)(\exists \sim i \in I \mid y \in \bigcap N_i \neq \emptyset), \quad (11)$$

where the expression $\exists \sim$ can be translated as: "there are as many instances as the structure of the network allows for" and this completes the definition of neighborhood.

It is worth mentioning that we obtained (11) as a result of relaxing the requirement (5). To put it another way, formula (10) can be seen as a negation of (4).

Note, that as a result we obtain a strongly overlapped structure (Fig.3 (left)). If two or more neighborhoods overlap then they share a common region, while this cannot be the case for clusters, which must exactly 'touch' each other. It means that neighborhoods do not structure a set of WSN nodes into mutually exclusive subsets.

4. NEIGHBORHOODS WITH ADAPTATION ABILITY

A neighborhood abstraction, defined by a set of criteria for choosing neighbors and set of common resources to be shared, is very useful in almost all algorithms of routing protocols in WSN. Realizing distributed operation/tasks in which nodes communicate only within vicinity, sensor network draws on concept of neighborhood. It is worth pointing out that the neighborhood relation is of the great significance since the whole activity of every node of WSN is determined by the state of the node and its neighbors. The neighborhood

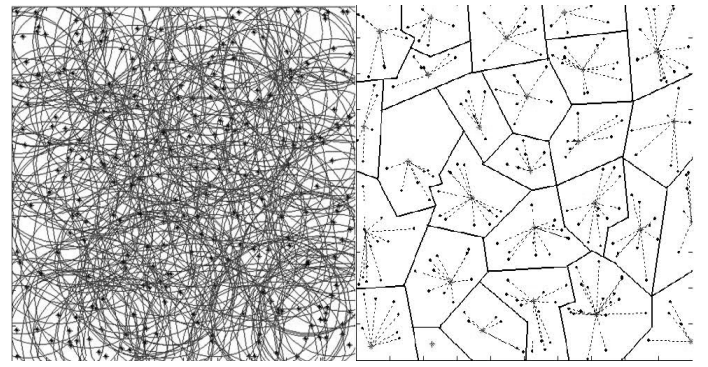


Figure 3. Two types of WSN structured topology: clusters (right) and neighborhoods (left)

is then used to perform local activities and to choose the best tactics that will be implemented in practice.

4.1. Relational attempt to network activity

Given our new primitive of the neighborhood, we can now start defining some new relations that exploit this abstraction. Lets introduce the following three new dyadic relations:

$$\text{Subordination } \pi = \{ \langle x, y \rangle; x, y \in Act \mid x \pi y \}. \quad (12)$$

The expression $x \pi y$ defines the action x which is subordinated to the action y or action y dominates over action x .

$$\text{Tolerance } \vartheta = \{ \langle x, y \rangle; x, y \in Act \mid x \vartheta y \}. \quad (13)$$

The expression $x \vartheta y$, states that the actions x and y tolerate each other,

$$\text{Collision } \varkappa = \{ \langle x, y \rangle; x, y \in Act \mid x \varkappa y \}, \quad (14)$$

and finally $x \varkappa y$ means the actions x and y are in collision to one another. The basic properties of mentioned above relations could be formulated succinctly as follows [8]:

$$\pi \cup \vartheta \cup \varkappa \subset Act \times Act \neq \emptyset, \quad (15)$$

$$\iota \cup (\pi \cdot \pi) \subset \pi, \quad (16)$$

where $\iota \subset Act \times Act$ is the identity on the set *Action*. Moreover:

$$\pi \cup \vartheta^{-1} \cup (\vartheta \cdot \pi) \subset \vartheta, \quad (17)$$

where ϑ^{-1} is the converse of ϑ . That is:

$$\vartheta^{-1} = \{ \langle x, y \rangle \in X \times Y \mid y \vartheta x \}. \quad (18)$$

For collision,

$$\varkappa^{-1} \cup \{ \pi \cdot \varkappa \} \subset \varkappa \subset \vartheta', \quad (19)$$

where ϑ' is the complement of ϑ i.e.:

$$\vartheta' = \{ \langle x, y \rangle \in X \times Y \mid \langle x, y \rangle \notin \vartheta \}. \quad (20)$$

$$\pi \cup \vartheta \cup \varkappa \subset Act \times Act \neq \emptyset, \quad (21)$$

and

$$\iota \cup (\pi \circ \pi) \subset \pi, \quad (22)$$

where ι is an identity relation on the set *Act*. Eq. (21) states that all three relations are binary on non-empty set of actions

(Act). Eq. (22) states that subordination is reflexive ($\iota \subset \pi$) and transitive ($\pi \circ \pi \subset \pi$). Further

$$\pi \cup \vartheta^{-1} \cup (\vartheta \circ \pi) \subset \vartheta \quad (23)$$

means that:

- subordination implies tolerance – if π holds for some $x, y \in Act$ then ϑ also holds for these,
- tolerance is symmetrical – if $x\vartheta y \Rightarrow y\vartheta x$,
- subordinated action tolerates all actions tolerated by the dominant – if $(x\pi y \wedge y\vartheta z) \Rightarrow x\vartheta z$.

For collision relation we have that

$$\varkappa^{-1} \cup \{\pi \circ \varkappa\} \subset \varkappa \subset \vartheta' \quad (24)$$

where ϑ' is the complement of ϑ :

$$\vartheta' = \{ \langle x, y \rangle \in X \times Y \mid \langle x, y \rangle \notin \vartheta \}. \quad (25)$$

Eq. (24) states that collision is symmetric and disjoint to tolerance. Moreover, all subordinated actions must be in collision with action being in collision with its dominant.

4.2. Modeling global network strategy

The main aim at the WSN, similar as for optimization problems (goal function, drainage function, constraints) related to communication in WSN, is defined globally within the scope of the whole network. Therefore, using relations we determine (globally) initial organization which remains static during a network lifetime.

Let us consider the node k and its neighborhood $\mathcal{N}(k)$. Any communication activity act_k that is performed by node k relates to some members of $\mathcal{N}(k)$ and the set of actions act_k within neighborhood $\mathcal{N}(k)$ can be defined as follows:

$$Act_{\mathcal{N}(k)} := \{ act_k \in Act \mid (\exists x \in \mathcal{N}(k)) (act_x \mathcal{R} act_k) \}. \quad (26)$$

The Cartesian product defined as:

$$IS_k := Act_{\mathcal{N}(k)} \times Act_{\mathcal{N}(k)} \subseteq \pi \cup \vartheta \cup \varkappa, \quad (27)$$

describes interaction space IS_k within $\mathcal{N}(k)$. Let us now consider a set of possible interactions fulfilled relation \mathcal{R} within neighborhood $\mathcal{N}(k)$ which can be expressed as:

$$\mathcal{R}_k := \{ y \in Act_{\mathcal{N}(k)} \mid \langle act_k, y \rangle \in IS_k \wedge k \mathcal{R} y \}. \quad (28)$$

Thus, for a given relation \mathcal{R} and a node k we define an intensity quotient within neighborhood $\mathcal{N}(k)$ as follows:

$$IR_k = Card(\mathcal{R}_k) / Card(IS_k), \quad (29)$$

where $Card(\mathcal{R}_k)$ is cardinality of set \mathcal{R}_k .

Let IS_{WSN} be a global interaction space consisting of all feasible actions in WSN and GS_{WSN} be a global strategy defined as a subset of IS_{WSN} . Notice that there is only one set IS_{WSN} while there may be many different ways GS_{WSN} can be chosen. However, for the simplicity let us consider only singleton $s = \{ \langle IR^\pi, IR^\vartheta, IR^\varkappa \rangle \}$:

$$s = \{ \langle 0.2, 0.54, 0.07 \rangle \} = GS_{WSN}^* \quad (30)$$

and initialize all WSN nodes with these intensity quotients as suggested intensity for relations π , ϑ , and \varkappa , within their neighborhoods. This implies that any communication activity within node neighborhood should be limited to this space and each node should retain demanded relational intensity within its neighborhood. This ensures that the local behavior is compliant with the global requirements in a full extent.

Finally, having a well defined neighborhood and required relational intensity for relations π , ϑ , \varkappa , we are ready to decompose the globally defined goal function and constraints into a uniform locally performed task assigned to each node in the network. It is not an easy task to cast all global dependencies from network area to the neighborhood. Moreover, the fact that neighborhood conditions for the network nodes might be, and usually are, quite dissimilar makes this issue even more difficult.

In the paper [12] we have proposed a concept to realize global/local task decomposition based on a Digital Terrain Model. Instead of globally formulated optimization tasks for network, we model 3D surface over the network area. It results in bare drainage surface that gives a basic reference frame which ensures data is send towards the BS. There are many possible surfaces, but for further consideration, let us assume that the drainage surface is based on node to base station hop-distance.

4.3. Modeling adaptive activity within neighborhoods

In the previous chapter we have described how to use three dyadic relations π , ϑ , \varkappa to ensure that all local activities are in the scope of desired global behavior. In WSN systems actions and decisions are taken by nodes based on their knowledge about the network. Due to limited communication range, nodes have to cooperate within the neighborhood to gain this knowledge.

Due to the global strategy GS_{WSN}^* (30) the activity of each node k is restricted to three subsets of neighbors:

$$\begin{aligned} N_\pi(k) &= \{ y \in N(k) \mid y\pi k \}, & N_\vartheta(k) &= \{ y \in N(k) \mid y\vartheta k \}, \\ N_\varkappa(k) &= \{ y \in N(k) \mid y\varkappa k \}. \end{aligned} \quad (31)$$

According to (30)

$$C(N_\pi(k)) / C(N_\varkappa(k)) / C(N_\vartheta(k)) = 0.2 / 0.54 / 0.07 \quad (32)$$

where $C(X)$ denotes cardinality of X . Intensity quotients of relation π , \varkappa , ϑ can determine number of the elements in the each of (31) sets. The decision of which node from the neighborhood belongs to which particular set must be taken locally, but the drainage function based on hop-distance will help to make the correct decision. The aforementioned distance is expressed in terms of number of hops required to reach the base station, moreover every node with distance $X+1$ can communicate with at least one node which distance to the BS equals X . Eventually any node k knows its hop distance ($dis^h(k)$) and distances of all of its neighbors ($dis^h(i)$ where $i \in N(k)$). Based on this information it is possible to split neighborhood $N(k)$ into three subsets:

$$\begin{aligned} N_{<}(k) &= \{ y \in N(k) \mid dis^h(y) < dis^h(k) \}, \\ N_{=}(k) &= \{ y \in N(k), y \neq k \mid dis^h(y) = dis^h(k) \}, \\ N_{>}(k) &= \{ y \in N(k) \mid dis^h(y) > dis^h(k) \}, \end{aligned} \quad (33)$$

which are mutually exclusive and collectively exhaustive $N(k)$ but they are not partition of a set $N(k)$.

When determining the sets (33) any node k can attach elements of sets (31) such that (32) is satisfied and retain a data-flow direction towards the BS. Neighbours that belong to a $N_\pi(k)$ set are selected from $N_{<}(k)$ set so it may consist of some nodes $i \in N_{<}(k)$. Based on $N_{>}(k)$ a set $N_\varkappa(k)$ is

being established and similarly $N_{\vartheta}(k)$ on $N_{=}(k) \cup N_{>}(k)$, so finally:

$$\begin{aligned} N_{\pi}(k) &\subset N_{<}(k); \quad N_{\vartheta}(k) \subset (N_{<}(k) \cup N_{=}(k)) \\ N_{\varkappa}(k) &\subset N_{>}(k). \end{aligned} \quad (34)$$

Such an idea of local communication tactics ensures compatibility with globally established strategy and retains a proper data drainage direction towards the BS. This works regardless of the fact that in most cases a node does not know the location of the BS.

The problem we now face is to order elements of the sets (31). For this purpose we use two indicators of quality connections: radio link quality indicator (*LQI*) and Received Signal Strength Indicator (*RSSI*). To order elements of $N_{\pi}(k)$ we use *LQI* values while to order the remaining two; $N_{\vartheta}(k)$, $N_{\varkappa}(k)$ we use *RSSI*. Ordering any of these sets we write down the indicator values in order of magnitude, beginning with the greatest.

Each node must locally undertake a decision, which node is the one to pass the packet to. It is vital that the packet from a node k traverses in a direction determined by slope of a bare drainage surface. This means that the next hop node has to be chosen from among the neighbors $N_{\pi}(k)$ which hop distance to the base station is smaller than $dis^h(k)$. In such a situation we choose the first node (e.g. this with the greatest value of *LQI* parameter).

Because, either the values of *LQI* or *RSSI* reflect up-to-date conditions in the neighborhood and vary in time, adaptability to current environment conditions within neighborhood is attributed to such techniques. When node detects changes in his vicinity or detects deviations from the normal neighbor behavior, it has wide choice of communication path. As a result, each multi-hop communication between any two nodes may use a different communication path.

5. FINAL REMARKS

The research work is still in progress on wireless sensors network and related formalisms based on the theory of sets and relations. Taking advantage of the relational approach, an innovative strategy is applied that involves the entire network, by sending triplets the strategy that is prepared for the whole network, by sending triples of intensity quotients (30). Moreover, the development of relations, its intensity quotients and metric, both globally – in network and locally – within neighborhood, results in routing adaptability.

Each node performs some communication activity, based on globally determined balance between subordination π , tolerance ϑ and collision \varkappa . So, this leads rather to governance than to control, what finally causes that particular action strongly depends on varied in time local/neighborhood conditions. The development of relation provided network not only with adaptability but also facilitate communication structure continuous revitalization.

The concept of neighborhood is essential and really native to WSN. This is because of radio communication; the radio link range constraint is an origin of neighborhood. Furthermore, there is no doubt that a high cardinality of any neighborhood set and its extensive overlapping makes the neighborhood more attractive than clusters or regions. A wide scope of choices is the reason for success. The approach allows for individual tactic selection (within node's

neighborhood), that takes into consideration conditions in the node's vicinity while at the same time retaining global strategy requirements.

The neighborhood relation is defined both for a single node of the network and for a group of nodes. It is worth pointing out that the neighborhood relation is of the greatest significance since the whole activity of every node of WSN is determined by the state of a node and its neighbors. The neighborhood is then used to perform local activities and to choose the best tactics that will be implemented in practice. As a result, the native neighborhood was advised as the most suitable form of the local range.

Acknowledgements

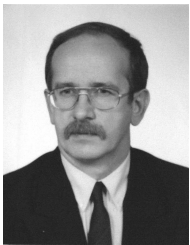
This work was supported by E.U. Regional Development Fund and by Polish Government within the framework of the Operational Programme - Innovative Economy 2007-2013. Contract POIG.01.03.01-02-002/08-00, Detectors and sensors for measuring factors hazardous to environment - modeling and monitoring of threats.

References

- [1] Cohn, A.G., Bennett B., Gooday J.M., Gotts N.M., 1997. Representing and Reasoning with Qualitative Spatial Relations about Regions. In: Cohn, A.G., Bennett B., Gooday J.M., Gotts N.M, eds. Spatial and Temporal Reasoning, Dordrecht, Kluwer, 97-134.
- [2] Cerpa A., Estrin D. ASCENT: Adaptive Self-Configuring Sensor Networks Topologies, IEEE Transactions On Mobile Computing, vol. 3, no. 3, Jul-Sep 2004.
- [3] C. Chevally, R. E. Van Dyck, and T. A. Hall. Self-organization Protocols for Wireless Sensor Networks. In Thirty Sixth Conference on Information Sciences and Systems, March 2002.
- [4] Chaczko Z., Ahmad F.: Wireless Sensor Network Based System for Fire Endangered Areas, ICITA 2005, Sydney, 2005.
- [5] Cohn, A. G. and Gotts, N. M.: 1994b, A theory of spatial regions with indeterminate boundaries, in C. Eschenbach, C. Habel and B. Smith (eds), Topological Foundations of Cognitive Science.
- [6] Crandall J.W., Goodrich M.A.: Experiments in adjustable autonomy, in IEEE International Conference on Systems, Man, and Cybernetics, vol.3, Tucson, USA, 2001, 1624-1629.
- [7] Descartes R., Lafleur L.: Discourse on Method and Meditations, New York: The Liberal Arts Press., 1960.
- [8] Jaro J.: Systemic Prolegomena to Theoretical Cybernetics, Scient. Papers of Inst. of Techn. Cybernetics, Wroclaw Techn. Univ., no. 45, Wroclaw, 1978
- [9] Lin Ch.R., Gerla M.: Adaptive Clustering for Mobile Wireless Networks, IEEE Journal On Selected Areas In Communications, vol. 15, no. 7, Sep 1997
- [10] Nikodem, J., 2008. Autonomy and Cooperation as Factors of Dependability in Wireless Sensor Network, Proceedings of the Conference in Dependability of Computer Systems, DepCoS - RELCOMEX 2008, 406-413. June 2008, Szklarska Poreba, Poland
- [11] Nikodem J., Klempous R., Chaczko Z., Modelling of immune functions in a wireless sensors network. W: The 20th European Modeling and Simulation Symposium. EMSS 2008, Campora S. Giovanni, Italy, 2008

- [12] Nikodem J., Klempous R., Nikodem M., Woda M.: Directed communication in Wireless Sensor Network based on Digital Terrain Model. 2nd International Symposium on Logistics and Industrial Informatics (LINDI), Linz, Austria, [Witold Jacak et al eds., pp. 87-91, Piscataway, NJ : IEEE
- [13] Scerri P., Pynadath D., Tambe M.: Towards Adjustable Autonomy for the Real World, Journal of Artificial Intelligence Research, vol.17, 2003
- [14] Sohrabi K., Gao J., Ailawadhi V., Pottie G.J.: Protocols for Self-Organization of a Wireless Sensor Network, IEEE Personal Communications, Oct 2000
- [15] Stojmenović, I., editor: Handbook of Sensor Networks Algorithms and Architectures, John Wiley and Sons Inc., 2005.
- [16] Veyseh M., Wei B., Mir N.F.: An Information Management Protocol to Control Routing and Clustering in Sensor Networks, Journal of Computing and Information Technology - CIT 13 (1) 2005, 53-68
- [17] Younis O., Fahmy S.: HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks, IEEE Transactions On Mobile Computing, vol. 3, no. 4, Oct-Dec 2004

AUTHORS BIOGRAPHY



data transmission.

Jan Nikodem received the B.Sc. in electrical engineering, M.Sc. in artificial intelligence in 1979 and Ph.D. degree in computer science in 1982 from Wrocław University of Technology (WUT), Poland. Since 1986, he is an Assistant Professor in the Institute of Technical Cybernetics, WUT. Since 2005 in the Institute of Computer Engineering, Automatics and Robotics (ICEAR). His current research is focused on the area of complex and distributed systems, cybernetics, wireless sensor networks and digital



of Computer Engineering, Control and Robotics, Faculty of Electronics at WUT.

Maciej Nikodem received a M.Sc. in Computer Science in 2003 and a M.Sc. in Control and Robotics in 2005 from the Wrocław University of Technology in Poland. In 2008 he completed Ph.D. studies in Computer Science at Faculty of Electronics, Wrocław University of Technology. For last 5 years Maciej Nikodem has worked on Countermeasures to Fault Analysis, Boundary Scan Security as well as security aspects of Wireless Sensor Networks. Maciej Nikodem is an Assistant in the Institute



Ryszard Klempous holds a M.Sc. in Automation (1971) and Ph.D. in Computer Science (1980) from Wrocław University of Technology (WUT). Since 1980 he has been an Assistant Professor in the Institute of Computer Engineering, Auto-matics and Robotics, WUT. Senior member of IEEE and NYAS, has already published over 90 papers in Optimization Methods and Algorithms, Simulation and Data Processing and Transmission



An Experiment in Delivering CBL Materials, VI FP Integrated Project FP6-IST 26600 DESEREC Dependability and Security by Enhanced Reconfigurability). He is the author of about 30 scientific articles and conference papers.

Marek Woda is an Assistant Professor in the Institute of Computer Engineering, Control and Robotics at Wrocław University of Technology. In 2001 he graduated at WUT. In 2007, he received Ph.D. degree in Computer Science from the Faculty of Electronics WUT. His research interests focus on multi-agents systems, e-learning, Internet technologies. He participated in international projects sponsored by European Union (e.g. PL96-1046 INCO-COPERNICUS project Multimedia Education:



work protocols and system software middleware. Mr Chaczko is a Senior Lecturer in the Information and Communication Group within the Faculty of Engineering at UTS.

Zenon Chaczko completed a B.Sc. in Cybernetics and Informatics in 1980 and a M.Sc. in Economics in 1981 at the University of Economics, Wrocław in Poland, as well as completed MEng in Control Engineering at the NSWIT 1986 and Ph.D. in Engineering at UTS, Australia. For over 20 years Mr Chaczko has worked on Sonar and Radar Systems, Simulators, Systems Architecture, Telecommunication network management systems, large distributed Real-Time system architectures, network protocols and system software middleware.

BUSINESS PROCESS SIMULATION FOR MANAGEMENT CONSULTANTS: A DEVS-BASED SIMPLIFIED BUSINESS PROCESS MODELLING LIBRARY

Igor Rust^(a); Deniz Cetinkaya^(b); Mamadou Seck^(c); Ivo Wenzler^(d)

^{(a)(b)(c)}Systems Engineering Group; Faculty of Technology, Policy, Management

Delft University of Technology; Jaffalaan 5, 2628BX, Delft, THE NETHERLANDS

^(d)Accenture Nederland; Gustav Mahlerplein 90, 1082 MA, Amsterdam, THE NETHERLANDS

^(a)i.j.rust@student.tudelft.nl, ^(b)d.cetinkaya@tudelft.nl, ^(c)m.d.seck@tudelft.nl, ^(d)ivo.wenzler@accenture.com

ABSTRACT

Business process simulation enables analysis of business processes over time and allows to experiment with scenarios (like for instance redesigning business processes) before implementing them into an organization. This research aims at providing an easy way of business process modelling and simulation for management consultants whose core competence is not simulation model development. During the design and development process, management consultants are actively involved following a user-centred design approach. The outcome of this research is a library of DEVS-based business process modelling elements implemented in Java and using the DSOL simulation library to provide the simulation capabilities.

Keywords: business process modelling, business process simulation, component based modelling, DEVS

1. INTRODUCTION

To stay competitive and to operate effectively, an organization needs to improve its process efficiency and its quality by adapting its strategy, structure, management and operations to its changing business environment. Management consultants provide expertise and recommendations to improve their clients' business performance. To support organizational decisions, a good understanding of the business processes is essential.

A business process is a series of activities that produces a product or service for a customer. Business Process Modelling (BPM) is the activity resulting in a representation of an organisation's business processes so that they may be analyzed and improved (Weske 2007).

A distinction can be made between static modelling and dynamic modelling of business processes (Bosilj-Vuksic, Ceric and Hlupic 2007). Static modelling tools often provide a graphical process representation, for example simple flowcharts, IDEF0 or BPMN diagrams to depict business processes. Business Process Simulation (BPS) tools, provides the possibility to simulate and evaluate the dynamic behaviour of business processes.

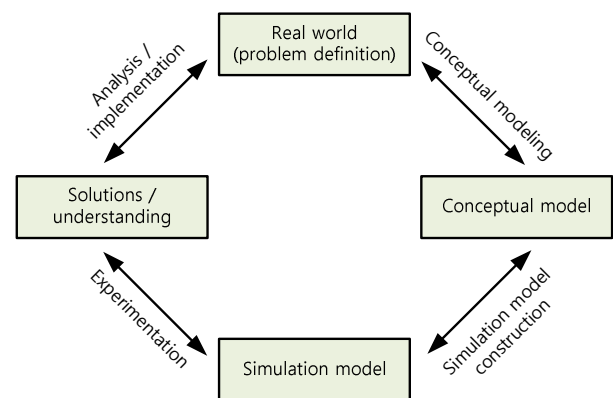


Figure 1: Simulation Project Life Cycle.

Figure 1 depicts the main phases and products of a simulation study. The organization, for which an analysis is undertaken, is part of the “real world”. First, a conceptual model is developed, often in a graphical form, which contains the essential aspects of the problem situation (Banks 1998). A conceptual model helps to build credibility for and guides the development of a simulation model. Next, a simulation model of the business process is developed based on this conceptual model. The executable model can be manually programmed by the modeller, or constructed through a visual interface (Pidd 2004a).

After the development of the simulation model, experiments can be set up and the simulation model is executed by simulation software to analyse the output. A simulation software generally consists of a simulation engine (or simulation executive) and an application program (Pidd 2004b). The engine keeps track of the state changes which occur at some moment in time and reminds the application when a state change is due. How the experiment is set-up and which output parameters are of interest, depend on the business case. Based on the presented outputs, more experiments can be performed or changes may be implemented in business processes.

Although the usefulness of business process simulation was proven by many authors and various simulation tools are available, still many consultants

and business analysts rely on simple static process mapping methods (Bosilj-Vuksic et al. 2007; Melão and Pidd 2003). Some reasons for the lack of adoption are that much experience is needed to develop valid simulation models and simulation model development is time consuming and costly (Van Eijck and De Vreede 1998). More specifically, there is a lack of business process simulation tools which supports an easy and quick approach of modelling and analysis of business process by consultants and business analysts.

This paper presents a business process simulation method to support management consultants to model, simulate and analyze business processes in a well defined manner. Next section provides background information about business process consultancy which is extracted from the interviews with the consultants. Section 3 discusses the design process of our research which is based on a user-centred design approach. Section 4 and 5 present a DEVS component library for business process modelling and its usage. Finally, conclusions and future work are given in Section 6.

2. SUPPORT FOR BPS BY CONSULTANTS

2.1. The Consultants' View on Business Processes

Harington (1991) defines a business process as a group of logically related tasks that use the resources of the organization to provide defined results in support of the organization's objective. Consultants see business processes more specifically as *"a series of activities and decisions which are performed by resources and which influences the flow and state of the entities"*.

An entity (or passive entity) is an abstract object which can represent anything that undergoes activities in a business process. The entity arrives at an organization, "flows" through the business process(-es), and then leaves the organization. What the entity actually represents is called the entity type. Examples are an order that is received by a company and needs to be processed, an insurance claim or a contract cancellation e-mail. During a business process, an entity undergoes state changes as a result of the activities that are performed by resources. The state of an entity may for instance be "received", "processed" or "finished".

A resource is responsible for making decisions and performing activities on entities and is considered as the leading part of a business process. A resource can be a human or a machine. Resources are typified by their capability, role, capacity, availability and state. The capability of a resource depicts whether a resource is able or allowed to perform a certain activity. Based on for instance the experience that a resource has, the resource may be able to perform more complicated activities. The collection of all capabilities of a resource is called the resource's role. It is possible that within an organization, multiple resources have the same capabilities and thus share the same role. In that case, a role has a certain capacity: the number of resources that are available to perform a determined set of activities.

The availability of a resource depicts when or for how long a resource is available to commence activities. For instance, a full time employee has an availability of 1 FTE (Full Time Equivalent), which depicts that during a complete working day the resource is available to perform activities. The state of a resource relates to whether a resource is currently busy (or active) with performing a certain activity, or is available for new work.

An activity is a piece of work performed by one or more resources which requires a certain amount of time. An activity can be a task or a grouping of task, called a sub-process. A task is a piece of work that cannot be subdivided into smaller pieces of work to be performed by a resource, or is not crucial for the purpose of describing and analyzing a business process. There are three options how an activity can be performed, namely 1) one resource starts and finishes an activity on its own; 2) a resource hands over the entity to another resource who will perform one or more activities; or 3) the amount of work is divided over two or more resources. In the first case, the flow of an entity through activities is called sequential. In the second case, a resource will hand over the entity to another resource that will perform his activities. The third case is called a parallel activity. After the work is split up, two or more resources will perform their activities independently. At some moment the work is synchronized again and some resource will continue performing activities.

In a business process decisions are made that influence the choice and order of activities to be undertaken by a resource. A decision can depend on the attribute of an entity (e.g. entity type or state), or the state of the system (e.g. what are other resources doing, how many entities are waiting to be processed, etc.).

2.2. Conceptual Modelling

Pidd (1996) defined a model in the context of operations research and management sciences as *"an external explicit representation of part of reality as seen by the people who wish to use that model to understand, to change, to manage and to control that part of reality"*. In other words, a model can be used as a representation of reality (like for instance an object, idea or an organization), to support someone who wants to understand that part of reality, and possibly wants to make decisions which will influence reality.

There exist various modelling languages that support the representation of business processes in a model in a standard and consistent way. Some examples are BPMN, Flow Chart, Gantt Chart, IDEF0, IDEF3, and UML (Aguilar-Saven 2004). Each of these languages has different characteristics (semantics, representation notation, ability to include decomposition and hierarchy of processes, etc.).

In order to support consultants with a new modelling approach, a conceptual modelling language should be chosen or developed that is able to represent the consultants view on business processes as described in the previous section. The notation should also be

understandable to enable correct interpretation by other consultants, as well as domain experts of the modelled organization.

2.3. Simulation Model Development

As mentioned in the previous section, various languages exist to represent conceptual business process models. However, most of these languages present an abstract way of thinking and don't provide the possibility to include all details needed to enable direct translation into an executable simulation model, nor direct execution of these models by a simulation engine. Due to the lack of possibilities for a conceptual modeller to include all details in a conceptual model, different simulation models can be developed based on the same conceptual model. If for instance the simulation model is developed by someone who was not part of the conceptual modelling stage, the risk increases that a final simulation model does not represent the business processes as was intended by the conceptual modeller. Cetinkaya, Verbraeck and Seck (2010) concluded in recent research that there is a large semantic gap between the conceptual modelling stage and the simulation model construction stage.

With regard to the actual development of an executable simulation model, various formalisms exist to support the formalization of simulation models, like for instance Discrete Event System Specification (DEVS) (Zeigler, Praehofer and Kim 2000) and Petri Nets (Peterson 1981). Developing an executable simulation model using one of these formalisms requires a deep understanding of the underlying concepts, as well as programming experience.

Reusing parts of simulation models or reusing even complete simulation models is suggested to decrease the complexity and time needed to develop models. Different forms of reuse are: code scavenging (reusing existing code of a simulation model), function reuse (reusing predefined functions that provide specific functionalities), model reuse (reusing a complete simulation model for a different situation) and component reuse (reusing an encapsulated simulation module with well-defined interfaces) (Pidd 2002). Usage of components for simulation modelling is considered to be a fruitful concept to increase the efficiency of hierarchical model construction (Cetinkaya, Verbraeck and Seck 2010).

3. DESIGN PROCESS

When we consider business process simulation from a consultant's point of view, the three main activities that he/she is interested in performing, are: 1) developing the business process model (conceptual model) and specifying the model parameters; 2) experimenting with a simulation model; and 3) interpreting the results. How the translation of a conceptual model into an executable simulation model can be done is important, but also irrelevant to the consultant (as long as the simulation model leads to results as how the consultant intends it to do). Because our goal is to support consultants with

business process modelling and simulation, we follow a user-centred design (UCD) approach.

The main goal of a UCD approach is to increase the likelihood that a designed and developed artefact is found usable by its end-users. User-centred design approach is concerned with incorporating the end-users perspective during the design and development process to achieve a usable system (Maguire 2001). Because management consultants are the end-user of our new business process modelling method, they are placed at the centre of the design process. To incorporate consultants in this research, a series of design and evaluation rounds are held (in the form of workshops) with management consultants of a large international management consultancy firm. These workshops are intended to get (among other things) an understanding of the consultant and his/her view on business processes (as it is already discussed in Section 2.1); to decide upon and evaluate an understandable modelling language of the consultants view and to evaluate the usability of the proposed modelling approach.

4. A DEVS COMPONENT LIBRARY FOR BPM

4.1. Modelling Elements

The modelling representation that was the outcome of the design research process is based on the Business Process Modelling Notation (BPMN). BPMN is an industry-wide standard for modelling of business processes and was chosen because of various reasons, like: 1) the way consultants see business processes, corresponds closely to how business processes are described in the BPMN specification; 2) many management consultants have experience with modelling of business processes through the use of "swimlane diagrams" (a method that resembles BPMN to some extent, but more simplified); 3) BPMN was used for conceptual modelling of business processes in past simulation projects and was found to be useful by the involved consultants; and 4) the BPMN is becoming a standard language to model business processes throughout industries, which increases the likelihood that clients are already familiar with the notation and the models.

A set of modelling elements were determined (see Figure 2), which allow modelling of business processes by consultants as how they actually perceive business processes (as mentioned in Section 2.1). A resource (or group of resources which share the same capabilities) is represented by a *Swimlane*, which is a graphical modelling element that can contain activities (like tasks and sub processes) and decision elements (gateways).

The arrival of entities in a business process is represented by a *Start Event*. The Start Event is placed outside a Swimlane, to depict that arriving entities are coming from outside the organization (or business unit) and no resources are busy at that arrival time. The *End Event* represents the end of a business process, namely when an entity leaves the organization, and is also placed outside a Swimlane.

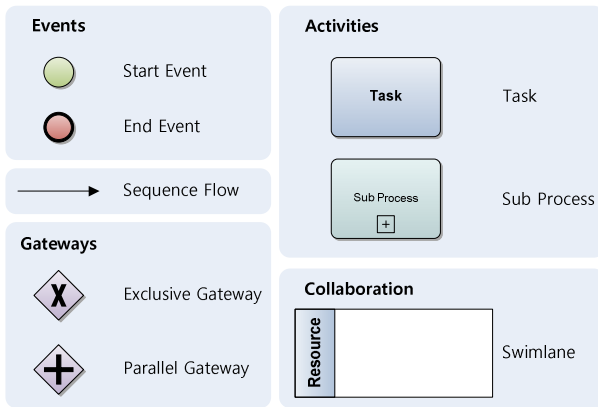


Figure 2: Overview of the BPM Elements.

After the entity enters through a Sequence Flow a Swimlane, a resource “picks up” the entity in a pre-specified manner (e.g. on a random basis, following some pattern or based on a certain priority rule) and performs one or more activities. A *Task* modelling element is an activity which represents work that is performed by a resource and consumes a certain amount of time. *Sub-process* modelling element is included to support hierarchical modelling.

Parallel Gateways are included to enable activities to be performed in parallel by multiple resources and are used in pairs: one gateway is used at the start of a parallel activity. It duplicates an entity and forwards the entities to the activities that are performed by different resources in parallel. A second *Parallel Gateway* is used to synchronize a parallel activity again after both activities are completed.

Exclusive Gateways provide the functionality to resemble business decisions. The flow of an Entity is directed based on the evaluation of a condition. This condition can be either the evaluation of an entity-specific attribute (e.g., entities of a specified type/state move in one specified direction, other entities move in another direction), or based on probability (e.g., 70% of the arriving entities move in one direction, the other 30% move in the other direction).

4.2. Formalization of Modelling Elements

We use DEVS to specify our simulation models. In DEVS, a system can be represented by two types of models: atomic and coupled models. Atomic DEVS models describe the behaviour of a system, whereas coupled models describe the composed structure of a system. Atomic models can be integrated into coupled models, and coupled models can be integrated in higher level coupled models. This way, a model is decomposed in a hierarchical manner.

The suggested BPM elements and some supplementary elements are matched to DEVS components. For every element a state-diagram was developed and validated. Figure 3 shows the state diagram for Task element. Since a Swimlane represents either a resource or a group of resources, a Task needs to check for waiting entities.

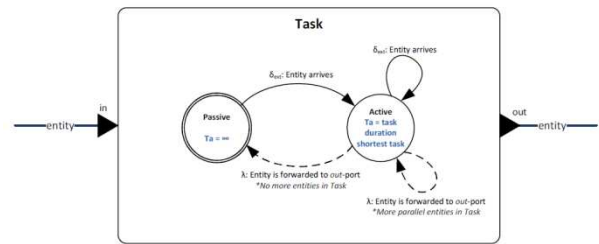


Figure 3: State Diagram for Task Element.

The supplementary components are developed to support some of the needed simulation functionality as discussed by the consultants. For instance, when an entity enters a swimlane, the entity is placed in a queue and a resource is requested. When available, the resource is attached to the entity after which the entity leaves the swimlane entry. For this purpose, the *Swimlane Entry* component was designed. When an entity leaves a swimlane, the resource occupied with the entity should be made available for other (possibly already waiting) entities. This is done by the *Swimlane Exit* component. To organize the assignment of resources to waiting entities, a partial de-central approach was chosen, namely by implementing a *Resource Manager* component which is part of each Swimlane.

The Resource Manager (RM) receives signals from amongst other the Swimlane Entry and Swimlane Exit that a new entity arrived and was placed in an entry queue, or that an entity is leaving a Swimlane. Based on a certain rule as specified by the modeller, the RM evaluates the states of all resources and queues within a Swimlane, and directs if possible a resource to a queue (by sending a signal to the appropriate queue containing information about the available resource and the destination queue).

4.3. Implementation with DEVSDSOL

DSOL, which stands for “Distributed Simulation Object Library”, was selected to provide the simulation and execution functionalities (Jacobs 2005). DSOL is a proven multi-formalism simulation library which can be considered as a generic purpose simulation tool. It is written in the Java programming language and has been used effectively in various simulation projects. DSOL also supports execution of simulation models based on the DEVS formalism through the DEVSDSOL library (which is compatible with hierarchical DEVS) (Seck and Verbraeck 2009). The choice for DSOL was made because of its flexibility and functionality regarding simulation, and its support for the DEVS formalism.

Each DEVS component has been implemented in Java and these components are executable with DEVSDSOL simulation library. Some implemented components are BPMNStartEvent, BPMNEndEvent, BPMNTask, BPMNExclusiveGateway, BPMN-ParallelGateway, etc. Figure 4 shows the DEVS internal transition function of the Task element.

```

1 @Override
2 protected void deltaInternal() {
3     if (this.phase.equals(Active)) {
4
5         if (this.parallelTaskList.isEmpty()) {
6             prt("Task " + myID + "] dInt, phase = " + this.phase + ",
7             parallelTaskList() = 0; no more entities in the taskList");
8             this.phase = this.passive;
9             this.sigma = this.phase.getLifetime();
10            prt("dInt - new sigma: " + sigma);
11            return;
12        }
13        else if (this.parallelTaskList.size() > 0) {
14            prt("Task " + myID + "] dInt, phase = active, parallelTaskList() >
15            0; (some) entities are still in the task list, namely " +
16            this.parallelTaskList.size() + " entities with ID " + this.
17            parallelTaskList.get(0).getID());
18            this.phase = this.active;
19            this.sigma = this.parallelTaskList.get(0).getTimeRemainingInTask();
20            prt("dInt - sigma is set to: " + sigma);
21            return;
22        }
23        else {
24            prt("dInt-EXCEPTION");
25        }
26    }
27 }

```

Figure 4: Sample Code from Task Component

In order to provide a higher level abstraction mechanism to our library, we applied the model driven development framework presented in (Cetinkaya, Verbraeck and Seck 2011). Next section gives brief information about the framework and then explains how it is performed in this work.

5. APPLYING THE MDD4MS FRAMEWORK

MDD4MS is a model driven development framework for modelling and simulation. The framework suggests an M&S life cycle with five stages (Problem Definition, Conceptual Modelling, Specification, Implementation and Experimentation), metamodel definitions for different stages, model to model (M2M) and model to text (M2T) transformations for the metamodels and a tool architecture for the overall process. MDD4MS presents a sample prototype implementation which is developed in Eclipse. The MDD4MS prototype provides:

- a BPMN metamodel and a BPMN editor,
- a DEVS metamodel and a DEVS editor,
- a DEVSDSOL metamodel and a DEVSDSOL editor,
- a BPMN 2 DEVS M2M transformation,
- a DEVS 2 DEVSDSOL M2M transformation,
- a DEVSDSOL 2 Java M2T transformation.

In this study, we used the BPMN editor to draw our business process models. A sample business process model is shown in Figure 5. Since the MDD4MS prototype provides generic model transformation rules for BPMN, we rewrote some rules for BPMN2DEVS M2M transformation. In this way, we directly transformed the visual modelling elements to the components that we implemented in our library.

For example, Figure 6 shows the rule to transform a Swimlane to a coupled model. We added the part that generates a Resource Manager with one input and one output port for each Swimlane.

The auto generated DEVS model via model transformations is shown in Figure 7. Once we have the DEVS model, the MDD4MS prototype automatically generates the DEVSDSOL model and the java code for

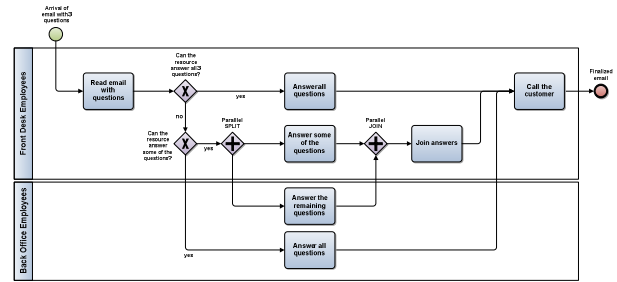


Figure 5: Sample Business Process Model.

```

1 rule BPMNSwimlaneToDEVSCoupled_inRoot {
2     from
3         s: CM_Metamodel!BPMNSwimlane (s.isInRoot())
4     to
5         t: SM_Metamodel!DEVSCoupledComp (
6             SMParentModel <- s.CMParentModel,
7             DEVSComponents <- s.BPMNActivities,
8             ...
9         ),
10        ...
11    resourceM : SM_Metamodel!DEVSAAtomicComp (
12        DEVSParentComp <- t,
13        Name <- 'RM' + s.Name,
14        DEVSComponentType <- 'ResourceManager',
15        ImplementationLink <- 'this, "ResourceManager", ' + s.getQueueMode()
16    ),
17    outRM: SM_Metamodel!DEVSOOutputPort (
18        Name <- 'out',
19        DEVSPortParent <- resourceM
20    ),
21    inRM : SM_Metamodel!DEVSIInputPort (
22        Name <- 'in',
23        DEVSPortParent <- resourceM
24    )
25    do {
26        thisModule.DefineResourceManager(t, resourceM, Sequence(outRM, inRM));
27    }
28 }

```

Figure 6: A sample rule from BPMN_2_DEVS.atl.

coupled components that uses the implemented classes for BPM modeling elements in our library. In other words, visual business process models, drawn by the BPMN editor, are transformed to executable Java code and they can be simulated with DSOL.

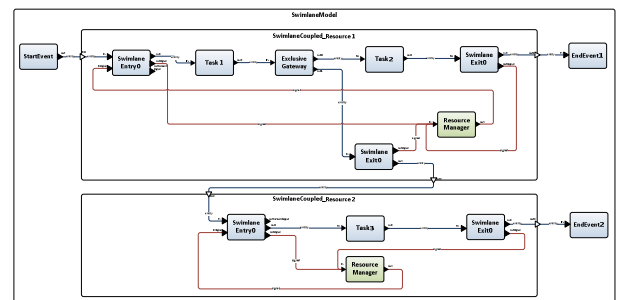


Figure 7: Auto generated DEVS Model.

6. DISCUSSION

This work proposed a new modelling approach for consultants to model and analyse business processes based on a proven theory, industry-wide standards and active end-user involvement during the design process.

A library of DEVS-based BPMN modelling elements is implemented with Java that uses the DSOL simulation library to provide the simulation capabilities. It provides an easy way of dynamic modelling for consultants with limited knowledge of simulation model development. As a future work, the credibility of our approach will be evaluated.

REFERENCES

- Aguilar-Savén, R.S., 2004. Business Process Modelling: Review and Framework. *International Journal of Production Economics*, 90(2), 129-149.
- Banks, J., 1998. *Handbook of Simulation - Principles, Methodology, Advances, Applications, and Practise* (p. 849). New York: John Wiley & Sons, Inc.
- Bosilj-Vuksic, V., Ceric, V. and Hlupic, V., 2007. Criteria for the evaluation of business process simulation tools. *Interdisciplinary Journal of Information, Knowledge, and Management*, 2, 73–88.
- Cetinkaya, D., Verbraeck, A. and Seck, M. D., 2010. Applying a Model Driven Approach to Component Based Modeling and Simulation. *Proceedings of the 2010 Winter Simulation Conference* (pp. 546-553). Baltimore, MD.
- Cetinkaya, D., Verbraeck, A. and Seck, M. D., 2011. MDD4MS: A Model Driven Development Framework for Modeling and Simulation. *Proceedings of the Summer Computer Simulation Conference 2011*. Den Haag, NL.
- Harrington, H., 1991. *Business Process Improvement: The Breakthrough Strategy for Total Quality Productivity*. New-York: McGraw Hill.
- Jacobs, P. H. M., 2005. *The DSOL Simulation Suite*. Thesis (PhD). Delft University of Technology.
- Maguire, M., 2001. Methods to support human-centred design. *International Journal of Human-Computer Studies*, 55(4), 587-634.
- Melão, N. and Pidd, M., 2003. Use of business process simulation: A survey of practitioners. *Journal of the Operational Research Society*, 54(1), 2-10.
- Peterson, J. L., 1981. *Petri Net Theory and the Modeling of Systems* (p. 290). New Jersey: Prentice Hall.
- Pidd, M., 1996. *Tools for Thinking: Modelling in Management Science*. Chichester, John Wiley & Sons, Inc.
- Pidd, M., 2002. Simulation Software and Model Reuse: A Polemic. *Proceedings of the 2002 Winter Simulation Conference*. (p. 772-775)
- Pidd, M., 2004a. Computer Simulation in Management Science (5th ed., p. 311). West Sussex, England: John Wiley & Sons, Inc.
- Pidd, M., 2004b. Simulation Worldviews – So What?. *Proceedings of the 2004 Winter Simulation Conference, 2004*. England: John Wiley & Sons, Inc.
- Seck, M.D. and Verbraeck, A., 2009. DEVS in DSOL: Adding DEVS Operational Semantics to a Generic Event-scheduling Simulation Environment. *Proceedings of the Summer Simulation Multiconference 2009*. Istanbul, Turkey
- Van Eijck, D. T. T. and De Vreede, G.-J., 1998. Simulation support for organizational coordination. *Proceedings of the 1998 Hawaiian Conference of Systems Sciences* (Vol. 1, p. 633–642). Los Alamitos: IEEE Computer Society.
- Weske, M., 2007. *Business Process Management: Concepts, Languages, Architectures* (p. 368). New York: Springer-Verlag.
- Zeigler, B. P., Praehofer, H. and Kim, T. G., 2000. *Theory of Modeling and Simulation* (Second Edi.). Academic Press.

AUTHORS BIOGRAPHY

IGOR J. RUST is an M.Sc. graduate in Systems Engineering, Policy Analysis and Management and received his degree in 2011 from Delft University of Technology. In 2011 he received his B.Sc. at the faculty of Technology, Policy and Management, also from Delft University of Technology. His B.Sc. thesis was nominated for the Dutch Logistics Thesis Award 2011. His interests include discrete event simulation and mathematical programming. His e-mail address is <i.j.rust@student.tudelft.nl>

DENIZ CETINKAYA is a Ph.D. student at Delft University of Technology. She is in the Systems Engineering Group of the Faculty of Technology, Policy and Management. She received her M.Sc. in Computer Engineering from the Middle East Technical University, Turkey in 2005. She received her B.Sc. with honours in Computer Engineering from the Hacettepe University, Turkey in 2002. Her research focuses on component based modelling and simulation. Her e-mail address is <d.cetinkaya@tudelft.nl>

MAMADOU D. SECK is an Assistant Professor in the Systems Engineering Group of the Faculty of Technology, Policy, and Management of Delft University of Technology. He received his Ph.D. degree from the Paul Cezanne University of Marseille and his M.Sc. and M.Eng. degrees from Polytech Marseille, France. His research interests include modelling and simulation formalisms, dynamic data driven simulation, human behaviour representation and social simulation, and agent directed simulation. His e-mail address is <m.d.seck@tudelft.nl>

IVO WENZLER is part of Accenture's Global Consulting Experts Group and was until recently a Senior Executive within Accenture's Talent and Organization Service Line. For one day per week he also holds a position of Associate Professor at Delft University of Technology where he teaches a master's course in simulation game design. His e-mail address is <ivo.wenzler@accenture.com>

RESEARCH ON CO-SIMULATION TASK SCHEDULING IN CLOUD SIMULATION PLATFORM

Chen Yang^(a), Bo Hu Li^(b), Xudong Chai^(c)

^(a)Beijing University of Aeronautics and Astronautics, Beijing, China

^(b)Beijing University of Aeronautics and Astronautics, Beijing, China

^(c)Beijing Simulation Center, Beijing, China

^(a)wzhyoung@163.com, ^(b)bohuli@moon.bjnet.edu.cn, ^(c)xdchai@263.net

ABSTRACT

In order to address the problem of single collaborative simulation task scheduling, considering the characteristics of simulation resource encapsulated with virtualization technology, the author first proposed the unified model describing the co-simulation system for co-simulation task, which is the basis of task scheduling; Secondly, based on the unified model, the virtualization-based supporting system, in cloud simulation platform (CSP), of dynamic construction of the co-simulation system were introduced. Thirdly, the scheduling procedure, namely the dynamic construction of the co-simulation system, was discussed. Finally, the primary application example and conclusion was presented.

Keywords: cloud simulation platform, HLA federation, co-simulation task scheduling, virtualization technology

1. INTRODUCTION

Recently, as an effective tool for understanding and reconstruction of the objective world, M&S theory, methodology, and technology have been well developed, form its own systematic discipline, and their application area are increasingly expanded. At the same time, they are developed toward “digitization, virtualization, networking, intelligence, integration, and collaboration”, which are considered as the characteristics of the modern trend.

As the application field of M&S continues to expand, the size and complexity of simulation application system are greatly increased, which poses a severe challenges to M&S technology. High level architecture provides a general framework and corresponding software engineering method “FEDEP” for developing large-scale distributed simulations, which can promote the reusability and interoperability of simulation model. However, HLA does not inherently take into account of resource management and task scheduling for co-simulation, especially the simulation resource is statically bound together with federates before simulation start, in which case the automatic scheduling is lacked. The combination of M&S and grid computing gives birth to simulation grid,

which realizes the dynamic share and reusability, collaborative interoperability, dynamic optimization for simulation execution, of different resource. Simulation grid has resource management and task scheduling for certain degree, but due to the heterogeneity of OS and software environment, the large variance of performance, of nodes in simulation grid, in addition to unstableness of network, the simulation grid can not effectively and quickly execute the large-scale simulation, and the effect of co-simulation task scheduling methods is decreased severely with little advantage. Moreover, simulation grid can not show efficient support for fine granular resource (e.g. CPU, storage, software in nodes) share, multi-user, fault-tolerance, etc. For example, in simulation grid, distributed computing nodes have different kinds and versions of OS and software, which will lead to the limited nodes the task can be scheduled to, such as the federates programmed and compiled to run in Windows OS is hard to scheduled to execute in Linux servers without any adaption.

Due to the unsolved problems in simulation grid, People in Beijing Simulation Center introduce the notion of “cloud computing”, and further with integration of virtualization, pervasive computing, and high performance computing technology, propose a networkized M&S platform-Cloud Simulation Platform (CSP) to enhance the ability of simulation grid. CSP employs virtualization technology to encapsulate the computing resource, software, simulation models, etc as virtualized simulation resource, masking the heterogeneity of resources. Based on encapsulated resource, virtualization technology can provide enabling technology for dynamic setting up of simulation execution environment, or even simulation system, which will benefit the co-simulation task scheduling and federates’ deployment.

Co-simulation task scheduling referred to in this paper focuses on one single task scheduling, in which several subsystems collaboratively run to accomplish the task, and the scheduling here is mainly about the resource scheduling process to construction virtualization-based simulation system. The scheduling in simulation runtime is not discussed in this paper.

Considering the requirement of co-simulation task in CSP, the author first proposes the top-level description model of collaborative simulation system for task, and then the supporting system for co-simulation task scheduling, and the whole procedure of scheduling are then presented. Lastly, the conclusion and further work are given. Without special declaration, the scheduling object in this paper refers to HLA-based co-simulation.

2. THE TOP-LEVEL DESCRIPTION MODEL FOR CO-SIMULATION SYSTEM

Co-simulation system is the execution entity for fulfilling the co-simulation task. The efficiency of its running directly determines the effect of simulation, which is the reason of co-simulation task scheduling. For the purpose of high efficient scheduling, we first build the unified top-level model for describing co-simulation system. The model contains the essential properties of co-simulation system, which determines the constitution of federation and interoperability between federates. The model provides solid foundations for dynamic construction of federation, which is the process of the scheduling.

Firstly, HLA-based co-simulation system can be described as follows by unified model:

Definition 1. *Co-simulation system* = $\langle Fed, Intera, Env, Comm, Comp, Num, StpCrit \rangle$

Definition 2. *Fed* = $\langle FM_i \mid 0 < i \leq Num \rangle$ is defined as the model of federation, where *FM* = $\langle DomainOnt, FuncDesc \rangle$ is defined as the model for describing federates. And $|Fed| = Num$ defines the number of federates in federation. *DomainOnt* is domain ontology-

based description of federate. *FuncDesc* denotes the semantic description of federate.

Definition 3. *Intera* = $\langle Src, Dest, Info \mid Src, Dest \in Fed \rangle$ is defined as the model for interactions between federates. In which *Src* refers to source federate generating information, *Dest* denotes the destination federate for interactive information, and *Info* is the corresponding information to be exchanged.

Definition 4. *Env* = $\langle EnvDepend(FM_i) \mid 0 < i \leq Num \rangle$ defines the running environment of each federate, including the requirement of OS and software sets. *EnvDepend*(*FM_i*) = $\langle OS, SoftSet \rangle$, in which *OS* refers to the type of operating system such as Windows XP, and *SoftSet* denotes the software list.

Definition 5. *Comm* = $\langle N_{net}(FM_i, FM_j) \mid 0 < i, j \leq Num, i \neq j \rangle$ defines the interaction requirement for communication capability of network between federates.

Definition 6. *Comp* = $\langle N_{compute}(FM_i) \mid 0 < i \leq Num \rangle$ defines the computation requirement of each federate.

Definition 7. *StpCrit* is defined as the condition for stopping simulation running. It can be the times of simulation execution, or the time of simulation running, or the bound of variables in simulation.

In fact, the process of dynamic construction of virtualization-based co-simulation system is the realization of the co-simulation task scheduling process. So, the supporting system for scheduling is given as follows.

3. THE VIRTUALIZATION-BASED SUPPORTING SYSTEM FOR DYNAMIC CONSTRUCTION OF CO-SIMULATION SYSTEM

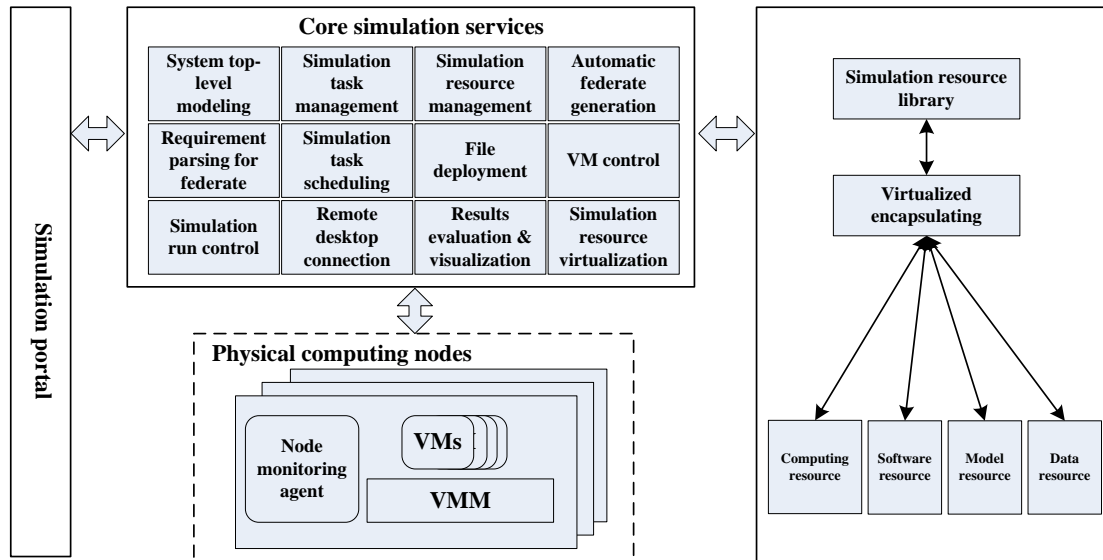


Figure 1: the framework of the supporting system

CSP employs virtualization technology to encapsulate the computing resource, software, simulation models, etc as virtualized simulation resource, which is stored in simulation resource library. Using encapsulated resource, based on the state information of each physical computing node collected by node monitoring

agent, the core simulation services in CSP can support the dynamic construction of co-simulation system.

CSP together with the virtualized simulation resource constitute Cloud Simulation System (CSS). The supporting system in CSS includes the logical functional modules shown in Figure 1. It consists of

three main parts: simulation portal, core simulation services, and simulation resource library.

- Simulation portal

Simulation portal is the entrance point of simulation activities for users. It supports the simulation activities based on internet or desktop, in which users can submit co-simulation task, acquire the results, etc.

- Core simulation services

Core simulation services are composed of services: system top-level modeling, simulation task management, simulation resource management, automatic federate generation, requirement parsing for federate, simulation task scheduling, file deployment, Virtual machine control, simulation run control, remote desktop connection, results evaluation and visualization, simulation resource virtualization, etc.

System top-level modeling service can provide the top-level modeling service for users based on internet or desktop, support to decompose the complex multidisciplinary tight coupled system into several sub-systems, and help to describe the models of federation and interactions between federates.

Simulation task management service enables the management of co-simulation task. Specifically, the services can support distributed simulation workers to accomplish the simulation task collaboratively.

Simulation resource management services support semantic-based simulation resource searching services, and also the downloading, uploading, revising, deleting, registering, publishing service, etc of simulation resource with certain permission.

Automatic federate generation services can generate the framework of federate according to the description model of federation. Moreover, it provides user-friendly interface to customize the framework, then user can finish functional entity development in the federate on the basis of the framework.

Requirement parsing for federate services can parse and acquire the configuration file of the federate, which includes the semantic description of the federate running environment.

Simulation task scheduling services provide the function to monitor the running state of physical machine, and use certain method to choose suitable physical computing nodes for the execution of virtual machines.

File deployment services give support to remotely deploy simulation related files to specified file path of the computer with certain IP address. These files together with the running environment constitute the executable simulation system.

VM control services lend support to the operation of the virtual machine: start, shutting down, suspending, and resuming the virtual machine. Further, it provides the user to access the virtual machines remotely by desktop connection for performing simulation activities.

Simulation run control services support the execution management of simulation federation, which includes creating federation execution, unified start of federate execution, monitoring federates' state,

synchronizing the logical time of federate. And the function of pausing, resuming, resigning, destroying federation execution, etc is also supported.

Simulation resource virtualization services include services of creating different kinds of virtualized simulation resource. Here, virtualized simulation resource mainly refers to virtual machines, into which the simulation resources are encapsulated.

- Simulation resource library

Simulation resource library is employed to store and manage the virtualized simulation resource, especially the management of VM pools, such as the searching of suitable VMs in library with semantic information.

4. THE PROCEDURE FOR THE DYNAMIC CONSTRUCTION OF CO-SIMULATION SYSTEM

The dynamic construction of co-simulation system is composed of three stages: system modeling stage, simulation model development stage, and simulation model deployment and assembling stage. Strictly the former two stages should not be included. But the former two stages are the basis for the scheduling, and for the aim of easy understanding, we list them out. The three stages will specify the content of each factor in unified model of simulation system. We do not pay much attention to the details of realization, and just focus on the principles in the procedure of federation dynamic construction.

The steps are as follows:

- System top-level modeling

Through the requirement analysis of the simulation task, the decomposition of complex research object is a must. The simulation system can be decomposed into several sub-systems in different domain. The decomposition principle is like this: after decomposition, the sub-systems each should have tight coupling inside and loose coupling with entities outside in order to wipe the bottleneck of communication brought by the irrational decomposition.

Using the system modeling services in simulation portal, based on graphical interface, users finish the decomposition of system, and build the description model of interactions between sub-systems. Then, according to certain semantic template, users should accomplish the description of federates (semantic-based domain ontology and function description). Users referred here are mainly chief simulation technology officers. Then, they create the co-simulation task by simulation task management services, with which they upload the files generated in system modeling process.

In this step, the factors *Fed*, *Intera*, *Num*, and *StpCrit* in unified description model are instantiated.

- Searching for simulation models based on the description of federates

System high-level modeling gives the description of different domain federates, including ontology based domain description and function description. Using

simulation task management services, simulation practitioners in different domains first get the federate description. Then according to the description, practitioners search for the professional models of entities using simulation resource management services (or develop professional models on their own). The models are downloaded. However, this is not enough, because normally, the professional models cannot be used directly in HLA-based collaborative simulation.

- Automatic federate generation services for HLA-based federate development

The professional practitioners employ the system top-level model to generate the framework of each federate using automatic federate generation services. Then based on the federate framework and the professional models, the complete federate which is ready for simulation execution can be developed. Finally, the professional practitioners build the requirement description of federate in these aspects: the running environment, computation and communication requirement, by referring to the description of professional models.

In this step, the factors *Env*, *Comm*, and *Comp* in unified description model *are* instantiated.

- Submitting the co-simulation task

The practitioners who are responsible for the subsystem development, upload federates and description files to corresponding containers in simulation task management services. The top-level model of co-simulation system, federate models and respective description models together constitute the basis of co-simulation system. After checking that all the subsystems have been well finished, the chief simulation technology officers will submit the simulation task using simulation task management services.

- Parsing the description files of the simulation task

The requirement parsing for federate services can acquire the running environment of each federate from the configuration files of respective federate, in which the requirement of executing federate is contained, namely:

$EnvDepend(FM_i) = \langle OS, SoftSet \rangle$, where FM_i denotes federate, *OS* refers to the type of operating system, *SoftSet* represents the software list needed. For the requirement of the whole federation, *Env* contains the requirement information of all federates.

The running environment information of each federate is used to searching for virtual machine image for each federate.

- Searching for the virtual machine image based on the information of federate running environment

The virtual machine image pool, which is a part of simulation resource library, is a collection of virtual machine images. The virtual machine image pool is located on shared storage. Users can create virtual machine images via simulation resource virtualization services, and upload them to the virtual machine image

pool, at the same time, register and publish the semantic description of them via simulation resource management services. Each virtual machine image is a kind of resource that can be shared and reused with different running environment. So, a large number of virtual machine images in the pool can meet the demand of most users. If not, users can create their own ones based on their special requirement.

Simulation resource management services can search virtual machine pools for suitable one in accordance to the semantic description of each federate's requirement. If there exist more than one virtual machines in the result, then users can click the virtual machines listed to check the detailed information to choose the most suitable ones.

- Start virtual machines on suitable physical computing nodes

The simulation scheduling services support the state monitoring service of physical computing nodes, such as the CPU utilization, the available Memory, and the bandwidth, the network delay between them. Further, these services can help gather the statistics of historic monitoring information and forecast the load of each physical computing node.

In order to realize the load forecasting function, the exponential smoothing algorithm is employed to forecast the load of physical computing nodes, then select suitable nodes to start virtual machines on. The simulation task scheduling services can optimally choose several physical nodes using different algorithms to guarantee enough computing and fast communication capability. Or users can search for physical nodes according to their configuration description, check the historical statistics of monitoring information of them, and then choose several ones manually.

VM control services can support the control function of virtual machines on physical nodes, for example start, powering off, suspending, and resuming virtual machines. VM control services first automatically configure the CPU, memory, storage, and network of virtual machines. And then start the virtual machines in chosen physical nodes.

- Remote connection to get the desktop of virtual machines

After the construction of virtual computing environment, the following demands show that the remote desktop connection services are needed.

1) Before the simulation system running, special requirement leads to changing the configuration of simulation software, or even the OS.

2) According to the demand of interactions between users and simulation system, users need the desktop of virtual machines to implement more sophisticated control on simulation process.

3) In order to guarantee the correctness of simulation results, not only the state monitoring of simulation system is need, but also whether the exceptions are thrown should be pay attention to, because some kinds of exceptions will not lead to the

crackdown of simulation system, but will affect the correctness of simulation results.

The remote desktop connection services can give support to getting the virtual desktop of virtual machines. After the start of virtual machines, based on these services, users can use the virtual desktop just like it is the local desktop of the physical machine. In other words, users can start and configure simulation software, and build the running environment of federates. Users can also check whether federates have thrown exceptions.

- Deploying model files and start the execution of simulation system to accomplish the dynamic construction of simulation system

File deployment services can support deployment of files to specified path of machines with designated IP address. Simulation system consists of simulation models and simulation running environment. After the dynamic construction of simulation running environment, the only need is to deploy and configure simulation models (federates). Then the simulation run control services are utilized to create federation execution, and then start federates deployed in virtual machines to join the federation execution. Other simulation run control services such as monitoring the state of federation, control the execution of federation are also provided to uniformly perform the simulation experiment.

5. APPLICATION EXAMPLE

The aforementioned virtualization-based dynamic construction of co-simulation system has been primarily applied in multidisciplinary virtual prototype, large-scale system collaborative simulation, and high performance simulation areas. This paper presented an application example of aircraft landing gear virtual prototype collaborative simulation system. The main steps are as follows:

- System top-level modeling

Using desktop virtualization technology, the virtual user interface of software can be acquired through select the desired software, which is shown in Figure 2. Simulation practitioners first get the requirement of simulation task, and then build the system top-level description model. Specifically, the aircraft landing gear simulation system can be decomposed into undercarriage control model, undercarriage multi-body dynamics model, undercarriage hydraulics model, etc, as shown in Figure 3. Each model here is one federate in the federation. And the interactions between these subsystems are built. Then the HLA FOM files and the ontology-based and function description files of professional models are generated. Simulation practitioners use the simulation task scheduling services to create the simulation task by submitting the system top-level description files, as shown in Figure 4.

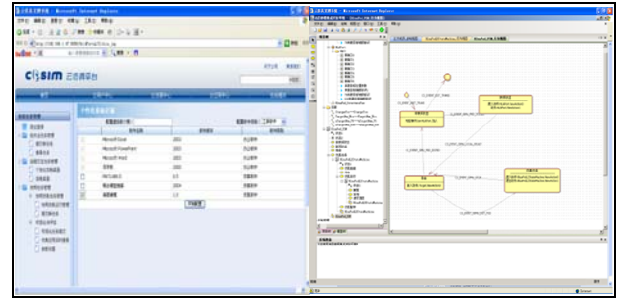


Figure 2: system top-level modeling

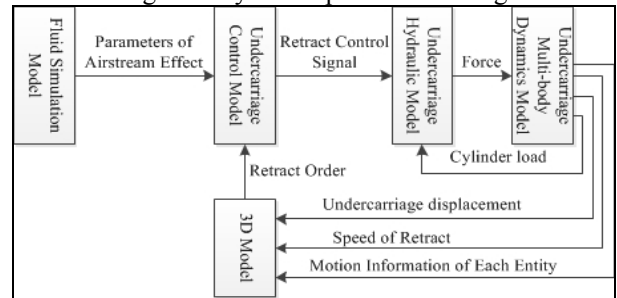


Figure 3: the decomposition of aircraft landing gear simulation system

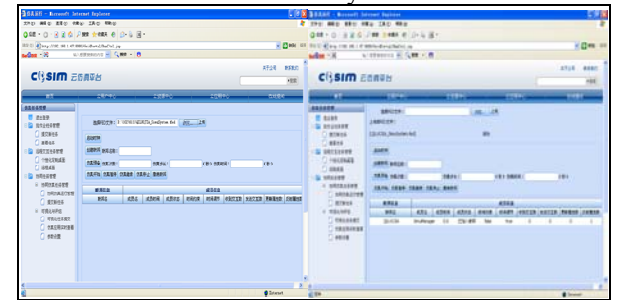


Figure 4: creation of co-simulation task

- Completing the development of federates on the basis of searching for domain simulation models

Domain simulation practitioners can obtain the system top-level description model files via the simulation task management services. The suitable models are listed by searching the domain models in simulation resource library. Then according to the detailed semantic description of listed domain models, domain simulation practitioners download the ones which meet the requirement. With the help of automatic federate generation services, all federates are then developed using these domain models. Finally, referring to domain model descriptions of required running environment, computing and communication capability, the requirement of federates in co-simulation system are demonstrated as follows in table 1.

Table 1: The requirement of federates

No.	Simulation model (federate)	Software Environment	Computing Environment	Performance Demand
1	Fluid Simulation Model	Fluent	Redhat	CPU 8core,

				memory 4GB
2	Undercarriage Control Model	Matlab	Windows	CPU 1core, memory 1GB
3	Undercarriage Hydraulic Model	Easy5	Windows	CPU 1core, memory 1GB
4	Undercarriage Multi-body Dynamics Model	Adams	Windows	CPU 1core, memory 1GB
5	3D Model	CATIA	Windows	CPU 2core, memory 2GB

In which, Gigabit Ethernet and 80GB storage space is fixed for each node.

- Submitting the simulation task and constructing the simulation running environment

Federates developed by domain simulation practitioners are submitted using simulation task management services. The chief simulation technology officers submit the simulation task. The supporting system for dynamic construction of co-simulation system in CSP will parsing the requirement information of each federate, search for five suitable virtual machines images (realized using Xen middleware), in which different simulation software are included. CSP will select five physical computing nodes simply using weight sorting method fusing the factors of computation and communication capabilities. Then the virtual machine images are deployed to physical computing nodes according to the same weight sorting method and started with the demand configuration of federates. Figure 5 is the interface of simulation task submitting and dynamic building of simulation running environment.

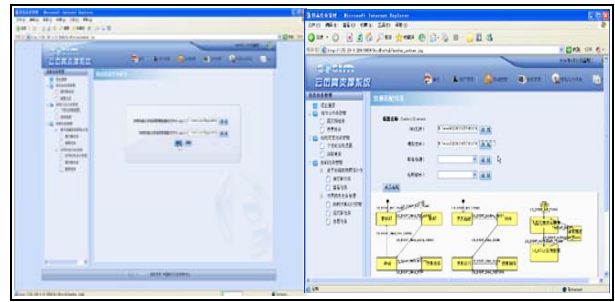


Figure 5: submitting simulation task and dynamic building of simulation running environment

- The dynamic construction of co-simulation system

After the creation of simulation running environment, the chief simulation technology officers can acquire the virtual desktop of virtual machines by remote desktop connection services. The virtual desktop of virtual machine with “matlab” software inside is shown in Figure 6. The files of each federate are then deployed to certain path in virtual machines. With the help of remote virtual desktop, users can configure OS and software. Here we just let the local RTI component point to central component with certain IP address. Finally, simulation run control services are used to start the co-simulation system and monitoring its running state, as shown in Figure 7.



Figure 6: remote virtual desktop

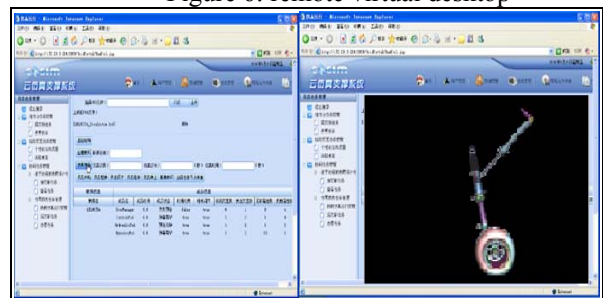


Figure 7: monitoring and controlling the co-simulation system

6. CONCLUSION AND FURTHER WORK

Through introducing virtualization technology, the author presents the supporting system and procedure for simulation task scheduling, and the primary application example. The primary application shows that: virtualization-based dynamic construction of federation can address the constrained scheduling problem caused by the tight coupling between the simulation system and physical computing resources, which will make the co-

simulation task automatically scheduled to certain degree.

Future work includes as follows:

1) Further research on the semantic-based unified description model and share mechanism of virtualized simulation resource.

2) Further research on high performance collaborative simulation technology to support high efficient execution of simulation system.

3) Further research on the scheduling method when facing a large number of co-simulation tasks.

ACKNOWLEDGMENTS

This paper is supported by National Defense Key Lab, the National 973 Plan (2007CB310900) and National Defense Pre-research Foundation of China. And authors should like to express the sincere thanks to all colleagues for their help and valuable contribution.

REFERENCES

- Bo Hu Li, et al., 2009. A network modeling and simulation platform based on the concept of cloud computing --- "Cloud Simulation Platform", *Journal of System Simulation*, 12 (17), 5292-5299.
- Hai Jin et al., 2008. *Computing system virtualization---theory and application* (in Chinese). Tsinghua press.
- Bo Hu Li, Xudong Chai, Baocun Hou, 2009. Cloud Simulation Platform. *Proceedings of the 2009 International Summer Simulation Conference*, Turkish, Istanbul.
- Z. Li, W. Cai, S. J. Turner, and K. Pan, 2007. Federate migration in a service oriented HLA RTI, *Procs of international Symposium on Distributed Simulation and Real-Time Applications*, 113-121
- Bo Hu Li, Xudong Chai, Yanqiang Di, et al., 2005. Research on service oriented simulation grid. *International Symposium on Autonomous Decentralized Systems (ISADS 05)*, pp. 7 – 14.
- Bo Hu Li, Xudong Chai, Baocun Hou, 2006. Research and Application on CoSim (Collaborative Simulation) Grid. *The Proceeding of MS-MTSA'06*.
- Bo Hu Li, Xudong Chai, Wenhai Zhu, 2004. Some focusing points in development of modern modeling and simulation technology. *Journal of System Simulation*, 16 (9), 1871-1878(in Chinese).
- IEEE, 2000. Standard 1516 (HLA Rules), 1516.1 (Federate Interface Specification) and 1516.2 (Object Model Template).
- S.A.Herrod, 2006. The Future of Virtualization Technology, *Keynotes of ISCA 2006*, <http://www.ece.neu.edu/conf/isca2006/docs/Herrod-keynote.pdf>.

AUTHORS BIOGRAPHY

Chen Yang was born in 1987. He received his B.S. degree in Beihang University. He is currently a Ph.D. candidate of Beihang University, Beijing, China. Research focuses on advanced distributed simulation,

distributed computing, HLA-RTI, theory and practice of modeling, etc.

Xudong Chai was born in 1969. He is a researcher and deputy director at Beijing Simulation Center of Second Academy of Aerospace Science & Industry Co. and council members of Chinese System Simulation Association and National Standardization Technical Committee. His research interests include automatic control and simulation.

Bo Hu Li was born in 1938. He is a professor at School of Automatic Science and Electrical Engineering, BeiHang University, and Chinese Academy of Engineering, and the chief editor of "Int. J. Modeling, Simulation, and Scientific Computing". His research interests include multi-disciplinary virtual prototype, intelligent distributed simulation and cloud manufacturing.

AN OPTIMAL NON-BLOCKING DISPATCHING IN FREE-CHOICE MANUFACTURING FLOWLINES BY USING MACHINE-JOB INCIDENCE MATRIX

Ivica Sindičić ^(a), Stjepan Bogdan ^(b), Tamara Petrović ^(b)

^(a) ABB, Bani 72, 10000 Zagreb, Croatia

^(b) LARICS - Laboratory for Robotics and Intelligent Control Systems, Department of Control and Computer Engineering, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia

^(a)ivica.sindicic@gmail.com, ^(b)stjepan.bogdan@fer.hr

ABSTRACT

In this paper we extend a method for resource allocation in particular class of flexible manufacturing system, namely, free-choice multiple reentrant flowlines (FMRF), which is based on MJI matrix. The proposed method is a solution to the problem on how to allocate jobs to resources and how to allocate resources to jobs. A solution is in a form of repeatable resource sequence over the set of resources available for particular choice job. As addition proposed in this paper, the solution is enhanced by the procedure that provides an optimal utilization of resources based on operation price and balanced use of all resources. Although efficiency of the proposed methods have been demonstrated on examples involving manufacturing workcells, the method can be used for other discrete event systems as well, as long as the system under study belongs to free-choice multiple reentrant flowlines class.

Keywords: dispatching, manufacturing systems, optimal control

1. INTRODUCTION

The first step in the supervisory controller design is modeling of the system and investigation of its structural properties. There are many approaches to modeling and analysis of manufacturing systems, including automata [1], Petri nets [2, 10], alphabet-based approaches, perturbation methods [3], control theoretic techniques, expert systems design, and so on.

One way to model relations between tasks in an FMS is in form of Steward sequencing matrix [4], also referred to as design structure matrix (DSM). DSM is a square matrix containing a list of tasks in rows and columns. The order of tasks in rows or columns indicates the execution sequence. Although very useful in production planning, DSM lacks of information related to the resources required for execution of tasks. This aspect of an FMS is captured by the resource requirements matrix [5], also known as the machine-part incidence matrix (MPI). Each column of MPI represents one resource, while rows represent part types processed by the system. The most common usage of MPI is in the field of manufacturing cells design by implementation

of various clustering methods [6]. In [17] we proposed construction of machine-job incidence matrix (MJI) which can be obtained from MPI and DSM matrices.

Efficient procedures for determination of simple circular waits (CWs) [7, 8] as well as other important structural properties (which are responsible for stability in the sense of absence of deadlock), such as critical siphons and critical subsystems [9, 11], based on MJI, are presented in [13]. It should be noted that MJI matrix can be straightforwardly transformed in matrix model described in [9].

Generally, scheduling requires a) allocation (dispatching) of available resources to predetermined operations (tasks), and b) definition of sequences in order to provide stable behavior of the system. Usually, supervisory controller not only stabilizes the system (in a sense of deadlock and bounded buffers) but in the same time optimizes some performance criteria.

Herein we extend a method for resource allocation in particular class of FMS, namely, free-choice multiple reentrant manufacturing systems (FMRF), which is based on MJI matrix, described in details in [17]. A solution is in a form of repeatable resource sequence over the set of resources available for particular choice job. As addition, proposed in this paper, the solution is enhanced by the procedure that provides an optimal usage of resources based on price and balanced use of all resources.

2. SYSTEM DESCRIPTION AND PROBLEM FORMULATION

We make the following assumptions that define the sort of discrete-part manufacturing systems: No pre-emption – once assigned, a resource cannot be removed from a job until it is completed, Mutual exclusion – a single resource can be used for only one job at a time, Hold while waiting – a process holds the resources already allocated to it until it has all resources required to perform a job. Furthermore, we assume that there are no machine failures. Multiple reentrant flowlines (MRF) class of systems, investigated herein, has the following properties: a) each part type has a strictly defined sequence of operations, b) each operation in the system requires one and only one resource with no two

consecutive jobs using the same resource, c) there are no choice jobs, d) there are no assembly jobs, e) there are shared resources in the system.

2.1. System description

Let Π be the set of distinct types of parts produced (or customers served) by an FMS. Then each part type $P_k \in \Pi$ is characterized by a predetermined sequence of job operations $J^k = \{J_1^k, J_2^k, J_3^k, \dots, J_{L_k}^k\}$ with each operation employing at least one resource. (Note that some of these job operations may be similar, e.g. J_i^k and J_j^k with $i \neq j$ may both be drilling operations.) We uniquely associate with each job sequence J^k the operations of raw part-in, J_{in}^k , and completed product-out, J_{out}^k .

Denote the system resources with $R = \{r_i\}_{i=1}^n$, where $r_i \in R$ can represent a pool of multiple resources each capable of performing the same type of job operation. In this notation $R^k \subset R$ represents the set of resources utilized by job sequence J^k . Note that $R = \bigcup_{k \in \Pi} R^k$ and

$J = \bigcup_{k \in \Pi} J^k$ represent all resources and jobs in a particular FMS. Since the system could be re-entrant, a given resource $r \in R^k$ may be utilized for more than one operation $J_i^k \in J^k$ (*sequential sharing*). Also, certain resources may be used in the processing of more than one part-type so that for some $\{l, k\} \in \Pi$, $l \neq k$, $R^l \cap R^k \neq \emptyset$ (*parallel sharing*). Resources that are utilized by more than one operation in either of these two ways are called *shared resources*, while the remaining are called *non-shared resources*. Thus, one can partition the set of system resources as $R = R_s \cup R_{ns}$, with R_s and R_{ns} indicating the sets of shared and non-shared resources, respectively. For any $r \in R$ we define the *resource job set* $J(r)$. Obviously, $|J(r)| = 1 (> 1)$ if $r \in R_{ns}$ ($r \in R_s$). Resource r , with its job set $J(r)$, comprise *resource loop* $L(r)$, $L(r) = r \cup J(r)$.

We define a *job vector* $\mathbf{v} : J \rightarrow \mathfrak{N}$ and a *resource vector* $\mathbf{r} : R \rightarrow \mathfrak{N}$ that represent the set of jobs and the set of resources corresponding to their nonzero elements. The set of jobs (resources) represented by \mathbf{v} (\mathbf{r}) is called the *support* of \mathbf{v} (\mathbf{r}), denoted $sup(\mathbf{v})$ ($sup(\mathbf{r})$); i.e. given $\mathbf{v} = [v_1 \ v_2 \ \dots \ v_q]^T$, vector element $v_i > 0$ if and only if job $v_i \in sup(\mathbf{v})$. In the same manner, given $\mathbf{r} = [r_1 \ r_2 \ \dots \ r_p]^T$, vector element $r_i > 0$ if and only if resource $r_i \in sup(\mathbf{r})$. Usually, index i is replaced with job (resource) notation, for example, r_{MA} stands for the component of resource vector r that corresponds to resource MA. The definitions of job and resource vectors imply that the job and resource sets should be ordered.

MRF class is a special case of FMRF - systems with jobs that do not have predetermined resources

assigned. That is, several resources might be capable and available to perform a specific job (MRF property c is not valid, i.e. there are jobs with choice). We define $R(J_i^k)$ as a set of resources that could be allocated to choice job $J_i^k \in J^k$.

An example of FMRF workcell is given in Figure 1. with $J = \{RP1, BP, MP, RP2\}$ and $R = \{M1, M2, B, R\}$. From buffer (job BP), part proceeds to machine M1 or machine M2 (choice job MP). Hence, vector representation of resources that could be allocated to choice job MP is $\mathbf{r}_{MP} = [1 \ 1 \ 0 \ 0]^T$ and $R(MP) = sup(\mathbf{r}_{MP}) = \{M1, M2\}$.

2.2. Problem formulation

Since the system contains shared resources and choice jobs, the scheduling problem discussed in the paper is twofold: i) for a given choice job $J_i^k \in J^k$ define *allocation sequence* of resources in $R(J_i^k)$, and ii) for a given shared resource $r_s \in R_s$ with resource job set $J(r_s)$, define *dispatching policy*. Both solutions, allocation sequence and dispatching policy, should be such that overall system is stable in a sense of deadlock.

A solution of problem i) offers an answer on how to *allocate resources to jobs*. For that purpose we propose a result in a form of *repeatable resource sequence* over the set of resources available for particular choice job.

On the other hand problem ii) is related to the number of active jobs in a particular parts of FMRF systems, called *critical subsystems*. In the chapters that follow we show why critical subsystems are important, how they can be determined from MJI matrix, and how their content (number of active jobs) can be controlled. In fact, solution to problem ii) describes how to *allocate jobs to resources*.

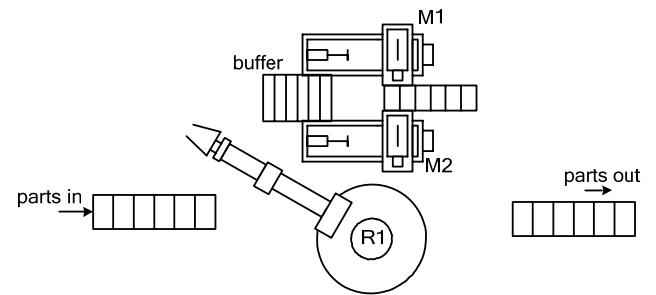


Figure 1: An example of FMRF class of FMS

3. MACHINE-JOB INCIDENCE MATRIX (MJI)

As we stated in Introduction, DSM is a square matrix containing a list of tasks in rows and columns with matrix elements indicating an execution sequence. The second matrix used for the system description is MPI. It captures relations between resources and parts processed by the system. Since the sequence $J^k = \{J_1^k, J_2^k, J_3^k, \dots, J_{L_k}^k\}$ represents part P_k processing

order, by combining DSM and MPI matrices, we get a general form of machine-job incidence matrix Λ for an FMRF system [17]. In case job i is performed by resource j , matrix element (i, j) is equal to '1', otherwise is '0'. For an MRF system, each operation in the system requires only one resource (there are no choice jobs), hence, exactly one element '1' would appear in each row of MJI matrix. On the other hand, column representing shared resource comprises multiple entries of '1'.

$$\Lambda = \begin{matrix} & R_1 & R_2 & \dots & R_q \\ \begin{matrix} J_1^1 \\ J_2^1 \\ \vdots \\ J_L^1 \\ J_1^2 \\ \vdots \\ J_L^2 \\ \vdots \\ J_1^m \\ \vdots \\ J_L^m \end{matrix} & \begin{vmatrix} \mathbf{0/1} & \mathbf{0/1} & \dots & \mathbf{0/1} \\ \mathbf{0/1} & \mathbf{0/1} & \dots & \mathbf{0/1} \\ \vdots & \vdots & & \vdots \\ \mathbf{0/1} & \mathbf{0/1} & \dots & \mathbf{0/1} \\ \mathbf{0/1} & \mathbf{0/1} & \dots & \mathbf{0/1} \\ \vdots & \vdots & & \vdots \\ \mathbf{0/1} & \mathbf{0/1} & \dots & \mathbf{0/1} \\ \vdots & \vdots & & \vdots \\ \mathbf{0/1} & \mathbf{0/1} & \dots & \mathbf{0/1} \\ \vdots & \vdots & & \vdots \\ \mathbf{0/1} & \mathbf{0/1} & \dots & \mathbf{0/1} \end{vmatrix} \end{matrix}$$

Machine-job incidence matrix can be defined separately for each part type in an FMS. In that case overall MJI matrix can be written as:

$$\Lambda = \begin{bmatrix} {}^1\Lambda^T & {}^2\Lambda^T & \dots & {}^m\Lambda^T \end{bmatrix}^T \quad (1)$$

For the system given in Figure 2. MJI attains the following form:

$$\Lambda = \begin{matrix} & M1 & M2 & B & R \\ \begin{matrix} RP1 \\ BP \\ MP \\ RP2 \end{matrix} & \begin{vmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{vmatrix} \end{matrix}$$

It can be seen that robot R is shared resource as the corresponding column has two elements equal to '1'.

As we demonstrate in [17] one of the benefits provided by MJI matrix is reduction of computational complexity in FMRF system analysis and simulation.

4. MJI AND RESOURCE SEQUENCING

As already mentioned, in addition to the assumptions made at the beginning of Chapter II, a general class of FMRF systems has the following nonrestrictive capabilities: i) some jobs have the option of being machined in a resource from a set of resources (allocation of jobs), ii) job/part routings are NOT deterministic, iii) for each job there exists a material handling buffer (routing resource) that routes parts.

In this Chapter we are interested to determine repeatable resource sequence over the set of resources available for execution of each choice job in the system. The sequence should prevent conflicts and deadlocks simultaneously. In the analysis that follows we consider one part-type regular FMRF with each buffer capable of holding one part at a time.

Let MJI matrix of the system is given as

$$\Lambda = \begin{matrix} & \begin{matrix} J1 \\ JB1 \\ J2 \\ JB2 \\ J3 \\ JB3 \\ J4 \\ JB4 \\ J5 \end{matrix} \\ \begin{matrix} M1 & M2 & M3 & M4 & M5 & B1 & B2 & B3 & B4 \end{matrix} & \begin{vmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{vmatrix} \end{matrix} \quad (2)$$

From the matrix we see that a part production requires sequence of 5 jobs, and system is composed of 5 machines and 4 buffers. For choice jobs J2, J3 and J4, we have $R(J2)=\{M2, M3\}$, $R(J3)=\{M3, M4, M5\}$, and $R(J4)=\{M2, M4\}$. All machines, except M5, are shared resources. There are many part routes that complete the required job sequence - to mention just few of them:

$\sigma_1=\{M1 \rightarrow M2 \rightarrow M3 \rightarrow M2 \rightarrow M1\}$,
 $\sigma_2=\{M1 \rightarrow M3 \rightarrow M5 \rightarrow M4 \rightarrow M1\}$,
 $\sigma_3=\{M1 \rightarrow M3 \rightarrow M3 \rightarrow M4 \rightarrow M1\}$, and so on. We partition set of part routes, denoted Σ , into two disjoint sets, $\Sigma = \Sigma_R \cup \Sigma_{NR}$, with Σ_R comprising all part routes with multiple use of resources allocated to choice jobs, and Σ_{NR} containing all part routes without multiple use of resources. In our example σ_1 and σ_3 belong to Σ_R , while σ_2 is an element of Σ_{NR} .

It is obvious that conflicts might occur in $\sigma_i \in \Sigma_R$ since same resource is used for execution of more than one job. In σ_1 resource M2 is used for jobs J2 and J4. On the other hand part routes in Σ_{NR} are inherently conflict free. We define resource sequences as combination of several routes from Σ_{NR} . As a result, the structure of repeatable resource sequences over the set of resources available for execution of choice jobs would, by itself, provide not only conflict free but also deadlock free behavior of the system.

The number of all possible part routes, $N = |\Sigma|$, is

$$\text{defined as } N = \prod_{i=1}^n \left(\sum_{j=1}^m \Lambda_{i,j} \right) \quad \text{where } n \text{ and } m$$

correspond to the number of rows and columns of Λ , respectively. In our example one has $N = 1 \cdot 1 \cdot 2 \cdot 1 \cdot 3 \cdot 1 \cdot 2 \cdot 1 \cdot 1 = 12$. For each part route σ_i we define *MJI sub-matrix* Λ^i in a way that in case of multiple entries '1' in row (choice job), sub-matrix comprises only the one that corresponds with resource belonging to the part route. For $\sigma_2=\{M1 \rightarrow M3 \rightarrow M5 \rightarrow M4 \rightarrow M1\}$ MJI sub-matrix attains a form

$$\mathbf{\Lambda}^2 = \begin{matrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} & \begin{matrix} J1 \\ JB1 \\ J2 \\ JB2 \\ J3 \\ JB3 \\ J4 \\ JB4 \\ J5 \end{matrix} \\ \begin{matrix} M1 & M2 & M3 & M4 & M5 & B1 & B2 & B3 & B4 \end{matrix} \end{matrix}$$

Having defined sub-matrices, search for sequences in Σ is in fact search for all MJJ sub-matrices that are characterized by single entry '1' in each row and multiple or no entry '1' in each column. It is easy to show that there exists a procedure of complexity $O(N)$ for calculation of such matrices.

Resource sequences are related to choice jobs, for this reason, we introduce *reduced form* of MJJ sub-matrices, denoted $\mathbf{\Lambda}^{*i}$, such that encompass only rows corresponding to those jobs, and without columns related to the buffers. Reduced form of $\mathbf{\Lambda}^2$ from previous example is given as

$$\mathbf{\Lambda}^{*2} = \begin{matrix} \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} & \begin{matrix} J2 \\ J3 \\ J4 \end{matrix} \\ \begin{matrix} M1 & M2 & M3 & M4 & M5 \end{matrix} \end{matrix}$$

Now, let us suppose that resource allocation policy requires that resources M3 and M2 should be used repetitively, one after the other, for execution of job J2. This repeatable resource allocation sequence can be written in a form of matrix

$$\mathbf{S}_{J2} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix},$$

$M1 \quad M2 \quad M3 \quad M4 \quad M5$

or in general form

$$\mathbf{S}_j = \begin{bmatrix} s_j^1 \\ s_j^2 \\ \vdots \\ s_j^w \end{bmatrix}. \quad (3)$$

Resource allocation sequence matrix should be defined for each choice job in the system. Our goal is to find set $\mathfrak{S} = \{\mathbf{S}_j\}$ (where $w = |\mathfrak{S}|$ equals to the number of choice jobs in the system) such that system has no conflicts and it is deadlock free. In [17] it has been proven that elements of such set, *i.e.* sequence matrices, are formed of rows of MJJ sub-matrices.

As an example, let us examine 3-step sequence for system (2), defined by the following resource allocation sequence matrices,

$$\mathbf{S}_{J2} = \begin{bmatrix} s_{J2}^1 \\ s_{J2}^2 \\ s_{J2}^3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{S}_{J3} = \begin{bmatrix} s_{J3}^1 \\ s_{J3}^2 \\ s_{J3}^3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\mathbf{S}_{J4} = \begin{bmatrix} s_{J4}^1 \\ s_{J4}^2 \\ s_{J4}^3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

In order to get a better insight in the system behavior, Figure 2. presents how parts pass through the line. It should be noted that sequences are executed in a way that new allocation of resources (new step) is performed after all jobs are finished and parts reside in the buffers. First part enters the system and it is processed by M3 (J2M3-'1' in the first row of \mathbf{S}_{J2}), than, it proceeds to job J3 on M4 (J3M4-'1' in the first row of \mathbf{S}_{J3}), and than to job J4 on M2 (J4M2-'1' in the first row of \mathbf{S}_{J4}). The second part enters the system and it is processed by M3 (J2M3-'1' in the second row of \mathbf{S}_{J2}), than, it proceeds to job J3 on M5 (J3M5-'1' in the second row of \mathbf{S}_{J3}), and so on. Conflict occurs for the first time at $k+2$ as J4 on part 1 is planned for M2 while, in the same time, the third part, that just entered the system, requires the same machine (J2 on M2).

This example clearly demonstrates that repeatable resource sequences can lead the system in conflict. In [17] it has been shown how to determine conflict-free set $\mathfrak{S} = \{\mathbf{S}_j\}$. Furthermore, we proved that usage of conflict-free set of resource allocation sequences not only resolves possible conflicts in the system, but also has direct consequence on the system stability, *i.e.* it provides deadlock-free behavior of the system.

part \	k	$k+1$	$k+2$	$k+3$	$k+4$	$k+5$...
1	J2M3	J3M4	J4M2				
2		J2M3	J3M5	J4M2			
3			J2M2	J3M3	J4M4		
4				J2M3	J3M4	J4M2	
...					

Figure 2: Presentation of parts passing through the line.

4.1. Sequence optimization

In order to determine an optimal sequence we introduce cost matrix $\mathbf{\Lambda}_w$. The cost matrix, with the form identical to MJJ matrix $\mathbf{\Lambda}$, captures cost of execution of i^{th} job by using j^{th} resource. By using cost matrix we can extract cost of each MJJ sub matrix as

$$C^p = \sum_{i=1}^n \sum_{j=1}^n (\mathbf{\Lambda}_w^p)_{ij}, \quad (4)$$

where $\Lambda_w^p = \Lambda_w \cdot (\Lambda^p)^T$, $\Lambda^p \in \Sigma_{NR}$, $p = 1..k$, is an MJI submatrix. If one assumes that sequence matrices are comprised of ω rows (sequence of ω repeatable steps), where each MJI submatrix Λ^p will be used ω_p times, *i.e.* $\omega = \sum_{p=1}^k \omega_p$, $\omega_p \in \mathbb{N}$, $0 \leq \omega_p \leq \omega$, then the total cost generated by those sequences is

$$C^{tot} = \sum_{p=1}^k \omega_p C^p. \quad (5)$$

It is clear that minimization of the total cost, defined as (5), is trivial problem – one should use only Λ^p with the smallest C^p in order to achieve minimal cost. However, in that case it might happen that utilization of the system resources would be highly unbalanced or even some resources would not be used at all. Hence, the cost function should be extended with a relation that captures resources utilizations. For each MJI sub-matrix one can define a *resource utilization vector* as

$$\mathbf{u}^p = \mathbf{1}^T \cdot \Lambda^p, \quad (6)$$

where $\mathbf{1}_{m \times 1}$ is vector with all elements equal to 1. As a result resource utilization vector \mathbf{u}^i is a binary vector with j^{th} element equal to 1 if corresponding resource participates in execution of the sequence containing rows of Λ^i sub-matrix. Finally, an integer vector that represents overall usage of the system resources is determined as

$$\mathbf{u} = \sum_{p=1}^k \omega_p \mathbf{u}^p. \quad (7)$$

Now, the second objective, balanced usage of the system resources, can be defined in the following form

$$\frac{(1-\varepsilon)}{(1+\varepsilon)} \leq \frac{u_i}{u_j} \leq \frac{(1+\varepsilon)}{(1-\varepsilon)}, \quad \forall i, j = 1..q, \quad (8)$$

where u_i and u_j are i^{th} and j^{th} component of \mathbf{u} , ε is design parameter such that $0 \leq \varepsilon < 1$, and q is the number of resources that should be balanced (for $q = m$ all resources in the system shall be included in optimization). Fully balanced utilization of resources is achieved if $\frac{u_i}{u_j} = 1, \forall i, j = 1..q$. However, this goal

might be very difficult (in some cases even impossible) to obtain, which depends on the system structure and executed sequence. Hence, by introducing parameter ε one is able to relax rigorous balancing requirement – for $\varepsilon \approx 1$ the system could become unbalanced, while for $\varepsilon = 0$ one requires full balance of the system resources exploitation.

Minimization of (5) by varying ω_p , $p = 1, \dots, k$ under conditions (8) with predefined ε and ω is a

mixed integer linear programming problem which can be solved using standard algorithms.

4.2. Case study

The proposed method has been tested on the system presented in [15] and [16]. Although, this example has only two choice jobs and it is comprised of MRF and FMRF sub-systems, it has been chosen so that the proposed method can be compared with various control techniques already implemented on this particular manufacturing system. The system's PN model is shown in Figure 3. The system has 3 part types, P1, P2 and P3, 4 machines M1-M4, and 3 robots R1-R3. Part routes for P1 and P2 are predetermined (MRF), while P3 has choice jobs (FMRF). All resources, except for M1, are shared (only utilization of shared resources will be optimized).

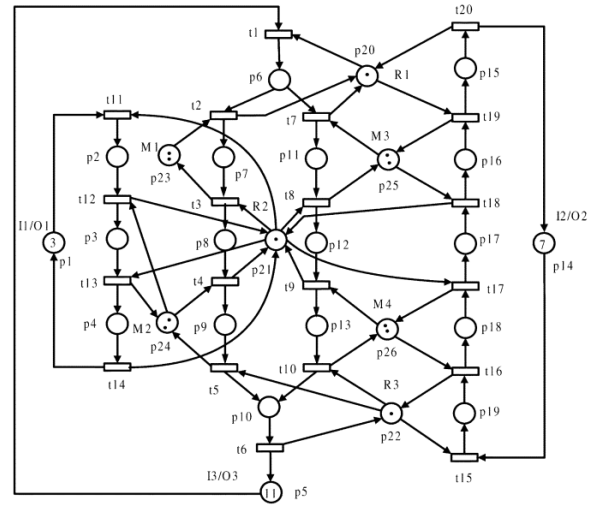


Figure 3: PN model of the system used for the case study [15].

The goal is to find the optimal sequence that includes all shared resources in the system for $\omega = 6$ and $\varepsilon = 0.2$.

Such value of ε gives $0.8 \leq \frac{u_i}{u_j} \leq 1.2$, *i.e.* it is required

that usage of resources is balanced. The following costs have been used in optimization:

$$c_{M11}=6, c_{M31}=4, c_{M22}=5, c_{M42}=9.$$

Three MJI sub-matrices have been used for construction of sequences such that $\mathbf{u}^1 = [1 \ 1 \ 0]^T$, $\mathbf{u}^2 = [0 \ 0 \ 1]^T$ and $\mathbf{u}^3 = [0 \ 1 \ 1]^T$, where components correspond with resources M2, M3 and M4. This gives overall usage of shared resources as

$$\mathbf{u} = \omega_1 \mathbf{u}^1 + \omega_2 \mathbf{u}^2 + \omega_3 \mathbf{u}^3 = [\omega_1 \quad \omega_1 + \omega_3 \quad \omega_2 + \omega_3]^T$$

. Optimization yields to the following values: $\omega_1 = 3$,

$\omega_2 = 2$, $\omega_3 = 1$ with total usage of resource within the

sequence equal to $\mathbf{u} = [3 \ 4 \ 3]^T$ as it is shown in Figure 5. Results obtained by simulation with MJIWorkshop, software tool presented in [12], are shown in Figures 5

and 6. It can be seen from Figure 6 that system is deadlock free, *i.e.* flow of the parts is uninterrupted.

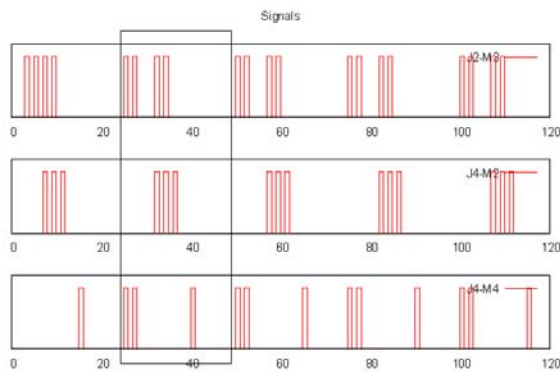


Figure 5: Utilization of M2, M3 and M4 within the sequence.

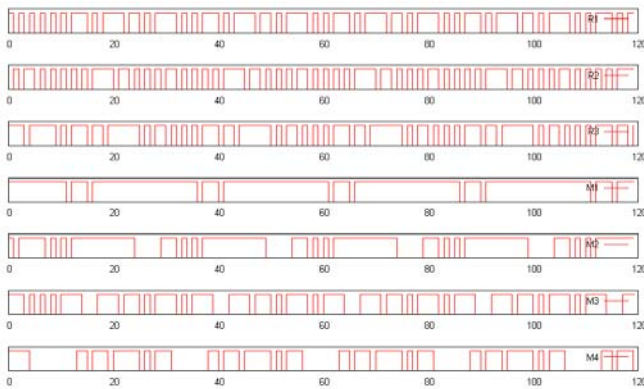


Figure 6: Utilization of all resource in the system.

5. CONCLUSION

In a finite-buffer flexible manufacturing systems, any dispatching policy for interrupted part flow has to essentially take into account the composition of the interconnection between jobs and resources. The proposed optimal non-blocking dispatching policy is based on machine-job incidence matrix (MJ), obtained from Steward sequencing matrix and Kusiak machine-part incidence matrix, and explained in details in [17].

Since FMRF systems contain shared resources and choice jobs, a solution to allocation of resources to jobs is determined in a form of repeatable resource sequence over the set of resources available for particular choice job. Obtained sequences not only stabilize the system but provide an optimal utilization of resources based on price and balanced use of all resources.

Efficiency of the proposed method has been demonstrated on an example involving multi-part type manufacturing system.

REFERENCES

[1] W.M. Wonham, *Supervisory Control of Discrete Event Systems*, Lecture notes, 2005.
 [2] T. Murata, Petri nets: properties, analysis and applications, *Proc. IEEE*, 77, 4, 1989, pp. 541–580.

[3] Y.C. Ho, X.R. Cao, *Perturbation Analysis of Discrete Event Dynamic Systems*, Kluwer Academic Publishers, 1991.
 [4] D.V. Steward, The Design Structure System: A Method for Managing the Design of Complex Systems, *IEEE Transactions on Engineering Management*, 28, 1981, pp. 71-74.
 [5] A.Kusiak, J. Ahn, Intelligent Scheduling of Automated Machining Systems, *Computer-Integrated Manufacturing Systems*, 5, 1, 1992, pp. 3-14.
 [6] T.R. Browning, Applying the Design Structure Matrix to System Decomposition and Integration Problems: A Review and New Directions, *IEEE Transactions on Engineering Management*, 48, 3, 2001, pp. 292-306.
 [7] M.D. Jeng, F. DiCesare, Synthesis Using Resource Control Nets for Modeling Shared-Resource Systems, *IEEE Trans. Rob. Autom.* RA-11, 1995, pp. 317–327.
 [8] R.A. Wysk, N.S. Yang, S. Joshi, Detection of Deadlocks in Flexible Manufacturing Cells, *IEEE Trans. Rob. Autom.* 7, 6, 1991, pp. 853–859.
 [9] S. Bogdan, F.L. Lewis, J. Mireles, Z. Kovacic, *Manufacturing Systems Control Design: a matrix based approach*, Springer, 2006.
 [10] M.V. Iordache, P.J. Antsaklis, *Supervisory Control of Concurrent Systems: A Petri Net Structural Approach*, Birkhauser, Boston, USA – 2006.
 [11] M.C. Zhou, M.P. Fanti, *Deadlock resolution in computer-integrated systems*, Marcel Dekker/CRC Press, New York 2005.
 [12] I. Sindičić, T. Petrović, S. Bogdan, Modeling and Simulation of Manufacturing Systems based on Machine-Job Incidence Matrix, *Proc. Int. Conference on Mathematical Modelling*, Vienna, 2009.
 [13] T. Petrović, S. Bogdan, I. Sindičić, Determination of Circular Waits in Multiple-Reentrant Flowlines based on Machine-job Incidence Matrix, *Proc. European Control Conference*, Budapest, 2009.
 [14] S. Lee, D.M. Tilbury, Deadlock-Free Resource Allocation Control for a Reconfigurable Manufacturing System With Serial and Parallel Configuration, *IEEE Trans. on SMC-part C*, 37, 6, 2007, pp.1373-1381.
 [15] ZhiWu Li, MengChu Zhou; Two-Stage Method for Synthesizing Liveness Enforcing Supervisors for Flexible Manufacturing Systems Using Petri Nets, *IEEE Transactions on industrial informatics*, Vol. 2, No. 4, November 2006, pp. 313-325
 [16] I.Sindičić, S.Bogdan, T.Petrović; Dispatching in Free-choice Multiple Reentrant Manufacturing Flowlines by using machine-Job Incidence Matrix; 6th IEEE Conference on Automation Science and Engineering – CASE 2010, p617-623, Toronto, Canada, 22-24. August 2010
 [17] I.Sindičić, S.Bogdan, T.Petrović; Resource Allocation in Free-choice Multiple Reentrant Manufacturing Systems Based on Machine-job Incidence Matrix; *IEEE Transactions on Industrial Informatics*, Vol. 7, No. 1, 2011, pp. 105-114.

SIMULATION OF VASCULAR VOLUME PULSATION OF RADIAL INDEX ARTERY

Pichitra Uangpairoj^(a), Masahiro Shibata^(b)

^(a,b) Department of Bio-science and Engineering, College of Systems Engineering and Science,
Shibaura Institute of Technology, Japan.

^(a)m710506@shibaura-it.ac.jp, ^(b)shibatam@sic.shibaura-it.ac.jp

ABSTRACT

This paper presents an application of finite element simulation with the analysis of arterial stiffness. The influences of the intravascular pressures on arterial wall which behaves like hyperelastic material was investigated by using Mooney-Rivlin hyperelastic constitutive model with the finite element solver of LS-DYNA. The results were obtained in the forms of nonlinear pressure-diameter relationship. Moreover, the diameter variation of arterial model corresponds to the pulsatile blood pressure. But the distensibility of artery reduces when the level of pulsatile blood pressure increases. These numerical results are expected to clarify an assessment of the arterial stiffness using photoelectric plethysmograph.

Keywords: arterial stiffness, vascular volume change, finite element simulation

1. INTRODUCTION

Arterial stiffness is associated with the development of cardiovascular risk factors. It is one of the indices which are used to diagnose the pathophysiology of cardiovascular system in both of research and clinical applications.

Arterial stiffness is typically investigated by monitoring the arterial motion in the circumferential direction. The instantaneous change of vessel circumference corresponds to the arterial pressure pulse which can be seen in the form of pressure-diameter relationship or

pressure-volume relationship from many in vitro tests (Cox 1978a-c; Carew, Vaishnav and Petal 1968). They investigated the volume change of blood vessel by increasing the intravascular pressure and the transmural pressure. The intravascular pressure was measured by using pressure transducer. Meanwhile, in vivo tests, the pressure-volume relationship can be obtained from photoelectric plethysmographic (PPG) technique (Kawarada et al. 1986). PPG system is compatible with the clinical application. It is a non-invasive measurement system and easy to use. To evaluate the arterial blood volume change, the other tissues are considered to be incompressible. The venous system is collapsed by the exertion of external pressure. At the same time, the changes of arterial blood volume can be controlled by the decrease in external pressure and the increase in transmural pressure. With these assumptions and the Lambert-Beer's Law, Kawarada et al. (1986) and Ando et al. (1991) could investigate arterial elasticity by detecting the arterial volume change at any changes in transmural pressure from DC signal of PPG.

In Biomechanics, arterial stiffness associates with the mechanical properties of arterial wall which consists of three main layers; tunica intima, tunica media and tunica adventitia. It is believed that elastin and smooth muscle cell in media layer assist artery to resist high loads in the circumferential direction. At the same time, the thick bundles of collagen fibres in adventitia also contributes significantly to protect artery from overextension and

rupture when artery is exerted by force from blood pressure. With the structure of arterial wall, it makes artery exhibits hysteresis under cyclic loading, stress relaxation under constant extension and creep under constant loads. This behavior can identify artery to be viscoelastic. However, arteries are frequently considered simply as hyperelastic material and all inelastic phenomena are neglected. Therefore, the constitutive models of arterial wall have been developed regarding the hyperelasticity and the distribution of collagen fibres which reflects an orthotropic property of arterial wall in both of microstructure (Bischoff, Arruda and Gresh 2002, Zhang et al. 2005) and macrostructure (Gasser, Ogden and Holzapfel 2006).

With the high performance of current computers, the constitutive models for arterial wall and finite element method have been widely implemented to observe the responses of arterial wall to various types of load. These observations have been utilized in many clinical application (Xia, Takayanagi and Kemomochi 2001; Zhang et al. 2005, 2007; Zhao et al. 2008).

In this study, the responses of arterial wall to the loads have been observed by coupling anisotropic hyperelastic constitutive models with the finite element method. These numerical results are expected to clarify the assessment of the arterial stiffness using PPG.

2. MATERIAL AND METHODS

2.1 The finite element model of artery

A radial index artery was considered in order to apply this study with PPG application. A tube with three layers of arterial wall; tunica intima, tunica media and tunica adventitia, was combined into one-layer to simplify anatomical structure of artery. The outer diameter of the tube was 1.54 mm (Bilge et al. 2006). The ratio of total wall thickness to outer diameter was 0.189 mm (Holzapfel et al. 2005). The length of the tube was specified to be 10 mm. This model was discretised into

15,360 hexahedron solid elements by using LS-Prepost version 2.1. All elements were assumed to be the constant stress solid element in order to avoid volumetric locking effect as shown in Figure 1.

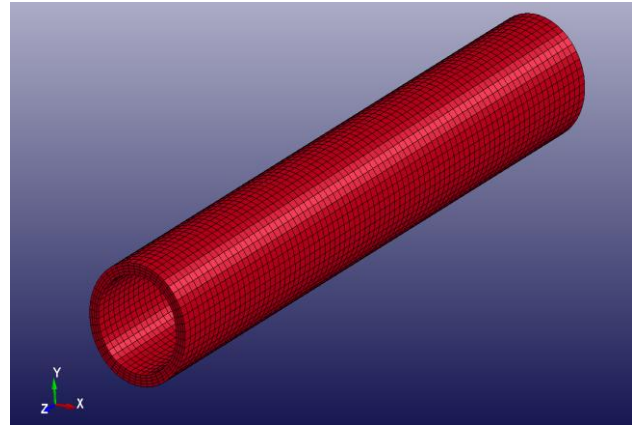


Figure 1: The geometrical model of radial index artery

2.2 The finite element method

The commercial explicit finite element solver of LS-DYNA 971 was employed to investigate the responses of arterial wall on various types of pressure which are given in the following form:

- Intravascular pressures which represented mean blood pressure were assigned to be 10, 20, 30, 40, 50, 70, 90, 100, 120, 140, 160 and 180 mmHg at the inner wall of the tube as the ramped loads. The pressure raised from 0 to the maximum pressure of each level within 100 ms. This was done to obtain the pressure-diameter relationship of the artery.
- Pulsatile loads which represented pulsatile blood pressures were given in the form of sinusoidal function to simplify the arterial pressure pulse.

$$P(t) = \bar{P} + P_{amp} \sin(2\pi ft) \quad (1)$$

where $P(t)$ is the instantaneous pressure (mmHg), \bar{P} is mean level of the pressure pulse (mmHg). The level of \bar{P} was assigned to be 70, 100 and 120

mmHg which represented low blood pressure, normal blood pressure and high blood pressure conditions, respectively. P_{amp} is the amplitude of pulsatile pressure (mmHg) which relates to systolic pressure and diastolic blood pressure. In this study, the systolic/diastolic blood pressures were 90/50, 120/80 and 140/100 mmHg for low blood pressure, normal blood pressure and high blood pressure conditions, respectively. f is frequency (Hz) and t is time (s). Pulsatile pressures were also applied at the inner wall of the tube to investigate the responses of arterial wall on pulsatile pressure.

- External pressure was assume to be zero in this simulation.

These specifications were considered as traction boundary conditions of boundary value problem. Meanwhile, the displacement boundary condition was applied at the annulus of the tube which constrained movement in all direction.

The numerical models were based on solving the momentum balance equation and boundary conditions which are given as follows:

The momentum balance equation:

$$\nabla \cdot \boldsymbol{\sigma} + \rho_0 \mathbf{f} = \rho_0 \mathbf{a} \quad (2)$$

where $\boldsymbol{\sigma}$ is the Cauchy stress, \mathbf{F} is the deformation gradient, ρ_0 is the mass density, \mathbf{f} is the body force and \mathbf{a} is the acceleration.

The traction boundary condition:

$$\mathbf{N} \cdot \boldsymbol{\sigma} = \bar{\mathbf{T}} \quad (3)$$

where $\bar{\mathbf{T}}$ is traction force and \mathbf{N} is the unit normal vector.

The displacement boundary condition:

$$\mathbf{u} = \bar{\mathbf{u}} \quad (4)$$

where \mathbf{u} is the displacement.

When $x^+ = x^-$ The contact discontinuity:

$$\mathbf{N} \cdot (\boldsymbol{\sigma}^+ - \boldsymbol{\sigma}^-) = 0 \quad (5)$$

The Cauchy stress in the momentum balance equation related to the strain energy function through the constitutive equation.

Constitutive Equations for incompressible hyperelastic material can be expressed as follows:

$$\boldsymbol{\sigma} = \mathbf{F} \frac{\partial W}{\partial \mathbf{F}} - p \mathbf{I} \quad (6)$$

where W is the strain energy function, \mathbf{F} is the deformation gradient tensor, p is the Lagrange multiplier and \mathbf{I} is the identity tensor.

The strain energy function for arterial wall was defined by a Mooney-Rivlin hyperelastic constitutive model as follow:

$$W = C_{10}(I_1 - 3) + C_{01}(I_2 - 3) + C_{20}(I_1 - 3)^2 + C_{11}(I_1 - 3)(I_2 - 3) + C_{30}(I_1 - 3)^3 \quad (7)$$

where C_{10} , C_{01} , C_{20} , C_{11} and C_{30} are hyperelastic coefficients used for artery. I_1 and I_2 are invariants which can be expressed as

$$I_1 = tr \mathbf{B} \quad (8)$$

$$I_2 = \frac{1}{2} [I_1^2 - tr(\mathbf{B}^2)] \quad (9)$$

where \mathbf{B} is the left Cauchy-Green deformation tensor.

The hyperelastic coefficients of arterial wall which were reported by Loree et al. (1994) are shown in Table 1

Table 1: hyperelastic coefficients of arterial wall

	C_{10} [KPa]	C_{01} [KPa]	C_{20} [KPa]	C_{11} [KPa]	C_{30} [KPa]
Artery	708.416	-620.042	0	2827.33	0

2.3 Analysis of results

In this simulation, the change in diameter of artery was obtained from the average movement of radial position of the inner wall at every 1 mm length in axial direction using post-processing of LS-DYNA software.

3. RESULTS AND DISCUSSIONS

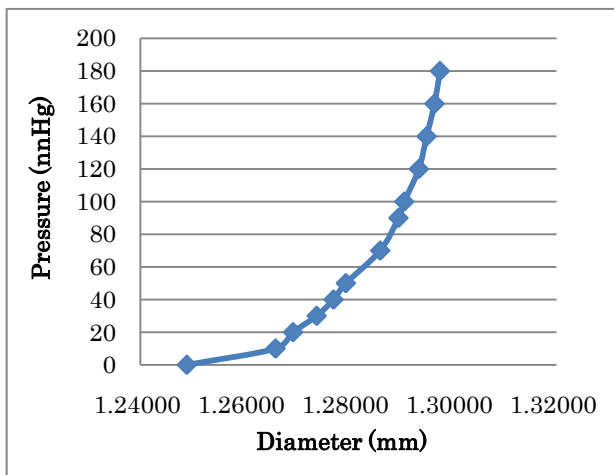
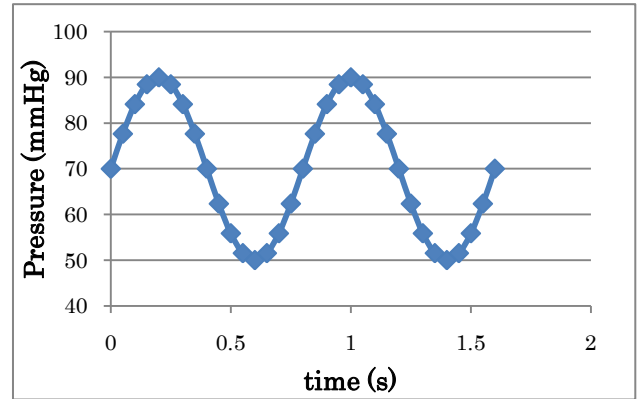


Figure 2: Pressure-diameter relationship of artery

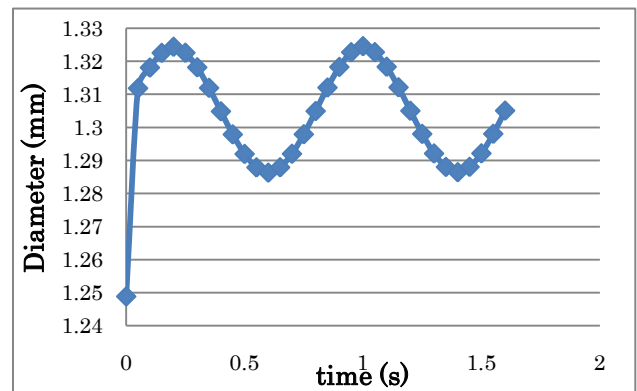
Figure 2 shows the relationship between intravascular pressure and diameter of arterial tube which was obtained by varying the level of intravascular pressure as in section 2.1. This curve shows that using Mooney-Rivlin hyperelastic model generates non-linear pressure-diameter relationship. At the lower level of intravascular pressure, 10-70 mmHg, the distensibility of arterial model is higher than the upper level pressure, 90-180 mmHg. This shows that arterial model is stiffer at higher pressure level. This result corresponds to the relationship between the transmural pressure and volume elastic modulus of rabbit artery from the experiment of Kawarada et al. (1986) that the elastic modulus of artery

increases with transmural pressure nonlinearly.

This result confirms that the simulation model of artery behaves similarly to the real artery, it is also suitable to apply this model with the investigation of the influence of pulsatile pressure on arterial wall.

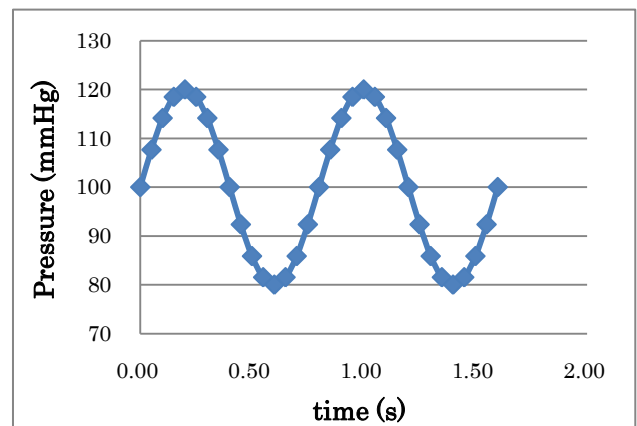


(a)

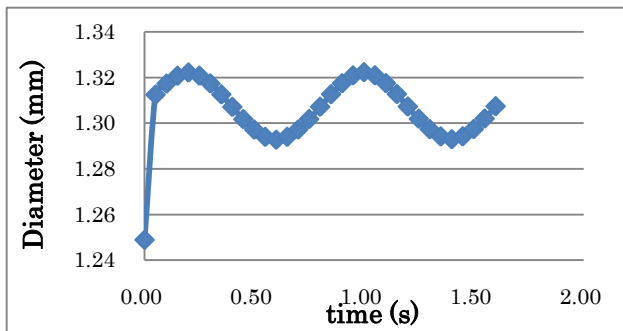


(b)

Figure 3: (a) Pressure-time relationship and (b) Diameter-time relationship of low blood pressure condition.

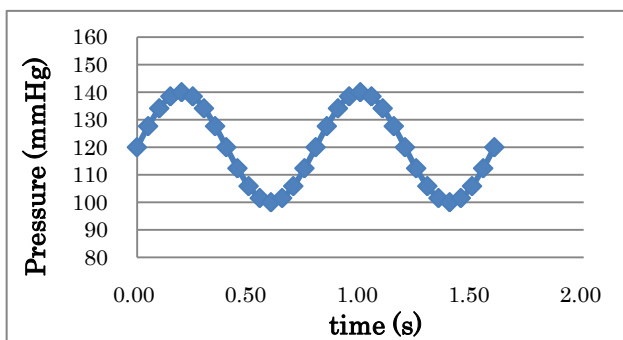


(a)

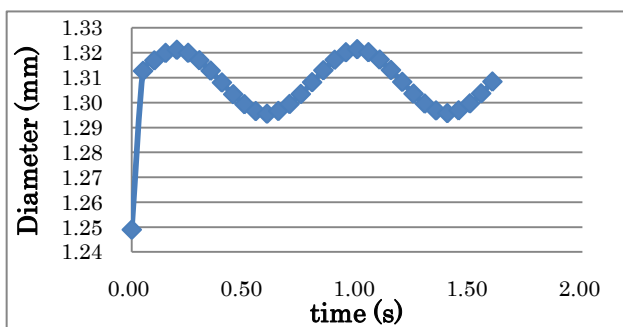


(b)

Figure 4: (a) Pressure-time relationship and (b) Diameter-time relationship of normal blood pressure condition.



(a)



(b)

Figure 5: (a) Pressure-time relationship and (b) Diameter-time relationship of high blood pressure condition.

Figure 3, 4 and 5 show the responses of arterial diameter on pulsatile pressure in low blood pressure, normal blood pressure and high blood pressure conditions, respectively. The results show that arterial diameters vary with pulsatile pressure in all conditions. The difference of arterial response in each blood pressure condition is the amplitude of diameter variation. Within the same level of

the amplitude of pulsatile pressure, the amplitude of diameter variation in high blood pressure condition is the lowest one (1.14% of mean diameter). Meanwhile, the amplitude of diameter variation in normal blood pressure condition (1.28% of mean diameter) is also lower than the amplitude in low blood pressure condition (1.62% of mean diameter). These are because the arterial distensibility decreases when the level of intravascular pressure increases which correspond to the pressure-diameter relationship.

The investigation of the responses of arterial diameter on pulsatile pressure is just only the first step to apply finite element method with the analysis of arterial elasticity. To apply the simulation results with the analysis of arterial elasticity using PPG. The further simulations need to be carried on for the future work. For example, the pulsatile pressure should be modified to be similar to the real pulsatile blood pressure. Moreover, the influence of external pressure (or cuff pressure for PPG) on diameter of arterial model also need to be studied.

REFERENCES

- Ando, J., Kawarada, A., Shibata, M., Yamakoshi, K. and Kamiya, A., 1991. Pressure-volume relationships of finger arteries in healthy subjects and patients with coronary atherosclerosis measured non-invasively by photoelectric plethysmography. *Japanese Circulation Journal* 55: 567-575.
- Bilge, O., Pinar, Y., Özer, M. A. and Gövsa, F., 2006. A morphometric study on the superficial palmar arch of the hand. *Surg Radiol Anat* 28: 343-350.
- Bischoff, J. E., Arruda, E. A. and Grosh, K., 2002. A microstructurally based orthotropic hyperelastic constitutive law. *Transactions of the ASME* 69: 570-579.
- Carew, T. E., Vaishnav, R. N. and Patel, D. J., 1968. Compressibility of the arterial wall. *Circulation Research* 23: 61-68.

- Cox, R. H., 1978a. Passive mechanics and connective tissue composition of canine arteries. *Am J Physiol* 234: H533-541.
- Cox, R. H., 1978b. Differences in mechanics of arterial smooth muscle from hindlimb arteries. *Am J Physiol* 235: H649-656.
- Cox, R. H., 1978c. Regional variation of series elasticity in canine arterial smooth muscle. *Am J Physiol* 234: H542-551.
- Gasser, T. C., Ogden, R. W. and Holzapfel, G. A., 2006. Hyperelastic modelling of arterial layers with distributed collagen fibre orientations. *J R Soc Interface* 3: 15-35.
- Holzapfel, G. A., Gasser, T. C. and Ogden, R. W., 2000. A new constitutive framework for arterial wall mechanics and a comparative study of material models. *J Elasticity* 61: 1-48.
- Holzapfel, G. A., Sommer, G., Gasser, T. C. and Regitnig, P., 2005. Determination of layer-specific mechanical properties of human coronary arteries with nanotherosclerotic intimal thickening and related constitutive modeling. *Am J Physiol Heart Circ Physiol* 289: H2048-H2058.
- Kawarada, A., Shimazu, H., Yamakoshi, K. and Kamiya, A., 1986. Noninvasive automatic measurement of arterial elasticity in human fingers and rabbit forelegs using photoelectric plethysmography. *Med. & Biol. Eng. & Comput* 24: 591-596.
- Loree, H. M., Grodzinsky, A. J., Park, A. J., Gibson, L. J., Lee, R. T., 1994. Static and circumferential tangential modulus of human atherosclerotic plaques. *Circulation* 27: 195-204.
- Xia, M., Takayanagi, H. and Kemmochi, K., 2001. Analysis of multi-layered filament-wound composite pipes under internal pressure. *Composite Structures* 53: 483-491.
- Zhang, Y., Dunn, M. L., Drexler, E. S., McCowan, C. N., Slifka, A. J., Ivy, D. D. and Shandas, R., 2005. A microstructural hyperelastic model of pulmonary arteries under normo- and hypertensive conditions. *Annals of Biomedical Engineering* 33(8): 1042-1052.
- Zhang, Y., Dunn, M. L., Hunter, K. S., Lanning, C., Ivy, D. D., Claussen, L., Chen, S. J. and Shandas, R., 2007. Application of a microstructural constitutive model of the pulmonary artery to patient-specific studies: validation and effect of orthotropy. *Journal of Biomechanical Engineering* 129: 193-201.
- Zhao, A. R., Field, M., Digges, K. and Richens, D., 2008. Blunt trauma and acute aortic syndrome: a three-layer finite element model of the aortic wall. *European Journal of Cardio-thoracic Surgery* 34: 623-629.

AUTHORS BIOGRAPHY

Pichitra Ungpaibroj received the B.Sc. (2007) in Food Technology and M.E. (2010) in Mechanical Engineering from Suranaree University of Technology, Thailand. She is the PhD student in the Department of Bioscience and Engineering, Shibaura Institute of Technology, Japan. Her current interests include the applications of numerical simulation in arteries.

Masahiro Shibata received the BS in Applied Physics and his PhD in Biomedical Engineering from Hokkaido University, Japan. Since 2008 he has been with the Department of Bio-Science and Engineering, Shibaura Institute of Technology, where he is a Professor of System Physiology. His research interests include oxygen dynamics in microcirculation.

CODING TCPN MODELS INTO THE SIMIO SIMULATION ENVIRONMENT

Miguel Mujica^(a), Miquel Angel Piera^(b)

^(a,b)Autonomous University of Barcelona, Faculty of Telecommunications and Systems Engineering,
08193, Bellaterra, Barcelona

^(a)miguelantonio.mujica@uab.es, ^(b)miquelangel.piera@uab.es

ABSTRACT

The coloured Petri net formalism has been widely used by scientific community to perform not only research and behavioural analysis of models but also as a simulation tool to carry out systems' analysis. Its characteristics allow a better understanding of the causal relationships present in systems. On the other hand discrete event system simulation software has evolved in order to reduce the efforts needed in simulation projects; the developers have improved the developing paradigm, graphical interface and analysis tools among other aspects. Unfortunately when dealing with big projects the simulation paradigms hinder the understanding of the causal relationships of systems. In this paper a way to integrate the timed coloured Petri net modelling formalism with the SIMIO simulation software is presented in order to overcome this problem.

Keywords: Discrete event systems, scheduling, decision support systems, timed coloured Petri nets, Simulation, Simio.

1. INTRODUCTION

Simulation is a very well recognized methodology which possesses a high descriptive level. Most commercial simulators have several modules to analyze data, implement the relationship between elements and to perform simulation experiments. In recent years discrete-event system simulation developers have put focus on improving the characteristics of the simulation software in order to reduce the efforts needed to develop a simulation project. Some simulators such as PROMODEL(www.promodel.com) or Witness (www.lanner.com) have developed graphical modules in order to improve the graphical representation of systems, but unfortunately its original source code remains the same. SIMIO simulation software (www.simio.com) was developed by the creators of the very well known ARENA (www.arenasimulation.com) software. They have developed a novel simulation program that improves several aspects of the simulation software; they combine the processes approach with the object oriented paradigm (OOP), 2D-3D visualisation among other features that makes it a very powerful tool. With the use of the OOP the developing phase of the simulation project is improved taking advantage of the characteristics of this approach (encapsulation, inheritance, polymorphism). The developing and analysis characteristics present in SIMIO can be further

improved if some underlying logic is added in order to achieve a better understanding of the modelled system.

The timed coloured Petri net formalism (TCPN) has characteristics that allow modelling true concurrency, parallelism or conflicting situations present in industrial systems (Jensen 1997, Moore & Gupta 1996, Mušič, 2009). Unfortunately when the TCPN formalism is used with the purpose of systems analysis it lacks of tools that can be used to perform statistical analysis by a user who is not expert in the TCPN field (industrial engineers, process engineers, managers, etc). Furthermore if the model or the results are going to be presented to decision makers within a firm the graphical representation results difficult to understand when the people is not familiar with the formalism. In the best cases the token game is the one that can be obtained which is the case of CPNtools (www.daimi.au.dk/~cpntools/) or Petrisimm (<http://seth.asc.tuwien.ac.at/petrisimm/>).

SIMIO simulation software has the processes paradigm that allow to code the TCPN firing rules using simple steps thus the casual relationships between the generated events can be governed by the TCPN semantic rules. The advantage of integrating TCPN models to govern some activities of the SIMIO model is that it is possible to develop models using the formalism and at the same time take advantage of the characteristics and the graphical potential that SIMIO possesses. The article is organized as follows. Section 2 presents the TCPN modelling formalism, section 3 discusses some characteristics of SIMIO; section 4 describes a way to code TCPN models using some elements of SIMIO for governing some events of the modelled system. Section 5 discusses briefly the graphical and analytical capabilities of SIMIO and section 6 gives the conclusions of the article.

2. TIMED COLOURED PETRI NETS

Coloured Petri Nets (CPN) is a simple yet powerful modelling formalism, which allows the modelling of complex systems which present an asynchronous, parallel or concurrent behaviour and can be considered discrete event dynamic systems (Jensen & Kristensen 2009). The formalism allows developing models without ambiguity and in a formal way. It is possible to model not only the dynamic behaviour of systems but also the information flow, which is an important

characteristic and very useful in industrial systems modelling and decision making.

In order to investigate the KPI's (Key Performance Indicators) at which the industrial systems operate under different policies, such as scheduling, resource occupancy, costs and inventory among others it is convenient to extend CPN with a time concept. This extension is made by introducing a global clock for the model, time stamps for the entities and a time delay for the model transitions; the nets that use this extension are known as *timed coloured Petri nets*(TCPN). When using TCPN the global clock represents the model time, and the time stamps describe the earliest model time at which the entities of the model, graphically represented by dots (tokens), can be used for the transition evaluation process (Jensen 1997). A token is *ready* if the correspondent time stamp is less than or equal to the current model time. If the token is not ready, it can not be used in the transition enabling procedure.

The transitions of the TCPN are used to model the activities of the real system and a time delay is attached in order to simulate time consumption of a certain activity.

It is a common convention to use the sign @ to denote time in the elements of the model. When it is attached to transitions, it specifies the time consumption.

The formal definition of TCPN is the following one.

Definition 1. The non-hierarchical TCPN is the tuple:

$$\text{TCPN} = (P, T, A, \Sigma, V, C, G, E, D, I) \text{ where}$$

1. P is a finite set of places.
2. T is a finite set of transitions T such that $P \cap T = \emptyset$
3. $A \subseteq P \times T \cup T \times P$ is a set of directed arcs
4. Σ is a finite set of non-empty colour sets.
5. V is a finite set of typed variables such that $\text{Type}[v] \in \Sigma$ for all variables $v \in V$.
6. $C: P \rightarrow \Sigma$ is a colour set function assigning a colour set to each place.
7. $G: T \rightarrow \text{EXPR}$ is a guard function assigning a guard to each transition T such that $\text{Type}[G(T)] = \text{Boolean}$.
8. $E: A \rightarrow \text{EXPR}$ is an arc expression function assigning an arc expression to each arc a , such that:

$$\text{Type}[E(a)] = C(p)$$

Where p is the place connected to the arc a

9. $D: T \rightarrow \text{EXPR}$ is a transition expression which assigns a delay to each transition.
9. I is an initialization function assigning an initial timed marking to each place p such that:

$$\text{Type}[I(p)] = C(p)$$

EXPR denotes the mathematical expressions associated to the elements of the formalism (variables, colours, logic conditions) where the syntax can vary when coding the formalism in a programming language. The $\text{TYPE}[e]$ denotes the type of an expression $e \in \text{EXPR}$, i.e. the type of values obtained when evaluating e . The set of free variables in an expression e is denoted $\text{VAR}[e]$ and the type of a variable v is denoted $\text{TYPE}[v]$.

The formalism can be graphically represented by circles which represent the place nodes and rectangles or solid lines that represent the transition nodes. The place nodes are used to model resource availability or logic conditions that need to be satisfied. The transition nodes can be associated to activities of the real system. The developed models can be analyzed with the help of available academic software such as CPNtools or PetriSimm. These types of software programs are commonly used by the scientific community in order to carry out analysis of the model to verify behaviour or systems' performance. This analysis can be performed through simulation of the system making use of the token game or performing state space analysis (Christensen et al. 2001, Mujica & Piera 2011, Mujica et al. 2010).

3. SIMIO SIMULATION SOFTWARE

The most common discrete event system (DES) simulators have been coded using the *activity scanning* or the *event scheduling* approach (Shannon 1997). These approaches have been developed to reproduce the behaviour of systems whose states change in discrete instants of time. Most of the commercial tools are very efficient in modelling DES using one of these approaches or even combinations of them.

With the development of fast CPU processors, powerful graphic cards, efficient statistical analysis tools etc., simulation has become more important than before. Some years ago most of the effort during a simulation project was put on the development phase of the model but with the development of more efficient software programs this effort has been considerably reduced. The available tools allow analysts to spend more time in the analysis phase of the simulation project. In order to reduce the lead time of the simulation project some developers have invested time in implementing new paradigms in their simulation products; this is the case of SIMIO. This software has been coded merging the OOP together with the processes paradigm in order to reduce the number of blocks needed to develop complex models. Based on these programming paradigms the developers coded SIMIO as a collection of objects that are instantiated from classes. These classes were designed using the principles of abstraction, encapsulation, polymorphism, inheritance and composition (Pegden 2007).

Making use of the few available objects it is only necessary to add new functionalities (processes) to the original ones in order to have additional behaviour or logic (inheritance) or even overriding the original one.

Since the objects in SIMIO follow the encapsulation principle their implementation is sealed from the outside world. The composition principle allows building new classes combining the existing ones. These characteristics allow great flexibility when developing a model.

One important aspect of the simulation project is the implementation phase which depends strongly on how the results are presented to decision makers.

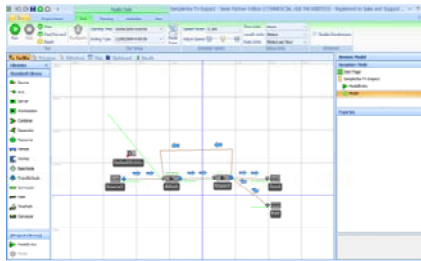


Figure 1. Example of a SIMIO model in 2D

In this sense, the graphical interface of SIMIO has been developed in such a way that it is very easy to have very good-looking results in short time. It can switch between 2D and 3D depending on which kind of task is being performed (development or validation). Figure 1 presents a typical 2D view of the SIMIO model and Figure 2 the 3D view of the same model.

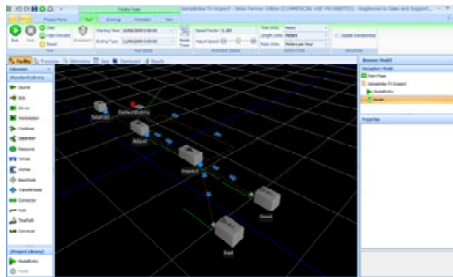


Figure 2. SIMIO model in 3D view.

The switch between both views is as easy as clicking the mouse. It is important to mention that SIMIO comes with a basic graphical library but it can be extended with graphical models from the GOOGLE 3D warehouse (<http://sketchup.google.com/3dwarehouse/?hl=en>).

4. INTEGRATING THE TCPN MODELS IN THE SIMIO ENVIRONMENT

Due to the dual developing paradigm used in SIMIO (process/object) it is possible to extend the functionality of the SIMIO objects with a sequence of steps which fit the purpose of coding the TCPN rules. These rules can be easily implemented using some elements of SIMIO such as the following ones.

Objects:

- SOURCE
- SINK
- SEPARATOR

- TRANSFER NODE

Steps:

- Decide
- Search
- Destroy
- Transfer
- Assign

Elements:

- Station

Other objects can be also added after the development of the model in order to improve the visual aspect of the model or to make the simulation more detailed.

The model of Figure 3 will be used to illustrate the implementation of the TCPN rules into SIMIO.

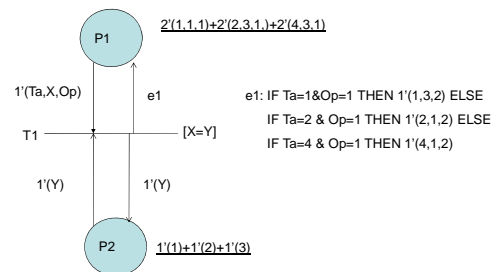


Figure 3. TCPN model

4.1. Modelling the Place nodes

The place nodes are modelled using the *Station* element which has been developed with the objective of storing entities.

The stations are *Elements* that need to be defined in the *Definitions* area of the software. Figure 4 illustrates the station definition.

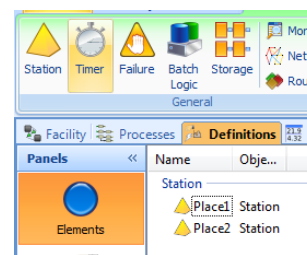


Figure 4. Station definition

In this figure Place1 and Place2 represent the place nodes of Figure 3.

Graphically a *Station* is not related to any predefined object of the *Standard Library* in SIMIO, but its graphical visualization can be performed associating a *detached QUEUE* to the *Station* element. The entities of SIMIO can have any 3D appearance which allows exploiting the graphical potential of SIMIO. Figure 5 gives an example of a detached queue associated to the place (Place1) in the SIMIO model.

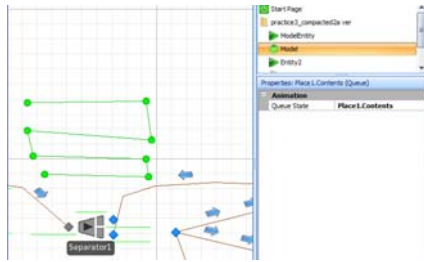


Figure 5. A detached queue for the Place1

The QUEUE is just added into the *facility area* and the association is performed in the *Properties* window:

Queue State| Place1.Contents.

Depending on the type of model the resource availability can also be modelled using another predefined object which can be used as a resource such as the SERVER or WORKSTATION; moreover with the use of a predefined object the graphical representation can be enhanced by taking advantage of the object's behaviour. In the example presented here the two place nodes will be modelled making use of two *Station* elements.

4.2. Defining the Token Colours

The entities that represent the tokens are generated using the SOURCE object and they are destroyed using the SINK object or the **Destroy** step in the *Processes* area of SIMIO.

The entities in SIMIO can also be extended with attributes. These attributes work as the fields of a record in any simulation language and they are associated to the entities in the *States Definitions* area. The states in SIMIO have been conceived as variables whose values change during the simulation run. They can be modified to include as many attributes as colours present in the token. The states can be of different types: Real, Integer, Boolean, Date, or Strings.

4.3. Modelling the TCPN Restrictions

The dynamics of the TCPN models are governed by the input/output flux of tokens that takes place when a transition is fired. The *Processes* area of SIMIO is used to code the TCPN logic which evaluates the constraints to unchain the processes modelled by the transitions.

In order to enable a transition the following conditions must be fulfilled:

- the number of tokens in the input place nodes are greater than or equal to the arc weight
- the colours of the correspondent tokens must have the particular value that is stated in the arc inscription
- The colour binding must satisfy the boolean expression stated by the *Guards*

The evaluation of the restrictions imposed by the arcs and guards can be performed using a combination of

Decide, Search, Assign, and Destroy steps such as the one depicted in Figure 6:

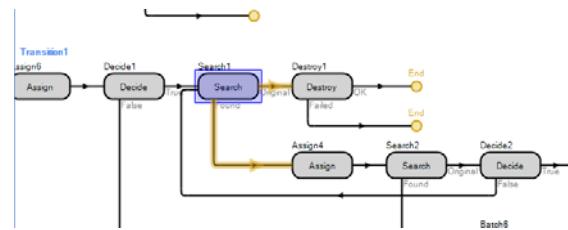


Figure 6: Evaluating the restrictions

- **Decide1**: the **Decide** step works like an IF..THEN..ELSE instruction in any programming language. The *Decide Type* attribute (Figure 7) must be put in *condition based* in order to state the expressions that need to be satisfied. The attribute *Expression* is used to specify the logic conditions that must be fulfilled by the entities of the TCPN model. The expression for the example model can be written in the following way.

Place1.Contents>0&Place2.contents>0

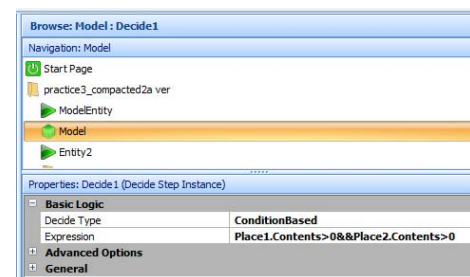


Figure 7: Decide condition

The **Decide1** is used to verify that both stations (place nodes) have at least the number of entities (tokens) imposed by the arc weights.

A similar expression can be used if one place is modelled using a SERVER or any other object of SIMIO (Resource is the object's name).

PLACE1.Contents>0&Resource.ResourceState==0

- **Search1**: The **Search** is used to perform searches within the list of elements of the stations or queues under particular restrictions. If the arc expressions have known information such as constant values, this condition is stated using the *match condition* attribute of the Search step. When the Search step has found one entity that satisfies the restrictions imposed by the arc, this entity goes out from the *found* side of the step and continues through the rest of the flowchart to check the conditions of the remaining place node. The latter is performed in the *Search1* of the flowchart of Figure 6. If the *Search1* finds an entity that satisfies the restriction, it leaves the Search step through the *found* side of the step. If it does not found any entity that fulfills the

restrictions then it leaves the step and goes to the *Destroy1* step. If it finds an entity that satisfies the restriction, it flows through the *Found* side of the search step and continues to the *Assign4*.

- *Assign4*: is used to bind the value of the entity to the variable which will be evaluated by the Guard (X variable)
- *Search2*: This step is used to perform searches in the Station2 (Place2) under the restrictions imposed by the value of the variable (X=Y). If the search obtains one entity that satisfies the restrictions it flows through the *found* side of the step and continues with the animation activities. If no entity is found then the original one continues to the *Decide2* step.
- *Decide2*: This step checks whether the last search step found or not an entity that satisfied the restrictions, if not it sends the entity to perform the cycle again testing another combination of entities from Station1 and Station2.

4.4. Sending the entities to the facility

In order to animate the capturing of resources, the entities can be sent to the *facility* window where the animation can be performed depending of the resource that is used (colours of tokens from place P2). Figure 8 illustrates the flowchart section where the successful entity is sent via *transfer step* to the correspondent node within the *facility* window based upon the attribute value of the correspondent token. In this example the attribute value can have three different values depending on the resource used (1,2,3) therefore three possible outcomes are included in the flowchart (two different *Decide* steps).

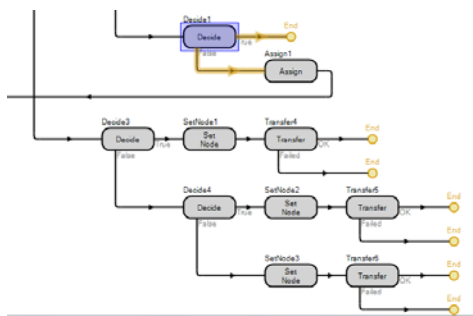


Figure 8. Sending the entities to the correspondent nodes

The **Decide** steps are used only to evaluate the colour attribute that specifies the kind of resource being used and based on the result of the evaluation the entities are sent to particular locations within the *Facility window*.

The continuous evaluation of Transition T1 is performed in the facility window making use of a SEPARATOR object after each *Server* has been released. It makes a copy of the entity that comes from the *Server* and it is sent to the node where the transition T1 is evaluated.

4.5. Time Consumption

The time consumption is modelled in a straightforward way using the time attributes available in every SIMIO object (these attributes can be found in the properties window). In most of the objects it is possible to associate time to the activities performed by the objects (entering the object, exiting the object, seize, delay etc). In the case of a predefined object such as the *SERVER* the time consumption is modelled by the *delay* or any other time-consuming activities available in the object (Figure 9). The advantage of using these properties is that it is possible to model activities which consume time in a deterministic or stochastic way.

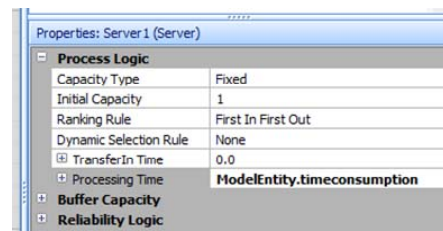


Figure 9: Defining the time consumed by an activity

4.6. Attaching a Transition to the SIMIO model

Finally, in order to govern the flow of entities in the SIMIO model, it is necessary to attach the logic (coded in the process section of SIMIO) of the TCPN to an event of the SIMIO model. All the objects in SIMIO have the *Add-On Process Triggers* which enumerate all the possible events associated to the object. These *triggers* are used to call the user-defined processes when a particular event occurs. The processes (transition evaluations) can be called at almost any point of the SIMIO model within the *facility* area.

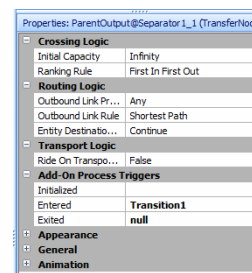


Figure 10. Adding the transition

Figure 10 illustrates the *Properties* window associated to an object. In this case *Transition1* is called every time the entity *enters* the object in run-time mode. When this event happens, the entity behaviour is governed by the logic defined by the flow diagram of the process used to model *Transition1*.

4.7. Modelling the output arc

The last element to be modelled in the example of Figure 3 is the output arc whose output function assigns the values of the colours for the output tokens.

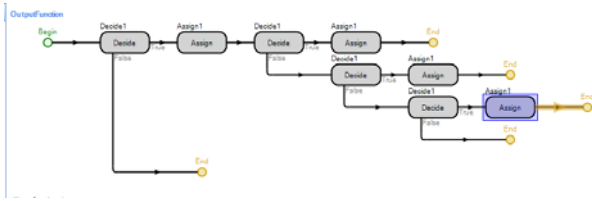


Figure 11. Coding the output function

This function is coded in SIMIO in a simple way (Figure 11) making use of nested combinations of **Decide-Assign** steps in order to evaluate the values of the attributes of the entities and afterwards based upon those evaluations the new ones are updated with the **Assign** steps before sending the entities (via a **Transfer** step) to the facility window.

5. ANALYSING THE SYSTEM

Once the TCPN rules have been defined within the *Process* area and attached to an event of the SIMIO model the simulation can be performed in the typical way.

5.1. Experimenting with the TCPN/SIMIO model

As it has been mentioned, one advantage of integrating CPN models in SIMIO is that it is possible to use the analytical tools that are integrated within SIMIO. Using those tools it is possible to obtain KPI's that allow the decision makers to evaluate the best engineering decisions. This analysis is performed making use of the *Experiment* tool (Kelton et al. 2010) which allows performing experiments (replications) of the model in order to gather information from the model. After performing the experiments a report or a matrix called *Pivot Grid* are generated and they can be used to analyze and filter the information obtained from the experiment.

Performance indicators such as resource utilization, average number of entities in the stations, average processing time, etc. are the kind of information that can be obtained from the analysis performed.

5.2. Improving the graphical view of the model

One great attribute of SIMIO is the graphical interface. Since it can switch easily from 2D to 3D and the graphical models can be downloaded directly from GOOGLE 3D Warehouse the resulting models can be graphically improved without effort.



Figure 12. Final view of a model

The model can be constructed starting with the available objects in SIMIO and afterwards download the figures that will represent the actual objects in the system. Figure 12 shows a view of a manufacture model that has been graphically enhanced using the Google 3D Warehouse models.

6. CONCLUSIONS

A way of implementing TCPN models making use of the analysis and graphical potential of SIMIO software has been presented. This approach can be used for developing simulation projects taking advantage of the characteristics of the TCPN formalism and the graphical and analytical tools available in SIMIO. In addition the GUI allows the user a better understanding of the real system and it allows giving a better aesthetic appearance if the simulation model is used as a tool for decision making.

REFERENCES

- Christensen, S; Jensen, K; Mailund, T; Kristensen, L.M., 2001."State Space Methods for Timed Coloured Petri Nets", in Proc. of 2nd International Colloquium on Petri Net Technologies for Modelling Communication Based Systems, pp. 33-42, Berlin, 2001.
- Jensen K; Kristensen L.M.; 2009."Coloured Petri Nets: Modelling and Validation of Concurrent Systems", Springer,2009.
- Jensen K.; 1997."Coloured Petri Nets: Basic Concepts, Analysis Methods and Practical Use", Vol.1 Springer-Verlag. Berlin, 1997.
- Kelton, W.D.; Smith, J.S.; Sturrock, D.T.; Verbraeck, A.; 2010."Simio & Simulation: Modeling, Analysis, Applications", McGraw-Hill, Boston, 2010.
- Moore, K.E.; Gupta, S.M.;1996." Petri Net Models of Flexible and Automated Manufacturing Systems: A Survey", International Journal of Production Research, Vol. 34(11), pp. 3001-3035, 1996.
- Mujica, M.A.; Piera M.A.; 2011. "A Compact Timed State Approach for the Analysis of Manufacturing Systems: Key Algorithmic Improvements", International Journal of Computer Integrated Manufacturing, Vol.24 (2), February 2011.
- Mujica, M.A.; Piera, M.A.; Narciso M.; 2010. "Revisiting state space exploration of timed coloured petri net models to optimize manufacturing system's performance", Simulation Modelling Practice and Theory, vol.8(9), pp. 1225-1241, Oct. 2010.
- Mušič G., 2009."Petri Net Based Scheduling Approach Combining Dispatching Rules and Local Search", in Proceedings of the I3M2009 Multiconference, Tenerife, Spain, 2009.
- Pegden, D., 2007."SIMIO: A new simulation system based on intelligent objects", in Proc. of the 39th winter simulation conference, 2007.
- Shannon, R. E. , 1997."Systems Simulation, the Art and Science", Englewood Cliffs, N. J., Prentice Hall, 1975.

Developing a Simulation Training Tool from a Medical Protocol

Catherine M. Banks, Ph.D., John A. Sokolowski, Ph.D.,

Virginia Modeling, Analysis and Simulation Center

Old Dominion University, 1030 University Blvd, Suffolk, VA 23435, USA

cmbanks@odu.edu jsokolow@odu.edu

ABSTRACT

This paper discusses the multidisciplinary effort for the development of a web-based simulation training tool that incorporates a medical protocol of patient blood management for a surgical procedure. The significance of the type of simulation tool development lies in the fact that medical simulation is able to execute training in a multiplicity of modes, it can house large digital libraries for a breadth of experiences, and it can accommodate a repetition of exercises to reinforce learning. This simulation training tool is built upon engineering and mathematical modeling. The tool is populated with simulations are drawn from actual patient case studies. The targeted trainees (users) are skilled anesthesiologists and surgeons in need of an initial introduction to this medical protocol via an expedient means to train. The tool is developed for web-based access with continuous simulation capability and hands-on exercises.

1. INTRODUCTION

Simulation training facilitates a *permission to fail environment* whereby the practitioner is taught in a multiplicity of modes [1]. It can house a *large digital library* of case studies that allows for random access to training scenarios and it can accommodate a repetition of exercises. All simulation training facilitates learning from errors and this is especially important in the field of medical training as the training takes place on a patient image, not the patient himself. The breadth of training enables a medical practitioner to methodically move from novice to master. In sum, *simulation training* incorporates both fundamental tasks and new tasks that serve to advance exposure to various patient cases and and develop expertise. It can support *re-training* that is, training experts who know what to

do, but who are learning something new. That re-training capability is core to this project, *Developing a Simulation Training Tool from a Medical Protocol*.

This paper describes the methodology, modeling paradigms, and programming development to create a web-based simulation tool to train anesthesiologists and surgeons on a patient blood management medical practice. Part 2, *Why This Tool, Why This Protocol, Who to Train* discusses why the tool is needed, who will benefit from this medical practice, and who is the targeted trainee audience. Part 3, *Who to Train, What to Train* describes the trainee and the philosophy and the practice of patient blood management techniques. Part 4, *How to Build It* speaks to the methodology, modeling paradigms, and computer programming steps taken to develop the web-based tool. Part 5, *Future Work*, provides concluding comments and further tool development opportunities. It explains how the methodology, design, and tool itself adheres to the unequivocal fundamentals of modeling and simulation (M&S): verification and validation.

2. WHY THIS TOOL, WHY THIS PROTOCOL, WHO TO TRAIN

There is a strong case endorsing the use of M&S in training among medical institutions and research centers—the sheer need for a larger body of health-care professionals who are trained in an effective and expedited manner leads that discussion.¹ Additionally, anesthesiologists and surgeons expert in the field of bloodless surgery have informed the developer community that a tool of this sort does not exist in a form they prefer. Essentially,

¹As of 2008 the American College of Surgeons certification requires three student categories who are to be taught using simulation

we believe a foremost, unmet need in medical simulation tools exists. Moreover, the medical sub-field of patient blood management as a whole is growing exponentially and simulation training instrument development is wide-open and timely.

Findings in medical literature have proven that many once excepted treatments, including blood transfusions, often carry more risk than benefit. Blood transfusions carry the risk of blood-borne illnesses including HIV, Hepatitis B, Hepatitis C, Human Lymphocytotropic Virus, Cytomegalovirus, West Nile, sepsis, and others. One study showed that transfusion of greater than 4 units of blood increased the risk of peri-operative infection by a factor of 9.28 [2]. Furthermore, the cost in dollars of transfusion is greater than once thought.

One study compared the various costs of one unit of blood ranging from \$522 to \$1183 [3]. The hospitals cited in that study had an annual expenditure on blood and transfusion related activities limited to surgical patients ranging from \$1.62 to \$6.03 million per hospital. Conversely, by implementing a blood management program Englewood Hospital (Englewood, New Jersey) was able to reduce blood use by 42% over the course of 4 years. Doing so has significantly lowered first and foremost patient morbidity and mortality and it has affected hospital financial costs [4] [5]. This tool serves as a training instrument for anesthesiologists and surgeons in the field of patient blood management techniques.

It was apparent that the targeted audience, anesthesiologists and surgeons, possessed an elevated proficiency requiring a sophisticated tool that would be readily accessible and user-friendly for very busy medical professionals. The tool should comprise exercises that require the anesthesiologist and surgeon to make medical decisions relating to blood management during the three phases of patient management with each phase containing appropriate decision points.

3. WHAT TO TRAIN

As with research required for any modeling task, the modelers needed to have a reasonable comprehension of the blood management philosophy *vis-à-vis* standard medical surgical procedure during the three surgical phases: pre-, intra-, and post-

operative. These three phases lay the foundation for the patient blood management practice.

3.1 The Blood Management Philosophy

The blood management philosophy can be expressed in terms of three pillars executed throughout three phases of patient care [6]. Below is a brief explanation of each pillar.

Pillar 1 Optimize Formation of Blood Cellular Components (haemopoiesis) This is done by producing or encouraging conditions in the body to generate healthy levels of blood cellular components.

Pillar 2 Lessen Blood Loss

This ranges from identifying and managing the risk of blood loss to the mechanical aspects of surgery to avoiding secondary hemorrhage.

Pillar 3 Channel / Optimize Patient's Tolerance of Anemia

Important to this pillar is the realization of patient's actual blood loss with his tolerable blood loss.

Various aspects of these three pillars are employed during three designated phases of patient care:

Phase 1 Pre-operative – During this phase simple measures can be taken to satisfy Pillar 1 such as detecting and treating anemia. Pillar 2 might include obtaining the autologous blood donation for hemodilution. Pillar 3 might see the implementation of a patient-specific management plan using appropriate blood conservation modalities. In short, the pre-operative phase is premised on *patient-readiness for the surgery*.

Phase 2 Intra-operative – Timing of the operation is key to addressing Pillar 1 and that requires optimizing the formation of blood cellular components. During the surgery meticulous care is taken, aka the mechanics of surgery, to satisfy the requirement of Pillar 2 in lessening blood loss. Pillar 3 focuses on optimizing the

patient's ability to tolerate anemia and in intra-operative phase this can be done through appropriate ventilation and oxygenation. In the intra-operative phase minimizing bleeding is core to the patient blood management philosophy.

Phase 3 Post-operative – This phase is a critical period for the patient. Pillar 1 necessitates much care be taken to note drug interactions that can increase anemia. Pillar 2 calls for monitoring post-operative bleeding and avoiding secondary hemorrhage. Pillar 3 concludes patient blood management by determining any post-operative anemia prescriptions. This phase pays close attention to *blood composition and volume*.

3.2 Individualized Strategies and Decisions

To develop the prototype tool it was decided that 12 case studies would be sufficient for providing a range of experiences. These case studies facilitate an opportunity to execute the trainee's individualized strategy / decisions for each procedure. These cases would be *elective surgeries* (as opposed to urgent or emergent) because elective surgeries offer greater opportunity for patient / physician interaction and preparation for the surgery in the pre-operative phase.

Information necessary to fully represent (model) a de-identified case study would include: the pre-operative note, any blood management orders/forms, and the post-operative note. Incorporating as much patient data as available would result in a complex characterization of the procedure.

With an understanding of the blood management philosophy, introduction to the blood management techniques employed, and the determined number and variety of patient case studies needed to provide a breadth of training experience, the developers made a few broad, preliminary assumptions on what the tool would entail:

- 1) it will be constructed as a web-based interface (for accessibility) and it will contain real patient case studies with an assessment capability (for lessons learned)

- 2) the simulations will unfold in real-time (for decision-making experience)

- 3) there would be an end-of-the session assessment whereby the trainee could compare his decisions / actions to the actual case study.

4. HOW TO BUILD IT

Developing a tool premised on a patient blood management protocol requires diverse modeling skills. For example, included in the model is the mapping soft or fuzzy data (human factors) such as patient subjective data and procedural decision-making on the part of the surgeon. Grounding the soft data characterization is the mathematical modeling used to accurately chart physiological changes of the case study based on a range of things like patient response to a procedure and unexpected or inadvertent bleeding.

Model development began by employing a *system dynamics modeling* paradigm as a means of crafting a visual representation of the factors and their correlative and/or causal relationships. This enabled the developers to take a holistic approach to tool design. This modeling methodology was followed by significant mathematical modeling to ensure precise measurements of patient vital signs like blood pressure, blood volume (loss), and other physiological reactions such as heart rate, oxygen saturation (SPO₂), and respiratory rate.

4.1 System Dynamics Modeling

The developers started by mapping out a system dynamics model to provide an explanation and validation of the tool. And because the model is all about blood management, the dependent variable would be blood volume. Independent variables include drugs, fluids, surgery, ventilation or oxygenation – all factors that could affect blood volume. As with any system dynamics model the first step is to craft a causal loop diagram. Mapping the causal loops led to representing the relationships between and among variables needed into stock and flow representations. These representations establish the feedback loops which serve to indicate the dynamic system that captures how the body functions from a mathematical standpoint. Figure 1 is the initial attempt at the system dynamics model.

Finally, in the Post-operative Phase the training tool reflects the same exercises as those provided in the pre-operative phase. Figure 2 below is a screen shot of the web interface.

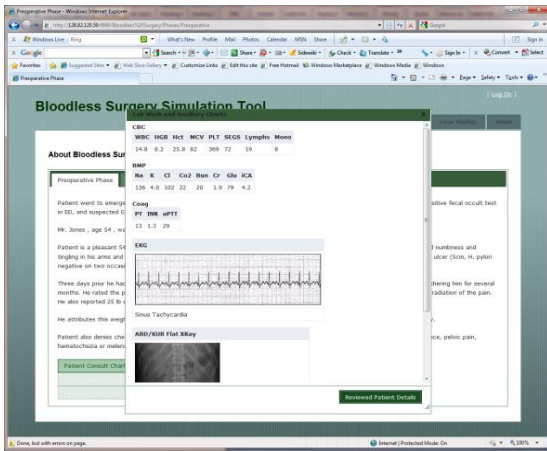


Figure 2 Screenshot of the Web Interface

The prototype tool has proven a viable mechanism for training blood management philosophy and techniques to the specified audience, trained anesthesiologists and surgeons. In actuality, the tool more literally serves to re-train with a view to better blood management practice to be used in the pre-, peri-, and post-operative phases of patient care. As a process or philosophy the developers have taken great measures to integrate and align the training experience with blood management techniques by linking this information with existent surgical decision making and practice. The medical body of literature details a knowledge gap in practice and this tool serves as a way to lessen that gap [16].

The tool design is an XML model that incorporates the comprehensive system dynamics model and the mathematical models. It mirrors the specific procedures used in the patient blood management practice by dividing the training into pre-, peri-, and post-operative phases as these phase act out the various aspects of the 3 practices. The encoded case studies drive the behavior of the tool without the need to re-implement the software. As a result the tool is designed in a scalable fashion and allows the registration of a large number of case studies.

The software that enables this tool is designed with three primary goals.

1. Keep the entry barrier for the medical professional to a minimum
2. Keep development and associated maintenance costs of the software to a minimum

3. Develop a training tool software that is scalable and extendable

The entry barrier was kept to a minimum by designing the tool as a web-based service, with potential trainees being able to access the service via web browsers, e.g., Internet Explorer, Mozilla Firefox, Google Chrome, etc. This also precluded the installation of proprietary software for the trainee.

The software is developed using a standard programming language, C#, available as part of Microsoft's .NET environment. For an interactive user interface, JavaScript was chosen as a language for producing visual artifacts, which range from showing patient demographic information to displaying patient medical status via ECG waveforms, and capturing user input, which includes prescription choices, etc. JQuery, a free-to-use software released under MIT License, is used for conforming to cross-browser JavaScript requirements.

The interactive flows of the training span the three operative phases. These flows are driven by the information specified in the case studies. As a result, depending on the case study selected the specifics of the trainee experience are updated. The tool also allows specifying a large number of case studies, each of which is captured using a data model defined as part of the software build. That data model for capturing case studies can be used for specifying new case studies or modifying existing case studies. Once case study models are registered with the tool, the tool makes them available for training purposes requiring no change either to the software or the deployed tool. This approach to drive the software behavior via case study models meets the extensibility goal and the capability to store a large number of such models meets the scalability goal.

5. CONCLUSION

The initial testing of the tool proved it to operate / function with great ease moving from decision point to decision point and phase to phase. The ability to progress at will or allow the simulation to provide error comments requiring corrective actions allows the physician to train at his own discretion. The assessment values at the close of the exercise allow the trainee to evaluate his own performance.

It is significant to note the multidisciplinary nature of the development of this tool. Expertise was called on from engineering, computer science, and social science disciplines. Integral was the role of medical subject matter experts in to include MDs, RNs (registered nurses), CRNAs (certified registered

nurse anesthetists) who shared in the development of the interface, explanation and application of the blood management practice, and testing and evaluation of the tool itself.

Medical simulation training is fast becoming a necessary modality in healthcare education. This tool is the means to providing a comprehensive, effective, and time-sensitive learning experience.

6. FUTURE WORK

As mentioned at the outset, simulation training for medical professionals is an acknowledged means of effective and efficient training as it provides a learning environment with depth and breadth. And although medicine is an evidence-based exercise, the blood management is relatively novel and goes contrary to standard practice. For a variety of reasons, many skilled medical practitioners are uninformed, unable, or unwilling to engage blood management techniques.

The purpose of this training tool is to facilitate that learning exercise for closing the learning gap through simulation re-training of new information and new technique. The fact that blood management practices have a dramatic impact on hospital finances and patient outcome itself serves as valid reasons why this area of medical simulation needs to be developed [7].

Future work along the lines of blood management simulation training can take many directions. First, there are divergent paths in the operating room setting as the medical professionals there, the anesthesiologist, the surgeon, and the nurses, have different roles and as such there can be a mental disconnection between / among them. Developing a tool that represents the roles of each to reflect the knowledge of who carries out different tasks and how to train to those tasks as a multidisciplinary team. A second potential training scenario is a tool that focuses on the mechanical aspects of surgical procedure in the intra-operative phase to include such things as precision with surgical incision, ANH and cell-salvage implementation, patient autologous blood donation for hemodilution (how much and when).

REFERENCES

- [1] Kyle RR, Murray WB, eds. *Clinical Simulation: Operations, Engineering, and Management*. Amsterdam: Elsevier, 2008.
- [2] Dunne Jr, MD, Tracy JK, Gannon C, Napolitano IM, (2002), "Perioperative anemia: an independent risk factor for infection, mortality, and resource utilization in surgery." *Journal of Surgical Research* Feb;102(2):237-44.
- [3] Shander A, et.al, (2010). "Activity Based Costs of Blood Transfusions" *Transfusion* 50 (4): 753-765.
- [4] Moskowitz DM, Klein JJ, Shander A, et al, (2004). "Predictors of transfusion requirements for cardiac surgical procedures at a blood conservation center." *Annals of Thoracic Surgery*; 77: 626-634.
- [5] Thomson A, Farmer S, Hoffmann A, Isbister J, Shander A, (2009). "Patient Blood management- a new paradigm for transfusion medicine?" *ISBT Science Series* 4, 423-435.
- [6] Shander A., et.al. *Perioperative Blood Management: A Physician's Handbook*. 2nd edition. Bethesda: AABB, 2009.
- [7] Seeber P, Shander A. *Basics of Blood Management*. Hoboken: Wiley-Blackwell Publishers, 2007.

AUTHOR BIOGRAPHY

Catherine M. Banks PhD, is Research Associate Professor at VMASC. Her focus is on qualitative research among the social science disciplines to serve as inputs into various modeling paradigms: game theoretical, agent-based, social network, and system dynamics. Dr. Banks' research includes models representing humans and human behavior to include the translating / mapping of data for quantitative representations, modeling states and their varied histories of revolution and insurgency, political economy and state volatility, and medical simulation. She has authored and edited books and journal articles on these topics and is contributor and co-editor of *Modeling and Simulation in the Medical and Health Sciences* (Wiley Publication to be released April 2011).

John A. Sokolowski PhD, is the Executive Director for VMASC, supervising 50 researchers and staff with an annual funded research budget of \$10 million. He supervises research and development in Transportation, Homeland Security, Defense, Medical M&S, Decisions Support, Business & Supply Chain, and Social Science (real-world) M&S applications. He is contributor and co-editor of Modeling and Simulation in the Medical and Health Sciences (Wiley Publication to be released April 2011).

MODEL SYNTHESIS USING A MULTI-AGENT LEARNING STRATEGY

Sebastian Bohlmann^(a), Arne Klauke^(b), Volkhard Klinger^(c), Helena Szczerbicka^(d)

^{(a)(d)}Department of Simulation and Modeling, Leibniz University Hannover, 30167 Hannover, Germany

^{(b)(c)}Department of Embedded Systems, FHDW Hannover, 30173 Hannover, Germany

^(a)bohlmann@sim.uni-hannover.de, ^(b)arne.klauke@fhdw.de, ^(c)volkhard.klinger@fhdw.de, ^(d)hsz@sim.uni-hannover.de

ABSTRACT

In this paper we give an overview of our multi-agent based model identification framework. We are identifying functional relationships in process data. We do this by using multi-agent based heuristic algorithms. Moreover we give a proof of concept concerning the abilities and performance of our system.

Keywords: model synthesis, agent-based evolutionary computation

1. INTRODUCTION

Manufacturing systems are one of the largest application areas for modelling and simulation. In particular the pulp and paper industry is one instance of a large-scale production processes (Bohlmann and Klinger, 2007). We have brought up a framework for modelling and simulation of those process environments in (Bohlmann, Klinger and Szczerbicka, 2009), (Bohlmann, Klinger and Szczerbicka, 2010b) and (Bohlmann, Klinger and Szczerbicka, 2010c).

This paper focuses on the identification procedure. We consider time series extracted from process data. These time series are subdivided in input and output series. The problem treated here, is to find a functional relationship between the input and output series. At the beginning it is unknown which of the input series are actually used in that relationship. Our approach to solve this problem is a multi-agent based learning strategy (Bohlmann, Klinger and Szczerbicka, 2010b).

In figure 1 the system identification overview is shown. It consists of two basic steps, the preprocessing and the multi-agent based optimization. The process data input (PData) is used to generate an appropriate process model (Law and Kelton, 2000).

To verify this identification procedure we have to evaluate the different steps very carefully not only to its technically correct function but on its performance behaviour.

$$\begin{aligned} f_1((x_1)_t, \dots, (x_m)_t) &= (y_1)_t, t \in \mathbb{N} \\ &\vdots \\ f_j((x_1)_t, \dots, (x_m)_t) &= (y_j)_t, t \in \mathbb{N} \end{aligned} \quad (0)$$

The verification strategy is based on a set of data sequences $(x_1)_t, \dots, (x_m)_t, t \in \mathbb{N}$, called Input Sequences and Output Sequences $(y_1)_t, \dots, (y_j)_t, t \in \mathbb{N}$, which are related to the Input Sequences by a functional relationship $f: \mathbb{R}^m \rightarrow \mathbb{R}^j$ (formula 0), illustrated in figure 2.

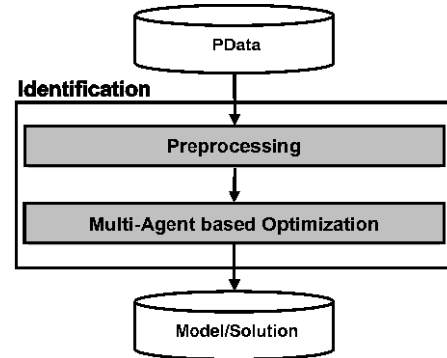


Figure 1: Function block view

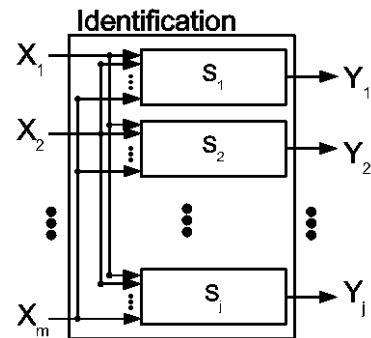


Figure 2: Input/ Output sequence

In figure 3 an example for $m=10$ and $j=1$ is shown. The problem we are solving is to identify this function f , only knowing values of $(x_1)_t, \dots, (x_m)_t$ (thin lines) and $(y)_t$ (thick line) for a limited set $T \subset \mathbb{N}$ of time indices, which may differ for each sequence. In this paper we are treating only problems with $j=1$.

Our approach for this challenge is formed by the identification framework used for process model identification and it uses the data management framework presented in (Bohlmann, Klinger and Szczerbicka, 2010c).

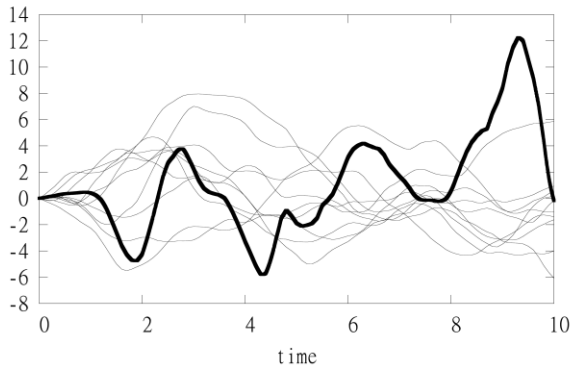


Figure 3: Input and output data series

The pre-processing is followed by a Multi-Agent-based Learning Strategy (section 3.1.2) using memetic evolutionary algorithm.

In the next sections the identification framework is explained in detail. It consists essentially of two parts (see Figure 1): A data pre-processing unit and our evolutionary algorithm. Finally some examples, proving the functioning of the framework, are discussed.

2. DATA PREPROCESSING

In this section the preprocessing of the raw data is explained in detail. figure 4 presents the basic function blocks.

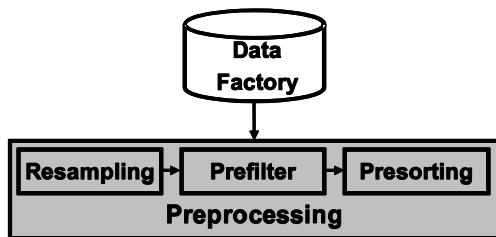


Figure 4: Data Preprocessing

2.1 Data Factory

The data used for the identification process is either extracted from various sensors in a real system or they may be synthetic. In the first case the data is usually collected over a long time period and saved in a data archive. This raw data then runs through different preprocessor units, explained below, to build usable data series for the identification framework. In the second case the raw data is produced synthetic using random data streams and is then passed to the preprocessing units.

2.2 Resampling the Data

In this initial unit the weaknesses of the data-recording by the sensors are remedied. The sensors distributed over the system are usually not working periodic or even synchronously. Moreover they might sometimes measure values, which are obvious wrong. These error values are simply removed from the data set. The actual task in this module is to produce a time series of equidistant data samples, each of which describes the state of the whole system at a moment.

To achieve this, at first the data from each sensor is linearly interpolated and then smoothed by a convolution. After that the new data sequences are equalized with the original values from the sensors. Finally equidistant values from each sensor are picked and combined to data samples, which describe the whole system, as desired.

2.3 Prefiltering the Data

The environment built for the evolutionary algorithm, explained in the next section, needs a predetermined number of data samples. In general the number of data samples delivered by the resampling unit is too large. Moreover it may contain redundant data samples, containing no information. This may happen if the state of the real system does not change for a time period. In this unit the samples, which contain the most information, i.e. these with the highest entropy, are chosen. This is implemented in different prefilter modules.

1. *No Prefilter*: The simplest way is to choose just the last 729 samples.
2. *Random Prefilter*: The samples to be passed onto the planets are chosen randomly.
3. *Weighted Random Prefilter*: The samples to pass on are chosen randomly, but with different probabilities. This probability corresponds to the angle between the input values of the current sample and its predecessor and successor.
4. *k-means Prefilter*: In this method we are using a cluster algorithm to choose the samples to be passed on (Kanungo, Mount, Netanyahu, Piatko, Silverman and Wu, 2002). The data set delivered by the resampling unit is subdivided in blocks of a fixed size. In each of these blocks we build a fixed number of clusters and only the centres of these clusters are passed to the planets. We have decided to use k-means clustering because one can choose the number of centres from start. Furthermore the clusters generated by this algorithm are formed spherical, what seems to be the most suitable form for our purpose.

2.4 Presorting the Data

After the samples are selected they need to be arranged on the planets in a useful way. The simplest method is to keep them in their current order. But there are concepts, which can improve the system behaviour.

One approach to order the samples in a more useful way is the so called TSP Filling (Travelling Salesman Problem). The samples are arranged in a way that approximately minimizes the sum of the distances of neighbouring samples, with respect to a chosen metric. This concept can be generalized by not just taking the direct neighbours into account, but the next n neighbours in both directions for each sample.

3. MULTI-AGENT-BASED OPTIMIZATION

The algorithm uses an evolutionary approach to find the functional relationship in the process data. We have created an environment which offers aliments to the creatures living in it. These creatures own a genotype, which they are trying to adapt to their location in the environment by building children with a changed genotype. A creature which is well fitted to its location has a better ability to absorb aliments from it. Aliments are used to perform various evolutionary operations to build a child. The genotype passed to the child can be mutated, crossed with the genotype of another creature, and enhanced in several ways. The creatures can move within their environment and interact with other creatures.

The environment is representing the data set, the local view is just a subset of it. The creatures are software agents and their genotype is a model function, approximating the functional relationship. We can rate how well an agent is fitted to its location by calculating the error of his model function using the local data and a search metric.

In the following section we will give an overview of the system architecture, depicted in figure 5.

3.1 System architecture overview

The architecture is organized in four stages, the data processing and local and global agent environment. Each is arranged in functional levels, for data management, agent behaviour control, supervision, execution and synchronization, regarding the overall management. The basic level is called MPU (Multi Processor Unit) and represents the system thread representation.

The optimization starts with the initialization of the preprocessed data, managed in the data processing stage. The splitter in the supervision level supplies the evaluation units in the other stages with data samples.

To illustrate the agent based algorithm, we give a detailed description of the environment, the individuals live in, of the individuals themselves and of their behaviour.

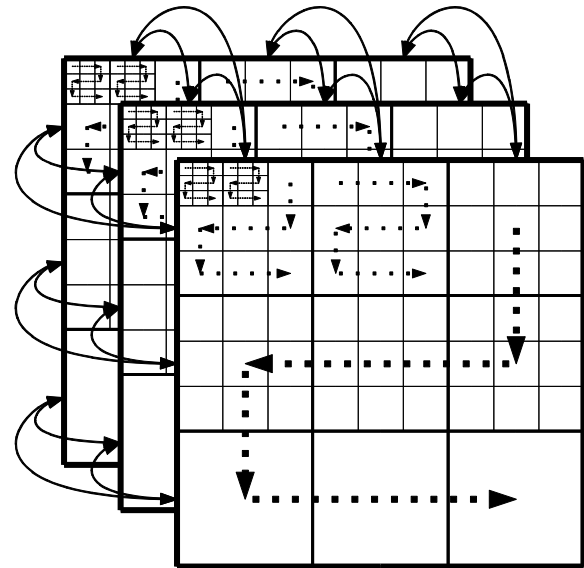


Figure 5: Planet Surface

3.1.1 Agent Environment

The environment of the agents consists of areas. Each area contains one data sample and can hold one agent. Moreover areas can be linked to other areas, called neighbours. These neighbours serve for two purposes. At first the agent held by one area can be moved onto one of the areas neighbouring areas. Secondly the links between the areas are used to build the set of data samples for the local learning.

The areas are aggregated in planets. The surface of a planet is a torus, represented by a quadratic field of areas.

This surface is build in a recursive pattern of squares containing nine elements, filled meander like. This method leads to the planet size $9^3 = 729$ and is illustrated in figure 6. Each area on a planet is linked to its four neighbours to the left, right, top and bottom.

Finally the planets build the universe, which is controlled by the universe supervisor. All planets are controlled by so called planet supervisors. Some of the areas on each planet are marked as beam areas. When a certain number of iterations have passed, the planet supervisor sends copies of all agents, which are placed on a beam area to a randomly chosen area on a randomly chosen planet.

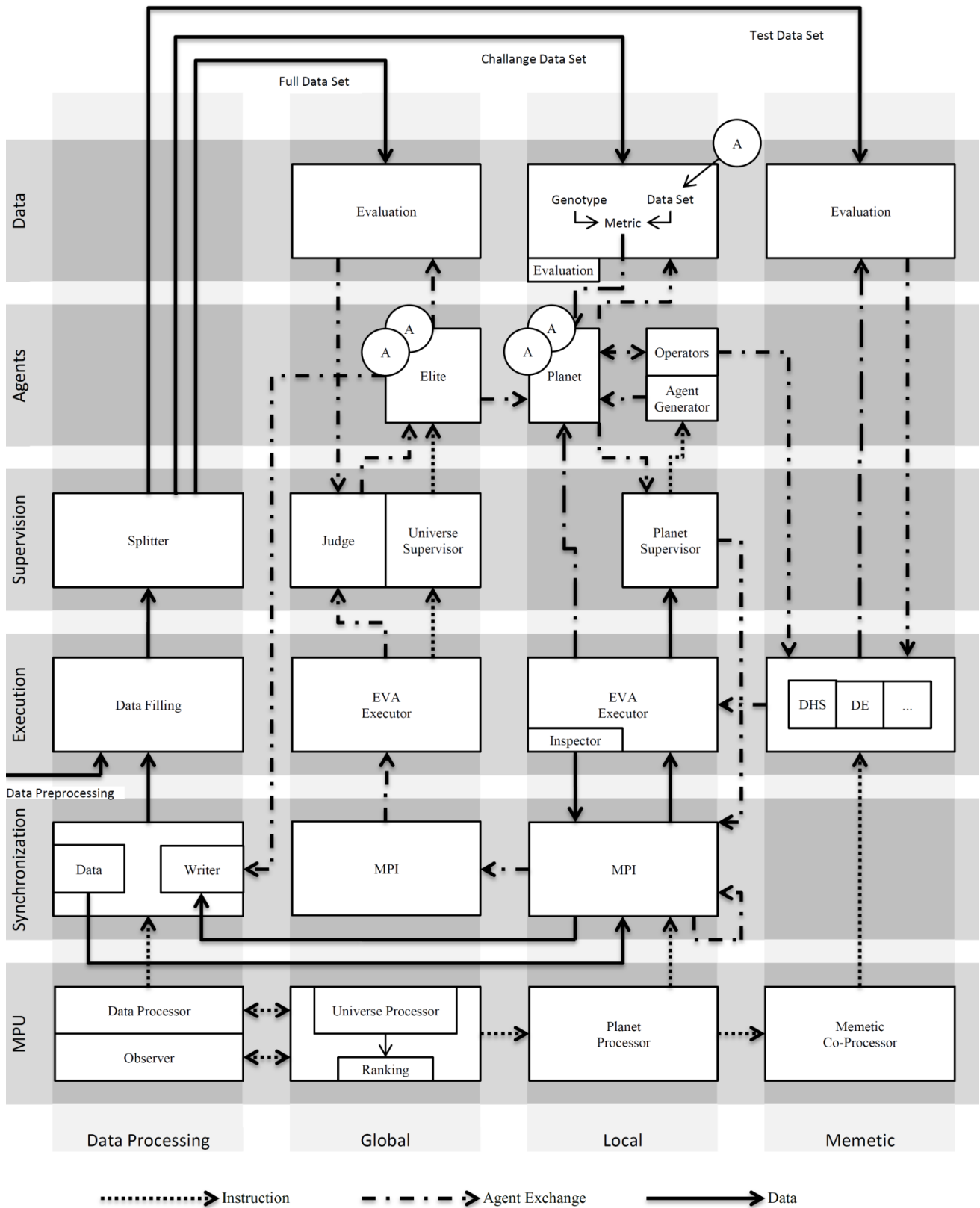


Figure 6: System Architecture

3.1.2 Software Agents

The software agents are operating in the environment described before. Each agent owns a model function and is placed onto an area. The model function is stored in a tree representation and is build using elementary operations like \sin , $+$, $*$, $/$ the variables x_1, \dots, x_m and a set of parameters (Schmidt and Lipson, 2007). This concept is shown in figure 7. Moreover agents may build a child, to pass on their information. We have implemented in the following evolutionary operations:

Replication: The individual produces a copy of himself. This is the most expensive operation.

Cross: Two individuals interchange randomly chosen parts of their model functions.

Mutate: A part of the model function is replaced by a randomly build function.

Enhance: The structure of the model function is simplified, if possible.

Short Learn: The parameters of the model function are fitted to the local learn data using a simple, but fast algorithm.

Learn: A more sophisticated algorithm is used to calibrate the parameters of the model function. This operation is implemented in a memetic coprocessor.

All of these operations have an energy effort, which is subtracted from the agents energy, if the operation is performed. Furthermore the agents can measure the error of their model function using different kind of search metrics.

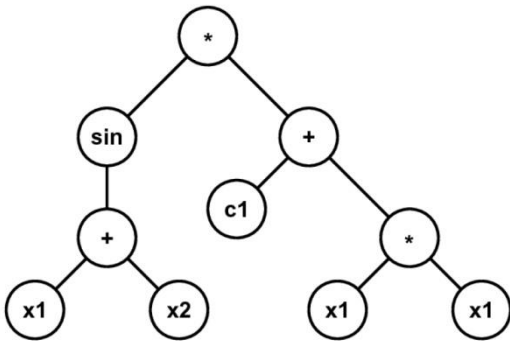


Figure 7: The tree representation of the model function

3.1.3 Agent Behaviour

Each iteration starts with an update of the agents properties. The age of the individual is increased. After that the energy level is recalculated. This happens as follows: The individual calculates the error of its model function using one of the search metrics described in section 3.4. If this error is too big the energy level is not increased. Else, the individual is allowed to absorb the energy offered by the area it is placed on and a value depending on the complexity of its model function is subtracted from its energy level. When these operations

are done we decide whether the agent may live for one more iteration or not.

If the energy level is not positive the agent is removed. If the agent had a child, it is placed onto the area.

If the energy level is positive the agent tries to move to a randomly chosen neighbouring area. When the chosen area is empty, the agent just moves. If the area is already occupied by another agent, the procedure depends on the agent's energy level. If it is not high enough to perform a cross operation, the agent does not move. Else a cross operation is performed. If the cross operation yields a new agent it is saved as the child of the original agent and the agents do not move. After the move operation, the individual tries to build a child, if none is present. Depending on the agent's energy level, the agent performs a Replication, Mutate or Enhance operation to build his child.

In the next step the agent may perform a Learn operation, if he is not adult (without loss of energy), or if his energy is high enough it is decided randomly if the agent is allowed to learn or not.

Now the individual performs a short learn operation, if its age is appropriate.

Finally, if the agent has moved in this iteration and owns a child, he tries to place it on his last area.

3.2 Optimization stages

In the last section the agents have been introduced. Here we map this agent based algorithm on the system architecture. According to the planet setup, there is the global stage (universe) and the local stage (one planet). The fourth stage is formed by memetic coprocessors, assigned to the planets. These stages form separate execution loops, which run in parallel. The listing below gives an overview of these three loops.

1. *Global Stage*: This loop is used to manage the elite population. Agents nominated in the local loop are passed via the MPI (Message Passing Interface, synchronization level) to the EVA Executor (execution level) and then to the Judge (supervision level), which decides, if the agent is added to the elite population (agent level). The agents in the elite population are evaluated on the full data set (data level). If one of them fulfils the termination criterion the algorithm stops. The universe processor (MPU level) returns the ranking of the best individuals if the algorithm terminates.
2. *Local Stage*: In this stage the agents are generated and put onto the planets. Once they are placed on an area, the agents start their life cycle, described in subsection 3.1.3. In the data level the agents evaluate their model function, using a search metric and a subset of the challenge data set, called local learn data. In the MPI (synchronization level) agents are exchanged between the planets.

3. *Memetic Stage*: In this loop agents, which were passed from the Operators (agent level, local stage) to the memetic unit (execution level), are optimized by more expensive algorithms, like downhill simplex (DHS) (Nelder and Mead, 1965). To evaluate the error of the model function, a small test data set is used. The agents are then passed back to the EVA Executor in the execution level of the local loop.

3.3 The Elite Population

There are two opportunities for an individual to get nominated for the elite population. The first is when an individual is removed from the planet and has lived for at least 200 iterations. Furthermore an individual is nominated for the elite population after every 250 iterations. The universe supervisor chooses the best 25 individuals from all nominated agents.

The elite population is used to check the termination criterion: The model functions in the elite population are evaluated not just using the local learning set, but the data from the whole universe. If one of them has an error under a predefined border the algorithm terminates and returns this model function.

3.4 Search Metrics

The agents are able to use different search metrics to estimate the error of their model functions. For each data sample in the local learn set, which is build using the neighbouring areas, the agent calculates the difference between the original output and the output calculated by the model function. For this step the agent can use the euclidean or the absolute distance. In the next step these values are aggregated by building the

mean value or the maximum over all samples. They might perform these steps using not all, but a randomly chosen subset of their local learn data, in these cases the metric is called partial.

4. PROOF OF CONCEPT

To proof the basic usability of our framework we executed 8 different experiments with different problem complexities to be solved by the agent system. Each experiment is repeated 30 times to reduce statistical deviation. For each run we take the runtime beginning with the first agent generation and ending at the first successful detection. A successful detection is defined by a mean absolute error of 0.001 over the whole dataset. Each run has a maximal runtime of 1800 seconds. If no valid solution is found in time, the execution is aborted and counted as an unsuccessful execution. To proof the parallelisation concept we repeated 8 times 30 experiments with 1 execution core (one Planet) and 8 cores.

4.1 Experimental Setup

This section specifies the used configurations and dependencies. All experiments are executed on a Dell PowerEdge R815 with in total 4 AMD Opteron 6174 processors (each providing 12 cores with respectively 128KByte L1-cache, 512 KByte L2-cache and common 12 MByte L3-cache) and an overall RAM configuration of 128 GByte. For the parallelization evaluation this platform provides a scalable hardware environment.

The challenge to be solved by the agent system is generated synthetically. Because randomness in the generation has huge impact on the detection system the same dataset is used in all 30 runs per experiment. We generated a five dimensional time series input stream

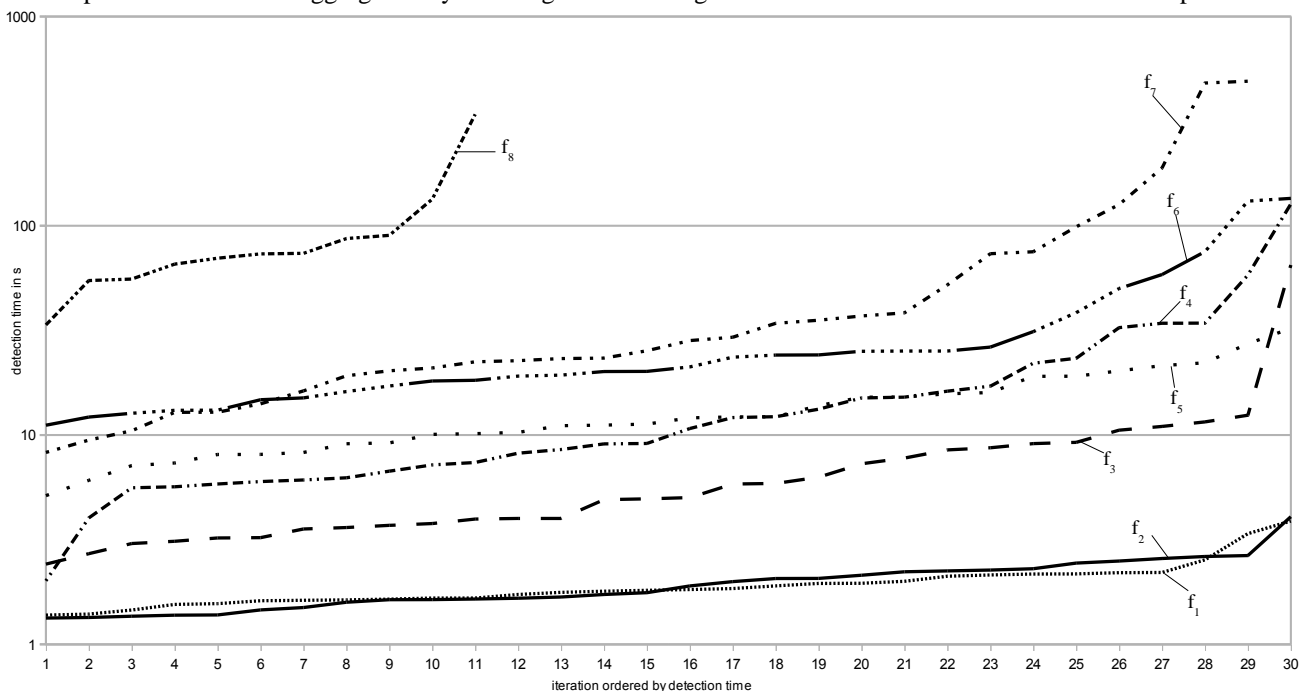


Figure 8: Measured ordered runtime

with 10^5 samples each. Further a result stream with the same length is generated using one of the 8 functions in formula (1) to (8). To measure the quality of the result a metric calculating the mean absolute error is used.

$$f_1(x_1, \dots, x_5) = x_1 \cdot x_2 \quad (1)$$

$$f_2(x_1, \dots, x_5) = \sin\left(x_1 + \frac{\pi}{4}\right) \quad (2)$$

$$f_3(x_1, \dots, x_5) = 42 \cdot x_1 \cdot x_1 + 22 \cdot x_2 \quad (3)$$

$$f_4(x_1, \dots, x_5) = \sin(x_1 \cdot x_2 + 3) + 4 \quad (4)$$

$$f_5(x_1, \dots, x_5) = x_3 + x_4 + 4.3 \cdot \sin(x_1 + 7) + 1.65 \quad (5)$$

$$f_6(x_1, \dots, x_5) = x_1 \cdot x_1 + x_2 \cdot x_2 + x_3 \cdot x_3 + 22 \quad (6)$$

$$f_7(x_1, \dots, x_5) = -3.23 \cdot \sin(x_1) \cdot \sin(x_2) + 2.43 \cdot \sin\left(x_1 + \frac{\pi}{4}\right) \quad (7)$$

$$f_8(x_1, \dots, x_5) = -122 \cdot x_1 + 2.3 \cdot x_2 + 0.2 \cdot x_4 + x_3 \cdot \sin(0.1 \cdot x_4) \quad (8)$$

4.2 Results

Figure 8 illustrates the results for the parallelised agent system with 8 planets. Each line type represents the different successfully ended runs for one of the generator functions ordered by runtime. The y-scale is logarithmic. As expected the mean runtime is higher if the function is more complex. This value can also be found in table 1. Some of the data lines do not have 30 samples, because not all runs resulted in a valid solution. For function 1 to 6 the agents evolution always leads to the correct structure. Especially for the last two functions the highest runtimes increase strongly.

Table 1: Statistics for different functions

	Mean	Std. Dev.	Detection Rate
f_1	1.8 s	0.5 s	100 %
f_2	1.8 s	0.6 s	100 %
f_3	5.0 s	11.2 s	100 %
f_4	9.9 s	23.9 s	100 %
f_5	11.7 s	6.4 s	100 %
f_6	20.6 s	31.0 s	100 %
f_7	25.3 s	121.6 s	97 %
f_8	73.4 s	84.4 s	37 %

This effect caused a deadlocked evolution, if the overall diversity of the agents is low. When the overall number of agents and the separation by multiple planets is increased this probability decreases. This is indicated if we compare the total number of detections and the speedup as done in table 2. Here the functions five to eight are not listed because the detection rate is too low to acquire adequate measurements for a non-parallelized system. Speedup for complex challenges (3-5) is effective. As positive detection rates increase and detection runtime decreases the positive effect is higher than the expected factor 8. At the moment more planets

do not lead to a better system performance on the used machine because the other system components (see figure 6) consume the remaining system resources. We conclude that also the agents only use heuristics for interaction and learning the combined execution is target-oriented.

Table 2: Statistics for parallelization

	Mean x1	Mean x8	Speedup	Detection Improvement
f_1	5,6 s	1.8 s	309 %	100 %
f_2	6,5 s	1.8 s	356 %	100 %
f_3	28,0 s	5.0 s	561 %	500 %
f_4	82,0 s	9.9 s	826 %	230 %
f_5	59,0 s	11.7 s	505 %	272 %

5. SUMMARY

Modeling and simulation of non formalized system behavior still is a grand challenge for science and engineering. As we demonstrated it is possible to implement a machine learning system to help modeling specialists to gain knowledge from the data produced by the original system. In this scenario it is required to formulate the produced recommendations in a human comprehensible form. Differential equations (and simple equations, as in this concept paper) are one possible knowledge representation. And in difference, from knowledge e.g. learned by a neuronal net, knowledge is not encapsulated. Engineers have a huge tool kit to continue processing such a result. As done for a simulation system in (Bohlmann, Klinger and Szerbicka, 2009) such a agent based modeling support system can easily connected to real word data sources and could be helpful to enhance or generate complex models for simulation environments (Zeigler, Praehofer and Kim, 2000). As a result the complexity to model a complex process is simplified by using the analytic strength of the modeling engineer and the knowledge compression strength of an agent based machine learning environment.

6. FURTHER WORK

The further work has two key aspects of activity: Increase the parallelization to be able to use more agents. As mentioned before the memetic co-processor cores use the majority of our machine resources. All used memetic algorithms are suited for SIMD coprocessors and would scale the system to about 40 planets. The second work to be done is to reduce the number of problem specific parameters by the help of control loops.

Finally the framework is written as generic as possible. There are only few dependencies e.g. the problem has to be dividable into local challenges. So we like to formulate solvers for different known problems in the area of modeling and machine learning.

REFERENCES

- Bohlmann, S. and Klinger, V. (2007)
Modellbildung für kontinuierliche Produktionsprozesse in der Papierindustrie. Forschungsberichte der FHDW Hannover (ISSN 1863-7043), 08:1–20.
- Bohlmann, S., Klinger, V., and Szczerbicka, H. (2009)
HPNS - a Hybrid Process Net Simulation Environment Executing Online Dynamic Models of Industrial Manufacturing Systems. In Proceedings of the 2009 Winter Simulation Conference M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.
- Bohlmann, S., Klinger, V., and Szczerbicka, H. (2010a)
System Identification with Multi-Agentbased Evolutionary Computation Using a Local Optimization Kernel. In Submitted to ICMLA 2010 (International Conference on Machine Learning and Applications).
- Bohlmann, S., Klinger, V., and Szczerbicka, H. (2010b)
System identification with multi-agent-based evolutionary computation using a local optimization kernel. In The Ninth International Conference on Machine Learning and Applications, pages 840–845.
- Bohlmann, S., Klinger, V., and Szczerbicka, H. (2010c)
Co-simulation in large scale environments using the HPNS framework. In Summer Simulation Multiconference, Grand Challenges in Modeling & Simulation. The Society for Modeling and Simulation.
- Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. (2002)
An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24:881–892.
- Law, A. M. and Kelton, W. D. (2000)
Simulation Modeling and Analysis. McGraw-Hill
- Nelder, R. and Mead, J. (1965)
A simplex method for function minimization. Computer Journal, 7(4):308–313.
- Schmidt, M. and Lipson, H. (2007)
Comparison of tree and graph encodings as function of problem complexity. In GECCO '07: Proceedings of the 9th annual conference on Genetic and evolutionary computation, pages 1674–1679, New York, NY, USA. ACM.
- Zeigler, B. P., Praehofer, H., and Kim, T. G. (2000)
Theory of Modeling and Simulation: Integrating Discrete Event and Continuous Complex Dynamic Systems. Academic Press, San Diego, USA, 2 edition.

AUTHORS BIOGRAPHY

SEBASTIAN BOHLMANN is a Ph.D. candidate at Department of Simulation and Modelling - Institute of Systems Engineering at the Leibniz Universität at Hannover. He received a Dipl.-Ing. (FH) degree in mechatronics engineering from FHDW university of applied sciences. His research interests are machine learning and heuristic optimization algorithms, complex dynamic systems, control system synthesis and grid computing. His email address is <bohlmann@sim.uni-hannover.de>.

ARNE KLAUKE is a researcher at the university of applied science FHDW in Hannover. He received a Dipl.-Math. from the Gottfried Wilhelm Leibniz Universität at Hannover. His email address is <arne.klauke@fhdw.de>.

VOLKHARD KLINGER has been a full time professor for embedded systems and computer science at the university of applied sciences FHDW in Hannover and Celle since 2002. After his academic studies at the RWTH Aachen he received his Ph.D. in Electrical Engineering from Technische Universität Hamburg-Harburg. He teaches courses in computer science, embedded systems, electrical engineering and ASIC/system design. His email address is <Volkhard.Klinger@fhdw.de>.

HELENA SZCZEBICKA is head of the Department of Simulation and Modelling - Institute of Systems Engineering at the Leibniz Universität at Hannover. She received her Ph.D. in Engineering and her M.S in Applied Mathematics from the Warsaw University of Technology, Poland. She teaches courses in discrete-event simulation, modeling methodology, queuing theory, stochastic Petri Nets, distributed simulation, computer organization and computer architecture. Her email address is <hsz@sim.uni-hannover.de>.

SERVICE OPTIMIZATION FOR SYSTEM-OF-SYSTEMS BASED ON POOL SCHEDULING AND INVENTORY MANAGEMENT DRIVEN BY SMART SIMULATION SOLUTIONS

Agostino Bruzzone, Marina Massei, MISS DIPTM University of Genoa
Via Opera Pia 15, 16145 Genova, Italy
Email {agostino, massei}@itim.unige.it - URL www.itim.unige.it

Enrico Bocca, Simulation Team
Via Molinero 2, 17100 Savona, Italy
Email enrico.bocca@simulationteam.com - URL www.simulationteam.com

Keywords: *Simulation, Power Plants, System-of-Systems, Pool Management, Service, Maintenance, Decision Support System, Optimization.*

ABSTRACT

The aim of this research is to support service and maintenance of pools of System-of-Systems, such as power plants or vessel/aircraft fleets by using simulations model dynamically integrated with smart optimizer driven by Artificial Intelligence techniques.

The proposed methodology permit to create a framework to evaluate, optimize and test the service and maintenance policies (involving both inventory and scheduling); this framework is based on a simulator combined with an intelligent optimizer

The authors proposed a new metrics to evaluate the real performance of pool service management of the whole complex system and support the optimization processes.

INTRODUCTION

The work performed by the authors in this research is to verify the benefits of the integration of simulation models with Artificial Intelligence (AI) techniques in complex system service and maintenance; in fact in most of the case the complex systems rely and require very expensive and sophisticated maintenance/service support; in fact this problem become even more difficult when the complex systems due to their interaction evolve becoming a system-of-systems; in this case the use of simulation is usually the only reliable approach to face the difficulties related to the management of their service/maintenance; in particular in this context it is critical to consider some KPI (Key Performance Indexes) with special attention to availability, costs, resource utilization, readiness. From this point of view one interesting opportunity for improving these KPIs and for guaranteeing better overall performances is to develop pool management strategies able to generate synergies in

this complex framework; in fact in the real industrial and business case provide many opportunities to apply "pool management" to system-of-systems; in fact there are several frameworks where to apply this approach and authors have completed successfully several R&D projects for major companies on this subjects such as:

- Service for fleets of helicopters and resources devoted to provide Search and Rescue
- Service for fleets of resources (i.e. Buses, Metro, Trams, etc.) for a mass transportation companies
- Service for different fleets of tank vessels supporting different chemical industries
- Industrial Plants Service Pool Management

In fact along the years the authors have developed methodologies and simulation models to face these challenges and in particular they have developed LAPIS (Lean Advanced Pooling Intelligent optimizer and Simulator) suite, integrating M&S (Modeling & Simulation) and AI, to support decision making over this context; therefore this paper proposes an example applied to power plants service over a pool of different sites and units; the results obtained from the simulator are used for demonstrating the potential and benefits of the new methodology proposed as well as validation support.

This papers focus in fact on applying LAPIS to power plant pool management for optimizing their service over a wide spectrum of target fuctions.

In addition by this approach is possible to control and optimize different aspects of system-of-Systems maintenance, as power plant pools, in terms of different hypotheses in term of the balance between service quality (i.e. availability), cost estimation and constrain respect; the case study proposed represent an example of system-of-systems simulation able to manage all the aspect above mentioned and to guarantee good results in a real industrial case.. The research activities are synthesized in the LAPIS model description and the

validation of this approach is obtained by the experimentation on a real power plant case study application (the data proposed in the analysis are modified due to confidential reason).

SERVICE & MAINTENANCE FOR A POWER PLANT POOLS

A pool of power plants is a set of power plant sites where multiple units of different machines are operating; for instance today in most of the case the traditional power generation relies on combined cycle power plants: each plant have usually more than one combined cycle each one incorporating just a main machine a Gas Turbine, a Steam Turbine and two Generators; each machine is composed by several systems (i.e. main and secondary systems) as well as by several auxiliary systems (i.e. Aqua Demi, HVAC, firefighting etc); in several case the subsystems are even very complex (i.e. Digital Control System, Burning Control System etc.); in addition the maintenance for such components is driven by preventive actions related to their use (equivalent operative hours) and on failures; the first component is strongly related to the power demand and utilization modes of each plant that is strongly affected by exogenous stochastic factors, while the failures obviously are characterized by high complex statistical distribution combining different phenomena (i.e. basic failures and rare catastrophic events).

It results evident that to provide efficient service such set of power plants corresponds to define a pool management strategy for a system-of-systems; the paper focuses in fact in the identification, design and engineering of best service and maintenance policies for such system-of-systems; in the proposed case it is considered a group of combined cycle power plants (4 up to a dozen); this management need to operate considering available resources (i.e. personnel), scheduling and available timeframes (i.e. technical, commercial and contractual constraints), inventory (i.e. spare part storage and replacement policies), acquisition/refurbishment policies (i.e. for item subjected to regeneration such as several layers of Gas Turbine Blades).

The final aim is to define the preventive service and to support decisions able to optimize costs and power plant availability and, at the same time, to reduce the risks and to guarantee a robust management in case of unexpected breakdowns, It is interesting to note that today availability concept evolved and it is more important respect to the traditional availability a new estimator related to the plant profitability; therefore the availability it

is still a very important factor, but the choice to use just that parameter was even due even to the fact that estimating profitability was quite complex until few years ago due to data availability and format; today, due to the high variability in demand and especially in energy prices over time, the target function to be maximized is often the profit achievable by a plant evaluated by a combined estimation of capability and prices integrated over time; the goal is to have the unit operative as much as possible during the most profitable timeslots; special algorithms for estimating this kind of target function are proposed by the authors (Bruzzone , Madeo, Tarone F 2010).

The final solution should include the definition of schedule, inventory management by fixing compatible timeframes with the service time cycles of each item; this solution needs to optimize concurrently availability, profitability, costs and technical commercial constraints that are often defined based on the specific case and point of view (i.e. different from user to service provider); it is evident that such application represents a very hard problem that to be solved need to be approached by the innovative techniques combining simulation and optimization as proposed in this paper.

In fact within real applications the stochastic factors (i.e. failures, repairing time, spare part delivery times, item refurbishment lead times) combined with the complex processes (i.e. refurbishment processes, commercial procedures, technical constrains) have a strong influence to the overall performance of the maintenance solutions. By the innovative approach proposed in this paper the solution is defined by applying a DSS (Decision Support System) combining stochastic simulation and intelligent optimization; therefore the positive results achieved in several previous application suggest this as appropriate approach to solve this problem (Bruzzone & Simeoni, 2002)

POWER PLANT POOL MANAGEMENT

Due to high order of interactions among different entities, stochastic factors, different objects and a lot of target functions each Power Plants Service represent itself a pretty complex framework.

In fact, often, many machines (i.e. generators, gas turbines, steam turbines, boilers) in multiple sites need to be maintained concurrently both in term of preventive maintenance as well as in term of failure recovery; for this reason modern service strategies deal with pooling the power plants and generating synergies to compensate the high degree of complexity. In the power plant maintenance case propose there are many important elements (i.e.

rotors elements have a cost of a million USD each, hot gas parts require specific controls), therefore for sure Gas Turbines and, in particular, their blades represent the most critical element to be optimized in term of service due to their costs, lead times and sensibility to different operative modes.

Each machine have many components that need to be checked, substituted and/or refurbished; for many types of turbine blades, for example, is possible the refurbishment or re-coating: it corresponds to an hi-tech process devoted to rebuild blade surface of the blades; depending by the model of the blades the process can be repeated one, two or even three times before to require the substitution with new ones; in addition the refurbishment of used component is usually costing about 1/10 respect acquisition of new elements (and a layer of gas turbine blades cost about 1 million dollar, while each single turbine include several layers), that means that an optimized management able to rotate among different sites and machines a blade layer maximizing the refurbishment is able to guarantee very big savings; therefore it is necessary to consider that at least some percentage of the elements subjected to the refurbishment processes should need to be substituted each time this is applied by new ones due to the too deep damages on the material (this is usually defined as scraping percentage in power plant blades or other regenerating items).

In this context, typical stochastic phenomena are failures, scraping percentages and the quantity of components/items/subsystems to be substituted, duration of inspections as well as duration of minor and major revisions. There are complex constraints among the maintenance over different components in the same unit, site or for the same users (i.e. space for dismounting the machines or wiliness to concentrate the operation in the same time frame). Due to this fact optimizing the preventive maintenance scheduling is not enough to manage effectively this system-of-systems, but it is required to define inventory management policies in coordinated way and to plan refurbishment activities: so it becomes necessary to simulate long term scheduling to check mutual influence of different choices, to check transversal constraints. In fact the Preventive Maintenance of many components is regulated by Equivalent Operative Hours (EOH) able to consider not only the their use, but also special operational mode that are reducing component life cycles (i.e. startups, shot downs, etc.); the EOH value is evolve obviously as a not deterministic variable due to many factors; the authors defined a set of parameters to characterize each machine that combine solar working hours of the plants (related to power production), the

intensity of the use Power Plants (related to the demand variations) and the mode of use (related to the policy for managing the machines by the users); that parameters are affected by stochastic factors and the results generates a complex behavior of the components; in fact currently the LAPIS simulator is able to be integrated with DCS and keep up-to-date on the EOH of each system in the power plant pool.

Through this example is easier to understand the complexity of the service and the emerging of complex behaviors in this system-of-systems; to keep the system under control and to guarantee good performances in this case it is critical to understand that the optimization process should be related to different target functions defined based on the specific case; in general the two main components of the these target functions for power plant service are related to service costs, power generation profitability and plant availability.

These are obviously competing functions and requires a multivariable combined optimization.

From other point of view the degrees of freedom for controlling and improving the power plant service management are related to the following main elements:

- Power Plant Preventive Maintenance Scheduling
- Power Plant Component Inventory Management
- Refurbishment Component Planning & Sequencing

The goals of this research was to create a power plants manager, that interact with an intelligent decision support system to estimate correctly the plant performances and to optimize resources, inventory and scheduling.

The approach proposed in this paper permits different and interactive modes: automated optimization integrated with the simulation as well as what if analysis; so the different solutions and policies are simulated driven by the intelligent optimizer or by the users and estimates the KPIs; the optimizer proposed for this case is based on GAs (Genetic Algorithms) due to the high number of variable and the strong influence of stochastic factors that could drive to local minimum traditional optimization techniques (Bruzzone, Signorile 1998).

In addition the approach is very robust and reliable even for evaluating different performance in term of service and maintenance of Power Plants considering market change (Bruzzone, Giribone 1998).

Therefore Pooling Power plant maintenance is based on the idea of a concurrent collaborative planning and management of the service over a set

of plants; in fact combining different machines, sites and power plant users it becomes possible to create a pool of entities requiring service and to identify how to manage the pool of resources for satisfy them; the success and the benefits in this care are strictly connected to the possibility to revamp items dismantled from one machine or plants and to use on another one reducing the acquisition of new components; these results are achievable by defining an effective sequence of major and minor inspections and a coordinated schedule; in fact it becomes possible to optimize the reuse of elements without affecting the availability of the power plant acting on the schedule of the service operations, refurbishment actions and on the inventory.

In Power plants many kinds of components may be reused more than one time after a refurbishment process (usually considering the percentage of scraped item for each treatment); to reduce new item acquisition by increasing refurbishment component uses it is necessary to finalize an optimized management of the inventory and an effective scheduling; the service planning, for instance, should be adapted anticipating or delaying the inspection/revision in order to have refurbishment items from other plant available in time; this require to correctly major and minor revision schedule taking into account technical and commercial constrains (i.e. component life cycle vs. EOH or period of the year where it is not allowed to conduce maintenance operation due to the service contract); this activity should be based on a comparative analysis on cost reduction and profitability improvements over a stochastic scenario and in presence of risks.

This approach is even more effective in large sets of plants, that require service within the same timeframe; in this case the good results emerge by a collaborative management obtained by maximizing the sharing of refurbished and new items as well as the demand. Creating a common power plant pool is possible to optimize inventory management, component safety stocks in order to reduce costs and to improve the profitability and availability over all plants and to optimize shared resources. (Bruzzone, Bocca 2008).

M&S FOR POWER PLANT SERVICE & MAINTENANCE.

There are consolidated experience in managing service and maintenance in industrial power plants; modeling in simple case the service quality obtained from different management strategies the

results is that pooling always improve the service levels (Taragas 1989);

A reliable support for manager of service part inventories are demand pattern identifier based on statistical methods (Cerdea et al. 1997; Sugita et al. 2005; Paschalidis et al. 2004; Muckstadt 2005; Beardslee et al. 2006).

Several authors investigates the methodologies for optimize inventory management and service for similar cases (Cohen 1990; Silver 1991; Nahmias 1994; Harris 1997).

The use of statistical techniques is effective to support service and maintenance modelling and to validation and verification of the conceptual model (Hill 1997; Hill 1999; Aronis et al. 2004).

The demand of spare parts for supporting service and maintenance procedures is an critical variable to model and requires specific and accurate analysis (Grange 1998).

In the past was developed multi location inventory models combining simulation models and optimizers (Federgruen, 1993; Kochel 1998; Nielander 1999); to optimize inventory and transportation cost polynomial-time algorithms was used (Wang & Cheng 2007). Some Traditional algorithms applied to maintenance optimisation may be questionable (i.e. Hallefjord et al. 2003)

The combination of Linear and non linear approaches (Gupta, Zhang 2010) was investigated on specific target function as well as Scheduling approach in combination with inventory optimisation (Fan et.al 2009).

Complex techniques, such as genetic algoritms, permit to approach multi-objective optimization (Srinivas & Deb 1994).

In power plants maintenance the decision management about replace and order new item referred to limited life cycle components was analyzed by separate and join optimization (Armstronga 1996)

Supply and inspection of components must be considered by the mathematical models developed (Chen et al. 2009).

The Simulation Team DIPTEM researchers have long experience in these methodologies and techniques applied to this sector (Giribone, Bruzzone & Tenti 1996).

Some innovative approach was proposed, based on simulation, by the authors to support management strategies in the identification of best solutions considering multiple constraints and target functions (Bruzzone et al. 1998).

In fact it was also developed a methodology for supporting pooling strategies for multiple power plant service (Bruzzone, Mosca, Pozzi, Cotto, Simeoni, 2000) as an innovative approach for defining criteria for serving the sites by clustering

the machines in subsets able to guarantee compatible timeframes respect life cycles of spare parts, components and items; this approach leads to the optimization of availability, costs respecting technical commercial constraints.

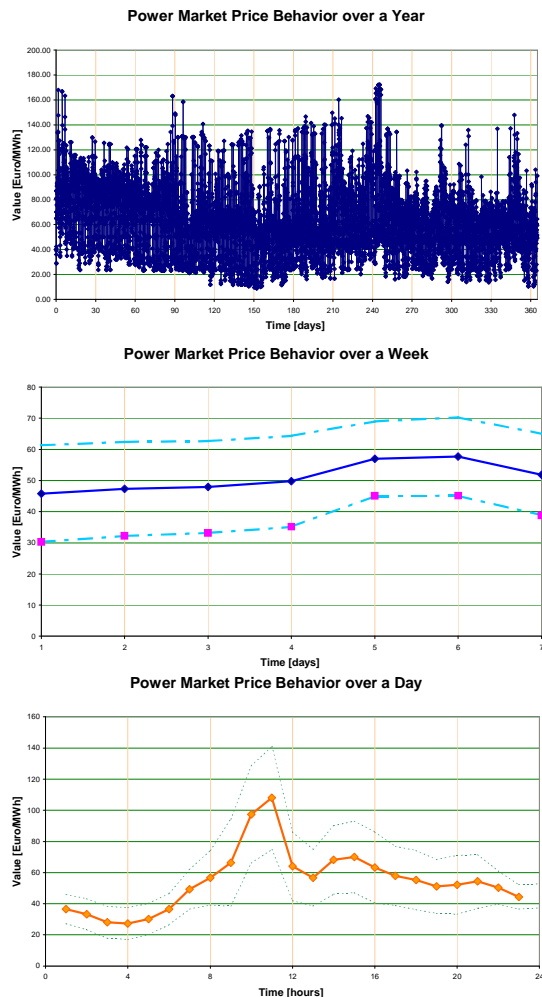


Figure 1 – Power Market Price Behaviours

Therefore the improvement of power plants performances in terms of service and maintenance, go through fast supply chain, flexible stock management, lean process that became key elements for a correct management DSS; so the integration of operation, administration, maintenance and business have a relevant importance on the power plant performances such as efficiency, reliability, availability, environmental impact and lifecycle cost (Bruzzone, Mosca Simeoni, Pozzi Cotto, Fracchia 2000).

The difficulties in the definition of a correct service and maintenance procedures in power plants need to be approached by modeling the revamping and refurbishment processes for specific components as

well the inspections, revisions and turnover strategies which permit to minimize spare parts acquisitions costs; increasing the number of machine (units) involved in the maintenance services it is critical to be able to track each component for each unit and to be able to process and elaborate this context (Bruzzone, Simeoni 2002)

NEW PERFORMANCE INDEXES

In order to identify the optimum in terms of service the authors proposed a specific metrics that permit to balance, in weighted way, cost and availability. The costs are defined considering refurbishment and acquisition of spares, the availability is defined in relation to the value of the market price for energy considering that in different days, hours and months (see fig.1), unavailability of the items (and then of the plants) correspond to profit lost of profit. Profit lost is different in different time frame.

$$SP = k_a VA + k_c VC$$

$$Oce = \sum_{i=1}^{np} Ace_i + \sum_{i=1}^{np} Rce_i + \sum_{i=1}^{np} Sce_i$$

$$na_{ij} = \text{int} \left[\frac{\Delta T_i}{LC_j (nre_j + 1)} \right] \quad nr_{ij} = \frac{\Delta T_i}{LC_j} - na_{ij}$$

$$Act_i = \sum_{j=0}^{nc} na_{ij} ca_j \quad Rct_i = \sum_{j=0}^{nc} nr_{ij} cr_j$$

$$VC = \frac{Act_i + Rct_i}{Oce}$$

SP	service performance
ka	availability importance factor
kc	cost importance factor
np	number of plant in the pool
nc	number of maintenance components
nre _j	maximum number of possible refurbishment for j-th component
ΔT _i	time frame for maintenance of i-th plant
Oce	Effective Overall Cost
Ace _i	Component Acquisition Cost for i-th plant
Rce _i	Component Refurbishment Cost for i-th Plant
Sce _i	Extra Cost due to stop for i-th Plant
ca _j	unit acquisition cost for j-th component
cr _j	unit refurbishment cost for j-th component
VC	cost performance index as ratio effective vs. minimum theoretical service costs

$$GT = \sum_{i=1}^{np} GT_i \quad GT_i = \int_{t_i0}^{t_i0+\Delta T_i} P_{N_i} f(t) dt$$

$$G'_i(\bar{t}^{i*}_{F_{isp}}, \bar{t}^{i*}_{F_m}, \bar{t}^{i*}_{F_M}) = \int_{t_i0}^{t_i0+\Delta T_i} P_{N_i} g_a^i(t, \bar{t}^{i*}_{F_M}, \bar{t}^{i*}_{F_m}, \bar{t}^{i*}_{F_{isp}}) f(t) dt$$

$$g_a^i(t, \bar{t}^i_{F_M}, \bar{t}^i_{F_m}, \bar{t}^i_{F_{isp}}) = g_{isp}^i(t, \bar{t}^i_{F_{isp}}) g_M^i(t, \bar{t}^i_{F_M}, \bar{t}^i_{F_m})$$

$$g_M^i(t, \bar{t}^i_{F_M}) = \begin{cases} 0 & \bar{t}^i_{F_M} \leq t \leq \bar{t}^i_{F_M} + \Delta t^i_{F_M} \\ 1 & t > \bar{t}^i_{F_M} + \Delta t^i_{F_M} \\ 1 & t < \bar{t}^i_{F_M} \end{cases}$$

$$g_m^i(t, \bar{t}^i_{F_m}) = \begin{cases} 0 & \bar{t}^i_{F_m} \leq t \leq \bar{t}^i_{F_m} + \Delta t^i_{F_m} \\ 1 & t > \bar{t}^i_{F_m} + \Delta t^i_{F_m} \\ 1 & t < \bar{t}^i_{F_m} \end{cases}$$

$$g_{isp}^i(t, \bar{t}^i_{F_{isp}}) = \begin{cases} 0 & \bar{t}^i_{F_{isp}} \leq t \leq \bar{t}^i_{F_{isp}} + \Delta t^i_{F_{isp}} \\ 1 & t > \bar{t}^i_{F_{isp}} + \Delta t^i_{F_{isp}} \\ 1 & t < \bar{t}^i_{F_{isp}} \end{cases}$$

$$R^e_i = \int_{t_i0}^{t_i0+\Delta T_i} P_{N_i} g^i_e(t) f(t) dt$$

$$\bar{t}^{i*}_{F_{isp}}, \bar{t}^{i*}_{F_m}, \bar{t}^{i*}_{F_M} / G'_i(\bar{t}^{i*}_{F_{isp}}, \bar{t}^{i*}_{F_m}, \bar{t}^{i*}_{F_M}) < G'_i(\bar{t}^{i*}_{F_{isp}}, \bar{t}^{i*}_{F_m}, \bar{t}^{i*}_{F_M}),$$

$$\forall \bar{t}^i_{F_{isp}}, \forall \bar{t}^i_{F_m}, \forall \bar{t}^i_{F_M}, \forall k,$$

$$|\bar{t}^i_{F_{isp}}(k) - \bar{t}^i_{F_{isp}}(k-1)| < \lambda_{isp} t_{F_{isp}l},$$

$$|\bar{t}^i_{F_m}(k) - \bar{t}^i_{F_m}(k-1)| < \lambda_m t_{F_ml},$$

$$|\bar{t}^i_{F_M}(k) - \bar{t}^i_{F_M}(k-1)| < \lambda_M t_{F_Ml}$$

$$Av_i = \frac{\int_{t_i0}^{t_i0+\Delta T_i} g^i_e(t) dt}{\Delta T_i} \quad Prod_p = \frac{1}{GT} \sum_{i=1}^{np} \frac{R^e_i}{\Delta T_i}$$

$$Prod_c = \sum_{i=1}^{np} \frac{R^e_i}{\Delta T_i \cdot G'_i(\bar{t}^{i*}_{F_{isp}}, \bar{t}^{i*}_{F_m}, \bar{t}^{i*}_{F_M})}$$

$$VA = \frac{\sum_{i=1}^{np} \frac{PN_i \cdot R^e_i}{G'_i(\bar{t}^{i*}_{F_{isp}}, \bar{t}^{i*}_{F_m}, \bar{t}^{i*}_{F_M})}}{\sum_{i=1}^{np} PN_i}$$

t_{i0} initial time for i-th plant

$f(t)$ Power Price at t time

g_a^i nominal operative state

$\bar{t}^i_{F_M}$ vector with time of m_1 Major revision for i-th plant

$\bar{t}^i_{F_m}$ vector with time of m_2 minor revision for i-th plant

$\bar{t}^i_{F_{isp}}$ vector with time of m_3 inspections revision for i-th plant

$\bar{t}^i(k)$ element k-th of the vector \bar{t}^i

$\bar{t}^{i*}_{F_M}$ optimal set of Major revision time for i-th plant

$\bar{t}^{i*}_{F_m}$ optimal set of Major revision time for i-th plant

$\bar{t}^{i*}_{F_{isp}}$ optimal set of Major revision time for i-th plant

$g_e(t)$ effective operative state based on decided

planning

GT Theoretical maximum revenues

GT_i Theoretical maximum revenues of i-th plant without any planned maintenance

G'_i Maximum revenues of i-th plant with optimal Planning maintenance

R^e_i effective revenues based on decided planning

$t_{F_{isp}l}$ technical/contractual interval between inspections

$t_{F_{ml}}$ technical/contractual interval between minor revisions

$t_{F_{Ml}}$ technical/contractual interval between major revisions

λ_{isp} technical/contractual tolerance between inspections

λ_m technical/contractual tolerance between minor revisions

λ_M technical/contractual tolerance between major revisions

$\Delta t_{F_M}^i$ theoretical/contractual duration of major review for i-th plant

$\Delta t_{F_m}^i$ theoretical/contractual duration of minor review for i-th plant

$\Delta t_{F_{isp}^i}$ theoretical/contractual duration of inspection for i-th plant

The aim of the application of the model is to find the best & feasible schedule and inventory management in order to optimize the general performance of the plants. The best result is evaluated by considering expected prices of power along days/weeks/months/years. To do this, and test different scenarios, is possible to use historical data of power consumption integrated with trend hypothesis.

Using Genetic Algorithms integrated with a simulation it is possible to estimate the fitness function in term of value and confidence band related to the different performance indexes considering the complex relationships among variables and parameters; the authors realized LAPIS framework in order to support the optimization and analysis of service and maintenance planning & policies as well as for supporting decision making in this framework,

MODEL VARIABLES

The main purpose of the simulation model defined by the authors is to be a DSS able to properly evaluate the KPIs over complex scenarios; the LAPIS is stochastic discrete event simulator integrated with an Optimizer based on Genetic Algorithms; in fact key performance indexes are affected by several variable such as:

- Effective Planning for each Plant
- Effective substitution/mounting Sequencing for components subjected to refurbishment
- Schedule Performances quantify the respect of time constraints such as :
 - Inspection/Revision exceeding the allowed time due to delays/problems (i.e. extra time for a revision due to spare part shortage)
 - Technical Times Interval among inspections/revision not respected (i.e. too many EOH before substituting some blade layer)
 - Dates not acceptable due to contract constraints (i.e. desire to avoid maintenance in some months with higher power prices or viceversa desire to concentrate all the maintenance within summer holiday break)
 - Too short interval among two sequential different machine inspection/revision on the same site respect desired value (i.e. desire to avoid to operate them concurrently inside the same power building due to interference and lack of space)
 - Too long interval among two sequential different machine inspection/revision on the same site respect desired value (i.e. desire to operate them concurrently inside the same power building creating synergies with service resources and kits)
 - Too few machine concurrently unavailable respect desired value (i.e. desire to distribute the maintenance to guarantee average power generation capability of the plant users)
 - Too long interval among two sequential different machine inspection/revision on the same site respect desired value (i.e. desire to operate them concurrently inside the same power building creating synergies with service resources and kits)
- Power Plant Number of Stops and Duration
- Inventory behavior for each Component
 - Warehouse Quantities
 - Stock-out Times, Importance and Quantity (i.e. how many spare parts of the component are missed, how critical it is the component for plant operations and how much it costs to acquire through unconventional channels)
 - Component Service Level
 - Component Rotation
 - Expected Final Status of the Component at the end
 - Quantities and values of spare parts of the component distinguishing among new ones and refurbished at the

different levels are mounted in the machines/plants at the beginning of the contract

- Quantities and values of spare parts of the component distinguishing among new ones and refurbished at the different levels are mounted in the machines/plants when service contract expires
- Quantities and values of spare parts of the component distinguishing among new ones and refurbished at the different levels are available in the warehouses when service contract expires Costs & Profitability detailed as:
 - New Spare Part Acquisition Costs
 - Refurbishment Costs
 - Warehouse Fees
 - Expediting Fees
 - Initial Costs for the defined Configuration
 - Plant Expected Profitability
- Risk Reports
 - Risks of Service and Maintenance Delays
 - Risk on Component Shortage
 - Risk of Power Plant Stops (Number and related duration)

The model implements many alternative management policies for power plants service as well as estimation criteria and these are defined inside the simulation model.

In fact are several simulation parameters that need to be set in order to properly estimate the performances such as:

- Replication Number for each scenario evaluation in order to estimate the stochastic factor influence
- Pseudo Random Number seeds (or automatic initialization)
- Simulation Duration
- Power Plant Pool Configuration
- Inventory initial configuration
- Initial Scheduling
- Operative Management Criteria
- Inventory Management Policies
- Policies for restoring of Safety Levels
- Policies for managing Expediting
- Policies for Interchanging compatible Components
- Policies for Cannibalization of Components in planned maintenance occurrences
- Policies for Cannibalization of Components due to failures

- Policies for processing Automatic Collected Data related to Power Demand and Plant use
- Policies for managing contract duration
- Definition of Technical Constraints
- Definition of Contractual Constraints
- Definition of Resource Constraints

In fact the initial conditions set in the model are used to start the simulation of a specific scenario; during the simulation run the model reproduce the operation in service and maintenance of the plants considering unexpected failures, managing initial schedule and inventory.

In order to verify and validate the model on all the events, costs and indexes the simulator generates a log file that contains all estimations of different stochastic components compared to initial planning and management strategies

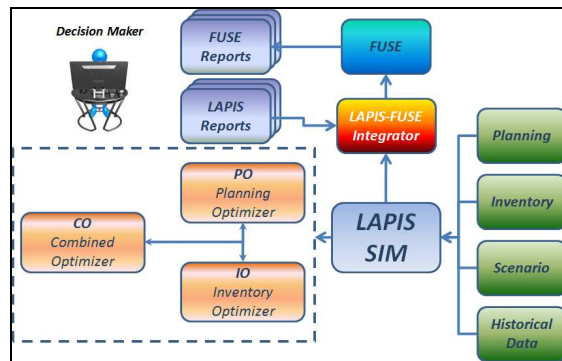


Figure 2 – LAPIS General Architecture

LAPIS simulator generates a lot of reports that represent, also in graphical mode, the temporal evolution of the following variables:

- Each Plant Unit EOH
- KPIs
 - Profitability
 - Costs
- Component Consumption
- Component Levels on the Warehouses
- Refurbishment Quantities
- Failures
- Inspection, Minor and Majors Revisions

LAPIS ARCHITECTURE & COMPONENTS

As anticipated, LAPIS solution is the combination of simulation and optimization (see figure 2), as shown in the architecture scheme the simulator is connected to each component (is the core of this proposed approach).

The box called “FUSE” is an interesting module of the system, where the application of Fuzzy Techniques is used to evaluate the interaction among technical (machine lifecycle), operational (interference among inspections), contractual (periods preferable for maintenance) and commercial factors such energy request from the market (Bruzzone et al. 2004).

In fact LAPIS is fully integrated with this fuzzy logic performance evaluator, developed in previous researches, devoted to evaluate the quality of the planning of maintenance operations by a hierarchical approach (Bruzzone & Williams 2005). Fuzzy Logic is very useful in this context due to the uncertainty on variables and constraints (Cox 1994).

Failures, planned maintenance events, critical time points such as shut downs, start-up, contract closure, item delivery and several other events are driving the time advance in LAPIS simulator; while the power demand behavior and unit EOH (Equivalent Operating Hours) are computed by the integration of the function between two consecutive events.

The stochastic variables are computed by using Montecarlo Techniques, the simulator for each time and in each run extract the value of the variable from distribution function.

The probability distribution of the variables was identified analyzed by statistical techniques (Test Chi² T and by the Subject Matter Expert in order to identify the best fitting of the real data with the known Probability Distributions.

For several reason (few historical data, short history, errors in records, confidential nature of the information etc.) in most of the cases without strong historical background the authors used Beta Distributions to model stochastic variables.

In fact Beta distribution allows to integrate easily the expert estimations with historical data in order to have consistent data.

There are three types of optimization modes supported by LAPIS architecture:

- Planning Optimizer (PO)
- Inventory Optimizer (IO)
- Combination of PO and IO

All the optimization models deal with the dynamic interaction among the GAs optimizer and the stochastic discrete event simulator.

The optimizers, as anticipated, use Genetic Algorithms (GAs) in order to find robust and cost effective solutions (Bruzzone A.G., 1995).

In this application the GAs are initialized from a set of solutions called “population” including proposals of the users; the genetic operators referred to fitness

function, obtained by running automatically the simulator, to guide the search and improving solutions (Bruzzzone, Bocca 2008); the genetic algorithms include:

- selection
- recombination (crossover)
- mutation

The optimizer elaborate the population and evaluate their fitness by running automatically the simulator and it recombines the solutions by the above mentioned algorithms for several generations; the parameters of the GAs could be set by the user as well as the weights of the fitness function to be used for each optimization.

LAPIS FRAMEWORK

Due to its complexity the model need a very high quantity of input data, it is so hard to modify this data to create different scenarios maintaining it consistent; in fact most of data needed for running the simulation (such as existing planning, technical data of items and spares, levels of the storage) are extracted by the company ERP System; therefore in LAPIS an easy interface is defined to quickly check and change hypotheses for creating different scenarios.

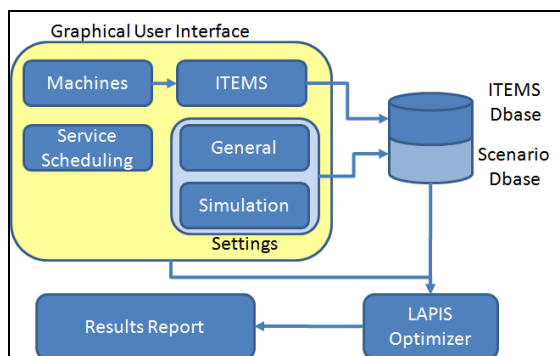


Figure 3 Lapis Function Configuration

Figure 3 show the LAPIS components, the user can interact with the model by using a Graphical User Interface (GUI) with whom is possible to modify the input parameters and the settings (of the model to change the characteristics of the scenario and also of the simulator) (Bruzzzone & Simeoni, 2002). The scheduling contains the data about the interventions planned by the user to make maintenance to each power units of the pool. The scheduling depends both of contracts and customer; it is not a fixed variable but the user must input it by using the GUI. Each complex plants is composed by different item with different characteristics and then different

needs in terms of maintenance. The analysis by the simulation of the maintenance policies of each component may be and hard and unproductive work, for this reason in combined power plants the maintenance procedures are driven by the most important item that is the Turbo Gas (TG)

The template of the sequence of events has the following order: I-I-PR-I-I-GR (I=Inspection, PR=Partial Revision and GR=General Revision).

All the characteristics of the Items are stored in the company Data Base (DB). All the scenario processed are stored in DB Scenario in order to collect a historical set of analysis performed (useful for future research).

In order to have the better interaction with the users the authors develop three different kind of output reports: a customer report (it is possible to send it directly to the customer), a pre-customer report (these reports need to be checked by the user before sending them), and a user-report (to control if the results are consistent with the related scenario).

Due to the high computational workload related to this very complex model, it was decided to implement it in C++ allowing to run optimization process within reasonable time (i.e. few minutes for simple case/partial optimization, few hours for complex scenarios); at the same the authors are used to consider the automatic optimization not as a stand alone solution, but as a procedure to be run by decision makers interactively while they test new hypotheses and ideas; in fact changing some hypotheses on exogenous factors the best solution change and the decision makers need to compare the results and evaluate the reliability of data and evaluation provided by experts.

So in order to guarantee an easy access to simulation and optimization results as well as an effective analysis tool the LAPIS report are post processed and handled by a module implemented using MS Office Suite, while FUSE model provides additional capabilities in evaluating the solutions.

By using the report carried out by the combined use of LAPIS and FUSE the decision makers are able to quickly evaluate, accept, modify or reject alternative proposed solutions.

The authors worked on optimization and simulation of pooling strategies for managing several power plants with different spare parts; therefore due to complexity of the system it was critical to guarantee an effective interface to make the use of the program easy for decision makers and to support all the functions provided by LAPIS.

LAPIS VV&A (Verification, Validation and Accreditation) was extensively applied even in term of dynamic analysis of simulation/optimizer over several complex scenarios. Several month of work, desktop review and dynamic testing in cooperation

with Subject Matter Experts (SME); in particular power plant service experts (i.e. project managers, supply chain management team member, planners) supported the validation phase. Several Statistical Test Technique was applied to LAPIS models such as Analysis of Variance (ANOVA), Mean Square pure Error (MSPe), Confidence Band, Statistical Comparison and Sensitivity Analysis (as presented in figure 4).

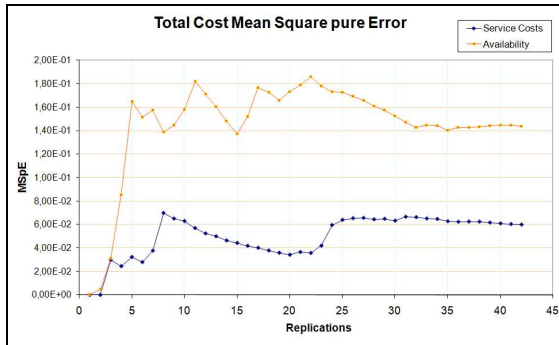


Figure 4 – LAPIS Verification and Validation

In the paper the results of sensitivity analysis based on Design of Experiments (DOE) are proposed as VV&A; the analysis allows to quantify the effects and the contrasts of selected parameters on the target functions as well as their interaction.

The case proposed for this analysis and optimization is related to a realistic scenario involving a collaborative service provided to nine different combined units (steam turbine plus gas turbine) located over different sites.

The aim of this analysis is to identify what, and in which way (direct or inverse proportionally), a controlled variable impact on a target function (i.e. availability of the pool and the overall cost); in similar way all the combined effect of different variable are estimate, in fact due to the high degree of not linearity of the problem under analysis the high level order effects are usually significant and cannot be neglected .

In the paper the analysis is focalized on some of the main parameters, according to the SME suggestions:

- Parameter A: Number of kits of Gas Turbine blades available (condition tested over the values: 3 kits or 6 kits).
- Parameter B: Enable or disabled the Cannibalization of new kits of blades.
- Parameter C: Scraping Percentage, that indicate the percentage of damaged blades that can't be refurbished at each step (based on company historical data and SME: from 1% to 5%)

- Parameter D: Computing/Neglecting the Residual Value of the item mounted in the machines at the end of the contract in the overall service cost KPI; this parameter is important because permit to consider the residual value of the material in the plants (especially if the customer don't renovate the contract).

As shows in Figure 5, considering the availability of the power plants the sensitivity analysis shows how, in all the scenario tested, the residual value of the plants don't affect the availability of the plants.

Analyzing the contrast graph shows that the scraping impact on the availability of the plants in inverted proportional way; unfortunately the scraping is not easy to control, in general sense scraping should be considered as an exogenous factor that evolves with the time, usually positively due to new technological material developments and operational experience; so this variable is able to produce future benefits that are estimated by LAPIS, but it is not subjected to a direct control by users or service providers.

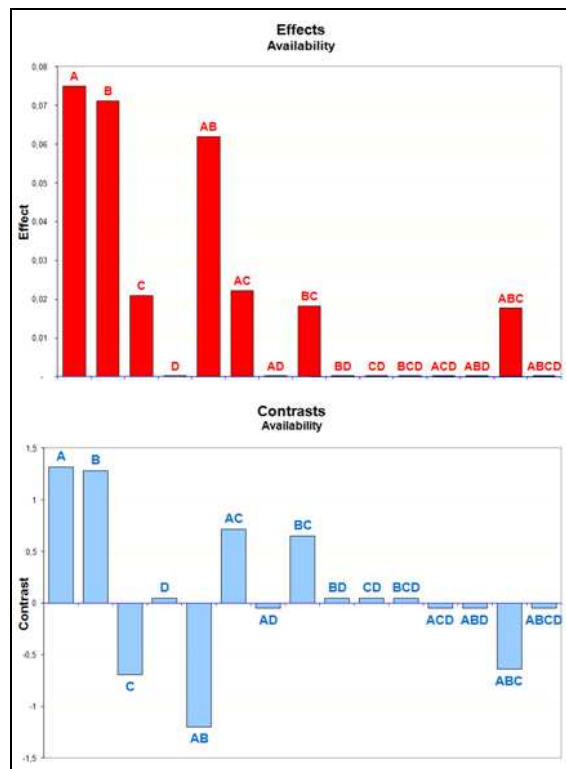


Figure 5 – Sensitivity Analysis Availability

Trough the analysis of Figure 6 related to a different target function (overall service costs) is possible to identify a particular behavior where the combination of 2 variable (A and B) is very big; in fact the influence on the related target function is

much more than the impact of the individual parameters if considered separately (please note that the scale is logarithmic); this is a classical confirmation of the complexity of the phenomena generating a very not-linear behavior.

In fact the possibility to cannibalize the new kits permits to optimize the maintenance operation and then to reduce total cost (as contrast shows).

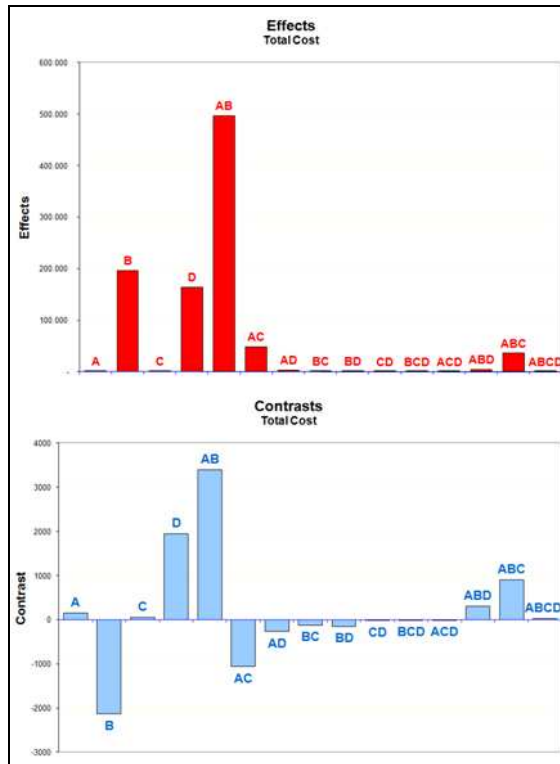


Figure 6 – Sensitivity Analysis Total Cost

The obtained results support to development of robust solutions able to consider inventory costs, stop costs, availability, contractual term respect and constraints in fact the simulator driven by the to genetic algorithms is able to provide useful service management solutions.

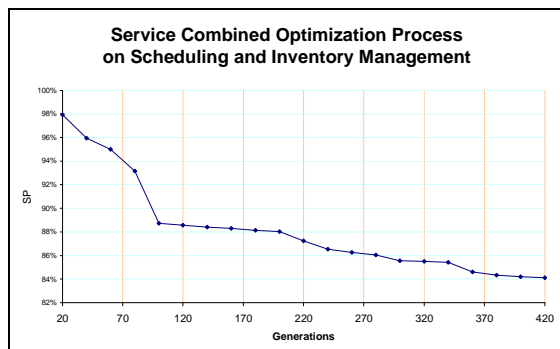


Figure 7 Lapis Optimisation Evolution

Figure 7 propose the evolution of the optimization process during an optimization; it is evident the significant saving and benefits provided by the integrated use of simulation and GAs.

CONCLUSIONS

The architecture of the model resulted very reliable and robust even when applied to pretty complex scenarios; currently this approach is proposed by the author to support decision making in terms of service and maintenance in system-of-systems service for major companies .

The evaluation of the profitability combined with availability related to the market prices represent an example of how it could be possible to redirect the service management to most efficient approaches by properly evaluating the overall effectiveness and efficiency; in power industry for instance this aspect have a growing impact and the use of such models could guarantee significant improvements and high level of competitiveness both for service providers as well as for users; in fact a major benefit of introducing these advanced models in the service of complex systems it is related to the sharing among providers and users of a common understanding of the scenario with possibility to achieve much better results and lean decision making process both in management and acquisition of service contracts. the introduction of this innovative approach in power plant service allows to define pooling management strategies able to evaluate more properly the real performances and to change the scheduling criteria for inspections/revisions as well as the policies in use (inventory management, safety stock, refurbishment services).

Due to its effectiveness LAPIS was effectively used to be integrated in the company decision making process for supporting complex power plant's service and maintenance.

REFERENCES

1. Armstronga M.J., Derek R. Atkinsa (1996) "Joint optimization of maintenance and inventory policies for a simple system" IIE Transactions, Volume 28, Issue 5 May 1996 , pages 415 - 424
2. Aronis P.K., I. Magou, R. Dekker and G. Tagaras, (2004) "Inventory control of spare parts using a Bayesian approach: A case study", European Journal of Operational Research 154, pp. 730–739

3. Beardslee E.A., Theodore B. Trafalis (2006) "Discovering Service Inventory Demand Patterns from Archetypal Demand Training Data" University of Oklahoma, USA
4. Bruzzone A.G. (1995) "Fuzzy Logic and Genetic Algorithms Applied to the Logistical and Organisational Aspects of Container Road Transports", Proc. of ESM95, Praha, June 5-7
5. Bruzzone A.G., Kerckhoffs (1996) "Simulation in Industry ", Genoa, Italy, October, Vol. I & II, ISBN 1-56555-099-4
6. Bruzzone A.G., Giribone P., Revetria R., Solinas F., Schena F. (1998) "Artificial Neural Networks as a Support for the Forecasts in the Maintenance Planning", Proceedings of Neurap98, Marseilles, 11-13 March
7. Bruzzone A.G., Giribone P. (1998) "Decision-Support Systems and Simulation for Logistics: Moving Forward for a Distributed, Real-Time, Interactive Simulation Environment", Proceedings of the Annual Simulation Symposium IEEE, April
8. Bruzzone A.G., Signorile R. (1998) "Simulation and Genetic Algorithms for Ship Planning and Shipyard Layout", Simulation, Vol.71, no.2, pp.74-83, August
9. Bruzzone A.G., Mosca R., Pozzi Cotto S., Simeoni S. (2000) "Advanced Systems for Supporting Process Plant Service", Proc.of ESS2000, Hamburg, Germany, October
10. Bruzzone A.G., Mosca R., Simeoni S., Pozzi Cotto S., Fracchia E. (2000) "Simulation Systems for Supporting Gas Turbine Service Worldwide", Proceedings of HMS2000, Portofino, October 5-7
11. Bruzzone A.G., Simeoni S. (2002) "Cougar Concept and New Approach to Service Management by Using Simulation", Proc/of ESM2002, Darmstad Germany June 3-5
12. Bruzzone A.G. (2002) "Supply Chain Management", Simulation, Volume 78, No.5, May, 2002 pp 283-337 ISSN 0037-5497
13. Bruzzone A.G., Simeoni S., B.C. (2004) "Power Plant Service Evaluation based on advanced Fuzzy Logic Architecture" Proceedings of SCSC2004, San Jose'
14. Bruzzone A.G., Williams E. (2005) "Summer Computer Simulation Conference", SCS, San Diego, ISBN 1-56555-299-7 (pp 470)
15. Bruzzone G. A., Bocca E. (2008) "Introducing Pooling by using Artificial Intelligence supported by Simulation", Proc.of SCSC2008
16. Bruzzone A.G., Madeo F., Tarone F. (2010) "Pool Based Scheduling And Inventory Optimisation For Service In Complex System" Proc. of the 16th International Symposium on Inventories, Budapest, August 23-27
17. Cerda R., C. B., and Espinosa de los Monteros F., A. J., (1997) "Evaluation of a (R,s,Q,c) multi-item inventory replenishment policy through simulation", Proceedings of the 1997 Winter Simulation Conference, pp. 825-831.
18. Chen SC, Yao MJ, Chang YJ (2009) "A Genetic Algorithm for Solving the Economic Lot and Inspection Scheduling Problem in an Imperfect Production/Inventory System" 8th International Conference on Information and Management Sciences, Kuming, China, July
19. Cohen, Morris, Pasumarti V. Kamesam, Paul Kleindorfer, Hau Lee and Armen Tekerian, 1990, "Optimizer: IBM's Multi-Echelon Inventory System for Managing Service Logistics," Interfaces, vol. 20, no. 1 (Jan-Feb), p. 65-82.
20. Cox E. (1994) "The Fuzzy System Handbook", AP Professional, Chestnut Hill, MA
21. Fan, BQ; Chen, RJ; Tang, GC (2009) "Bicriteria Scheduling on Single-Machine with Inventory Operations" 3rd International Conference on Combinatorial Optimisation and Applications, Huangshan China, June 10-12
22. Federgruen, A., (1993) "Centralized planning models for multi-echelon inventory systems under uncertainty", in Graves, S., Rinnooy Kan, A., Zipkin, P. "Handbook in Operations Research and Management Science", Vol. 4, Logistics of Production and Inventory. North-Holland: Amsterdam.
23. Giribone P., Bruzzone A.G. & Tenti M. (1996) "Local Area Service System (LASS): Simulation Based Power Plant Service Engineering & Management", Proceedings of XIII Simulators International Conference SMC, New Orleans LA, April 8-11
24. Grange F. (1998) "Challenges in modeling demand for inventory optimization of slow-moving items", 30th Wintersim Washington, D.C., United States, November
25. Gupta S, Zhang N (2010) "Flexible Scheduling of Crude Oil Inventory Management", *Ind. Eng. Chem. Res.*, pp 1325–1332, Washington DC, USA, December
26. Harris, Terry, 1997, "Optimized Inventory Management," Production and Inventory Management Journal, vol. 38, no. 1, p. 22-25
27. Hallefjord Å., K. Jörnsten, S. Storøy and S. W. Wallace (2003) "Inspection and maintenance optimization methods" Michelsen Institute, Fantoft, Norway.
28. Hill R.M., (1997) "Applying Bayesian methodology with a uniform prior to the single period inventory model", European Journal of Operational Research 98, pp. 555–562.

29. Hill R.M., (1999) "Bayesian decision-making in inventory modeling", *IMA Journal of Mathematics Applied in Business and Industry* 10 (1999), pp. 165–176.
30. Köchel P. (1998) "A survey on multi-location inventory models with lateral trans-shipments" In: S. Papachristos and I. Ganas., Editors, *Inventory Modelling in Production and Supply Chains*, 3rd ISIR Summer School, Ioannina, Greece, pp. 183–207
31. Muckstadt, J.A., (2005) "Analysis and Algorithms for Service Parts Supply Chains" Springer, New York
32. Nahmias, Steven, 1994, "Demand Estimation in Lost Sales Inventory Systems," *Naval Research Logistics*, vol. 41, pp. 739-757.
33. Nieländer, U., 1999 "Simulation Optimisation of Kanban Systems using a non-standard Genetic Algorithm" *Proc. Of 4th ISIR Research Summer School on Inventory Modeling*, Exeter, UK.
34. Paschalidis, I. C., Liu, Y., Cassandras, C. G., Panayiotou, C. (2004) "Inventory control for supply chains with service level constraints: a synergy between large deviations and perturbation analysis", *Annals of Operations Research* 126, pp. 231-258.
35. Silver, Edward A., 1991, "A Graphical Implementation Aid for Inventory Management of Slow-Moving Items," *Journal of the Operational Research Society*, vol. 42, no. 7, pp. 605-608.
36. Srinivas N., K.Deb,(1994) "Multiobjective optimization using nondominated sorting in genetic algorithms", *Evol. Comput.*,vol.2, no.3, pp.221-248.
37. Sugita, K., and Fujimoto, Y. (2005) "An optimal inventory management for supply chain considering demand distortion", *Industrial Informatics*, 3rdIEEE International Conference on (INDIN), pp. 425-430.
38. Tagaras G. (1989) "Effects of Pooling on the Optimization and Service Levels of Two-Location Inventory Systems" *IIE Transactions*, Volume 21, Issue 3 September
39. Wang XL , Cheng TCE (2007) "Logistics scheduling to minimize inventory and transport costs" *International Conference on Industrial Engineering and Systems Management*, Beijing

MODELING OF OBESITY EPIDEMICS BY INTELLIGENT AGENTS

Agostino G. Bruzzone,
MISS DIPTM University of Genoa
Email agostino@itim.unige.it - URL www.itim.unige.it

Vera Novak
BIDMC, Harvard Medical School
Email vnovak@bidmc.harvard.edu - URL <http://www.bidmc.org/SAFE>

Francesca Madeo
M&S Net
Email francesca.madeo@m-s-net.org - URL www.m-s-net.org

Cecilia Cereda
Simulation Team
Email cecilia.cereda@simulationteam.com - URL www.simulationteam.com

KEYWORDS

Health Care, Simulation, Intelligent Agents, Human Behavior Modeling

ABSTRACT

The paper focuses on a large scale problem related to population health care with special attention to obesity. The authors present a proposal for modeling human behavior and its influence on the evolution of obesity epidemics, and its effects on social networks, infrastructures and facilities. This approach is based on Intelligent Agents tools developed for reproducing country reconstruction and human factors. These models represent the base that allows to add specific and complex aspects related to pathologies and correlated behaviors that allows to reproduce these phenomena.

INTRODUCTION

Today one of the main problems is to adapt health care to the existing challenges in the society. This means that both public and private health care system have to face the problem to obtain and properly allocate resources for prevention, treatment and rehabilitation of a large assisted population with limited assets. In fact, these challenges are expected to grow in coming decades, due to a variety of reasons in different world regions: i.e. population growth, changes in life-span expectation, aging of the population, social and economic evolution. To solve this problem it is necessary to improve the effectiveness and efficiency in allocating the available resources.

Therefore, a better understanding of these complex phenomena affecting population health aspects is currently very critical to properly define health policies, actions and to plan future infrastructures and services to be able to face such challenges.

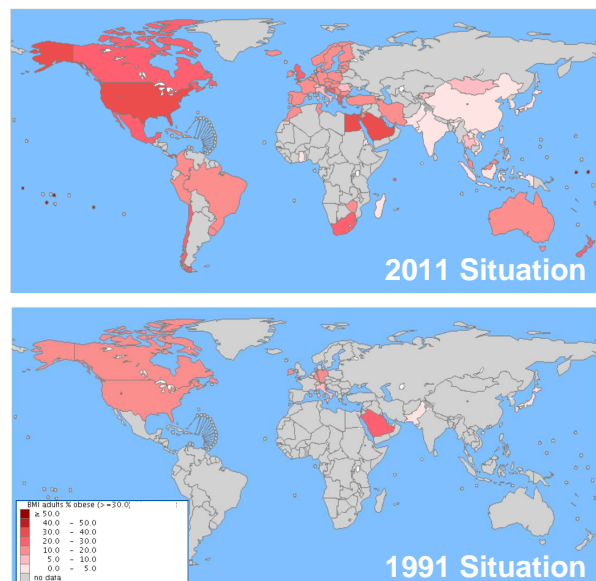


Figure 1. Obesity Epidemic evolution in last 20 years

It is evident that in health care more than in other sectors the high influence of stochastic factors and very complex correlation make difficult analysis on large scale without using modeling and simulation; so a very interesting aspect is obtain a better understanding of the phenomena generating/affecting current development of obesity epidemics that is dynamically and quickly evolving in many countries; and therefore, providing challenges today and threats for the future.

The authors propose to develop simulation models able to reproduce human behavior to address this problem and to provide support for decision makers; a first important benefit expected by these simulators should be the possibility to test and to validate the different existing hypotheses related to the mutual influence among different factors related to physiological and psychological issues, behavioral aspects and regional/ethnic/social/geographical and economical

factors. In fact, by conducting experiments using information available in larger samples, it will allow us to validate the consistency of these predictions within a population over time and also across diverse populations..

Once the models become validated, it will be possible to use a forecasting system, that would allow to conduct risk analysis and estimation of future development; the simulator, used in this way, reproduces scenarios and generates forecasts in term of resources and facilities required for treatments, as well as estimation of the impact of costs and demand on the country infrastructures and industry. It would also allow to predicting the impact of behavioral, social and economic interventions to prevent further development of obesity epidemics and to curtail its costs.

Finally such simulators could become a very strategic advantage to support decision and to evaluate the impact of actions and countermeasures of public and private institutions and organizations on such a critical sector as the health care. The authors decided to start a joint research on these issues by focusing on obesity due to its strong impact on country economies and to its very complex and dynamic evolution (as proposed in figure 1, source World Health Organization Statistics Reports); in fact obesity of a population evolve based on individual and social behaviors over time (i.e. depression due to some social problems, lack of mobility, socially acceptable overeating etc.). The use of intelligent agents in this area represents an innovative opportunity for research; currently the authors present the first development on this track based on some available data and some adaptation of their simulation frameworks to this new context. It is expected that such research will be further developed in support of specific R&D programs.

THE OBESITY EPIDEMIC

The obesity epidemic (Wolf et al. 2007) has been increasingly spreading worldwide in the past three decades, involving even the countries that never in the past showed obesity among their population. The United States has observed of the highest rate of obesity increase in the world (Wang et al. 2007), affecting people of all ages including children, both genders, different ethnic and racial backgrounds, and various socioeconomic groups. Adult overweight and obesity are defined using body mass index (BMI): normal weight < 25; overweight: 25-29.9; obesity > 30

BMI Body Mass Index

$$BMI = \frac{W}{H^2}$$

W Weight [kg]

H Height [m]

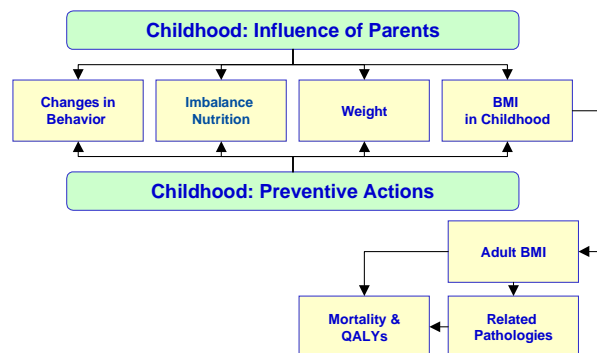


Figure 2. Basic Model of Obesity in Childhood

Children at risk for overweight are defined as the 85th and 95th percentiles of body mass index. Currently, (2010), no state has prevalence of obesity less than 20% and 26 states had a prevalence of 25% or more (Centers for Disease and prevention; www.cdc.gov); 66% of adults are overweight or obese; 16% of children and adolescents are overweight, and 34% are at risk of overweight. Minority and low-socioeconomic-status groups are disproportionately affected at all ages. By 2015, 75% of adults will be overweight or obese, and 41% will be obese.

Obesity is associated with increased risk of diabetes, hypertension, cardiovascular diseases, strokes and dementia, mobility dysfunction, cancer and mortality thus increasing significantly health care cost in the society.

Elevated body mass index is being increasingly recognized as a risk factor for stroke, cardiovascular disease, and cognitive decline (Falkstedt et. al. 2006; Cournot et. al. 2006). Diabetes epidemic follows obesity spread in the world, affecting countries that previous had a lower prevalence of diabetes.

It is expected that diabetes prevalence of people with diabetes will dramatically increase further between 2010 and 2030 i.e. India from 50.8 to 87 millions; China from 43.2 to 62.6; United States of America from 26.8 to 36.0; Pakistan from 7.1. to 13.6; Brazil from 7.6 to 12.7 (Wang et al. 2007).

A long-term population study with 27 years of the follow-up has shown prospectively that in the multiethnic population, midlife obesity increases the risk of dementia later in life.

Obese people (BMI ≥ 30) had a 74% increased risk of dementia (hazard ratio 1.74, 95% confidence interval 1.34 to 2.26), while overweight people (body mass index 25.0-29.9) had a 35% greater risk of dementia (1.35, 1.14 to 1.60) as compared with those of normal weight (body mass index 18.6-24.9) (Whitmer et al. 2005).

The increased risk for Alzheimer's disease and dementia later in life is independent even if adjusted for elevated blood pressure, smoking, socioeconomic status and genetic factors (Wolf et al.2007;Kivipelto et al.2005).

Obesity epidemic is associated with significant burden for the people, families and society and contributes significantly to increased health care cost, morbidity and mortality.

The exact cost of obesity to the society and people is not known.

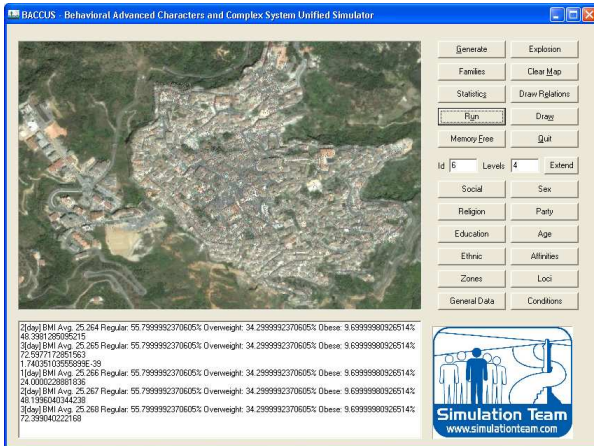


Figure 3. BACCUS: Behavioral Advanced Characters and Complex System Unified Simulator

However, a cost of diabetes could be used as an example; i.e. in 2010 -418 billions of international dollars, 8787 ID per person in the world and 55.7 billions of national income loss in China alone. It is expected that diabetes will increase death rates globally by 17% in 2030 and by 25-27% in Middle East, India and China (IDF Atlas 4th ed. International Federation on Diabetes). Therefore, health care cost will increase exponentially, as the obesity epidemics spreads into younger population and other countries.

Obesity is a multifactorial process that results from interactions among the individual health status, functional and social habits, social networks, education and other factors that cannot be predicted from a single variable (i.e. body mass) but requires nonlinear modeling of multiple variables and their interactions.

Although that reduced physical activity, increased food intake and social networking are commonly cited factors, the mechanism underlying obesity spread worldwide mechanisms remain poorly understood.

The mechanisms of obesity spread and their impact on the world's health, personal life and economy have not been well studied and therefore, the effective strategies to prevent obesity epidemics are lacking.

Impact of Obesity in Term of Costs

The social cost of obesity is related to several aspect (Cereda 2011); as first element it was demonstrated that in fact the obese individuals impact on national health care system costs exceed the average expenditure per capita of an individual's normal weight; therefore much of the social costs associated with obesity are due to different social interactions that obese individuals may

have, always considered in comparison to a normal weight individual. A recent study in Australia has estimated the intangible costs arising from obesity in adulthood, reaching a value between 13 to 18 billion AU\$; in addition to these aspects, many studies have proposed estimations of indirect costs of obesity highlighting the fact that their economic impact could exceed the direct health care costs, both in term of absolute and percentage of GDP (Magarey et al 2001).

For instance in the U.S.A., it is estimated a loss of production due to the obesity corresponding to 23 billion USD (1989-1990), while in Australia the loss for the country is estimated into 272 million AU\$ (Australian Bureau of Statistics).

Recently some researchers have carried out a study with respect to the Chinese context, the analysis was derived from two reference years, 2000 and 2025 (Popkin et al 2006); based on these estimates it results, for the year 2000, an economic impact of 49 billion USD (4.06% of GDP), of which 43 billion dollars in indirect costs (3.58% of GDP). The magnitude of the economic impact is expected to reach a size even more critical, both for increased health spending and the consequences on the labor market. The projections to the year 2025 describe an expected total cost overpassing 112 billion USD representing 9.23% of GDP, of which 106 billion USD (8.73% of GDP) is attributable to indirect costs; these researches estimated that, even in China, the largest component of costs was represented by the loss of productivity due to absences due to illness, which causes 75% of indirect costs. The issues related to sickness, early retirement and disability, have been investigated in Sweden, Finland and Denmark (World Health Organization 1997).

The results produced have highlighted the link between increased BMI and sickness absence in the long term. Furthermore, considering the child obesity as a risk factor for obesity in adulthood, a fraction of the costs, direct and indirect, previously mentioned, the adult is generated from the high number of adolescents in which obesity has persisted over time. In this direction, some researchers have investigated the effects expected along the course of life resulting from a weight reduction program implemented in American schools.

It was introduced the parameter QALYs to quantify the results obtained: the QALY (Quality Adjusted Life Years acronym) is a unit of measure used in cost utility that combines the life span with the same quality. One QALY equal to 1 corresponds to the expected life of one year in normal health, the value 0 corresponds to death (Pliskin et al. 1980).

The measurement scale is continuous and a few years of life may also be given negative values (if you have serious conditions and acute suffering of immobility). QALY is used as an index weighting in the assessment of increases in life expectancy associated with health interventions.

Thus, for example, whether the introduction of a new surgical technique allows the patient to survive on average 6 years older, but the conditions are such that after the operation be considered equal to 0.2 QALY (eg., Because of serious motor deficiencies and frequent pain), the intervention effect on life quality will be weighted for only 1.2 years. Cost-utility analysis performed in the above studies was a cost of 4305 USD/patient per QALY gained, while the sum of the avoided health care costs and productivity losses avoided, you get an expected benefit, net of implementation costs of 7313 USD/patient for each program implemented (Wang et al.2003).

MODELLING THE PHENOMENA

The idea to reproduce the phenomena related to obesity is to use intelligent agent reproducing the population and let them evolve based on their behavior and on the applied actions and scenario evolution (Bruzzone et al.2011); the behavioral models are defined inside the agents, defined IA-CGF (Intelligent Agent Computer Generated Forces); for instance an high level example is proposed in figure 2; a combined stochastic simulation engine manage the interactions among the agents; this framework defined as NCF (Non Conventional Frameworks) could be tailored over very different contexts and areas. Previous researches have been carried out by Simulation team to develop models able to reproduce human behavior over town or regions (Bruzzone et al.2008); in fact these models were used originally for epidemic evolution (Avalle et al.1999), for analyzing urban disorders (Bruzzone et al.2006) and for country reconstruction (Bruzzone & Massei 2010).

In this case it was decided to create an ad hoc NCF with full inheritance of IA-CGF Libraries. For the specific research it was possible to start the development of new conceptual model and to design a first shell of NCF defined BACCUS (Behavioral Advanced Characters and Complex System Unified Simulator) that introduce several additional parameters related to physiology, health status and behavior; these parameters includes:

- BMI
- Sport Profile
- Alcohol Profile
- Stroke
- Infarct
- Diabetes
- Cancer
- Hypertension
- Atrial Fibrillation
- Hyperlipidemia

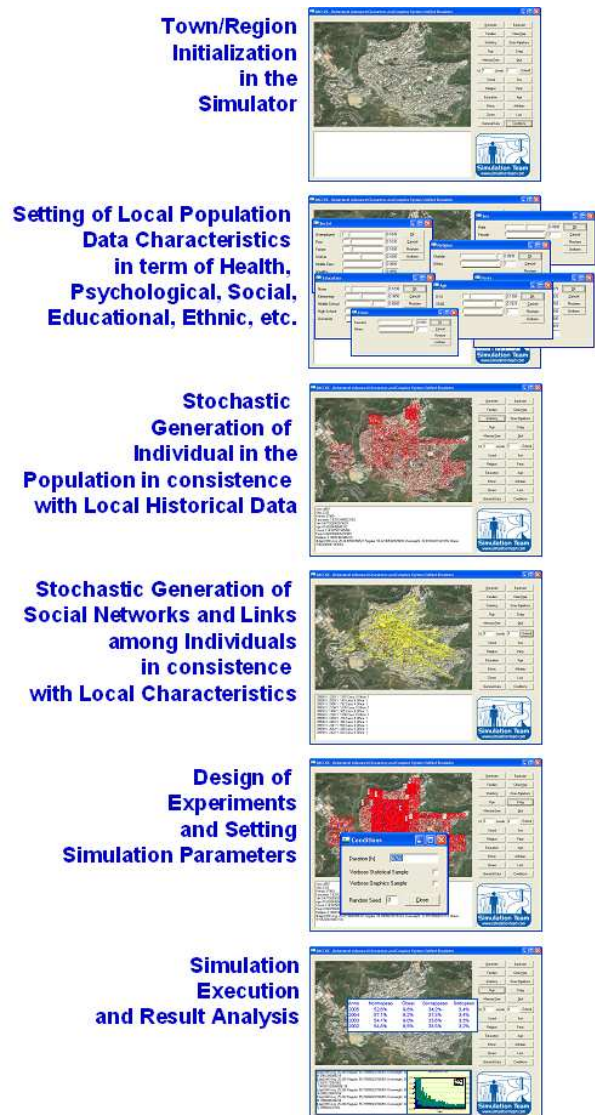


Figure 4. Investigation Process of the Obesity over a Region based on Modeling the Population Behavior

The BACCUS simulator is implemented using IA-CGF engine of the Simulation Team as anticipated and it is proposed in Figure 3. The development concept is based on a several step process; first phase is related to defining the population in term of PIG (people initialization groups); the PIGs represent the different groups present on a scenario; PIGs are defined in term of statistical distribution for their main factors (Social Level, Educational Level, Political Attitudes, Religion, Ethnic, Tribe, Gender, Age); based on these parameters the population is generated randomly respecting original statistics on the geographic region, but even considering the different groups characteristics with their specific structure (i.e. some ethnic group have different social status statistics respect other one living in the same region); this process generates the people agents; in addition inside the region it is possible to define Zones as

objects that have affinities with the main factors and so the people agents are distributed geographically in term of living and working locations in consistency with their affinities; it is possible to overlap zones with different affinities in order to represent inconsistent mixes of different groups of the population. In addition to the individual aspects, in the model the aggregation parameters are defined to regulate the generation of social networks based on stochastic distributions; by Montecarlo technique the people agents are associated in term of families and working connection generating the social network; the behavior of each agent is defined by models that regulates how it operates under regular (i.e. working days, holidays) or special conditions (i.e. natural disasters, sickness period); an overall presentation of the procedure is proposed in Figure 4.

Table I: Extract of Results based on non parametric rank correlation on available samples

BMI	Age	-0,1937	0,0001*
BMI	Years of school	-0,0563	0,3876
BMI	Alcohol (Dose/Week)	-0,0543	0,3093
BMI	Glucose (mg/dL)	0,1265	0,0196*
BMI	Cholesterol (mg/dL)	-0,1278	0,0225*
BMI	Triglycerides (mg/dL)	0,2704	<,0001*
BMI	HDL (mg/dL)	-0,2944	<,0001*
BMI	Cholesterol/HDL Ratio	0,1610	0,0030*
BMI	LDL (mg/dL)	-0,1808	0,0010*
BMI	Hb A1C%	0,4136	<,0001*
BMI	HR BP BASELINE	0,1784	0,0133*
BMI	SBP BASELINE	0,0166	0,8176
BMI	DBP BASELINE	-0,0041	0,9544
BMI	Walk speed (m/s)	-0,3111	<,0001*

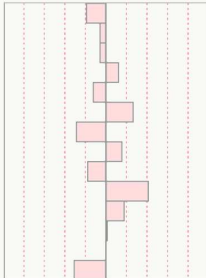


Table II: Some Correlation Extract among factors and BMI based on the available samples

BMI	BMI	1,0000
Age		-0,2019
Years of school		-0,1108
Alcohol (Dose/Week)		0,0278
Glucose (mg/dL)		0,2203
Cholesterol (mg/dL)		-0,1334
Triglycerides (mg/dL)		0,1971
HDL (mg/dL)		-0,3103
Cholesterol/HDL Ratio		0,1540
LDL (mg/dL)		-0,1712
Hb A1C%		0,4041
HR BP BASELINE		0,1973
SBP BASELINE		0,0270
DBP BASELINE		0,0017
Walk speed (m/s)		-0,2725

Obviously dealing with human modeling the verification and validation of the simulator is critical as well as the data collection and analysis; for the preliminary test the hypotheses on the conceptual model relating behaviors with obesity was based on current researches (Christakis et al. 2007) as well on the analysis of data available in Beth Israel Deaconess Medical Center (BIDMC) in Boston and Harvard Medical School affiliate. In fact some general data used for simulation were derived from previous studies obtained by the databases of associations (i.e. World Health Organization, the American Heart Association and the American Diabetes Association).

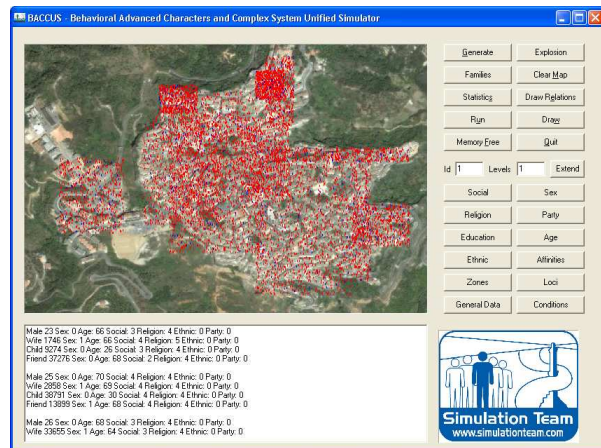


Figure 5. IA-CGF in BACCUS Simulator allows to evaluate the population behavior and its influence on the obesity evolution in the town

In addition it was used a sample related to 170 subjects in the Syncope and Falls in the Elderly Laboratory (Lab SAFE), active in the BIDMC; these 170 patients in the sample was volunteers, aged between 50 and 85 years, recruited from the SAFE Lab for four samples, conducted in the years 2006 to 2010.

All four samples considered issues related to diabetes; in fact the data were related to studies correlating diabetes with other pathologies, so the sample includes other diseases such as patients with myocardial infarction or stroke.

All samples have a quite balanced ratio between balance between male patients and female patients.

There are some limitations in using samples of existing populations that were collected or originated for different studies, and therefore do not provided sufficient information i.e. about frequency of measurements, include specific populations, or have missing values for certain parameters.

Therefore, this paper represents a first step forward for modeling and tuning simulators for investigating obesity, so the focus is mostly on creating the conceptual models, introducing consistent data and tuning the parameters in order to obtain reasonable results by the simulation runs.

Some examples, of the parameters used to check mutual influences are reported in table I e II; statistical analysis and ranking methodologies was applied in order to check data significance and their correlation.

The low level of correlation obtained was expected due to the reason above mentioned; therefore the analysis was useful to identify procedures and aspects to be investigated in future data collections and researches.

In order to proceed it was decided to implement some behavioral model and correlation algorithms among the factors based on author's hypotheses consistent with the data available.

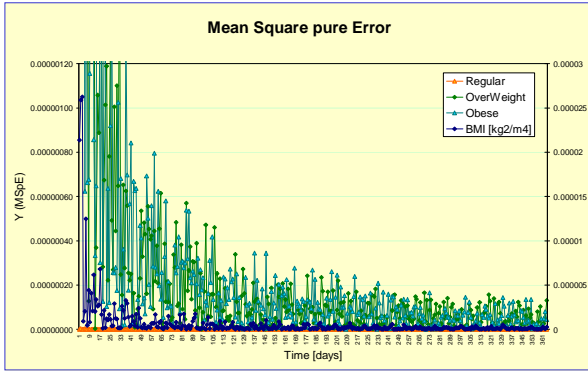


Figure 6. BACCUS VV&A based on Dynamic Analysis of Mean Square pure Error on Population Obesity Classes

The scenario used to the tests was a small town of about 15'000 inhabitants; target functions included the population sharing among the different obesity classes and the average BMI:

- Regular
- Overweight
- Obese
- Average BMI

In figure 5 it is proposed the BACCUS during execution of the proposed scenario, analyzing each single individual. The VV&A (Verification, Validation & Accreditation) of BACCUS is based analysis of MSPe (Mean Square pure analysis) as measure of the variance of the target functions among replicated runs over the same boundary conditions; by this approach it becomes possible to identify the number of replications and the simulation duration able to guarantee a desired level of precision; MSPe values in correspondence of these experimental parameters determines the amplitude of the related confidence band:

$$MSPe^m(t, n_0) = \frac{\sum_{i=1}^{n_0} \left[Sr_i^m(t) - \frac{1}{n_0} \sum_{j=1}^{n_0} Sr_j^m(t) \right]^2}{n_0}$$

$$CBA^m(t, n_0, \alpha) = \pm t_{\alpha, n_0} \sqrt{MSPe^m(t, n_0)}$$

t	simulation time
m	m-th target function estimated by the simulator
no	number of replications with same boundary conditions and different random seeds
Srk ^m _k (t)	m-th target function value at t time of the k-th replicated simulation
MspE ^m (t, n ₀ , α)	Mean Square pure Error at t time and with no replications for the m-th target function
α	percentile
CBA ^m (t, n ₀ , α)	Confidence Band Amplitude at t time, with no Replications for the m-th target function

In fact the MSPe allows to quantify the experimental error due to influence of the stochastic components; as presented in figure 6 the variance of all the target function reach steady state situation over a reasonable number of replications and over a time horizon of about 1 year. so this confirm that simulator provides consistent results on a stable situation with capability to define the confidence band for estimating the obesity target functions.

CONCLUSIONS

The use of agents and simulation to investigate large scale health care problems represent an important opportunity; obviously in these case it is critical to guarantee a multidisciplinary approach to the problem; in fact from this point of view the authors represents a good example of different skills and background with common interest.

The obesity epidemic represents a very important and interesting application framework that could be very useful to consolidate research in this area of M&S related to Medicine and Health Care.

The research highlighted the critical aspects related to collecting, mining and filtering the data to define the conceptual models related to such complex problems as well as to support parameter fine tuning and simulator VV&A. The model and the present results in this first phase are promising and the potential of using Intelligent Agents in this context is very great considering the impact of the obesity epidemic.

This presentation represents the first step on this research track, and currently the authors are working on some experimental analysis in term of impact on different industries (i.e. beverage and airlines) as well as in the further development of the models and their validation using datasets of a larger , longitudinal cohorts of diverse populations in different countries.

REFERENCE

- [1] Bruzzone A.G. Tremori A., Massei M., Adding Smart to the Mix, Modeling Simulation & Training: The International Defense Training Journal, 3, 25-27, 2011
- [2] Cereda C., Models and Analysis of Complex Systems for the Evaluation of Future Scenarios and of Related Infrastructural and Operational Needs, Genoa University Thesis, DIPTeM Press, 2011
- [3] Bruzzone A.G., Massei M., Intelligent Agents for Modelling Country Reconstruction Operation", Proceedings of AfricaMS 2010, Gaborone, Botswana, September 6-8, 2010

- [4] Bruzzone A.G., Scavotti A., Massei M., Tremori A., Metamodelling for Analyzing Scenarios of Urban Crisis and Area Stabilization by Applying Intelligent Agents, Proceedings of EMSS2008, September 17-19, 2008, Campora San Giovanni (CS),Italy, 2008
- [5] Wolf PA, Beiser A, Elias MF, Au R, Vasani RS, Seshadri S: Relation of obesity to cognitive function: importance of central obesity and synergistic influence of concomitant hypertension. The Framingham Heart Study. *Curr Alzheimer Res* 4:111-116, 2007
- [6] Wang Y, Beydoun MA: The Obesity Epidemic in the United States- Gender, Age Socioeconomic, Racial/Ethnic and Geographic Characteristics : A systematic Review and Meta-Regression Analysis. *Epidemiologic Reviews* 29:6-28, 2007
- [7] Christakis N.A., Fowler J.H., The Spread of Obesity in a Large Social Network Over 32 Years, *The New England Journal of Medicine*, 357-4, 370-379, 2007
- [8] Falkstedt D, Hemmingsson T, Rasmussen F, Lundberg I: Body Mass Index in late adolescence and its association with coronary heart disease and stroke in middle age among Swedish men. *International Journal of Obesity* 31:777-7783, 2006
- [9] Cournot M, Marquie JC, Ansiau D, Martinaud C, Fonds H, Ferrieres J, Ruidavets JB: Relation between body mass index and cognitive function in healthy middle-aged men and women. *Neurology* 67:1208-1214, 2006
- [10] Popkin B.M., Kim S., Rusev E.R., Du S., Zizza C., Measuring the full economic costs of diet, physical activity and obesity-related chronic diseases, *Obesity Review*, 7, 271-293, 2006
- [11] Bruzzone A.G., Bocca E., Rocca A., Algorithms devoted to Reproduce Human Modifiers in Polyfunctional Agents, Proc. of SCSC2006, Calgary, Canada, July 30-August, 2006
- [12] Whitmer RA, Gunderson EP, Barrett-Connor E, Quesenberry CP, Jr., Yaffe K: Obesity in middle age and future risk of dementia: a 27 year longitudinal population based study. *BMJ* 330:1360, 2005
- [13] Kivipelto M, Ngandu T, Fratiglioni L, Viitanen M, Kareholt I, Winblad B, Helkala EL, Tuomilehto J, Soininen H, Nissinen A: Obesity and vascular risk factors at midlife and the risk of dementia and Alzheimer disease. *Arch Neurol* 62:1556-1560, 2005
- [14] Australian Bureau of Statistics, National Health Survey 2004-05: Summary of results. ABS cat.no. 4364.0. Canberra, 2005
- [15] Wang L.Y., Yang Q., Lowry R. and Wechsler H.,Economic Analysis of a School-Based Obesity Prevention Program, *Obesity Research* 11, 1313-1324, 2003
- [16] Magarey AM, Daniels LA & Boulton JC, Prevalence of overweight and obesity in Australian children and adolescents: reassessment of 1985 and 1995 data against new standard international definitions. *Medical Journal of Australia* 174: 561-564, 2001
- [17] Avalle L, A.G. Bruzzone, F. Copello, A. Guerci, P.Bartoletti Epidemic Diffusion Simulation Relative to Movements of a Population that Acts on the Territory: Bio-Dynamic Comments and Evaluations, Proc. of WMC99, San Francisco, January, 1999
- [18] Pliskin, J., Shepard, D, Weinstein, M, Utility Functions for Life Years and Health Status". *Operations Research (Operations Research Society of America)* 28 (1): 206-224, 1980
- [19] Spearman C.E. "The Abilities of Man: Their Nature and Measurement", Macmillian, London, 1927
- [20] Foster G.D., What is a reasonable weight loss? Patients' Expectations and Evaluation of Obesity Treatment Outcomes, *Journal of Consulting and Clinical Psychology*, 65 (1): 79-85, 1997
- [21] WHO, World Health Statistics: Reports, 2011
- [22] WHO, Preventing and Managing the Global Epidemic of Obesity: Report of the World Health Organization Consultation of Obesity, Geneva, June 1997
- [23] National Institutes of Health, Clinical Guidelines on the identification, Evaluation and Treatment of overweight and Obesity in Adults: The Evidence Report, 1998

MARITIME SECURITY: EMERGING TECHNOLOGIES FOR ASYMMETRIC THREATS

Agostino Bruzzone, Marina Massei, Alberto Tremori,
MISS-DIPTEM, University of Genoa, Via Opera Pia 15, 16145 Genova, Italy
Email {agostino, massei, tremori}@itim.unige.it - URL www.itim.unige.it

Francesca Madeo, Federico Tarone
Simulation Team, via Molinero 1, 17100 Savona, Italy
Email {madeo, tarone}@simulationteam.com - URL www.simulationteam.com

Francesco Longo
MSC-LES, Mechanical Dept, University of Calabria, Via P. Bucci 44C, 87036 Rende, Italy
Email f.longo@unical.it - URL www.msc-les.org

ABSTRACT

This paper analyses the evolution of a complex scenario for security: maritime environment and in particular coastal areas and harbors the critical nodes in the whole system. In new technologies provide an effective support in this asymmetric framework for situation awareness and threat assessment. M&S, CGF, Data Fusion are techniques that allow the users to obtain efficient awareness on the general on-going situation in real time and to support decision over complex scenarios.

Keywords: *Maritime and Harbor Security, Human Behavior Modeling, Computer Generated Forces, Data Fusion*

INTRODUCTION

This paper provide an overview on a combined approach using M&S (Modeling & Simulation) and Data Fusion techniques to analyze complex scenarios involving asymmetric marine environments; the idea to use intelligent agents (IAs) as driver for Computer Generated Forces is a very critical aspect for modeling scenarios where many entities interact (i.e. commercial and nautical traffic around a port); in order to succeed in this sector it is critical to identify the requirements for such combined solution; the authors focus this paper on the following aspects

- To provide a quick Overview on the Modern Complex Scenarios and to identify the related Challenges
- To present the Potential of R&D within this Framework
- To Outline innovative enabling Technologies, Methodologies and Solutions for succeeding
- To present actions, investments on the R&D Tracks to support these activities
- To outline R&D potential Outcomes
- To present Examples and Approaches in this context

SECURITY IN MARITIME: AN EVOLVING SCENARIO

Today Maritime Security is a very critical aspect on Marine Framework introducing the concept of Asymmetric Marine Environment with new special attention to Threats such as:

- **Piracy**
- **Conventional Terrorism**
- **CBRN (Chemical, Biological, Radiological and Nuclear) Threats**

Some important aspects are expected to increase over Next Years their impact in General as well in Marine Framework increasing on Asymmetric Threats such as:

- Movement of European Region Social Economic Center of Gravity to South increasing maritime traffic with North Africa
- Stabilization and Normalization Processes and Country Reconstruction Initiatives Overseas
- Overseas Developing Areas Growth, Production/Demand & Sustainability Issues Technologies
- Easier access to New Dimensions for preparing and creating critical threats (i.e. Cyberspace)
- Multiple opportunities to Access to Resources to develop WMD (i.e. smallpox, RDD)
- IT & Web empowering the potential of individuals and small groups (i.e. C2 capabilities)
- Increasing new reachable targets such as Oil Platform, Environmental Threats, Social Service Political Issues
- Political Instability on Critical Regions (i.e. Africa)
- Evolution of Principle of Nations and Populations (i.e. Commercial States)
- Evolution of new critical issues requiring rational on joint Defense and Homeland Security Budgets (i.e. natural resource issues: water)

The Real World: Multi Dimension and Multi Layer Resolution

Asymmetric warfare is a very complex framework and modeling and simulation need to properly address all the related issues; in fact this context is:

- A Real World on 5 Dimensions:
 - Surface
 - Underwater
 - Air
 - Space
 - Cyber
- A Multi Layers & Resolutions Frame
 - Fleets and Parties
 - Ships and Commercial Traffic
 - Crew & People Accessing Ports/Vessels
 - Services & Infrastructures

As explanatory Example from the new challenges in this context it could be useful to consider the evolution over the years; in fact today Modeling is critical to evaluate Strategies in Threat Identification, Decision Making & Evolution Prediction based on their behavior much more respect on their features:

- Once upon Time people were used to identify threats based on Platform Detection, Identification and Classification
- In Some case the same Platform is in use on multiple sides by different actors, someone friendly and someone "extremely foe".
- In some case the Platform is becoming a Menace just based on own it is operating

It is sufficient to consider the case of piracy to realize that the problem is not to detect the kind, class or even name of a ship based on their silhouette or using EMS, but to identify suspect behaviors that suggest presence of pirates inside a fisherman boat.

Port Protection and Asymmetric Naval Warfare

For facing threats in asymmetric naval warfare it is necessary to develop new Models and Solutions able to Interoperate with the critical components of such Scenarios such as:

- Non Conventional Operations
- Human Behaviors on (i.e. Crew, Stakeholders, Domestic Opinion)
- Services & Infrastructures
- Commercial Traffic & Yachting
- Port Infrastructures and Resources
- Joint Operations (i.e. Ship Inspections, Littoral Control, C5I2)

NEW ENABLING TECHNOLOGIES

The existing and new technologies have a great potential in this area; for instance communication infrastructures and mobile solutions allows today to distribute information as well as data collection, data processing and decision making over a large complex

network; in addition to physical technologies and infrastructures it is even more important the benefits provided by innovative soft computing techniques and methodologies.

Impact of Innovative Technologies such as IA, CGF and HBM in Marine Frameworks

In fact, innovative *IA (Intelligent Agents)*, *CGF (Computer Generated Forces)* & *HBM (Human Behavior Modeling)* represent a Strategic Issues in different application areas to be applied to asymmetric marine warfare; in particular it is possible to consider the following application area and related benefits provided by these innovative solutions:

- **Simulation Based Acquisition and Test & Analysis**
 - Capability to Proceed in Data Farming on Different Hypotheses on Vessel and System Design on Virtual Prototypes
- **Training and Exercise**
 - Reduction of human personnel for Training & Exercising
 - New Scenarios involving Dynamic Simulated Complex System vs. the old pre-defined scripts
- **Operational Planning**
 - Reducing Time for Planning Development due to the reduction of human experts employed in the different roles
 - Possibility to Experiments different Alternatives by replicated runs carried out in Automatic way
- **Mission rehearsal and conduct operations**
 - Capability to keep the simulation on-line and to conduct statistical experimental analysis

Therefore, in order to apply R&D to current and future Asymmetric Marine Framework, the authors identified the following innovation tracks to be investigated:

- **Cognitive Technologies**
 - Data Fusion (i.e. Situation Assessment)
 - Human Behavior Models
 - Intelligent Agents & CGF
 - Decision Support Systems (i.e. Web 3.0)
- **Modeling & Simulation**
 - Concept and Doctrine Development (i.e. Interoperable Simulation)
 - Simulation Based Acquisition (i.e. Virtual Prototyping)
 - Training (Mobile Training, Serious Games)
 - Serious Games used to create complex scenarios with multiple players and to investigate different strategies
- **Equipment & Devices**
 - Integrated Solutions (i.e. Mobile Tactical Control Systems)
 - Platforms (i.e. UAV, AUV, NMM)
 - Sensors (i.e. Through Wall Sensors)
 - Weapons (i.e. Non Lethal Weapons)

AVAILABLE EXPERIENCES

The authors have experience in using several of these techniques in different applications, for instance, currently, the authors are developing a solution (PANOPEA), to test complex scenarios related to piracy and to investigate different C2 solutions as well as strategies and technologies within this framework; obviously the authors have even long experience in traditional applications (i.e. data fusion over conventional air naval scenarios). In the paper some experience and simulation model are presented as example of the potential of these techniques within marine asymmetric scenarios.

An Example of Simulated Attack to a Port

As Example of Port Attack Simulation it is proposed an unclassified simulation scenario developed in cooperation among Bulgarian Academy of Sciences, Lockheed Martin Canada, MISS DIPTM Univ. Genova, CRTI, NATO PBIST Experts, Port Authorities, CUBRC within a NATO working group; the scenario is based on the following objects and hypotheses:

- Small boat (fishing or pleasure normally seen in the harbour, not regulated by ISPS code) filled with a mixture of explosive and CBRN;
- The boat is heading for a oil Terminal or tanker within a port
- No predicted pattern from the pleasure boat is available
- No clear strategic warning available until close to the attack
- Necessity to concentrate on attack assessment and response
- A Priori and HUMINT information are essential in the fusion process and need specific models

For this scenario it is necessary to model and simulate available technologies such as:

- Tracking small targets : Track continuity
- Sensors for CBRN detection
- Unusual maneuver detection for threat assessment
- Data fusion for Recognizing in a very short time the attacking boat
- Solutions for permanent and continuous observation and surveillance

The use of simulation is an important benefits to address critical questions that are necessary to clarify for port security assessment, for new equipment design and for security procedure definition; an example of this question is following: if no automated radiation detectors are available in an Harbor Area, when is the true nature of the threat discovered?

Information	Source
Ship track and maneuver	Sonar, HFSWR, optical, acoustic
Radiation signature	Radiation detector
Human Intelligence	Police, government agencies in order to identify the crew; unusual behavior of the crew

In the following figures the different fallout areas and contamination risks are summarized respect different kind of CBRN devices



Fig.1 Areas of contamination Case 1: 100Ci 241 Am Two oil-well logging sources



Fig.2 Areas of contamination Case 2: 20kCi 90Sr Russian RTG

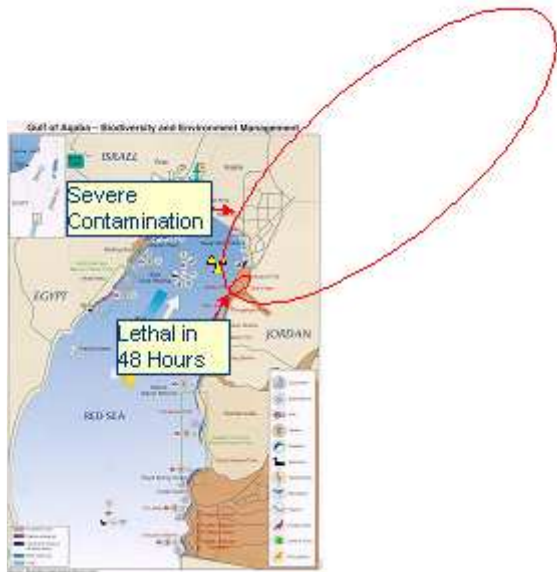


Fig.3 Areas of contamination Case 3: 10kT Nuclear Weapon

It is evident the importance to define best solution to face these challenges by considering the area impact of such devices as well as the implication of a port attack in term of economic costs and strategic issues. The use of M&S, CGF, and Intelligent Agent CGF is critical to test new algorithms to detect suspect behaviors or to develop the requirements for new security solutions; today the use of mobile networks provides very interesting opportunity to share info quickly and easily and to develop deployable netcentric solutions, therefore the use of M&S is critical to properly design the specific configuration and operative modes.

ASYMMETRIC THREATS: ASYMETRIC MODELING

The complexity of marine scenarios is due often to the involvement of many entities that generate a very challenging framework for detecting real asymmetric threats from false alarms or uncommon behaviors; in order to face this challenge it is necessary to create models able to reproduce complex behaviors such as that ones that characterize general cargo operations, commercial traffic, pleasure boats as well as the threat tactics. By this approach it is possible to create models that support the marine asymmetric threats assessment.

Human Factors and Marine Simulation

Most of the critical issues in generating and simulating large maritime scenarios is dealing with the necessity to model the humans factors that affect the activities of the vessels, boats, airplanes, coast infrastructures as

well as all the other elements present in a specific framework; in fact the complexity to Coordinate Humans in not-conventional operations for improving their coordination and capabilities to face complex challenges is a well known element in Navy. So considering asymmetric threats it is even more important to models these Human Factors both for the directing the threats, for reproducing the boundary elements as well as for being actors for our resources; this point is even more evident by a simple example: looking to each single Vessel it is evident that for Simulating its capabilities in reacting to threats it is very critical to model the crew and its human behavior modifiers (i.e. stress, fatigue, harmony).

IA-CGF & Human factors in maritime security

The authors have developed models for reproducing human factors and to represent intelligent agents able to direct objects within interoperable simulators; in fact the Simulation Team create a new generation of CGF, titled IA-CGF (Intelligent Agent Computer Generated Forces) for this purpose and some application within marine environment is already available and in the experimentation phase over complex scenarios.

These new IA-CGF are organized based on Modules that are interoperable in HLA Federation (High Level architecture) and they include:

- IA-CGF Units (i.e. commercial ships, contractors on the ship, special teams, fisherman boats, coast guard units)
- IA-CGF HBL Human Behavior Libraries (i.e. fatigue, stress, aggressiveness, trustiness)
- IA-CGF NCF Non-Conventional Frameworks devoted to reproduce specific scenarios (i.e. piracy)

The IA-CGF are available to support different aspects in the marine asymmetric threat simulation, such as:

IA Drive the General Traffic & Critical Entities

In fact the use of Intelligent Agents provide the capability to create large simulation frameworks where airplanes, yachts, ships, ground entities act in consistency with their nature and within the Scenario and react dynamically to the Simulation Evolution.

IA Direct the Port and Coast Protection within its 5 Dimensional Space

IA are able to direct actions of the different resources for port and coast protection, so it becomes possible to run extensive experimental campaigns by simulation for defining optimal protection solution and to assess the threats over a complex scenario; these results are achievable by testing and evaluating the effectiveness and efficiency of the all Naval Resources, including platforms, weapons, individual sensors, ground infrastructures, C2 and different information sources for protecting assets against new threats including not conventional use of civil resources.

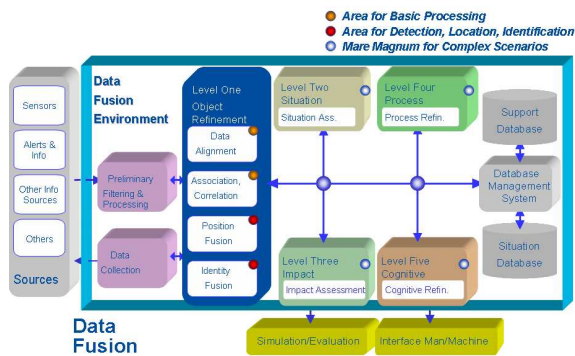


Fig.4 Asymmetric Data Fusion

In the following part of the paper, different Interoperable M&S solutions, developed by the authors for Marine Environment are presented as example of their potential in facing the above mentioned challenges.

PANOPEA

PANOPEA is a IA-CGF NCF that use IA for reproducing a complex framework related to piracy involving several thousands of vessels, plus all related activities (i.e. intelligence, ports, special forces, contractors, helicopters, UAV, etc.) In fact PANOPEA simulation allows to model Piracy activities; a specific study on-going by using this simulator is related to the evaluation of different strategies in NEC C2 M2 (Net-centric Command and Control Maturity Models) and in quantify benefits related to guarantee C2 agility. PANOPEA reproduces military vessels, helicopters, ground base units, cargos, as well as small medium boats, fishermen and yachts traffic as well as Pirates; all these entities are driven by Intelligent Agents and apply strategies for succeeding in their specific tasks.

PLACRA

The Placra simulator was developed by the authors in order to reproduce the crew activities and behavior on Oil Platforms as well as on vessels; in this case the simulator takes care of reproducing operative procedures, on-board micrologistics as well as the human behavior modifiers and their impact on crew efficiency; the model consider the workload, individual and team characteristics, their history and previous experiences as well as the platform infrastructures and equipment; the simulation evolve over a scenario where regular or critical events have to be handled.

MESA

MESA is an integrated environment, developed by the authors, to perform simulation and risk analysis in ports and maritime sector considering the evolution of emergencies; MESA combines the simulation with GIS to support safety and security assessment plans and operations; in fact MESA is devoted to support port

organizations, entities and operators in Emergency & Environmental Management. MESA is a modular system based on combined simulators running on PC able to export directly the results on WWW servers.

FLODAF

As Asymmetric Data Fusion example FLODAF framework was developed by the authors as tool devoted to support engineering and performance estimation of Data Fusion architectures and algorithms; this suite includes a Scenario Generator and a Simulator for analyzing the Data Fusion performances over complex Air-Naval scenarios including surface and underwater vessels, aircrafts.

ST-VP

ST-VP was by Simulation Team originally as a framework to support Training in marine environment; in fact the Interoperability of ST-VP simulators is based on HLA and guarantees in addition to traditional stand-alone training, even Concurrent Cooperative Training in complex Operations and Policies; ST-VP have a long experience in being applied within commercial ports.

In fact The ST-VP includes all the different port equipment and even other marine devices and platforms; ST-VP in addition to Operator Training supports even Safety and Security Training, Procedure Definition, Equipment Design and Virtual Prototyping; among ST-VP innovative capabilities the following aspects are interesting

- ST-VP is a fully containerized real-time distributed HLA Simulator reproducing Marine Environments and Ports. ST-VP is integrated within a 40' High Cube Container ready to be used on site immediately after arrival.
- ST-VP Simulator allows to operate all the different Equipment in a Virtual World by an immersive Cave (270 ° Horizontal and 150° Vertical), reproducing Sounds, Vibrations, Motion in all weather conditions
- ST-VP includes a Full-Scope Simulation for Training Operations & Procedures, an Integrated Class Room, the Instructor Debriefing Room, and secondary Interoperable Simulators of all the Port Cranes and a Biomedical Module for Safety, Ergonomic and Posture Enhancement.
- ST-VP World is customizable for each Port, Procedure and Equipment

An example of ST-VP federation applied to a marine security scenario is proposed in the following scheme:

In fact ST-VP is able to interoperate with other simulators (virtual and constructive) as well as with real equipment; among the others it is possible create connections with: ST_PT & ST_RS Simulators

(driving simulators), Seaports (Simulator of Terminal and Ports Security Procedures and Operations), TRAMAS (Simulation of the Logistic Network & Impact on the Town of port activities), KATRINA LIKE (regional scenario simulation reproducing a large scale crisis)

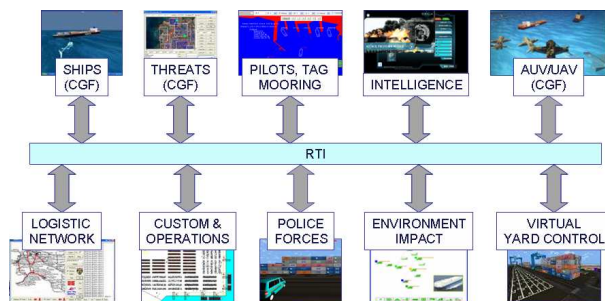


Fig.5 ST_VP FEDERATION

CONCLUSIONS

It is evident that Maritime Security is part of a wide Scenario and needs to be addressed by an integrated approach: due to the complexity of the framework the use of simulation is very effective and this paper proposes some of the critical methodologies and techniques to be used in this direction.

In addition, it is important to outline that Marine Asymmetric Warfare is fast evolving, introducing new issues and new threats affecting more and more subjects, so it is becoming very urgent to create capabilities in defining, evaluating and optimizing solutions to face these challenges; so it becomes evident that Simulation and Cognitive Technologies are the key issues for succeeding in this goal; in fact today it is very critical to proceed in research and investigation on these domains respect the new evolving threats and to develop of New models and simulators for supporting the development of Systems, Devices and Equipment.

In fact the importance of these aspects suggests that it is critical to create new capabilities in security for Maritime Scenario and to network with all international research centers operating in this context. In order to succeed it is critical to develop critical assessments as well as to establish connections with Agencies, Companies and Institutions operating in this area.

REFERENCES

- Alberts, D.S. (2008) "NATO NEC C2 Maturity Model Overview," Draft for Peer Review, SAS-065 Study Group, 2008.

- Alberts, D.S. (2007) "Agility, Focus, and Convergence: The Future of Command and Control." The Int. C2 Journal, Vol. 1, No. 1, 1-30
- Alberts David S., Hayes Richard E., (2006) "Understanding Command and Control", Washington: CCRP, fig. 11, p. 35
- Alberts D.S., Hayes R.E. (2003) "Power to the Edge", CCPR Publications, Washington D.C.
- Alberts D. S., (2002), "Information Age Transformation", revision June 2002, CCRP Publication Series
- Alberts David S., Gartska J.J., Stein F.P. (2000) "Net Centric Warfare", CCRP, Washington
- Anderson N. (2006) "Are iPods shrinking the British vocabulary?", Ars Technica On-Line Magazine, December 15
- Bocca E., Pierfederici, B.E. (2007) "Intelligent agents for moving and operating Computer Generated Forces" Proceedings of SCSC, San Diego July
- Bruzzone A.G., Massei M., Madeo F., Tarone F. (2011) "Simulating Marine Asymmetric Scenarios for testing different C2 Maturity Levels", Proceedings of ICCRTS, Quebec, Canada, June
- Bruzzone A.G. Tremori A., Massei M. (2011) "Adding Smart to the Mix", Modeling Simulation & Training: The International Defense Training Journal, 3, 25-27, 2011
- Bruzzone A.G., Tarone F. (2011) "Innovative Metrics And VV&A for Interoperable Simulation in NEC, Urban Disorders with Innovative C2", MISS DIPTM Technical Report, Genoa
- Bruzzone A.G., Massei M. (2010) "Intelligent Agents for Modelling Country Reconstruction Operation", Proceedings of Africa MS2010, Gaborone, Botswana, September 6-8
- Bruzzone A.G., Massei M. (2010) "Advantage of mobile training for complex systems", Proceedings of MAS2010, Fes, Morocco, October 13-15
- Bruzzone A.G. (2010) "CGF & Data Fusion for Simulating Harbor Protection & Asymmetric Marine Scenarios", Proceedings of SIM&SEA2010, La Spezia, June 8
- Bruzzone A.G., Cantice G., Morabito G., Mursia A., Sebastiani M., Tremori A. (2009) "CGF for NATO NEC C2 Maturity Model (N2C2M2) Evaluation", Proceedings of I/ITSEC2009, Orlando, November 30-December 4
- Bruzzone A.G. (2008) "Human Behavior Modeling: Achievement & Challenges", Invited Speech at SIREN Workshop, Bergeggi, Italy, June 6th

- Bruzzone A.G., (2007) "Challenges and Opportunities for Human Behaviour Modelling in Applied Simulation", Keynote Speech at Applied Simulation and Modelling, Palma de Mallorca
- Bruzzone A.G., Figini F. (2004) "Modelling Human Behaviour in Chemical Facilities and Oil Platforms", Proceedings of SCSC2004, San Jose'
- Bruzzone A.G., Viazzo S., Massei M., BC (2004) "Modelling Human Behaviour in Industrial Facilities & Business Processes", Proc. of ASTC, Arlington, VA, April
- Bruzzone A.G., Massei M., Simeoni S., Carini D., B.M. (2004) "Parameter Tuning in Modelling Human Behaviours by Using Optimization Techniques", Proceedings of ESM, Magdeburg, Germany, June
- Bruzzone A.G., Figini F. (2004) "Modelling Human Behaviour in Chemical Facilities and Oil Platforms", Proceedings of SCSC, San Jose
- Bruzzone A.G., Mosca R., Simeoni S., Massei M., B.M., B.C., (2004) "A Methodology for Estimating Impact of Different Human Factor Models on Industrial Processes", Proceedings of EUROSIM, Paris, France, September
- Bruzzone A.G., Procacci V., B.M., B.C. (2001) "FLODAF: Fuzzy Logic Applied to a Multi-Sensor Data Fusion Model", Proceedings of FLODAF2001, Montreal, August 7-10
- Bruzzone A.G., Rapallo S., Vio F., (1999) "MESA: Maritime Environment for Simulation Analysis", Tech.Report of ICAMES, ENSO, Bogazici University, Istanbul, May 15-21
- Bruzzone A.G., Page E., Uhrmacher A. (1999) "Web-based Modelling & Simulation", SCS International, San Francisco, ISBN 1-56555-156-7
- Bruzzone A.G., Giribone P. (1998) "Decision-Support Systems and Simulation for Logistics: Moving Forward for a Distributed, Real-Time, Interactive Simulation Environment", Proc. of the Annual Simulation Symposium IEEE, Boston
- Bruzzone A.G., Giribone P. (1998) "Quality Service Improvement by Using Human Behaviour Simulation", Proceedings of ESM, Manchester, UK, June
- Bruzzone A.G. (1996) "Object Oriented Modelling to Study Individual Human Behaviour in the Work Environment: a Medical Testing Laboratory ", Proc. of WMC, San Diego, January
- Cantice G. (2008) "Serious Games... Serious Experimentations?", Proc. of SeriGamex, November
- CTA (2002) "Agents for Net-Centric Warfare and Time Critical Targets", CTA Technical Report
- Fogel D. (2005) "Volutionary Computation-toward a new philosophy of machine intelligence", IEEE Press series on Computational Intelligence
- Goldstein J. (2007) "Trial in Absentia Is Ordeal for Veteran Who Was Cleared by U.S. in a Killing", NY Sun, July 16
- Krulak C.C. (1999) "The Strategic Corporal: Leadership in the Three Block War" Marines Magazine
- Ladner R., Petry F. (2005) "Net-Centric Web Approaches to Intelligence and National Security", Springer, NYC
- Ladner R., Warner E., Petry F., Katikaneni U., Shaw K., Gupta K., Moore P. (2009) "Web Services: Evolving Techniques in Net-Centric Operations", Proceedings of MTS/IEEE OCEANS,
- Liddy L. (2005) "The Strategic Corporal: Some Requirements in Training and Education", Australian Army Journal, Volume II, Number 2, 139-148
- Merkuriev Y., Bruzzone A.G., Novitsky L (1998) "Modelling and Simulation within a Maritime Environment", SCS Europe, Ghent, Belgium, ISBN 1-56555-132-X
- Molagh J. (2009) "How Afghanistan's Little Tragedies Are Adding Up", Time, May 26
- Moniz D. (2002) "Afghanistan's Lessons Shaping New Military", USA Today, October 7
- Mosca R., Viazzo S., Massei M., Simeoni S., Carini D., B.C. (2004) "Human Resource Modelling for Business Process Re-Engineering", Proceedings of I3M2004, Bergeggi, Italy, October
- Patton M.S. (2003) "ES2: Every Soldier is a Sensor", The Washington Post, November 5
- Ray D.P. (2005) "Every Soldier Is a Sensor (ES2) Simulation: Virtual Simulation Using Game Technology", Military Intelligence Professional Bulletin
- Reverberi A. (2006) "Human Behavior Representation & Genetic Algorithms Characteristics Fitting", Invited Presentation on Piovra Workshop, Savona, February 7
- Shahbazian E., Rogova G., Weert M.J. (2009) "Harbour Protection Through Data Fusion Technologies", Series: NATO Science for Peace and Security Series C: Environmental Security, Springer
- Warne L., Ali I., Bopping D., Hart D., Pascoe C. (2004) "The Network Centric Warrior: The Human Dimension of Network Centric Warfare", Tech.Report DSTO, Edinburgh, Australia

ON THE SHORT PERIOD PRODUCTION PLANNING IN INDUSTRIAL PLANTS: A REAL CASE STUDY

Agostino Bruzzone^(a), Francesco Longo^(b), Letizia Nicoletti^(c), Rafael Diaz^(d)

^(a) MIS-DIPTEM, University of Genoa, Italy

^(b) MSC-LES, University of Calabria

^(c) CAL-TEK S.r.l, Italy

^(d) VMASC, Old Dominion University

^(a) agostino@itim.unige.it; ^(b) f.longo@unical.it; ^(c) l.nicoletti@cal-tek.eu; ^(d) rdiaz@odu.edu

ABSTRACT

Due to the increased level of competition, nowadays production systems have to keep high performances ensuring customer satisfaction, cost-reduction and high product quality. The features of the actual competitive scenario drive to pursue even higher levels of efficiency in companies management. In this perspective production planning, with special regard to short period production planning, plays a key role. As a matter of fact, while the long period planning aims at the evaluation of production quantities for each product, the short period planning aims at the definition of an optimal schedule to achieve even higher system performances. As well known a scheduling problem encompasses a great complexity, this kind of problem can be seen as a double allocation problem where the allocation of the jobs to production resources and the allocation of the jobs in a specific time production horizon have to be defined. The complexity grows even further considering that many interacting and variables must be taken into account simultaneously and the stochastic system behaviour cannot be neglected.

This paper faces scheduling problems in a real manufacturing system proposing an approach based on genetic algorithms, dispatching rules and Modelling & Simulation.

Keywords: Shop Order Scheduling, Discrete Event Simulation, Genetic algorithms, Dispatching Rules.

1. INTRODUCTION

The short period production planning tackles the problem of assigning the arriving jobs to workers, machines, equipment and other resources over time. As stated in (Kiran, 1998), scheduling problems are concerned with the determination of which resources should be used and the determination of the completion and starting time for each operation of each order so that no constraint are violated and some scalar functions, measuring the effectiveness of a particular schedule, are maximized (or minimized). Getting a lower inventory level, a high plant efficiency (it means high machine and labor utilization), and respecting due dates, are some examples of scheduling criteria. (Riane

et al, 2001). The problems arising in production scheduling are notoriously very difficult and technically complex because they involve a large number of tasks and resources subject to different constraints and objectives; the complexity grows even further due to uncertainties in the manufacturing environment (Smith, 1992).

Note, in addition, that optimal allocation of the jobs to production resources over time is a combinatorial problem (Garey et al., 1976).

Scheduling problems can be formulated using analytical methods like mathematical programming or network theory. In this way, for small size problems, optimal solutions can be detected but, in most cases, the assumptions required for the analytical formulation are too restrictive so the resulting mathematical model may be not able to represent with accuracy the real problem (Son et al, 1999). In other words theoretical notions tend to oversimplify crucial factors of the actual production process proving that an analytic formulation and resolution is inadequate.

Many research works on scheduling problems have been carried out with analytical approaches but most of them consider only one or few constraints (e.g. setups, failures, blocking, etc.) at the same time and as often as not one scheduling objective (criteria) while multiple scheduling objectives subject to several constraints have to be considered in real manufacturing systems. Also the enumerative methods and in general exact methods (usually applied when analytic procedures are not available) are prohibitive to use because of their unrealistic computing requirements (Riane et al, 2001).

It is evident that the advances of theory have had a limited impact in practice but it does not mean that advances in scheduling theory have been a waste of time because they have provided interesting insights into the scheduling problem (Pinedo, 2008). An alternative approach to face this problem lies in the use of Modeling & Simulation, simulating reality by building a simulation model (Johtela et al, 1997). Modeling & Simulation allows to overcome the gap between theory and real-world scheduling problems thanks to the capability to represent real word systems and its constraints (Frantzen et al, 2011). Different

simulation modeling approaches taken in the literature about job-shop have been reviewed by (Ramasesh, 1990) providing a state-of-the-art survey of the simulation-based research on dynamic job shop scheduling.

In the literature, there are two major approaches to deal with simulation-based scheduling problems, namely:

- A simulation-based approach using dispatching rules;
- A simulation-based approach using meta-heuristic search algorithms.

The first approach allows to put in comparison dispatching rules establishing which one performs better (Andersson et al, 2008) .

Carri (1986) describes this approach as the experimentation of scheduling rules and the assessment of the effect of different rules on shop's ability to meet delivery dates and utilize machines. Experimentation with simulation models makes it possible to compare alternative scheduling rules, test broad conjunctures about scheduling procedures and develop greater insight into the job shop operation (Vinod and Sridharan, 2011).

Many research works about this approach can be mentioned.

Parthanadea and Buddhakulsomsirib (2010) develop a computer simulation for canned fruit industry and conduct computational experiment on the simulation model to determine a set of appropriate dispatching rules.

In Liu (1998) a two-stage simulation process has been presented: in the first stage, a number of dispatching rules are used as input parameters to generate candidate production schedules from a simulation model; in the second stage the performances of these production schedules are evaluated by another simulation model.

Goyal et al. (1995) have carried out a simulation study in order to analyze the scheduling rules for a flexible manufacturing system. Different combinations of scheduling rules have been applied evaluating their effect on system performances.

Huq and Huq (1995) have developed a simulation model, using a hypothetical hybrid job shop, to study the performance of different scheduling rules combinations with variations in arrival rates and processing times. Flow time, tardiness and throughput have been used as performance measures. They have found out that the rule combination performance varies with the performance criteria, and the combinations are sensitive to arrival rates and processing times.

Holthaus (1997) developed new scheduling rules by the combination of well known rules, and conducted a simulation-based analysis of those rules in the dynamic job shop environment. He concluded that the new scheduling rules are quite efficient.

Many other simulation studies have been carried out to evaluate the performances of dispatching rules: Holthaus and Rajendranb (1997),(Hicks and Pupong , 2006).

However, in general, this approach does not allow to find the optimal schedule.

The second approach mentioned above is based on the combined use of meta-heurist optimizer with simulation and allows to detect the optimal schedule (Andersson et al, 2008).

Among the meta-heuristic algorithms, genetic algorithms (GA) have been recognized as a general search strategy and an optimization method which is often useful for finding combined problems; for these reasons GA have been used with increasing frequency to address scheduling problems (Jeong et al, 2006).

The application of genetic algorithms to scheduling problem has been proposed by Bierwirth (1995), Syswerda (1991), Dorndorf and Pesch (1993), Yamada and Nakano (1992), Sakawa and Mori (1999) , Ghedjati (1999), Haibin and Wei (2001), Yun (2002), Vinod and Sridharan (2011) and many others.

The joint use of genetic algorithm and simulation are further proposed in Hou and Li (1991), Rabelo et al.,(1993), Ferrolho and Crisóstomo, (2007).

A comparison of these two approaches has been presented by Kim et al.,(2007) for job shop schedules of standard hydraulic cylinders and genetic algorithm were found to be better than dispatching rules (LPT, SPT, most work remaining MWKR, and least work remaining LWKR).

Similar results were found out in (Sankar et al., 2003) where the results obtained with GA are compared with the results obtained using six different dispatching rules including SPT, LPT, EDD, largest batch quantity (LBT), smallest batch quantity (SBQ) and highest penalty (HP). In this study it has been found out again that the solutions generated by GA outperform the solutions obtained by using Priority Dispatching Rules. PDRs and meta-heuristic optimizer can also be jointly used with good results as shown by (Andersson et al, 2008)

In this research work we present a study in which simulation is jointly used with genetic algorithms and dispatching rules to face stochastic scheduling problems in a real manufacturing system. The main goal of the present work is to provide a useful tool that can be integrated in the management system of the company and that can be profitably and efficiently used for short period production planning. The paper is structured as follows: Section 1 presents an accurate description of the system under study, Section 2 presents the steps that have been followed to built the simulation model, Section 3 deals with the verification and validation of the simulator, in section 4 the main results have been presented and finally the last section describes the main conclusions .

2. MANUFACTURING PROCESS DESCRIPTION

The project has been developed in collaboration with a small company, which produces high pressure hoses, under specific request of the company top management.

During the initial meetings and analyzing the initial collected data it had been evident the efficiency reduction due to the short period production planning. In particular the effective production was smaller than the target production and there were continuous delays in Shop Orders (here in after S.O.s) completion that caused the decrease of the customers' satisfaction level. So the purpose of this study is to create a decision making tool (specifically a simulator) that could be easily integrated and profitably used in the company management system to support the short period production planning. It is useful to give a brief description of the manufacturing process in order to provide a greater understanding of the steps carried out in the present research work.

Each product (see figure 1) is made up by a high pressure hose, two adapters and two hydraulic fittings.



Fig. 1- Hydraulic hoses

The production process is made up by 8 operations:

- *Preparation* : all the materials, needed for each Shop Orders, are taken from the warehouse
- *Fittings stamp* : the information required by customers are stamped on the hydraulic fittings
- *Cut* : hydraulic hoses are cut in order to obtain the right hose length
- *Hose skinning*: the external (internal) hose diameter is reduced (increased) in order guarantee an optimal junction between hose, adapters and fittings.
- *Assembly*: hoses, fittings and adapters are assembled.
- *Junction*: all the components are definitively joined
- *Test* : hydraulic hoses are opportunely tested to check the resistance to high pressures
- *Final controls and packaging*

These operations are performed in the same order in which they are described but the cutting phase and the fittings stamp operation can take place in parallel since they involve two different components not yet assembled. Further for the cutting phase, two different machines are available: manual and automatic; these machines have different setup times and working times so different levels of productivity.

3. MODELING & SIMULATION FOR THE MANUFACTURING PROCESS

This research work faces a *dynamic-stochastic* scheduling problem. It is dynamic because new S.O.s arrive during the scheduling horizon and the system allows the *passing* between jobs. Normal and priority S.O.s can enter in the system. Usually normal S.O.s are scheduled on a 2-weeks time window and each new S.O. enters in the last position of the queue. On the contrary, a priority S.O., depending on its priority level, can enter the 2-weeks queue in any position at any time- Each S.O. has a finite number m of operations, all the S.O.s entered into the system have to be necessarily completed.

The stochastic nature of the problem is due to the presence of stochastic numerical quantities. In effect set-up's time can be considered as stochastic variables each one with a specific statistical distribution. Further, during the scheduling period, some failures can occur reducing the availability of machines. In the present work failures have been modeled by using a negative exponential distribution for both the Mean Time To Failure (MTTF) and the Mean Time To Repair (MTTR), where MTTF expresses the average time between two consecutive machine failures and MTTR expresses the average time required for repairing the machine.

Once the main features of the problem and the production process have been described, the main steps of the research work can be presented. The simulation model development can be summarized as follows:

- initial analysis, data collection and distribution fitting;
- simulation model development;
- Verification, Validation and Accreditation (VV&A);
- Genetic Algorithms implementation to support Shop Order scheduling,
- simulator integration in the company management system as real time decision tool for short period production planning .

All the phases for the simulation model development are detailed in the following sections.

3.1. Initial data analysis, data collection and distribution fitting

The most important information were collected by means of interview and by using the company informative system.

Data collection is concerned with information regarding products, working methods, short period production planning and management, actual S.O.s scheduling rules, inventory management and company informative system. In particular the collected data regard: customers, production mix, bill of materials, work shifts, process times, stocks and refurbishment times, due dates, frequency of customers requiring orders, frequency of customer orders, number of S.O. for each customers, quantity of pieces for each S.O.

The most important information were collected by means of interview and by using the company informative system. In particular a key role in data collection has been played by the company informative system from which a database has been extracted. The database reports information regarding final products as: operation identifying number, worker name, Shop Order identifying number, number of pieces, operation competition date, operation competition time, drawing identifying number, hose description, adapters and fittings description. In the same database are also reported information regarding final products opportunely ranked for due date and S.O. identifying number.

All the stochastic variables have been analyzed in order to find out statistical distributions capable of fitting the empirical data with satisfactory accuracy. Figure 2 shows the histogram obtained putting in relation the time process observed for the junction operation with the frequency of occurrence. The same kind of histogram has been built for each operation which makes up the production process.

Is then possible to find out the most suitable statistical distribution.

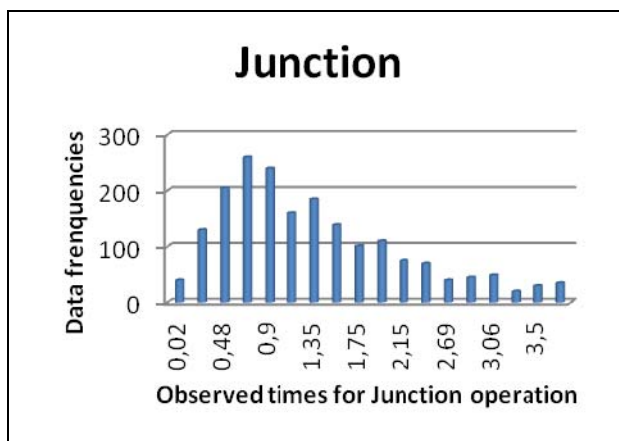


Fig 2: Histogram and Statistical Distribution of the Process Time for the Junction Operation

3.2. Simulation model development

Without doubt the most important step of a simulation study is the modeling phase. In this research work a new approach, quite different from traditional approaches, has been adopted; in the following there is a detailed description of the architecture used during the modeling phase.

The main requirement that has been taken into account was to develop a flexible and time efficient simulator. A flexible simulator is a simulator able to easily integrate new additional features over the time while a time efficient simulator is not time consuming in the execution of simulation run. So during the modeling phase we do not use the classic object oriented approach characterized by library objects and entities (that opportunely set and define the simulation model) but we propose a structural design completely based on programming code and tables to store the information.

In effect the simulator flexibility cannot be easily achieved by using library objects: each library object should represent a specific component/part of a real system but sometime such objects do not represent the real system with satisfactory accuracy. To overcome this problem a programming code must be used for the simulator development. In the present work classes and objects have been implemented by using Simple++ , a simulation language provided by eM-Plant. In this way classes can be accessed and modified at any time and also, if needed, used in other simulation models. So the use of a programming code in developing a simulation model ensures a great accuracy and offers the possibility to change it in the future according new emerging needs; as a consequence, high level of flexibility can be achieved.

Concerning the computational efficiency of the simulator and the time required for executing simulation runs, we should take into consideration how a discrete event simulation software works. In a discrete event system the state of the system changes at discrete event time points due to the flow of entities inside the system, for example at the end of an operation, at the arrival of a new shop order, etc. In other words entities with their actions change the state of the system. Usually entities are defined as classes instantiated inside the simulation model. So each entity can also have attributes in which specific information are stored. Note that the number of entities defined in a simulation model is strongly related to the computational load of a simulator: the higher is the number of entities flowing in the simulation model the higher is the computational load of the simulator. Consider that in most production processes thousands of components and products usually flow inside the system, it means thousands of entities flowing inside the simulation model and consequently a high computational load. To overcome this difficulty the approach used for developing the simulation model proposed in this paper is based on the idea to substitute the flow of entities with a flow of information opportunely stored in tables. The events generation is committed to specific objects (provided by the eM-Plant library) called event generators. Ad-hoc programmed routines manage the change of the state of the system due to the generation of an event; the information stored in the tables are updated by the programming code. By following this approach, two main advantages can be obtained: (i) a great gain in term of computational load of the simulator; (ii) reduction of the time required for executing simulation runs. Figure 3 shows an example of information stored in table for each entity (shop order) flowing into the simulator. The simulator main frame is called *model*. It contains 10 secondary frames (see figure 4).

In particular 8 frames are built to the recreate the operations described in section 2 (Preparation, Fittings stamp, Cut, Hose skinning, Assembly, Junction, Test, Final controls and packaging) whilst the remaining 2 frames are respectively:

ID CUSTOMER	ID SHOP ORDER	ID ITEM	S.O. ROUTING	QUANTITY	BIL OF MATERIALS	S.O DATE OF ENTRY
2008	2001	FG01	TABLE18	50	TABLE1	2011/05/01
5022	2002	FG02	TABLE19	70	TABLE2	2011/05/02
5895	2003	FG06	TABLE20	90	TABLE3	2011/05/03
1235	2004	FG07	TABLE21	85	TABLE4	2011/05/04
1568	2005	FG04	TABLE22	60	TABLE5	2011/05/05
5022	2006	FG03	TABLE23	12	TABLE6	2011/05/06
5022	2007	FG01	TABLE24	165	TABLE7	2011/05/07
5895	2008	FG09	TABLE25	145	TABLE8	2011/05/08
1235	2009	FG06	TABLE26	12	TABLE9	2011/05/09
2578	2010	FG08	TABLE27	123	TABLE10	2011/05/10
2578	2011	FG07	TABLE28	145	TABLE11	2011/05/11

Fig. 3: An example of information stored in table for each entity (shop order) flowing into the simulator

- the *Production Manager* (PM);
- the *Graphic User Interface* (GUI).

The PM generates the S.Os and the relative production planning, takes care of S.Os scheduling, resource allocation and inventory management. The graphic user interface allows the user to select the dispatching rule to be used for S.Os scheduling or to select S.Os scheduling based on the results of genetic algorithms.

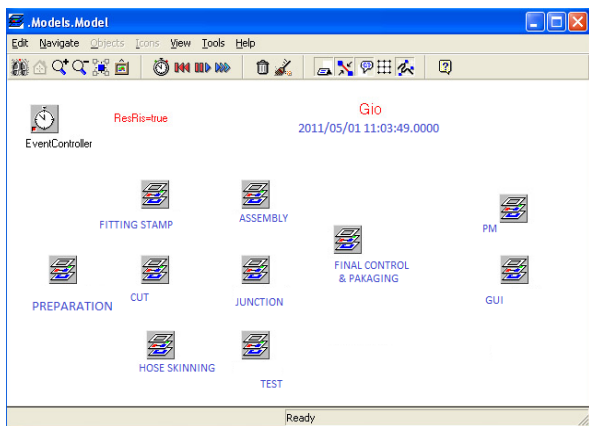


Fig. 4 – Simulator Main Frame

Furthermore the GUI provides the user with many commands as, for instance, simulation run length, start, stop and reset buttons and a Boolean control for the random number generator to reproduce the same experiment conditions in correspondence of different operative scenarios. The dispatching rules that have been implemented in order to study the Shop Orders scheduling are the Short Production Time (SPT), the Longest Production Time (LPT), Due Date (DD). Some performances indexes have been implemented in the simulation model to evaluate the S.Os scheduling:

- the average and the variance of the *Flow Time* (FT),
- the average and the variance of the *Lateness* (LT),
- the *Fill Rate* (FR).

The FT of the *i-th* S.O., as reported in equation 1, is the difference between the S.O. Completion Time (CT) and the S.O. Release Time (RT).

The LT of the *i-th* S.O. is the difference between the S.O. Completion Time and the S.O. Due Date (DD), as expressed by equation 2.

Finally the FR, as expressed by equation 3, is the percentage of S.Os meeting the due date.

$$FT_i = CT_i - RT_i \quad (1)$$

$$LT_i = CT_i - DD_i \quad (2)$$

$$FR_i = \frac{\sum_{i=1}^k S.O._i}{\sum_{i=1}^n S.O._i} \quad (3)$$

3.3. Simulation model Verification , Validation and Accreditation

In the course of a simulation study the accuracy and the quality are not guaranteed “a priori”, for this reason the verification, validation and accreditation processes have to take place to assess the goodness of the developed simulation tool (Balci 1998). Usually a conceptual model is an abstract representation of a real system; in a simulation study the conceptual model is required to build a computerized simulation model. The verification allows to verify if the translation of the conceptual model into the computerized simulation model is accurate and correct. Furthermore the simulator has to be able to reproduce the behaviour of the real system with accuracy since it will take over from the real system for the purpose of experimentation. The validation phase is devoted to assess the accuracy of the simulation model. *Accreditation* is “the official certification that a model or simulation is acceptable for use for a specific purpose.” (DoD Directive 5000.59).

For further details on simulation model Verification, Validation & Accreditation, refer to the American Department of Defence Directive 5000.59.

There are two basic approaches for testing simulation software: static testing and dynamic testing (Fairley, 1976). In static testing the computer program is analyzed to determine if it is correct by using such techniques as structured walk-throughs, correctness proofs, and examining the structure properties of the program. (Sargent, 2000). Dynamic techniques require model execution and are intended for evaluating the model based on its execution behavior.(Balci 1997)

The simulator verification has been carried out by using the Assertion Checking dynamic technique. Detailed information about this technique can be found in Adrion et al. (1982). We inserted global region and local assertion in order to check the entire model. In this way some errors, most about raw materials inventory management, were detected and corrected. The simulator validation has been carried out by using the Mean Square Pure Error analysis (MSPE). The MSPE is a typical technique devoted to find the optimal simulation run duration that guarantees the

goodness of the statistical results in output from the simulation model.

Considering the stochastic distributions implemented in the simulation model we can assert that the outputs of the simulation model are subjected to an experimental error with normal distribution, $N(0, \sigma^2)$. The best estimator of σ^2 is the mean squares error. The simulation run has to be long enough to have small values of the MSPE of the performance measures being considered. In other words, the experimental error must not “cover” the simulation results. Considering the Flow Time, we can write:

$$MSpE(t) = \sum_{h=1}^n \frac{(FT_h(t) - \overline{FT}(t))^2}{n-1} \quad (4)$$

- $FT_h(t)$, value of the Flow Time at instant of time t during the replication h ;
- $h=1, \dots, n$ number of replications.

Analogous equation can be written for the LT and the FR.

The simulation run length chosen is 200 days. Such time, evaluated with four replications, assures a negligible mean squares error for the Flow Time. The same analysis for the Lateness and the Fill Rate gives lower simulation run lengths.

The accreditation analysis has been carried out in the present work by monitoring the performance indexes (FT and LT).

The best results in terms of mean daily flow time can be obtained using the Longest Production times scheduling rule (LPT). Taking into account that the model proposed in theory is too simplified, this result can be completely accepted even if it is in contrast with theory. Concerning the impact of the different scheduling rules on the mean daily lateness, the difference between the scheduling rules is not so remarkable.

3.4. Genetic Algorithms implementation to support Shop Order scheduling

The modeling architecture has been opportunely programmed to be interfaced with genetic algorithms. So once tested the validity of the simulation model, further implementations were carried out to introduce Genetic Algorithms (GA) as support tool for short period production planning. The use of genetic algorithms goes through three fundamental steps:

- initial S.Os scheduling (proposed by the user);
- setting of genetic operators and algorithms initialization
- optimization.

The GA was implemented as a functional part of a particular tool called optimizer. This object aims at: optimising S.Os scheduling by means of GA, testing the proposed scheduling, monitoring the manufacturing system performances by using the Flow Time, the Lateness and the Fill Rate indexes. In the following part the problem (which has to be solved) and the optimizer have been described. Simulation tool is not the only way to solve stochastic shop orders scheduling

problems. A simulation tool allows to monitor the system performances under different S.Os scheduling but an optimization algorithm is required to improve the S.Os scheduling .

Interfacing the optimization algorithm with the simulation model it is possible to find out the most suitable solution (evaluated optimizing the scalar function chosen to measure scheduling goodness).

The interface between the simulation model and genetic algorithms was created through specific sub-routines written using the simulation language Simple++. In this way the optimization algorithms and the simulation model jointly work for the scheduling problem resolution: the former finds out acceptable solutions while the latter validates and chooses the best solutions.

4. SIMULATION RESULTS AND ANALYSIS

The research work faces the Shop Orders scheduling problem into a real manufacturing system devoted to hydraulic hoses production. The proposed approach is based on the use of Modelling & Simulation jointly used with dispatching rules and genetic algorithms. The system performances, under different dispatching rules have been tested, as well as the guidelines obtained by using genetic algorithms.

The scheduling rules (implemented in the simulator) being tested in the following analysis are:

- the Shortest Production Time (SPT);
- the Due Date (DD);
- the Longest Production Time (LPT).

The average values of the FT, LT and FR in correspondence of each scheduling rule are shown in Table 1. As it can be seen in table 1 the SPT rule guarantees the best performances in terms of Flow Time, while the DD rule allows to get the best performance in terms of lateness and Fill Rate. Table 2 reports the standard deviation values for each performance measure in correspondence of each scheduling rule.

	SPT	DD	LPT
Flow Time [days] (FT)	3,600	4,580	5,590
Lateness [days] (LT)	1,500	1,090	2,370
Fill Rate [%] (FR)	78,640	79,250	73,780

Table 1: Average values of the Performance Measures

	SPT	DD	LPT
Flow Time [days] (FT)	0,031	0,039	0,035
Lateness [days] (LT)	0,028	0,031	0,036
Fill Rate [%] (FR)	0,21	0,17	0,19

Table 2: Standard deviation of the Performance Measures for each Scheduling Rule

The S.Os scheduling has also been investigated by using genetic algorithms trying to minimize the FT, minimize the LT and maximize the FR. Table 3 reports the simulated FT in correspondence of each generation; for each generation are reported the best, the average and the worst FT values . After 25 replications the best, the average and the worst solutions converge to the value of 3.20 days. Note that such value is lower than best result obtained with the SPT rule (the improvement is about 9.17%). The optimization on the FT with genetic algorithms is also shown in the figure 5.

Generation	FT Best	FT Average	FT Worst
1	8,76	9,54	9,82
2	7,00	7,76	8,43
3	6,12	6,85	7,98
4	5,91	6,60	7,83
5	5,23	6,09	7,65
6	4,76	5,73	7,60
7	4,95	5,69	6,85
8	4,63	5,53	6,82
9	4,48	5,01	6,12
10	4,32	4,99	5,96
11	4,30	4,80	5,58
12	4,10	4,73	5,51
13	4,02	4,49	5,20
14	3,73	4,17	4,55
15	3,64	3,96	4,17
16	3,64	3,92	4,25
17	3,48	3,87	4,00
18	3,48	3,29	3,38
19	3,35	3,29	3,29
20	3,27	3,29	3,29
21	3,27	3,20	3,20
22	3,27	3,20	3,20
23	3,27	3,20	3,20
24	3,27	3,20	3,20
25	3,20	3,20	3,20

Table 3: Best, Average and Worst values of Flow Time obtained by GA

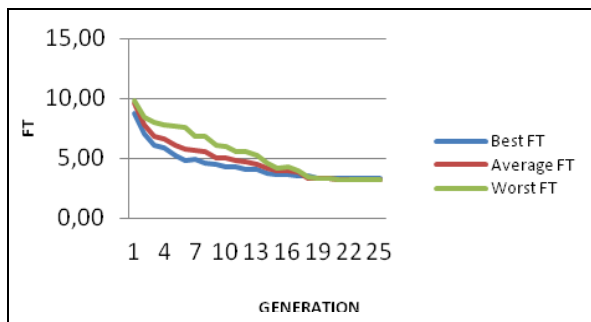


Fig 5: FT Optimization with Genetic Algorithms

The same approach has been applied for the Lateness optimization. The main results have been reported in table 4 and in figure 6. As in the previous case after 25 replications the best, the average and the worst solutions converge to 0.92 days. This value is still better than the result obtained with the DD dispatching rule, the improvement is about 16 % . Finally the table 5 and the figure 7 reports the optimization results for the FR. In this last case after 25 generations the algorithm converges with an improvement of about 1,4%.

Generation	LT Best	LT Average	LT Worst
1	3,85	4,25	5,61
2	3,78	4,03	5,49
3	3,62	3,95	4,99
4	3,46	3,82	4,67
5	3,19	3,67	4,28
6	2,92	3,46	4,05
7	2,68	3,24	3,91
8	2,20	2,97	3,72
9	2,05	2,86	3,35
10	1,90	2,51	3,27
11	1,79	2,12	3,16
12	1,65	2,07	3,06
13	1,49	2,00	2,94
14	1,26	1,83	2,29
15	1,19	1,64	2,03
16	1,13	1,48	1,82
17	1,09	1,29	1,76
18	1,02	1,14	1,58
19	0,99	1,03	1,32
20	0,95	1,00	1,20
21	0,92	0,99	1,14
22	0,92	0,95	0,99
23	0,92	0,92	0,92
24	0,92	0,92	0,92
25	0,92	0,92	0,92

Table 4: Best, Average and Worst values of Lateness obtained by GA

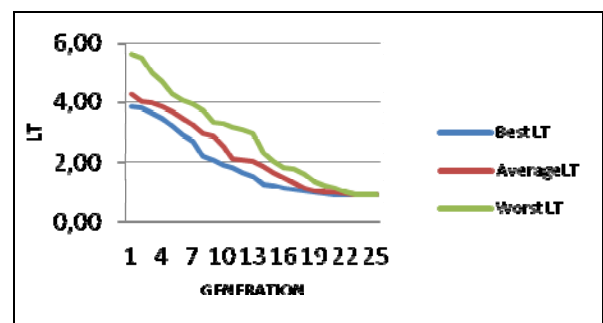


Fig. 6: LT Optimization with Genetic Algorithms

Generation	FR		
	Best	Average	FR Worst
1	82,52	80,20	77,99
2	83,70	81,62	79,64
3	84,45	81,37	78,40
4	85,43	81,92	78,52
5	86,21	83,20	80,30
6	87,04	84,32	81,71
7	88,08	85,43	82,89
8	89,39	86,44	83,60
9	90,38	87,66	85,04
10	91,31	87,81	84,42
11	92,16	89,63	87,21
12	92,91	90,31	87,82
13	93,48	91,15	88,93
14	94,21	92,01	89,91
15	94,75	92,00	89,35
16	95,09	93,03	91,07
17	95,43	94,35	93,38
18	95,75	95,01	94,37
19	96,09	95,39	94,79
20	96,24	95,65	95,16
21	96,34	95,87	95,51
22	96,38	96,14	96,00
23	96,42	96,36	96,36
24	96,42	96,42	96,42
25	96,42	96,42	96,42

Table 5: Best, Average and Worst values of Fill Rate obtained by GA

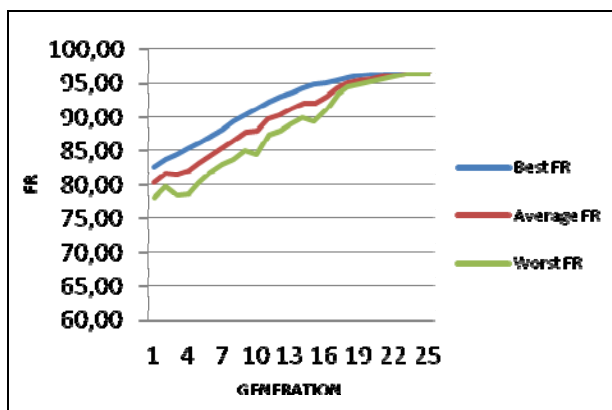


Fig. 7: Fill Rate Optimization with GA

5. CONCLUSIONS

The authors implemented a discrete simulation model by using an advanced modeling approach. The simulation model was developed with the purpose to study the behaviour of different dispatching rules and the potential of genetic algorithms for the S.Os scheduling within a manufacturing system devoted to produce hydraulic hoses. The simulator architecture is totally different from the traditional modeling approach proposed by the commercial discrete event simulation packages. Such architecture is completely based on a different modeling approach :

- all the objects have been modeled by means of code
- all the information have been stored in table;

These features allow to get high flexibility in terms of future changes and new tools implementation (for example genetic algorithms or neural network).

The behavior of three different scheduling rules was analyzed in terms of Flow Time, Lateness and Fill Rate. In addition, Genetic algorithms were also used to perform three different optimizations: FT, LT and FR . Comparing the results obtained in these two steps of the research work it was found out that the genetic algorithms are capable of finding better shop orders scheduling improving the results obtained by using the classical scheduling rule. Further, thanks to the high computational efficiency, the simulator has the potentials to be used real-time for short period production planning.

In conclusion, the approach proposed in this case study during the modeling phase has been useful for creating a decision and problem solving tool that can be profitably used by the integration in the company informative system and used real-time to support stochastic S.O.s scheduling.

REFERENCES

- Adrion, W.R., Branstad, M.A., & Cherniavsky, J.C., *Validation, verification, and testing of computer software*, Computing Surveys, 14 (2), pp. 159–192, 1982.
- Andersson M, Ng AHC, Grimm H. , Proceedings of 2008 winter simulation conference; 2008. p.2004–11.
- Balci O. *Verification, Validation And Accreditation Of Simulation Models*. Proceedings of the 1997 Winter Simulation Conference ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson
- Balci, O., *Verification, Validation and Testing*. In Handbook of Simulation, edited by J. Banks, pp. 335-393, New York: Wiley Interscience, 1998
- Bierwirth C., *A generalized permutation approach to job shop scheduling with genetic algorithms*. OR Spektrum 17(213):87–92, 1995
- Carri A., *Simulation of Manufacturing Systems* (Wiley, New York, 1986)
- Castilla i, Longo F. (2010). *Modelling and Simulation Methodologies, Techniques and Applications: a*

- State of the Art Overview. INTERNATIONAL JOURNAL OF SIMULATION & PROCESS MODELLING, vol. 6(1); p. 1-6, ISSN: 1740-2123
- Cimino A., Longo F., Mirabelli G., Papoff E. (2008). Shop orders scheduling: dispatching rules and genetic algorithms based approaches. In: Proceedings of the European Modeling & Simulation Symposium. Campora S. Giovanni (CS), Italy, 17-19 September, vol. I, p. 817-823, ISBN/ISSN: 978-88-903724-0-7.
- Curcio D, Longo F. (2009). Inventory and Internal Logistics Management as Critical Factors Affecting the Supply Chain Performances. International Journal of Simulation & Process Modelling, vol. 5(4); p. 278-288, ISSN: 1740-2123
- Dorndorf U., Pesch E., *Combining genetic and local search for solving the job shop scheduling problem.* APMOD93 Proc, pp 142–149,1993
- Fairley, R. E., *Dynamic testing of simulation software,* Proc. of the 1976 Summer Computer Simulation Conf., Washington, D.C., 40–46.
- Ferrolho A., Crisóstomo M., *Optimization of Genetic Operators for Scheduling Problems,* Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.11, No.9 pp. 1092-1098, 2007
- Frantzen M., Amos H. C. Ng, P. Moore, *A simulation-based scheduling system for real-time optimization and decision making,* Robotics and Computer-Integrated Manufacturing 27(2011)696–705
- Garey M.R, Johnson D.S., Sethi R., *The complexity of flowshop and jobshop scheduling,* Mathematics of Operations Research 1 (1976) 117}129.
- Ghedjati F., *Genetic algorithms for the job-shop scheduling problem with unrelated parallel constraints: heuristic mixing method machines and precedence.* Comput Ind Eng 37(1-2):39–42, 1999
- Goyal S. K., Mehta K., Kodkodali R., Deshmukh S.G., *Simulation for analysis of scheduling rules for a flexible manufacturing system,* Integrated Manufacturing Systems[ING], 6(5), pp. 21–26, 1995.
- Haibin Y, Wei L., *Neural network and genetic algorithm-based hybrid approach to expanded job-shop scheduling.* Comput Ind Eng 39:337–356, 2001
- Hicks C., Pupong P., *Dispatching rules for production scheduling in the capital goods industry,* Production Economics 104 (2006) 154–163
- Holthaus O., *Design of efficient job shop scheduling rules,* Computers and Industrial Engineering [CIE], 33(1,2), pp. 249– 252, 1997.
- Holthaus O., Rajendran C. Efficient dispatching rules for scheduling in a job shop, Production Economics 48 (997) 87-105
- Hou, E.S.H., Li, H.-Y., *Task scheduling for flexible manufacturing systems based on genetic algorithms,* IEEE International Conference on Systems, Man, and Cybernetics, 1991. 'Decision Aiding for Complex Systems, Conference Proceedings.,vol-1,pp- 397 – 402, 1991
- Huq F., Huq Z., *The sensitivity of rule combinations for scheduling in a hybrid job shop,* International Journal of Operations and Production Management [IJO], 15(3), pp. 59–75, 1995.
- Jeong S. J., Lim S. J., Kim K. S., *Hybrid approach to production scheduling using genetic algorithm and simulation,* Adv Manuf Technol (2006) 28: 129 136
- Johtela T., Smed J., Johnsson M., Lehtinen R.and O. Nevailainen O., *Supporting production planning by production process simulation,* Computer Integrated Manufacturing Systems 10 (1997) 193}203
- Kim I, Watada J, Shigaki T., *A comparison of dispatching rules and genetic algorithms for job shop schedules of standard hydraulic cylinders.* Soft Computing 2007;12:121–8.
- Kiran AS., *Simulation and scheduling.* Banks J, editor. Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice. New York: John Wiley and Sons, Inc; 1998. p. 677–717.
- Liu K. C., *Dispatching Rules For Stochastic Finite Capacity Scheduling,* Computers ind. Engng Vol. 35, Nos 1-2, pp. 113-116, 1998
- Longo F. (2010). Design And Integration Of The Containers Inspection Activities In The Container Terminal Operations. International Journal of Production Economics, vol. 125(2); p. 272-283, ISSN: 0925-5273, doi: 10.1016/j.ijpe.2010.01.026
- Longo F., Mirabelli G. (2008). An Advanced Supply Chain Management Tool Based on Modeling & Simulation. Computers & Industrial Engineering, vol. 54/3; p. 570-588, ISSN: 0360-8352, doi: 10.1016/j.cie.2007.09.008
- Longo F., G. Mirabelli, E. Papoff (2006). Modeling, Analysis & Simulation of Tubes Manufacturing Process and Industrial Operations Controls. In: Proceedings of the Summer Computer Simulation Conference. July 30th – August 03rd 2006, Calgary Canada, SAN DIEGO, CA: SCS, vol. 38, p. 54-59
- Parthanadea P., Buddhakulsomsirib J., *Simulation modeling and analysis for production scheduling using real-time dispatching rules: A case study in canned fruit industry,* Computers and Electronics in Agriculture 70 (2010) 245–255
- Pinedo M., *Scheduling: theory, algorithms, and systems,* 3rd Ed..New York, NY: PrenticeHall;2008.
- Rabelo L. et al., *Intelligent Scheduling For Flexible Manufacturing Systems,* 1993 IEEE International Conference on Robotics and Automation. Proceedings., vol.3, pp-810 – 815
- Ramasesh R., *Dynamic job shop scheduling: A survey of simulation research,* Omega, Volume 18, Issue 1, 1990, Pages 43-57
- Riane F., Artiba A., Iassinovski S. *An integrated production planning and scheduling system for*

- hybrid flowshop organizations*, Production Economics 74 (2001) 33}48
- Sakawa M, Mori T., *An efficient genetic algorithm for job-shop scheduling problems with fuzzy processing time and fuzzy due date*. Comput Ind Eng 36(2):325–341, 1999
- Sankar, S. S., Ponnambalam, S. G., Rajendram C., *A multiobjective genetic algorithm for scheduling a flexible manufacturing system*. International Journal in Advanced Manufacturing Technologies 22:229–236, 2003
- Sargent R. G. *Verification, Validation, And Accreditation Of Simulation Models*. Proceedings of the 2000 Winter Simulation Conference J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds.
- Smith S.F., *Knowledge-based production management: Approaches, results and prospects*, Production Planning & Control 3 (4) (1992) 350}380.
- Son Y. J., Rodríguez-Rivera H., and Wysk R. A., *A multi-pass simulation-based, real-time Scheduling and shop floor control system*, Transactions of the Society for Computer Simulation International - modeling and simulation in manufacturing Volume 16 Issue 4, Dec.1.(1999)
- Syswerda G., *Schedule optimization using genetic algorithms*. In: Davis L (ed) Handbook of genetic algorithms. Van Nostrand Reinhold, New York, pp 332–449, 1991
- Vinod V., Sridharan R., *Simulation modeling and analysis of due-date assignment methods and scheduling decision rules in a dynamic job shop production system*, Production Economics 129 (2011) 127–146
- Yamada T, Nakano R., *A genetic algorithm applicable to large scale job shop problems*, vol 2. Parallel problem solving from nature. North Holland Publishers, Amsterdam, pp 281–290, 1992
- Yun Y. S., *Genetic algorithm with fuzzy logic controller for preemptive and non-preemptive job-shop scheduling problems*. Comput Ind Eng 43:623–644, 2002

Authors' Index

- Acebes* 211
Adegoke 675
Affenzeller 448, 454
Aguilar 410
Aguirre 384
Ahmed 289
Aizstrauts 62
Altmann 272
Aquilar Chinaea 62
Arhip 380
Ávila 654
Azadeh 562
Azimi 17
Babarada 371, 380
Backfrieder 100, 111
Bagnaro 44
Banks 740
Bao 105
Barbey 25
Battista 325
Belmabrouk 375
Bendella 375
Benhari 220
Benkhedda 375
Bizub 78
Blanco 619
Blanco Fernandez 635
Blank 433
Bocca 755
Bogdan 722
Bohlmann 747
Bossomaier 314
Boudraa B. 460
Boudraa M. 460
Bracale 44
Brandstätter-Müller 488
Brandt 78
Bruzzone 755, 768, 775, 782
Bubevski 555
Burmeister 87
Cafaro 384
Callero 410
Capocchi 351
Carneros 72
Carraro 44, 465
Cartlidge 299
Castellanos 642, 654
Castro 384
Cereda 768
Cesarotti 423
Cetinkaya 709
Ceylan 278
Chaczko 703
Chai 508, 715
Choi 308
Chouhal 11
Coates 470
Codetta-Raiteri 545
Crespo Pereira 400, 626
Crouch 470
Dabrowski 659
D'Ambrosio 696
Davendra 593
De Nicola 201
De Prada 211
del Rio Vilas 400, 626
Dello Stritto 325
Di Gregorio 696
Diaz 782
Dorfer 433
Dowman 363
Drejja 62
Dreiseitl 176
Drouin 363
Dupont 241
EL Boukili 144
Eldemir 535
Elizondo Cortes 187
Faschang 433
Federici 351
Ferguson 465
Fernandez de Miguel 635
Fiebig 181
Filippone 696
Fischer 448
Flores 150
Fowler 264, 363
Franz 195
Friðgeirsson 44, 50, 87
García 648, 654
Gardeshi 17

Gargiulo 44, 50, 87, 465
Ghaemmohamadi 562
Ginters 62
Giordano 325
Giuiusa 423
Göbel 137
Gray 264
Griffin 264
Grundspenkis 62
Gunal 278
Guntinas-Lichius 87
Hamad 229
Hamadou 675
Han 308
Harrè 314
Haueisen 50
Hawe 470
Helgason B. 465
Helgason T. 44, 465
Helm 272
Hernandez-Romero 690
Hoare 333
Hoess 478
Hözlwimmer 488
Huerta Barrientos 187
Hunt 659
Iannone 325
Ighoroje 669
Ingvarsson 44, 465
Introna 423
Jacak 345, 454
Jasek 593
Ji 340
Jimenez-Macias 580, 587, 613
Kang 308
Karakaya 535
Kebabla 5
Kern T. 433
Kern H. 44, 465
Kesserwan 516
Khadem Geraili 220
Kim 319
Klancar 118
Klauke 747
Klompous 703
Klinger 747
Klingner 87
Kolstad 247
Kommenda 454
Kopp 264
Krieg 478
Kronberger 448, 454
Krzesinski 137
Kulczycki 488
Lamas Rodriguez 400
Latorre-Biel 580, 587, 613
Lauberte 62
Lazarova-Molnar 516, 526
Lecca 36
Lee D. K. 308
Lee J.O. 541
Lee K.S. 308
Lee T.E. 319
Lee Y.J. 308
Legato 93
León Samaniego 605
Leventhal 363
Li B.H. 508, 715
Li Q. 502
Lin 508
Lipovszki 394
Lirk 488
Liu 340
Longo 775, 782
López 72
Lowenstein 363
Lulu 127
Luo 105, 502
Madeo 768, 775
Mahdaoui 11
Maiga 669
Maio 496
Mandl 44
Manzo 201
Martínez Camara 605
Massei 755, 775
Mayr H 195
Mayr W. 44, 465
Mazaeda 211
Mazza 93
Medina-Marin 690
Mendez 384
Merazi Meksen 460
Merino 211

Merkuryev 62
Merzouki 241
Mizouni 516
Molnar 394
Mondal 363
Morvan 241
Mouss N. 5
Mouss L. H. 5, 11
Mouss M.D. 11
Mujica 357, 734
Music 118, 681
Neumann J. 478
Neumann G. 568
Nicoletti 782
Niculiu M. 254
Niculiu T. 254
Nieto de Almeida 626
Nikodem J. 703
Nikodem M. 703
Nissen 235
Novak 768
Novelli 599
Novitsky 62
Núñez 72
Onggo 333
Oplatkova 593
Or 439
Orsini 264
Osl 176
Özbaş 439
Özlem 439
Ozmen 55
Page 137
Pan 340
Parsapour 488
Perea 150
Pérez de la Parte 605, 619, 635
Pérez V. 654
Perez-Lechuga 690
Perna 696
Petrovic 722
Petz 433
Pfeifer 272
Piera 357, 734
Popa 1
Proell 345
Ramanan 283
Ramon 50
Rarità 201
Ravariu 371, 380
Rego Monteil 400, 626
Ren 105, 502
Reynisson 465
Rios Prado 400, 626
Rodriguez A. 211
Rodriguez D. R. 619
Rongo 696
Rust 709
Sabuncuoglu 283
Sáenz-Díez Muro 605
Saetta 166
Saft 235
Sakne 62
Schiraldi 325
Schoenberg 78
Schuler 272
Seck 709
Seck-Tuoh-Mora 690
Senkerik 593
Seo 541
Shah 289
Shen 340
Shibata 728
Silva 496
Simon-Marmolejo 690
Sindicic 722
Sodja 574
Sokolowski 740
Somolinos 72
Soyez 241
Spataro 696
Spingola 696
Sriram 299
Stekel 454
Strasser 272
Suk 541
Szczerbicka 747
Tao 105
Tarone 775
Tarvid 158
Tiacchi 166
Togo 675
Toma 351
Traoré 669, 675

Tremori 775
Uangpairoj 728
Ulbrich 478
Ulgen 30
Unnpórsson 465
Usher 247
Viamonte 496
Volk 87
Wagner 448, 454
Wellens 642, 648, 654
Wenzler 709
Williams 30
Wilson 470
Winkler 433, 448, 454
Woda 703
XU 418
Yang 715
Yilmaz 55
Zelinka 593
Zhang L. 105, 502
Zhang Y. 502
Zhongjie 127
Zottolo 30
Zupancic 574
Zwettler 100, 111