3rd INTERNATIONAL INDUSTRIAL SIMULATION CONFERENCE 2005

ISC'2005

EDITED BY

Jörg Krüger

Alexei Lisounkin

Gerhard Schreck

JUNE 9-11, 2005

BERLIN, GERMANY

A Publication of EUROSIS-ETI

Printed in Ghent, Belgium

3rd Industrial Simulation Conference 2005

BERLIN, GERMANY
JUNE 9-11, 2005

Organised by:
EUROSIS, The European Simulation Society
ETI

Co-Sponsored by

AEKI

CREAX

ENSAIT

Fraunhofer IPK

Ghent University

UPV

Hosted by

Fraunhofer-IPK
Berlin, Germany

EXECUTIVE EDITOR

PHILIPPE GERIL
(BELGIUM)

EDITORS

General Conference Chair

Jörg Krüger, Fraunhofer IPK, Berlin, Germany

General Program Chairs

Alexei Lisounkin, Fraunhofer IPK, Berlin, Germany
Gerhard Schreck, Fraunhofer IPK, Berlin, Germany

Track Chair Workshop om Modelling and Simulation in the Textile Industry

Vladan Koncar
GEMTEX-ENSAIT
Roubaix, France

INTERNATIONAL PROGRAMME COMMITTEE

Simulation Methodology
Carmen Bobeanu, Ghent University, Ghent, Belgium
Jan Lemeire, VUB, Brussels, Belgium
Beniamo di Martino, Seconda Universita di Napoli, Naples, Italy
Henrikas Pranevitchius, TU Kaunas, Kaunas, Lithuania
Georgios Theodoropoulos, University of Birmingham, Birmingham, United Kingdom
Hans Vangheluwe, McGill University, Montreal, Canada
Danny Van Welden, Oudenaarde, Belgium
Ghislan Vansteenkiste, Ghent University, Ghent, Belgium

Simulation in Manufacturing
Pascal Beruet, Universite Bretagne Sud, Lorient, France
Antoneta Iuliana Bratcu, Ecole Nationale Superieure des Mines de St.Etienne, France
Luis Camarinho-Matos, Universidade Nova de Lisboa, Lisbon, Portugal
Remy Dupas, Universite d'Artois, Bethune, France
Michel Gourgand, Universite Blaise Pascal, Clermont-Ferrand, France
Bernard Grabot, LGP-ENIT, Tarbes, France
Peter Lawrence, Monash University, Clayton, Australia
Marie-Ange Manier, UTBM, Belfort, France
Pascal Meyer, Forschungszentrum Karlsruhe, Karlsruhe, Germany
Harvey Millar, Saint Mary's University, Halifax, Canada
Henri Pierreval, IFMA, Aubiere Cedex, France

Simulation in Automotive Systems
Samarn Chantavarapan, University of Michigan, Dearborn, USA
Habtom Mebrahtu, APU, Chelmsford, United Kingdom
Matthias Reuter, CUTEC Gmbh, TU-Clausthal, Clausthal, Germany
Caroline Strange, APU, Chelmsford, United Kingdom
Steffen Strassburger, DaimlerChrysler AG, Germany
Rob Walker, APU, Chelmsford, United Kingdom
Edward Williams, University of Michigan, Dearborn, USA

Simulation in Robotics
Track Chair: Rolf Bernhard, Fraunhofer IPK, Berlin, Germany
Leopoldo Arnesto Angel, UPV, Valencia, Spain
George L.Kovacs, Hungarian Academy of Sciences, Budapest, Hungary
Martin Mellado, UPV, Valencia, Spain
Bogdan Raducanu, Computer Vision Centre, UAB, Barcelona, Spain

Simulation in Electronics, Computer and Telecommunications
Marwan Al-Akaidi, de Montfort University, Leicester, United Kingdom
Juan C.Guerri, UPV, Valencia, Spain
Clemens Heitzinger, Arizona State University, Tempe, USA
Ana Pajares, UPV, Valencia, Spain
Carlos Palau, UPV, Valencia, Spain
Renate Sitte, Griffith University, Gold Coast, Australia

Simulation in Logistics, Transport and Harbour Simulation
El-Houssaine Aghezzaf, Ghent University, Ghent, Belgium
Jaap Ottjes, TU Delft, Delft, The Netherlands
Rik van Landeghem, Ghent University, Ghent, Belgium
Hans Veeke, TU Delft, Delft, The Netherlands

Complex Systems Modelling
Alessandro Genco, Universita di Palermo, Palermo, Italy
Marielle Mokhtari, Defence R&D Canada, Valcantier, Canada
Alexandre Nketsa, Universite Paul Sabatier, Toulouse, France

Simulation in Aerospace
Martin Adelantado, ONERA-CT, Toulouse, France
Jorge Perez Mateos, Industria de Turbopropulsores SA, Madrid, Spain
Barbara Pralio, Politecnico di Torino, Turin, Italy
Martin Spieck, DLR, Goettingen, Germany
Tao Wan, University of Bradford, Bradford, United Kingdom

Simulation in Industrial and Product Design
Jose Luis Arjuria, European Virtual Engineering, Vitoria, Spain
Chiara Catalano, IMATI-CNR, Genoa, Italy
A. Chatzinikolaou, Athens, Greece
Yan Luo, NIST, Gaithersburg, USA

Simulation in Engineering Processes
Chrissanti Angeli, Technological Institute of Piraeus, Athens, Greece
Jose Barata, Universidade Nova de Lisboa, Lisbon, Portugal
Zdzislaw Bubnicki, Wroclaw University of Technology, Wroclaw, Poland

Simulation in Energy and Power Systems
Janos-Sebestyen Janosy, KFKI, Atomic Energy Research Institute, Budapest, Hungary
Mohammad Montakhab, University of West England, Bristol, United Kingdom

Simulation in Multibody Systems
Bernd Schaefer, DLR, Wessling, Germany

Simulation in Chemical, Petroleum and Mining Engineering
Klaus Hellgardti, Loughborough University, Loughborough, United Kingdom
Vahid Nassehi, Loughborough University, Loughborough, United Kingdom

Simulation in Military and Defense
Aleksander Kurec, US Army TACOM, Warren, USA
Mike Letherwood, US Army RDECOM, Warren, USA

Virtual Reality and Graphical Simulation in Industrial Simulation
Peter Astheimer, I-CAVE, University of Abertay Dundee, United Kingdom
Bedrich Benes, ITESM, Mexico City, Mexico
Anders Hast, University of Gavle, Gavle, Sweden
Fabrizio Lamberti, Politecnico di Torino, Turin, Italy
S.P. Mudur, Concordia University, Montreal, Canada
Bengt Mueck, Heinz Nixdorf Institute, University of Paderborn, Paderborn, Germany
Andrea Sanna, Politecnico di Torino, Turin, Italy
Mike Sips, University of Konstanz, Konstanz, Germany

Virtual Automation Systems
Cornelius Willnow, Fraunhofer IPK, Berlin, Germany

Verification, Validation and Accreditation
Track Chair: Alexei Lisounkin, Fraunhofer IPK, Berlin, Germany
Mike Rennie, Deimos Space SA, Tres Cantos, Madrid, Spain
Roberto Revetria, University of Genoa, Genoa, Italy
Andre Thomas, CRAN, Vandoeuvre-les-Nancy, France

Simulation and Training
Gerhard Schreck, Fraunhofer IPK, Berlin, Germany

Workshops

Workshop on Modelling and Simulation in the Textile Industry
Vladan Koncar, ENSAIT, Roubaix, France

Workshop on Intelligent Transport Systems
Andrzej Adamski, Autocom, Poland
Ana Almeida, Universidade Nova de Lisboa, Lisbon, Portugal
Paul Davidsson, Blekinge Institute of Technology, Ronneby, Sweden
Ruili Wang, Massey University, New Zealand

NANOSIM
Javier Marin, ETSI, University of Malaga, Malaga, Spain

INDUSTRIAL SIMULATION 2005

© 2005 EUROSIS-ETI

Responsibility for the accuracy of all statements in each peer-referenced paper rests solely with the author(s). Statements are not necessarily representative of nor endorsed by the European Simulation Society. Permission is granted to photocopy portions of the publication for personal use and for the use of students providing credit is given to the conference 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 and the conference.

All author contact information provided in this Proceedings falls under the European Privacy Law and may not be used in any form, written or electronic, without the written permission of the author and the publisher.

For permission to publish a complete paper write EUROSIS, c/o Philippe Geril, ETI Executive Director, Ghent University, Faculty of Engineering, Dept. of Industrial Management, Technologiepark 903, Campus Ardoyen, B-9052 Ghent-Zwijnaarde, Belgium.

EUROSIS is a Division of ETI Bvba, The European Technology Institute, Torhoutsesteenweg 162, Box 4, B-8400 Ostend, Belgium

EUROSIS-ETI Publication
ISBN: 90-77381-18-X

Preface

Modelling and simulation techniques have found extensive application in every industrial sector, as well as in administration, research, and human and social services. The use of this approach cannot be overestimated. Systematically exploring the world we live in, we are able to evaluate, supervise, and control target processes as structured systems with known and predictable interdependencies. A basic understanding of human environments, human activity, and human beings themselves can be set upon a formal background by means of modelling and simulation.

The increase in complexity and artificiality of our environment has occurred in step with our expanding knowledge of nature, production and social processes. Modelling and simulation methods and practises are constantly under development and improvement. In this context, the Industrial Simulation Conference 2005 (ISC-2005) is the third annual International Simulation Conference, the aim of which is to provide a forum for industrial simulation related research and a status report on advances to the present day for a broad range of researchers as well as industry.

The contributions focus on basic areas of modelling and simulation methodology, techniques, and tools, and cover a wide range of applications: robotics and automotive industry, factory and production management, traffic and supply, energy and electronic networks, chemical and process industry, textile industry, etc. We are glad that the contributions originate from many different areas of the world, from Europe, Asia, Australia, and America.

We at the Fraunhofer Institute for Production Systems and Design Technology are excited to welcome scientists and experts from all countries and continents to Berlin. Our institute possesses a high level of expertise related to the use of modelling and simulation technology in industrial design and production. With this in mind, we would like to offer experts an opportunity to exchange their experience and latest research findings in a fruitful environment.

Berlin is a wonderful city for science and culture. We decided to organise this event during the well-known "Lange Nacht der Wissenschaften" (The Long Night of the Sciences) which takes place on Saturday, June 11. Research institutions and universities of the city of Berlin offer citizens and guests a chance to get to know the scientific traditions and cutting edge research.

On behalf of the International Programme Committee of ISC-2005, we would like to welcome you to Fraunhofer IPK. We are certain that this conference will be an event of inspiration, new ideas and contacts for you.

Jörg Krüger, Alexei Lisounkin, Gerhard Schreck
Conference Chairs

CONTENTS

Preface ..... **IX**
Scientific Programme **1**
Author Listing **443**

SIMULATION ESTIMATION, VALIDATION AND OPTIMIZATION

Efficient Truncation Point Estimation for Arbitrary Performance Measures
Mirko Eickhoff, Donald C. McNickle and Krzysztof Pawlikowski **5**

Deploying Validation Tests in Systems Development
Juan Garbajosa and Pedro P. Alarcón **13**

SOM-Validation Method for Wastewater Treatment
Hilario López and Iván Machón **17**

Simulation based Framework for Maintenance Optimization
Thibaut Lust, Olivier Roux, Fouad Riane and Pierre Dehombreux **23**

Computing Simulation and Heuristic Optimization of the Marine Twin Shaft Gas Turbine
Josko Dvornik, Ante Munitic and Enco Tireli **28**

NEURAL NETWORK AND GENETIC ALGORITHM SIMULATION

Corti's Organ Physiologist-based Cochlear Model: Neural lateral Inhibition
Francisco Ríos, Raquel Fernández-Ramos, Jorge Romero-Sánchez,
Jose Francisco Martín, A.Ignacio Molina-Conde and Francisco Javier Marín .. **33**

Rule Representation and Initial Population in Genetic Learning
I. Bruha **37**

VISUAL SIMULATIONS

A Simplified Aerodynamic Model for Flight Simulation
H. Chen, T.R. Wan and R.A. Earnshaw **45**

3D Visualisation of Musical Content
Jacek Grekow **50**

CONTENTS

A Knowledge Representation System to support Medical Diagnostic Imaging
Edoardo Ardizzone and Daniele Peri **53**

Dynamic Generation of Hypermedia Presentations from Distributed Digital Archives
Chia-Hung Lin, Chia-Hung Yeh, Bai-Hsuen Chen and Jen-Shin Hong **58**

SIMULATION IN ROBOTICS

Using „Complete Kinematical Structure" in a Faster Method for Robot Kinematic Modelling
Viorel Stoian **63**

Simulation of mobile Robot Applications with *VirtualRobot*
Martin Mellado, Carlos Correcher, Rosa Mª de Molina, José Camacho and Ginés Benet **68**

Pathfinding with Kohonen Maps
Matthias Reuter and Steffen Harneit **73**

Extraction of System Parameters and Their Importance for Controlling
Matthias Reuter and Sabine Bostelmann **78**

ELECTRONICS PERFORMANCE ANALYSIS AND OPTIMIZATION

Clustering Procedure for Performance Analysis of Mobile Communication Networks
Dimitar Radev, Izabella Lokshina and Svetla Radeva **83**

A Measurement Based Simulation Model of a Web Cluster
Isabel Wagner, Kai-Steffen Hielscher and Reinhard German **88**

NAPT- A Simulation Tool for Network Analysis and Parameter Tuning
Mario Barbera, Andrea G. Busà, Alessandro G. Di Nuovo, Alfio Lombardo and Giovanni Schembra **93**

Optimizers with learning Distributions in Simulation: An Application All-Pass Disperser Compensator of Optic Filters
Mufit Ozden and Chi-Hao Cheng **101**

MANAGEMENT SIMULATION TOOLS

Blended Learning Applying Interactive Discrete Event Simulation in a Web Based Learning Management System
Wolfgang Kühn **109**

PLANSIM, A framework to build Automated Planners with Simulators
Ruy L. Milidiu and Frederico dos Santos Liporace **115**

A Virtual Real-Time Model for Control Software Development –Applied on a Sheet Metal Press Line
Bo Svensson, Fredrik Danielsson and Bengt Lennartson **119**

Enterprise Petri net based CAM software Opera-Topaz
Dmitry Zaitsev, ... **124**

Soleil: modeling a Health System
Ruben Gonzalez-Rubio and Lourdes Zubieta **129**

A Simulation System with Neural Data Interface for Shipbuilding Operation
Guodong Shao, Charles McLean and Swee Leong **137**

PRODUCTION AND DESIGN SIMULATION TOOLS

Virtual Enterprise Risk and Security Management
Tomas Feglar **145**

D-Fact Insight – Simulation of Huge Scale Models in Cooperative Work
Kiran Mahajan, Daniel Huber, Bengt Mueck and Wilhelm Dangelmaier **150**

MANUFACTURING OPTIMIZATION AND SET-UP

BetaSIM Extended Framework for General Discrete Event Simulation
František Huňka **159**

Achieving efficiency through Simulation in a declining Manufacturing Market
Rob Walker, Habtom Mebrahtu and Caroline Strange **162**

Visualizing Lean Effects with Simulation-Based Value Stream Mapping
Yang-Hua Lian and Hendrik Van Landeghem **168**

OEE Simulation in Production Ramp-up
Jürgen Fleischer; Gisela Lanza and Thomas Ender **173**

CONTENTS

PRODUCTION PLANNING AND JOB SHOP SIMULATION

The Impact of Human Performance Variation on the Accuracy of a Job Shop Simulation
Peter Lawrence **181**

A Simulation Application Framework for Production Planning and Scheduling
Hans P.M. Veeke and Gabriel Lodewijks **187**

An Event Driven Approach to Workflow Simulation
Agustín Yagüe, Pedro Pablo Alarcón, Juan Garbajosa, Jessica Díaz and José Luis Esteban **193**

AUTOMOTIVE PRODUCTION SYSTEM SIMULATION

Simulation validates Design and Scheduling of a Production Line
Vidyasagar Murty, Neelesh A. Kale, Rohit Trivedi, Omar M. Ulgen and Edward J. Williams **201**

Bezier Curves for Representing Clothoids, Circles and Straight Lines
N. Montés, J.Tornero and L.Arnesto **206**

Genetic Algorithm Based Optimization of Model Reference Adaptive Control Approach for Vehicle Active Suspension System
Ahmad Bagheri, Moussa Rezaee and S.Mehdi Hashemi **211**

Using Modelling Techniques For Fault Prediction In Hydraulic Drive Systems
C. Angeli and A. Chatzinikolaou **216**

Design of a Custom Plowing Process using Simulation
Allan Burke, Víctor Acosta, José Carlos Miranda and Juan de Dios Calderón **221**

FACTORY WAREHOUSE AND STORAGE MANAGEMENT SIMULATIONS

Greenhouse Automation: A Model for Design and Evaluation
Jaap A. Ottjes, Joan C. Rijsenbrij and Auke Groenenboom **227**

Pervasive Items for Factory Security
Alessandro Genco, Gaspare Scherma and Salvatore Sorce **231**

AutoTrans: Management and Transport Automation in Warehouses
Leopoldo Armesto and Josep Tornero **236**

Virtainer: Graphical Simulation of Container Storage Yard
Miguel Escriva, Marcos Martı, Jose Manuel Sanchez, Emilio Camahort, Javier Lluch and Roberto Vivó **242**

SUPPLY CHAIN SIMULATION

A Methodology for Supply Chain System Modeling and Process Evaluation
M.Chabrol, M.Comelli, P.Féniès and N.Tchernev **251**

A Modelling Framework for Supply Chain Planning Simulation under Uncertainty
François Galasso, Colette Mercé and Bernard Grabot **256**

The Impact of Various Delivery Prioritization Strategies in heterogeneous Supply Chain Environments
Thomas Thron, Gábor Nagy and Niaz Wassan **262**

Decision Support System for Plant Location Analysis Based on Fuzzy Model
Francesco Costantino, Giulio Di Gravio and Massimo Tronci **269**

TRAFFIC SIMULATION

Basic Aspects of Transport Demand Modelling
Nina Kudlackova and Alexander Chlaň **277**

Research Project Theory of Transport Systems
Karel Šotek and Alexander Chlaň **281**

Preparation of Digital Maps for Traffic Simulation: Part 1 Approach and Algorithms
Daniel Krajzewicz, Georg Hertkorn, Julia Ringel and Peter Wagner **285**

Urban Traffic Simulation Modeling for Reconfigurable Hardware
Anders Å. Hansson, Henning S. Mortveit Justin L. Tripp and Maya B. Gokhale **291**

Simulation of modern Traffic Lights Control Systems using the open source Traffic Simulation SUMO
Daniel Krajzewicz, Elmar Brockfeld, Jürgen Mikat, Julia Ringel, Christian Rössel, Wolfram Tuchscheerer, Peter Wagner and Richard Wösler **299**

CONTENTS

ENERGY SYSTEMS SIMULATION

Integration of a Hydrogen Economy into the German Energy System: an Optimising, Modelling Approach
Michael Ball, Otto Rentz and Martin Wietschel **305**

Feasibility Study of Static Field Converter
Anil Puppala, Mohammed Soliman, Mohammed Safiuddin and Andrew Abolafia **310**

Simulation of the Steam Condenser in MATLAB Environment
Milan Cepák **314**

Object-Oriented Database Management System for Industrial Combustion Processes
Nuray Kayakol and Marcus A. Nogueira **319**

Computer Simulation Model of the Marine Diesel Engine
Josko Dvornik, Luko Milic and Enco Tireli **325**

SIMULATION IN PROCESS INDUSTRY

Numerical Modelling and Analysis of Flow and Heat Transfer Processes in a Coke Oven
K. Hellgardt, K. J. Vyas, N. S. Hanspal and V. Nassehi **331**

Influence Spare Parts Supply on Modeling Mantenance Organization
Petar Stanojević, Vasilije Mišković, Vladimir Bukvić and Mirjana Misita **337**

Using Simulation to Analyse the Procurement of Addictives for Lubricants in the Oil Refinery of Porto, Portugal
J. Manuel Feliz de Teixeira and António E. S. Carvalho Brito **341**

Modelling of Porous Flow in Underground System
Naghmeh Keshavarzi-Roonizi, Rahim Ghorashi, Vahid Nassehi and Diganta Bhusan Das **348**

Architecture for Simulation and Knowledge Based Supervision for Process Facilities
Hans-Werner Schmidt, Ulrich Saffert, Gerhard Schreck, Alexander Sabov and Matthias Pleßow **353**

Tools for knowledge Acquisition, Modeling and Visualisation Applied to Process Supervision
Rainer Fröhling, Michael Pocher, Matthias Pleßow and Alexei Lisounkin **358**

CONTENTS

MODELLING AND SIMULATION IN THE TEXTILE INDUSTRY

FIBRE AND POLYMER SIMULATION

Development of a Mathematical Model for the Drying of Textile Fabrics
Pradeep K. Agrawal and Li Xue **367**

A Synergistic Approach by MRI and CFD for the Study of Microflow in Textiles
Johannes Leisen, Marco Märtin, Thomas Schipotinin, Elke Rodehüser, Karin Farber and Peter Farber **371**

Using Neuro-fuzzy for Prediction ring spun yarn Strength from Cotton Fiber properties
A.Shams-Nateri **376**

A New Technique to Measure Raleigh's Viscosity Parameters of Textile
Hatem Charfi and André Gagalowicz **380**

GA for Optimisation of Fibre Steering in Composite Laminates
Xavier Legrand, Don Kelly and Alan Crosky **388**

Electrical Properties of PANI-Dyneema conductive polymer Fibers – Fuzzy Logic Modeling
Bohwon Kim, Vladan Koncar and Claude Dufour **393**

Microscopic modelling of polymers and their interactions
Peter Ahlström, Kim Bolton and Mikael Skrifvars **398**

TEXTILE WEAVING SIMULATION

Computer Modeling of Electrostatic Web Formation
Yiyun Cai, Yong K. Kim and Abdel-fattah A. Seyam **405**

Neural Network Modeling of Fabric Pilling Evaluation
Xin Binjie and Jinlian Hu **410**

TEXTILE INDUSTRY INVENTORY AND SALES MANAGEMENT SIMULATION

Development of an Inventory Management Spreadsheet Simulator for a large Apparel Company
Jeffrey A. Joines, Kristin A. Thoney, and Ben R. Martin **417**

CONTENTS

Implementation of the Newsboy Method for the Sales Forecasting of the Apparel Industry
Sébastien Thomassey and Pierre Douillet **422**

LATE PAPERS

The Design, Fabrication and Image Processing of a 3 DOF Passive Biped Walking Robot
Ahmad Bagheri, Amir Hajiloo, Salar Basiri and
Rahman Yousefi Moghaddam **429**

Knowledge Modeling and Processing for Supervision of Process Facilities
Jörg Krüger, Alexei Lisounkin, Alexander Sabov, Gerhard Schreck and
Michael Pocher **435**

SCIENTIFIC PROGRAMME

SIMULATION ESTIMATION VALIDATION AND OPTIMIZATION

EFFICIENT TRUNCATION POINT ESTIMATION FOR ARBITRARY PERFORMANCE MEASURES

Mirko Eickhoff,
Donald C. McNickle,
Krzysztof Pawlikowski

University of Canterbury
Christchurch, New Zealand
E-Mail: m.eickhoff@cosc.canterbury.ac.nz

KEYWORDS

Sequential Steady State Simulation, Output Analysis, Truncation Point Detection, Arbitrary Performance Measures

ABSTRACT

In steady state simulation the output data collected during the transient phase often causes a bias in estimators of steady state performance measures. Common advice is to discard data collected during this transient phase. Finding an appropriate truncation point is a well-known and still not completely solved problem. In this paper we propose an improved algorithm for the determination of the truncation point for the output data sequence when its probability distribution reaches (approximately) steady state. The required run time of this algorithm is substantially improved without reducing the accuracy of the results. Because this algorithm is based on comparisons of empirical cumulative distribution functions, the truncation point is valid for an arbitrary steady state performance measure, such as the mean, variance or quantiles.

INTRODUCTION

In discrete-event simulation the sequence of output data represents a realization of a stochastic process $\{X_j\}_{j=1}^{\infty}$. The consecutive observations of this process are usually correlated and influenced by the initial state I of the system chosen by the analyst. Let $F_j(x|I) := \Pr[X_j \leq x|I]$ denote the cumulative distribution function of X_j. Assuming an ergodic system, $F_j(x|I)$ is converging towards $F(x) = \lim_{j\to\infty} F_j(x|I)$ which is called the marginal cumulative distribution function of the process $\{X_j\}_{j=1}^{\infty}$ in steady state. The primary concern of steady state simulation is to determine this distribution or its specific measures, such as e.g. the mean value.

In general the influence of I is significant in the beginning and decreases with increasing model time. If the interest is focused on the steady state behavior of the system, this initialization bias is obviously undesirable. A common way to reduce the influence of I is to truncate the "most" influenced part of the stochastic output process $X_1, \ldots, X_{l-1}$. Following this strategy the problem is to find an appropriate truncation point l. In the literature the steady state phase $\{X_j\}_{j=l}^{\infty}$ is described as a phase which is "relatively free of the influence of initial conditions" [Fishman, 2001] or by the statement that $X_l, X_{l+1}, \ldots$ "will have approximately the same (marginal) distribution" [Law and Kelton, 2000]. In practise there will often be an observation index l, such that

$$\forall j \geq l : F_j(x|I) \approx F(x) \tag{1}$$

is valid, unless the process $\{X_j\}_{j=1}^{\infty}$ is statistically unstable. Of course l should be finite, and should be the minimum of all indices, for which Equation (1) holds. Even though the estimation of $F(x)$ is the ultimate goal of steady state simulation, the expected value of the steady state random variable $E[X] = \lim_{j\to\infty} E[X_j]$ is often the only measure of interest. In this situation it is a generally accepted approach to replace Equation (1) by:

$$\forall j \geq l : E[X_j] \approx E[X] \tag{2}$$

In [Lee et al., 2000] it is shown that in the case of a $M/M/1/\infty$ system, the use of Equation (2) works even in quantile estimation. In general, however, the convergence of the mean is only a necessary condition for stationarity, and not a sufficient preliminary (see [Welch, 1983]). Therefore, Equation (1) can be applied in analysis of mean, variance, quantiles, or even rare event estimation, and Equation (2) should be used in mean value analysis only. In [Ghorbani, 2004] experimental investigations of methods based on Equation (2) are reported. Finding a truncation point on the basis of Equation (1) is not straightforward. It is therefore not very surprising, that the most common methods for detection of the truncation point are based on Equation (2) (see [Pawlikowski, 1990]) or on a visual inspection of the output data (see e.g. [Welch, 1983]). Completely algorithmic methods only give proper results under special conditions (see [Gafarian et al., 1978]) and are mostly based on one long simulation run. Nearly all methods use Equation (2), although the steady state phase should in general be defined by the convergence of $F_j(x|I)$ towards the steady state distribution $F(x)$.

In the MRIP scenario (multiple replications in parallel, see [Pawlikowski et al., 1994]) it is possible to collect a random sample of p independent and identical distributed realizations of each X_j, one from each replication. Let x_{ij} be the jth observation in the ith replication, with $1 \leq i \leq p$ and $1 \leq j < \infty$. Therefore, the empirical (marginal) cumulative distribution function $\tilde{F}_j(x|I)$ based on the order statistic

of $x_{1j}, \ldots, x_{pj}$ is an estimate of $F_j(x|I)$. This strategy requires a synchronization of the parallel replications so that the same number of observations is taken from each replication. Note, that this can lead to idle processors in an inhomogeneous computer network. The use of independent replications was considered in the case of mean value estimation in [Whitt, 1991] and [Alexopoulos and Goldsman, 2004]. The main source of error using independent replications is the initialization bias. If this source of error is eliminated, the estimates of independent replications are more accurate than e.g. ones obtained by means of methods based on batching. In [Eickhoff, 2002] an approach to reduce the initialization bias is proposed which is based on Equation (1) and uses the MRIP scenario. Its performance is examined and improved in [Bause and Eickhoff, 2002] and [Bause and Eickhoff, 2003]. An application of this approach is demonstrated in [Arns et al., 2003]. The results show that this approach is more reliable than methods which are based on Equation (2), especially if the transient behaviour is more complicated. Unfortunately, its computing time is quite large, possibly too excessive, as in many applications the computing time is as important as accuracy.

In the next sections an algorithm is described, which is based on Equation (1). In comparison to the algorithm described in [Bause and Eickhoff, 2003], the required run time of the new algorithm is substantially improved without reducing the accuracy of the results. This is examined in a later section by comparing the worst-case run time and the accuracy of both algorithms. In this paper only sequential methods and algorithms which are based on a dynamic set of data are considered. Therefore, the output analysis is performed online and guides the simulation experiment until estimates are statistically accurate.

IMPROVED ALGORITHM

The basic idea of the algorithm is given by Equation (1). The aim is to determine the first index l from which on all following probability distribution functions are (approximately) identical. Obviously the time horizon of every simulation experiment is limited. Therefore it is not possible to access "all" successive probability distribution functions. Accessible is only the observed part of the steady state phase. It is strongly necessary that this observed part of the steady state phase is reasonably large to avoid the determination of a misleading truncation point. Therefore the algorithm selects the size of the observed part of the steady state as a (constant) factor r of the size of the transient period, i.e. the size of the observed part of the steady state phase is always r-times larger than the so far selected transient phase. During each step of the algorithm the index l is increased by one. To assure the proper ratio between the two phases during each step $r+1$ new observation indices have to be accessed and included in analysis. For details on the basic idea of this approach see [Bause and Eickhoff, 2002]

Listing 1 shows a pseudo code of the improved algorithm, where for convenience some special notation is used. Let the sequence $\{y_{ij}\}_{i=1}^{p}$ be the order statistic of observations $\{x_{ij}\}_{i=1}^{p}$. Using the operators +, -, / and := in conjunction with sequences (see lines 0, 5, 8, 10 and 15) means to use these operators on each component of a given sequence separately. The operator $\approx$ in line 10 and 15 implements Equation (1) and is realized by the Kolmogorov-Smirnov two-sample test (KS-test). If the null hypothesis of equality is accepted, the operator $\approx$ (resp. the KS-test) results *true*. The procedure *observe()* collects one observation of each replication and the procedure *uniform(a,b)* delivers a uniform distributed integer random number between a and b used as index. The variable l points at X_l as the candidate for the truncation point, at the end of the algorithm it is the valid truncation point. The variable n represents the number of observations collected of each replication so far.

Listing 1: Pseudo code of the improved algorithm

```
0  int l := 0; int n := 0; int r := 10; {s_i}_{i=1}^p := 0;
   bool NoTestFailed := false;
   while (¬NoTestFailed){
      n := n + 1;
      observe({x_in}_{i=1}^p);
5     {s_i}_{i=1}^p := {s_i}_{i=1}^p + {y_in}_{i=1}^p;
      if (0 ≠ n mod (r+1)) continue;
      l := l + 1;
      {s_i}_{i=1}^p := {s_i}_{i=1}^p − {y_il}_{i=1}^p;
      NoTestFailed := true;
10    if (¬({y_il}_{i=1}^p ≈ {s_i}_{i=1}^p/(n − l)))
         NoTestFailed := false;
      for (int k := 1; k ≤ r; k := k + 1){
         if (¬NoTestFailed) break;
         int u := uniform(lk + 1, l(k + 1));
15       if (¬({y_il}_{i=1}^p ≈ {y_iu}_{i=1}^p))
            NoTestFailed := false;
      }
   }
```

The most time consuming factor of the algorithm described in [Bause and Eickhoff, 2003] is the increasing number of KS-tests executed during each step of the algorithm. The improved algorithm described in Listing 1 avoids these tests by using the calculated sequence $\{s_i\}_{i=1}^{p}$ which is an estimate of $F(x)$ based on the latest observations during each step of the algorithm. The sequence $\{s_i\}_{i=1}^{p}$ is the sum of all order statistics which are not part of the transient period. New observations are added whereas observations of the transient period are subtracted from $\{s_i\}_{i=1}^{p}$ (see lines 5 and 8). Dividing each component of $\{s_i\}_{i=1}^{p}$ by the number of addends results in an estimate of $F(x)$. This sequence is compared with the order statistic of the actual test sample $\{x_{il}\}_{i=1}^{p}$ (see line 10). Because $\{s_i\}_{i=1}^{p}$ is calculated over observations at different model times, a possible periodic behaviour could be overlooked and an unreliable estimate of the truncation point could be accepted (cf. [Bause and Eickhoff, 2003]). Therefore additional r randomly chosen sequences are used to avoid this trap. The observed part of the steady state phase is divided into r equally sized intervals. Each interval contains one randomly chosen sequence. In a loop all of these sequences are compared with the actual test sample $\{x_{il}\}_{i=1}^{p}$ (see lines 12 to 17). If the assumption of equality is rejected by the KS-test for $\{s_i\}_{i=1}^{p}$, or any of the randomly chosen test samples, the truncation point l is not adequate and more steps of the algorithm have to be performed. In contrary to the previous version of this algorithm, the number of needed KS-tests in each step is maximum $r+1$.

WORST CASE TIME COMPLEXITY

In [Bause and Eickhoff, 2003] it is demonstrated experimentally, having considered a number of different kinds of transient

behaviour, that the previous version of the algorithm described in Listing 1 is very accurate and has therefore a great advantage over some other commonly used methods for truncation point estimation. But the price for its accuracy is the length of its run time which is $O(n^2 p \log(p))$. This run time is possibly too large for practical implementations. In this section it will be shown that the algorithm described in the previous section is a substantial improvement, because its run time is only $O(np\log(p))$, without significant loss of accuracy, in conjunction with any kind of transient behaviour.

As before, let p denote the number of replications and n the number of observations in each single replication. Note, that at the end of the algorithm n is a multiple of $r+1$. The total number of observations is pn. Assume, that all basic arithmetic operations are in $O(1)$ (cf. [Cormen et al., 1994]). In the following the running times of algorithms are considered from the point of view of their worst-case time complexity.

Theorem
The worst case running time of the algorithm described in Listing 1 is $O(np\log(p))$.

Proof: Only KS-tests with random samples of size p are performed. The basis of the KS-test are two sorted random samples. Sorting can be done in $O(p\log(p))$. To determine the maximum distance in the compared samples a pointer has to be shifted through each sorted random sample. This can be done in $2p$ steps which leads to $O(p)$. To accept or reject the null hypothesis, the determined maximum difference has to be compared with a tabulated critical value. This can be done in $O(1)$. Therefore, the run time of one KS-test is $O(p\log(p)) + O(p) + O(1) = O(p\log(p))$.

The run time of a single execution of lines 3 and 6 is $O(1)$ and of lines 4 and 5 it is $O(p)$. Because the while-loop in line 2 is executed n times before the algorithm stops, the run time of this part of the algorithm is $O(np)$.

A single execution of lines 7, 9 and 11 can be done in $O(1)$; a single execution of line 8 can be done in $O(p)$; to execute line 10 a run time of $O(p\log(p))$ is needed, because a KS-test has to be performed. Because of the condition in line 6 this part of the algorithm is executed only $\frac{n}{r+1}$ times. Therefore, the run time of this part of the algorithm is $\frac{n}{r+1} \cdot O(p\log(p))$ which leads to $O(np\log(p))$ because r is a constant parameter.

A single execution of lines 12, 13, 14 and 16 needs only a minor run time of $O(1)$. The KS-test in line 15 can be done in $O(p\log(p))$. The for-loop is executed at maximum r times, therefore, the run time of one complete for-loop in each step of the algorithm is $r \cdot O(p\log(p))$. All in all $\frac{n}{r+1}$ for-loops have to be performed. Therefore, the run time of this part of the algorithm is $n \cdot \frac{r}{r+1} \cdot O(p\log(p))$ which leads to $O(np\log(p))$.

Combining all results, the run time of the algorithm is $O(np) + O(np\log(p)) + O(np\log(p)) = O(np\log(p))$ ■

Because p could be considered as a constant parameter and usually $p << n$ holds, the run time could be described by $O(n)$. This run time is linear and highly efficient, because each observation has to be processed at least once.

PERFORMANCE

To compare the accuracy of both versions of the algorithm the same artificial processes as in [Bause and Eickhoff, 2003] are used, except the starting condition of the ARMA process is changed, to create a more unusual initial state. The output data of these artificial processes is used to show the performance of the algorithm on many different kinds of transient behaviour. Let $\{\epsilon_t\}_{t=1}^{\infty}$ denote an independent Gaussian white noise process (see [Hamilton, 1994]). In all experiments we used $p = 100$, $r = 10$ and an α-level of 0.05 for the critical value of the KS-test. For details on this parameters the reader is referred to [Bause and Eickhoff, 2002]. Further more, in all of the performed experiments the pseudo random number generator described in [L'Ecuyer et al., 2002] is used. This generator is suitable for many parallel replications, because its period is reasonably large and can be divided into many substreams.

Process A: linear transient mean

$$Y_t^{(A)} = \begin{cases} \epsilon_t + x - t\frac{x}{l} & \text{if } t < l, \\ \epsilon_t & \text{else.} \end{cases}$$

with $x = 10$, $l = 100$. This process can be regarded as the easiest case, because all performance measures are influenced by the initial state and show a transient behaviour, especially the mean and all quantiles. At a well defined index l the influence of the initial condition disappears completely.

Process B: linear transient variance

$$Y_t^{(B)} = \begin{cases} \epsilon_t \cdot (x - t\frac{x-1}{l}) & \text{if } t < l, \\ \epsilon_t & \text{else} \end{cases}$$

with $x = 10$, $l = 100$. The probability distribution function of this process has a transient behaviour, which is not visible for Equation (2) because the mean (resp. the median) is constant right from the beginning. Therefore, in this case Equation (2) is not suitable for an estimation of any other measure than the mean.

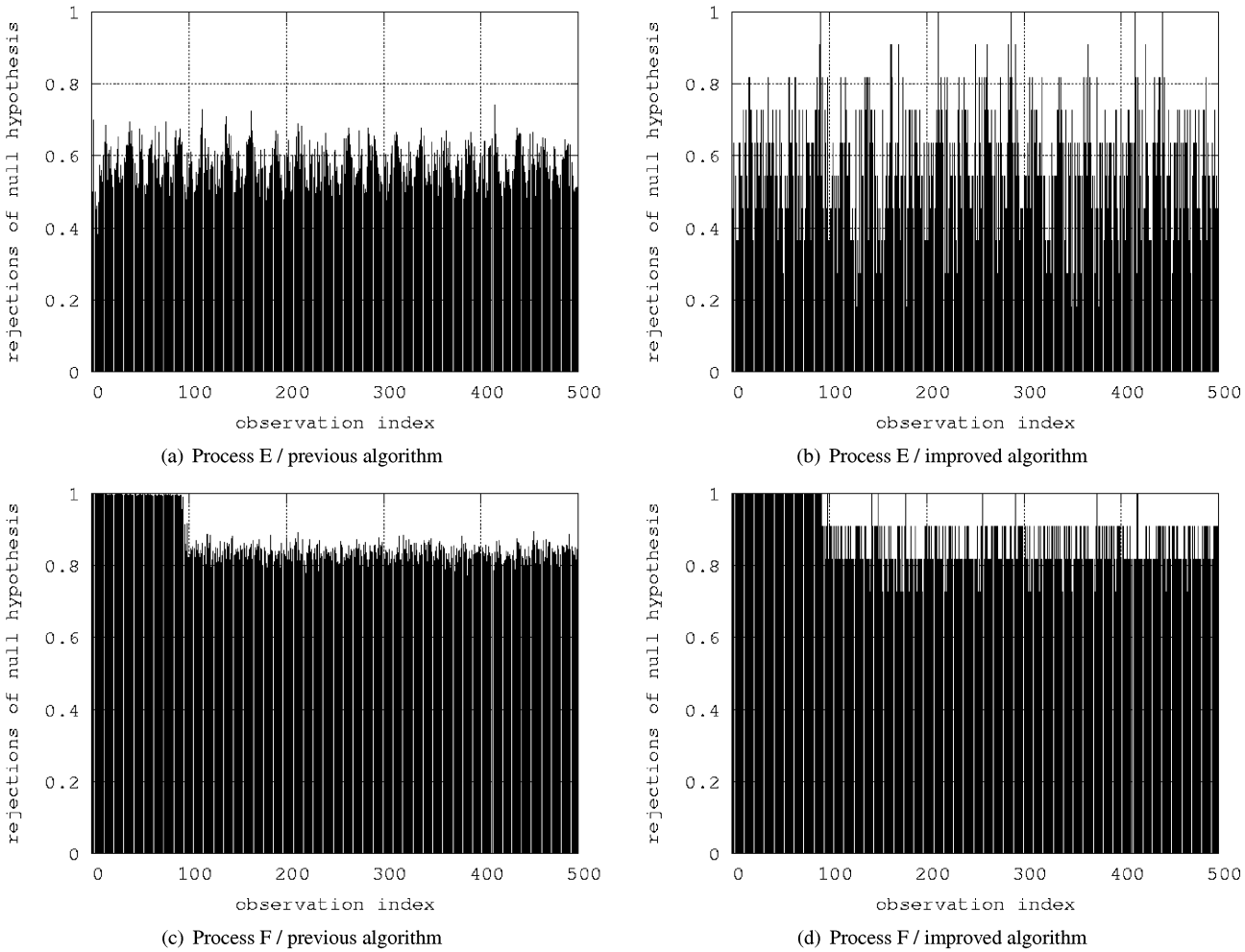
Process C: exponential transient mean

$$Y_t^{(C)} = \epsilon_t + x \cdot e^{(t\frac{\ln(0.05)}{l})}$$

with $x = 10$, $l = 100$. In this case every performance measure has a transient behaviour. However, the influence of the initial state is disappearing exponentially. When regarding a finite simulation horizon, the influence never disappears completely. The definition of this process implies that at index l the impact of the influence of the initial state is reduced to 5% of its impact at $t = 0$.

Process D: ARMA$(5, 5)$

$$Y_t^{(D)} = 1 + \epsilon_t + \sum_{i=1}^{5} \frac{1}{2^i}(Y_{t-i}^{(D)} + \epsilon_{t-i})$$



(a) Process E / previous algorithm

(b) Process E / improved algorithm

(c) Process F / previous algorithm

(d) Process F / improved algorithm

Figure 1: The number of rejections of the null hypothesis standardized by the number of all performed KS-tests. The values are plotted over model time, each peak represents one step of the algorithms.

with $Y_{-5}^{(D)} = Y_{-4}^{(D)} = Y_{-3}^{(D)} = Y_{-2}^{(D)} = Y_{-1}^{(D)} = 100$. Each value of this process depends on the 5 previous values. Therefore, the observations of this process are highly autocorrelated. A theoretical investigation shows that the expected value of this process in steady state is $E[Y_{\infty}^{(D)}] = 32$. The initial state has no influence on this expected value.

Process E: periodic

$$Y_t^{(E)} = \epsilon_t + b \cdot \sin(\omega t)$$

with $b = 1$, $T = \frac{2\pi}{\omega} = 50$. The transient behaviour of this process is governed by a sine oscillation and is not converging to a steady state distribution at all. A possible pitfall concerning output analysis of this process is that a batched mean seems to converge, if the batch size is equal to the period of the oscillation.

Process F: non-ergodic

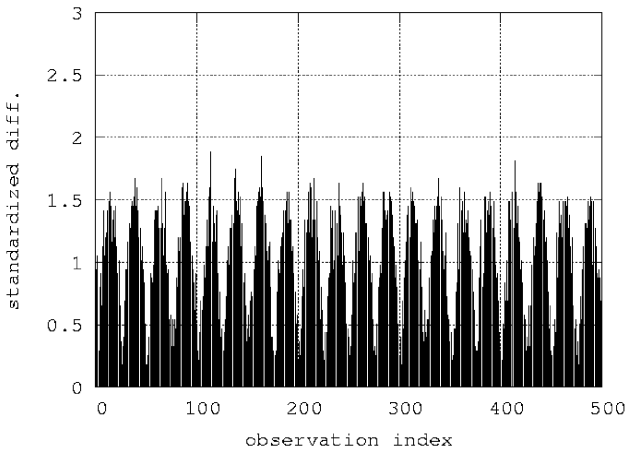
$$Y_t^{(F)} = \begin{cases} \epsilon_t(ct+1) + x - t\frac{x}{l} & \text{if } t < l, \\ \epsilon_t(ct+1) & \text{else.} \end{cases}$$

with $x = 10$, $l = 100$, $c = 0.01$. This process is governed by two different transient behaviours. The first one is obvious and is the same as in Process A. The second transient behaviour is more hidden and results in a non-ergodic behaviour. The pitfall concerning output analysis of this process is to overlook the second transient behaviour. This is especially likely when performing a visual inspection.

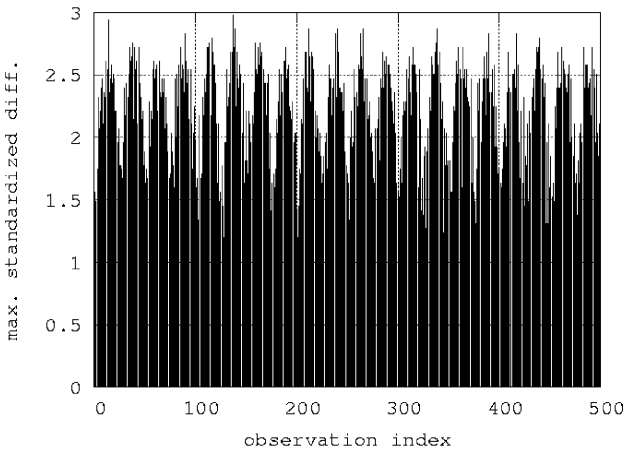
The processes A to D converge towards a steady state distribution. Therefore Equation (1) can be used to estimate the beginning of the steady state phase. To obtain a statistically accurate result simulation experiments with the algorithm described in [Bause and Eickhoff, 2003] and the algorithm of Listing 1 are performed 20 times. The average of all 20 results is listed in Table 1 with the halfwidth of its confidence interval and its relative statistical error.

using 20 runs	previous algorithm	improved algorithm
Process A	98.00 ± 0.43 (0.4%)	97.25 ± 0.57 (0.6%)
Process B	86.20 ± 1.01 (1.2%)	82.20 ± 1.87 (2.3%)
Process C	103.05 ± 1.76 (1.7%)	99.25 ± 1.81 (1.8%)
Process D	190.25 ± 6.53 (3.4%)	185.00 ± 6.55 (3.5%)
run time	$O(n^2 p \log(p))$	$O(np \log(p))$

Table 1: The average truncation points and their confidence intervals. Each result is based on 20 simulation experiments.



(a) Process E / Line 10: KS-tests with $\{s_i\}_{i=1}^{p}$



(b) Process E / Line 15: KS-tests with r random samples

Figure 2: Standardized statistics produced by the KS-tests executed in line 10 an line 15 respectively (see Listing 1) when applied to Process E. The null hypothesis is accepted when the standardized statistic is below 1.

The halfwidths of the confidence intervals are in all our experiments smaller than five percent of the mean value after fewer than 20 simulation experiments. This shows that both algorithms deliver a robust estimate with a small variance. Even though the results of both algorithms are comparable, the algorithm described in Listing 1 tends to estimate the truncation point a bit earlier. This shows that the fewer KS-tests of this algorithm are weaker than the large number of KS-tests in the previous version of the algorithm, which cause the longer run time.

The processes E and F do not converge towards a steady state distribution, and so there is no steady state phase at all. Therefore, both algorithms should not return a truncation point. To check the accuracy in these cases each step of the algorithms is observed by plotting the number of rejections of the null hypothesis (see Figure 1). The rejections should be on a high level so that Equation (1) will not be true for any tested value of l. To make the number of rejections comparable, they are standardized by the number of all performed KS-tests. Therefore, the value 1 means that all KS-tests reject the null hypothesis, and the value 0 means that all KS-tests accept the null hypothesis. To obtain results depicted in Figures 1(b), 1(d) and 2(b) the condition in line 13 of Listing 1 is ignored, to force the execution of all $r + 1$ KS-tests to obtain continuous plots.

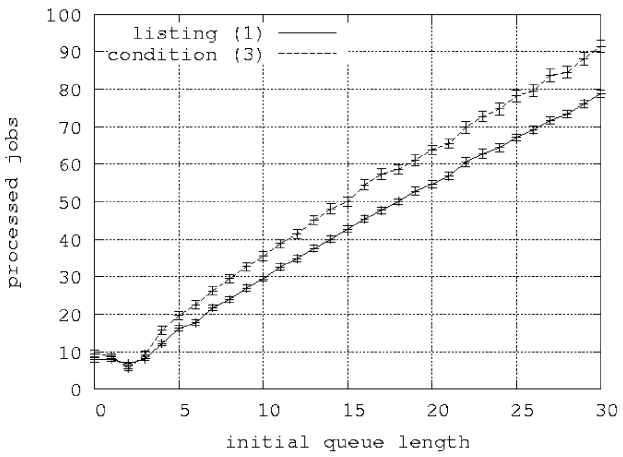
The plots in Figure 1(a) and 1(c), with the results for the original algorithm described in [Bause and Eickhoff, 2003] are quite smooth, because a large number of KS-tests are used to check Equation (1). Additionally these plots show that the KS-tests of the previous algorithm work reliably, because of the large number of rejected null hypotheses when analyzing Process E and F. Note, this algorithm gives an estimated length of the initial transient phase if the number of rejections is below a certain threshold, which is usually set at 0.05. The plots in Figure 1(b) and 1(d), for the modified algorithm, are not as smooth, because the number of executed KS-tests is much smaller. However, the number of rejections of the null hypothesis is high in both cases. Note, that the improved algorithm accepts an estimate only if all KS-test accept the null hypothesis. Therefore the improved algorithm works in the case of on Process E and F as reliably as the previous algorithm.

Furthermore, Process E is a good example to demonstrate that a truncation point estimation exclusively based on $\{s_i\}_{i=1}^{p}$ is not sufficient. The statistics produced by the KS-test are plotted in Figure 2. Figure 2(a) shows the results of the comparisons between $\{x_{il}\}_{i=1}^{p}$ and $\{s_i\}_{i=1}^{p}$ performed in line 10 of Listing 1. Figure 2(b) shows the maximum statistics of all KS-tests performed in line 15 of Listing 1 during the for-loop. In the second case, the maximum statistics of all r comparisons are plotted by disregarding the condition from line 13. To achieve comparability, the statistics are standardized by appropriate critical values of the KS-test. A standardized statistic below 1 means that the corresponding null hypothesis was accepted by the KS-test.

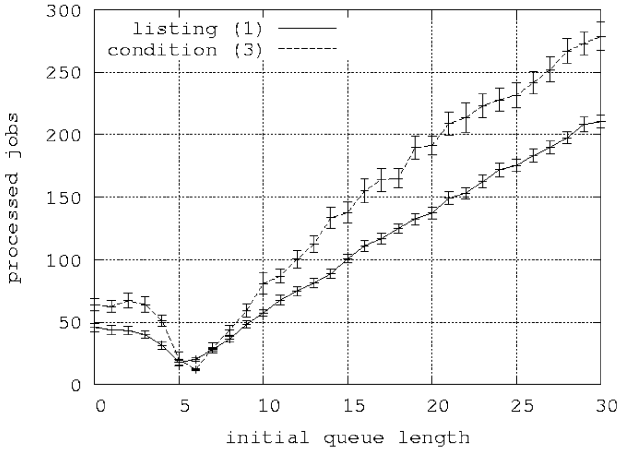
Figure 2(a) shows a periodical pattern of acceptance of the null hypothesis. The reason for this is, that $\{x_{il}\}_{i=1}^{p}$ is governed by the periodical behaviour of Process E. The situation is different in the case of $\{s_i\}_{i=1}^{p}$ since then many random samples of different model times are taken into account. Therefore, $\{s_i\}_{i=1}^{p}$ cannot reflect the periodical behaviour and represents the average test sample over many periods of the sine oscillation. When Process E is at an extreme value of the sine oscillation, the KS-test detects the difference between $\{x_{il}\}_{i=1}^{p}$ and $\{s_i\}_{i=1}^{p}$. If Process E is at a zero point of the sine oscillation, the KS-test is not able to recognize a difference between $\{x_{il}\}_{i=1}^{p}$ and $\{s_i\}_{i=1}^{p}$. These results in the same problem as shown for the method of Welch (see [Bause and Eickhoff, 2003] and [Welch, 1983]). However, the KS-tests with the randomly selected samples make sure that the improved algorithm works reliably (see Figure 2(b)).

TRANSIENT BEHAVIOUR OF THE M/M/1 QUEUE

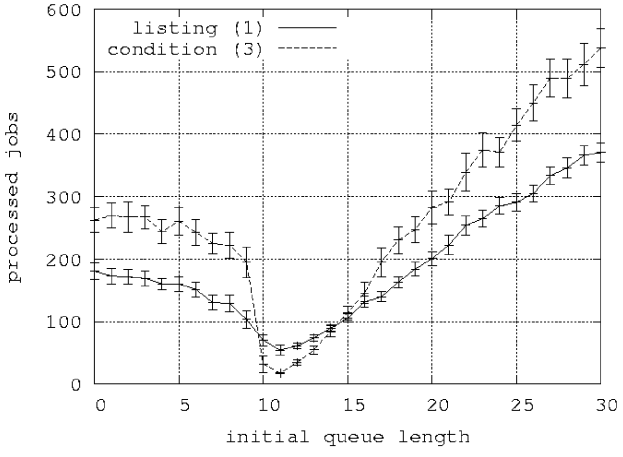
In [Kelton and Law, 1985] it is pointed out that true steady state behaviour of a M/M/1 queueing system is very difficult to observe by simulation experiments. The reason is that the dependence of the output data on the initial state can last very long. In [Jain, 1991] the most important properties of an M/M/1 queue are summarized. Only the mean arrival rate λ and the service rate μ is necessary to analyze the steady state behaviour of the



(a) traffic intensity $\rho = 0.5$



(b) traffic intensity $\rho = 0.8$



(c) traffic intensity $\rho = 0.9$

Figure 3: The estimated truncation points in response time analysis obtained according to Listing 1 and Condition (3).

M/M/1 queue. The traffic intensity is given by $\rho = \frac{\lambda}{\mu}$. The stability condition is $\rho < 1$. Let N denote the number of jobs in the system in steady state, in queue and in service. If the stability condition holds, the mean number of jobs in the system in steady state is $E[N] = \frac{\rho}{1-\rho}$. Let D denote the response time (or turnaround time) of the system in steady state, including the time in queue and in service. Then, the mean response time in steady state is given by $E[D] = E[N]/\lambda$. Let D_{ij} denote the response time of the jth job leaving the system in the ith replication. When using the MRIP scenario, the mean response time of the jth job, $E[D_j]$, can be estimated by

$$\bar{D}_j(p) = \frac{1}{p}\sum_{i=1}^{p} D_{ij} \quad .$$

The aim of the experiments reported here is to assess the results produced by the algorithm described in Listing 1 when analyzing the M/M/1 queue. A comparison with the previously published results of [Kelton and Law, 1985] is done. The authors of that publication used the same condition as a measure of convergence to steady state, as [Gafarian et al., 1978]. Namely:

$$|\bar{D}_j(p) - E[D]| \leq \epsilon \cdot E[D] \tag{3}$$

When analyzing the response time D, j should not be smaller than the initial number of jobs in the system. However, it should be the minimum index for which this condition holds. This condition is based on the estimated mean values $\bar{D}_j(p)$ and, therefore, implements Equation (2). Further more, this condition is only applicable, if the expected steady state value of $E[D]$ is known. Therefore, this rule has only theoretical importance and cannot be applied in practise.

The primary aim in [Kelton and Law, 1985] was to use that condition to assess the rate of convergence of the mean waiting time in the queue in different queueing models to steady state. Because of this, the results of the algorithm described in Listing 1 are not expected to be the same as those of Kelton and Law.

In all our experiments we set $\lambda = 1$ so that $E[D] = E[N]$ holds. The series of experiments are done for estimating the truncation point of the M/M/1 queue with a constant traffic intensity of $\rho = \{0.5; 0.8; 0.9\}$ and an initial queue lengths between 0 and 30. In this experiments we set the parameter $r = 1000$. This leads to a large ratio between the length of the transient phase and the observed part of the steady state phase when executing the modified truncation point detection algorithm. For each initial value of the queue length the experiment is repeated 100 times to estimate the average truncation point and its confidence interval. In parallel to the algorithm described in Listing 1, Condition (3) is being checked on the same set of output data. We set $\epsilon = 0.05$ for all experiments. The analyzed measure is the response time.

The results are plotted in Figure 3 for $\rho = \{0.5; 0.8; 0.9\}$. The shapes of the curves obtained under Condition (3) are similar to those obtained for the improved truncation algorithm. In all cases they start on a higher level and drop down at the point, where the initial queue length is (approximately) equal to $E[N]$ resp. $E[D]$. After this point all curves rise again. This shows, that the results of the improved algorithm conform the theory.

The convergence of the mean is a necessary condition for a process being in the steady state phase. Therefore, a truncation point chosen on the basis of the converging probability distribution is expected to be equal or larger than a truncation point chosen on the basis of the converging mean. However, in the implementation of Equation (1) and Condition (3) two different measures of convergence have been used, so the results are not directly comparable. One can see that the estimated value of the truncation point depends not only on the convergence of the observed measure, but also the measure of closeness which in

Condition (3) is represented by a scalar value ϵ, while in the implementation of Equation (1) the statistic of the KS-test is used.

The experiments with this M/M/1 queue show that the results of the algorithm described in Listing 1 conform to theory and are comparable to the results of [Kelton and Law, 1985]. In contrast to the previously published results, all simulation experiments were performed automatically without any a priori knowledge about the steady state distribution or steady state mean values. Note, that the large number of experiments at high traffic intensities is only feasible because of the efficient running time of the improved algorithm.

LIMITS AND REJECTED IDEAS

Because Equation (1) uses an approximation the estimate of the beginning of the steady state phase of Process A is always smaller than the theoretically best value which is in this example $l = 100$. However, using the equality instead of an approximation does not make sense in output processes such as Process C. Using the equality instead of an approximation would lead to an infinite value for l, because in Process C the influence of the initial state disappears exponentially.

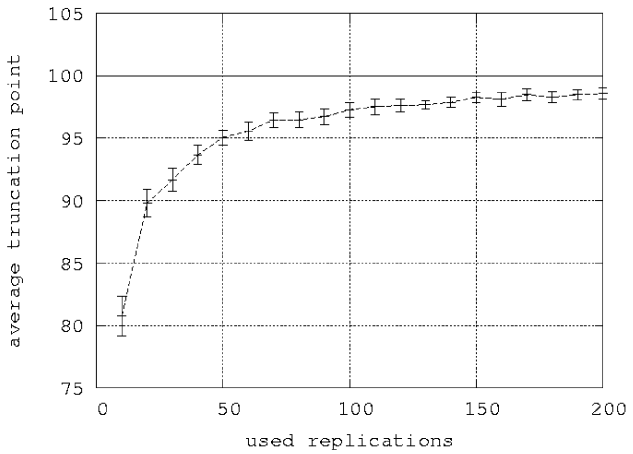


Figure 4: Truncation point analysis of Process A. The quality of estimates depends on the number of parallel replications.

Figure 4 depicts the results of simulation experiments for the Process A executed with different numbers of parallel replications. These experiments were executed as described in previous sections, with 20 independent simulation experiments for each plotted point. These results validate the assertion, that the estimated length of the initial transient phase is always smaller than its theoretically best value. Furthermore, another effect can be observed. The values of the estimates decrease with decreasing numbers of parallel replications. Note, that the critical values of the KS-test are defined for very small sample sizes, too. The realization of the approximation of Equation (1) based on the KS-test gets weaker for smaller numbers of parallel replications. Therefore, we recommend at least 30 parallel replications, and if possible more than 50 parallel replications should be used. This limit is valid for the algorithms described in [Bause and Eickhoff, 2003] and in Listing 1.

Before investigating the performance of the algorithm described in Listing 1, it was uncertain, whether the obtained results would be accurate or not. Therefore, it was an additional idea to perform a large number of KS-tests, as it is done in the previous version of the algorithm, whenever the $r+1$ KS-tests of the new version of the algorithm do not recognize a difference. This idea has been rejected because of two reasons. Firstly, the results obtained for both versions of the algorithm are convincingly close, see Table 1. Another reason for the rejection of that idea was, that the worst case time complexity would still be $O(n^2 p \log(p))$. In the worst case, the $r + 1$ KS-tests of the new version of the algorithm would always accept the null hypothesis and the large number of KS-tests of the previous version of the algorithm would always reject the null hypothesis. In this situation the worst case run time of the algorithm would be even worse than in the previous version.

CONCLUSIONS

We introduced an efficient algorithm for the estimation of the length of the initial transient phase in the sense of probability distribution; see Equation (1). The worst case time complexity of this algorithm is limited by $O(np \log(p))$ and is, therefore, substantially faster than the previous version of the algorithm (see [Bause and Eickhoff, 2003]). This improvement is achieved without reducing the accuracy of the estimates. Its performance has been experimentally assessed by conducting a number of simulations of artificial output processes with a variety of different types of transient behaviour.

We also investigated the newly introduced algorithm in M/M/1 queue simulation and compared them with previous results published in [Kelton and Law, 1985]. The results are as expected and conform to theory.

In the introduction it is pointed out that a truncation point estimation based on Equation (1) can be used for many different performance measures, such as mean values, variances, quantiles or even rare-events. Equation (2) is in general only useful in mean value analysis.

However, limits concerning Equation (1) are also discussed. For Process A is demonstrated that this equation does not always leads to the theoretically best truncation point. This is caused by the approximation used. Using the equality instead of an approximation is no alternative, because this would lead in some cases (e.g. Process C) to an infinite truncation point. Equation (1) can be used to reduce the initialization bias dramatically, but it cannot be used to eliminate it completely.

REFERENCES

[Alexopoulos and Goldsman, 2004] Alexopoulos, C. and Goldsman, D. (2004). To batch or not to batch. *ACM Transactions on Modeling and Computer Simulation*, 14(1):76–114.

[Arns et al., 2003] Arns, M., Eickhoff, M., Fischer, M., Tepper, C., and Völker, M. (2003). New Features in the ProC/B Toolset. In F. Bause (ed): Tools of the 2003 Illinois International Conference on Measurement, Modelling and Evalu-

ation of Computer-Communication Systems. Technical Report 781, Uni Dortmund, Fachbereich Informatik.

[Bause and Eickhoff, 2002] Bause, F. and Eickhoff, M. (2002). Initial transient period detection using parallel replications. *Proc. of the 14th European Simulation Symposium*, pages 85–92.

[Bause and Eickhoff, 2003] Bause, F. and Eickhoff, M. (2003). Truncation point estimation using multiple replications in parallel. *Proc. of the Winter Simulation Conference*, pages 414–421.

[Cormen et al., 1994] Cormen, T. H., Leiserson, C. E., and Rivest, R. L. (1994). *Introduction to Algorithms*. MIT Press.

[Eickhoff, 2002] Eickhoff, M. (2002). Statistische Auswertung und Erkennung der stationären Phase in der Simulation zustands-diskreter Systeme. Diploma Thesis, Uni Dortmund, Fachbereich Informatik.

[Fishman, 2001] Fishman, G. S. (2001). *Discrete-Event Simulation*. Springer.

[Gafarian et al., 1978] Gafarian, A. V., Ancker, C. J., and Morisaku, T. (1978). Evaluation of commonly used rules for detecting steady state in computer simulation. *Naval Research Logistics Quarterly*, pages 511–529.

[Ghorbani, 2004] Ghorbani, B. (2004). The issue of initial transient in sequential steady-state simulation. Master thesis, University of Canterbury, Department of Computer Science.

[Hamilton, 1994] Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

[Jain, 1991] Jain, R. (1991). *The Art of Computer Systems Performance Analysis*. John Wiley & Sons, Inc.

[Kelton and Law, 1985] Kelton, W. D. and Law, A. M. (1985). The transient behavior of the m/m/s queue, with implications for steady-state simulation. *Operations Research*, 33:378–396.

[Law and Kelton, 2000] Law, A. M. and Kelton, W. D. (2000). *Simulation Modeling and Analysis*. McGraw-Hill Higher Education, New York.

[L'Ecuyer et al., 2002] L'Ecuyer, P., Simard, R., Chen, E. J., and Kelton, W. D. (2002). An object-oriented random-number package with many long streams and substreams. *Operations Research*, 50(6):1073–1075.

[Lee et al., 2000] Lee, J.-S. R., McNickle, D., and Pawlikowski, K. (2000). Initial transient period detection for steady-state quantile estimation. *Proceedings of the Summer Computer Simulation Conference*, pages 169–174.

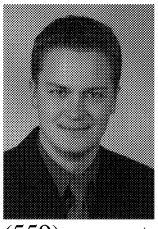
[Pawlikowski, 1990] Pawlikowski, K. (1990). Steady-state simulation of queueing processes: a survey of problems and solutions. *ACM Computing Surveys*, 22:123–170.

[Pawlikowski et al., 1994] Pawlikowski, K., Yau, V., and McNickle, D. (1994). Distributed stochastic discrete-event simulation in parallel time streams. *Proc. of the Winter Simulation Conference*, pages 723–730.

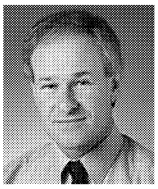
[Welch, 1983] Welch, P. D. (1983). The statistical analysis of simulation results. *In The Computer Performance Modeling Handbook, ed. S. Lavenberg, Academic Press*, pages 268–328.

[Whitt, 1991] Whitt, W. (1991). The efficiency of one long run versus independent replications in steady-state simulation. *Management Science*, 77(6):645–666.

AUTHOR BIOGRAPHIES



MIRKO EICKHOFF holds a Diploma degree in Computer Science from the University of Dortmund. His research interests are in the area of output analysis of discrete event simulation using multiple replications. His diploma thesis is part of the Collaborative Research Center "Modelling of Large Logistic Networks" (559) supported by the Deutsche Forschungsgemeinschaft. He worked for Delmia (Germany) in the area of workload balancing in manufacturing industry. In 2004 he received the targeted doctoral scholarship of the University of Canterbury and is currently a Ph.D. Candidate in Computer Science in the Simulation Research Group at this University. His e-mail address is m.eickhoff@cosc.canterbury.ac.nz.



DONALD C. MCNICKLE is an Associate Professor of Management Science in the Management Department at the University of Canterbury. His research interests include queueing theory, networks of queues and statistical aspects of stochastic simulation. He is a member of INFORMS and the Operational Research Society. His e-mail address is don.mcnickle@canterbury.ac.nz.



KRZYSZTOF PAWLIKOWSKI is a Professor in Computer Science at the University of Canterbury, in Christchurch, New Zealand. The author of over 130 research papers and four books; has given invited lectures at over 80 universities and research institutes in Asia, Australia, Europe and North America. Alexander-von-Humboldt Research Fellow (Germany) in 1983-84 and 1999. His research interests include performance modelling of telecommunication networks, discrete-event simulation and distributed processing. Senior Member of IEEE, member of ACM and SMSI. His e-mail address is krys.pawlikowski@canterbury.ac.nz.

DEPLOYING VALIDATION TESTS IN SYSTEMS DEVELOPMENT

Juan Garbajosa[1]
Pedro P. Alarcón[2]
System and Software Technology Group
http://syst.eui.upm.es
Universidad Politécnica de Madrid
Postal address: E.U. Informatica. Ctra. Valencia Km. 7 E-28031 Madrid
Spain
E-mail[1]: jgs@eui.upm.es
E-mail[2]: pcavero@eui.upm.es

KEYWORDS

System testing, software/system engineering environment, tool integration, component-based development, process model standards, TOPEN

ABSTRACT

This paper presents a way to deploy validation tests in system/software development supported by an automation approach. This automation approach is based on the integration of the SEE (ESA System/Software Engineering Environment) and TOPEN, an environment that can be used for defining and executing validation test procedures. A process model (ECSS) is underlying. To achieve the goals an operational relationship between requirements and tests is required, together with clear tool integration strategy.

1. INTRODUCTION

System testing, understood as acceptance/conformance testing, is a system/software lifecycle process defined in widely used standards such as (ISO/IEC 15288, 2002). However system testing may be used advantageously to get a better and more precise knowledge not only on the accomplishments of functionalities by the implemented system but also on the implementation status. For this, a process model and a proper automation approach become essential. This paper describes how the proposed automation approach is able to support the deployment of system tests. The first section, this one, is Introduction; section 2 presents the automation strategy; section 3 Test Data Integration and Management, and last section introduces a number of conclusions.

2. AUTOMATION STRATEGY

The automation strategy has a number of key issues. The first issue is to provide a proper support to system testing definition and execution. The second issue is to be able to manage tests and tests reports from the system requirements perspective. While (acceptance) system testing tools depend on the application domain, requirement engineering tools do not.

The acceptance testing tool, TOPEN, Test and Operation Environment, is an environment that has been envisaged to simplify the process of defining and executing tests of complex systems, and was firstly described in (Garbajosa et al, 2000). TOPEN supports a multi-site tele-testing environment for systems made of inter-related components. The system to be tested shall be named as Unit Under Test (UUT). Fig. 1 describes TOPEN generals for a specific application domain, a TV digital decoder. The test engineer can define, compile and execute test procedures, made of operation commands that are sent to UUT. TOPEN has, as components, a database (MIB), a Man machine interface and a code generator. A gateway communicates the UUT and TOPEN. TOPEN supports the definition and execution storage of test reports. This tests results storage can be used to perform system execution monitoring. This monitoring can be a simple, e.g. to verify if each test procedure result was that expected, or a system monitoring according to a complex model

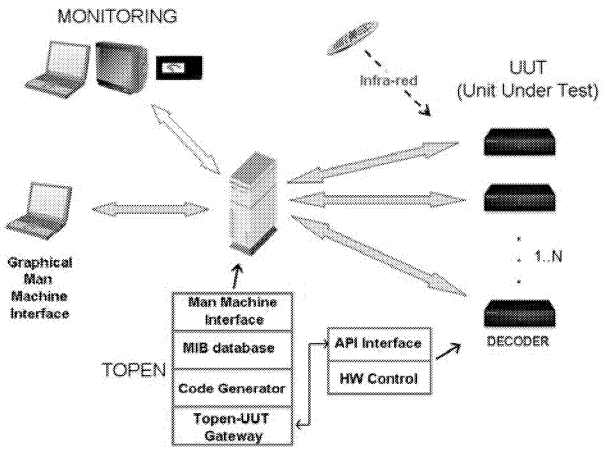


Figure 1. TOPEN and UUT

As a requirements engineering tool a subset of a SEE software/system engineering environment is used. SEE was produced in the context of an European Space Agency project with the objective of providing low-cost support for the development of critical systems and conform to the ECSS process model documents (ECSS e-40B, 2003). SEE is document-centric. Documents are defined in CASEML, a language derived from XML. CASEML has been devised

with the purpose of supporting tree-like process model documents in general and, specifically those of the ECSS process model. Some of the features of the SEE have been described in (Bollaín et al, 2003). Traceability to other document nodes is supported and managed. SEE has been implemented in Tcl/Tk and C languages in order to maximize portability issues.

SEE can use third parties tools. For the purpose described within this paper it was specially useful to integrate it with TOPEN. The integration of SEE and TOPEN environments is obtained by means of two additional components, the TOPEN-SEE Mediator (TSM) and the Data Sources Server (DSS). TSM allows us to specify validation tests, compile them, execute them and store execution results directly from the SEE. The results may be stored in TOPEN but also in SEE repository. Repositories are different but are kept consistent. Figure 2 shows a TSM window where test procedures are presented, and shared, regardless they are in SEE or TOPEN. TSM has been implemented in Tcl/Tk. DSS provides a transparent access to distributed data sources -e.g. a database- and has been described in (Alarcón et al., 2002). This architecture is described in figure 3.

TSM communicates with TOPEN and DSS components by means of sockets. When UUT interacts with TOPEN, TOPEN acts as a server. TSM has an option to configure the connection parameters: host name and communication port. Thanks to DSS capabilities it is possible to manage different TOPEN instances for testing one or more different UUTs or even projects.

To integrate TOPEN with SEE it has been required to define a new operation associated to a CASEML document, the Validation Test Specification (VTS) document type. This operation associates a CASEML node with the TSM invocation. VTS is defined by the ECSS e40B standard. The VTS CASEML schema is presented in figure 6. Two operation modes are available:

a) Standalone mode. TSM component is executed with no test procedure associated. Then user can write a new test procedure in TSM window or retrieve an existing test procedure from the TOPEN database (MIB).

b) Test procedure mode. TSM component is executed with a given test procedure. In this case, TSM receives a parameter with the CASEML subtree of the test procedure (TP) selected. It obtains the name (test procedure identifier) and the content of the TP from the subtree. Then it generates a query for obtaining a TP with the same name, and then tries to retrieve one stored in the TOPEN MIB if exists. This query is sent to Data Sources Server (DSS) component that executes it. The result is returned to the TSM. A document Validation Testing Report (VTR), see figure 6, is then updated, stored in SEE, managed by SEE as any other document. Similarly it will be possible to compile and/or to execute the TP from the TSM application. TSM requests TOPEN the execution of these operations. TOPEN returns the result of the compilation and/or execution of the specified TP that is visualized on the TSM window. The execution result of the TP can be stored in the specific document VTR of the SEE project. For that the result of the execution of the TP must be stored in the corresponding Caseml subtree within the VTR document. All these operations are summarised in figures 3 and 4.

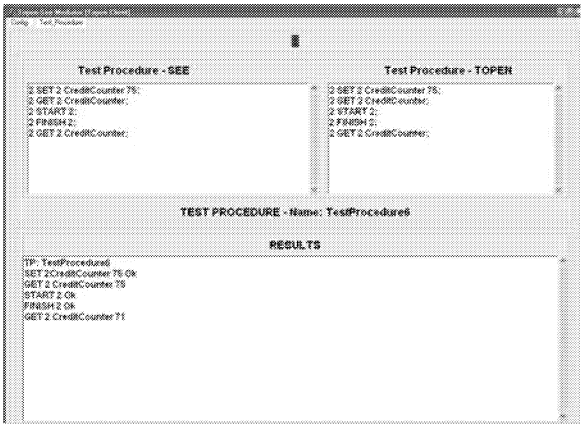


Figure 2. TOPEN-See Mediator main window

3. TEST DATA INTEGRATION AND MANAGEMENT

In principle there is no constrain in SEE to prevent it from supporting documentation based on any process model. ECSS has been the one actually implemented but SEE design does not prescribe that. SEE uses the Caseml language for describing the document structure (Caseml-schemas) and documents themselves (Caseml documents). An example of Caseml document can be seen in figure 6. Caseml-schemas not only define the valid structure for a type of project document. They also work as templates for documents. Thus, a type of document is created according to the template specified in the Caseml-schema. This way, elements and document structure (titles of chapters, sections, subsections, etc.) are created from the beginning with the initial default values.

SEE allows us establishing traces between different documents elements (nodes) within the same project. From systems validation point of view, possibility to establish traceability between requirements and validation tests is very interesting and rewarding. Thus, traceability can be used as development driver. Data integration between SEE and TOPEN can be achieved in two ways, as shown in figure 3:

1) Mediator invokes the TOPEN system. In this case TOPEN must be in execution.
2) Mediator gets access through DSS to the TOPEN MIB (database). TOPEN need not be executing, though MIB must be accessible. TOPEN will be in stand-by mode.

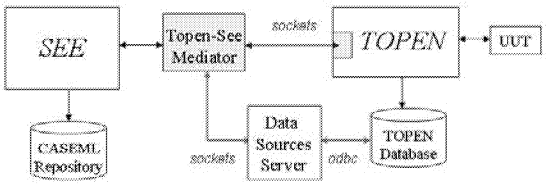


Figure 3. Data integration into SEE

For the first case, TSM is invoked from a test procedure element defined in the VTS document. TSM receives this test procedure as a Caseml subtree and scans this subtree obtaining the identifier and content of the test procedure. Then the user can compile and/or execute the TP. Figure 4 shows this operation. TSM sends the request to TOPEN by means of a socket based communication. TSM supports the configuration of the port and host values for the TOPEN in execution. After compiling or executing a TP, TOPEN sends the results to TSM in the same way. TSM generates a Caseml subtree from the received results, and inserts them within the SEE VTR Caseml tree.

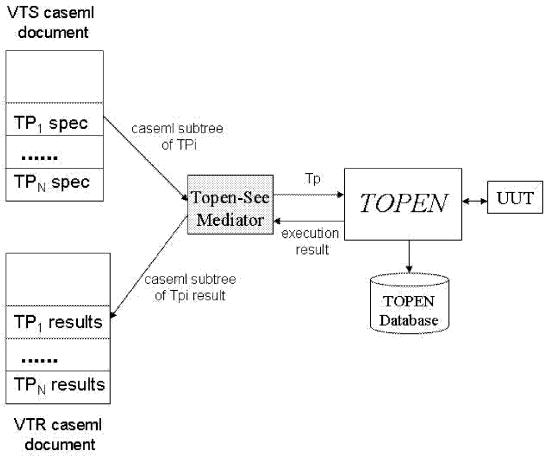


Figure 4. Test data integration using TOPEN execution

An interesting feature is that SEE can manage more than one project, and each project may require to test one or more UUTs. In the context of a project it could be of interest to execute validation tests on different UUTs. This scenario is supported. Each project can use an instance of TOPEN for the UUT validation.

The UUT can be a real system or a simulator. Simulators can be used as elements within the development process. The combination of really implemented systems and simulators can be used advantageously. In terms of TOPEN for each UUT to be tested a TOPEN instance will be created. To facilitate the process SEE documents may contain some configuration information such as the name of the data source. This data source must be registered in the repository of the DSS, and will correspond to a database of a concrete instance of TOPEN.

For the second case of data-integration, with TOPEN in standby, the TSM can retrieve stored test procedures from the TOPEN MIB database, and store them into the SEE VTS document. In the same way, it can retrieve the execution results of test procedures from the MIB database and store them into the corresponding SEE VTR document. Figure 5 summarises the process.

DSS supports obtaining data from the database of any instance of TOPEN. Furthermore, it provides total independence between a SEE project and a specific TOPEN database. DSS has a repository with registered data sources. For each data source it stores an alias that will be used from the DSS clients (in this case from TSM), the name of data source -usually the name of the database management system (DBMS) and the name of its ODBC interface are required-, a user and authorized password that provide the access.

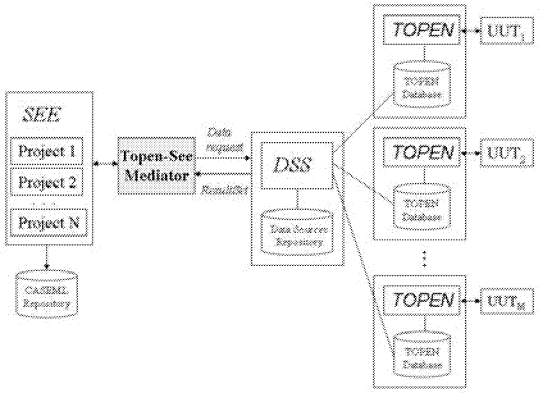


Figure 5. Test Data Access with TOPEN in standby

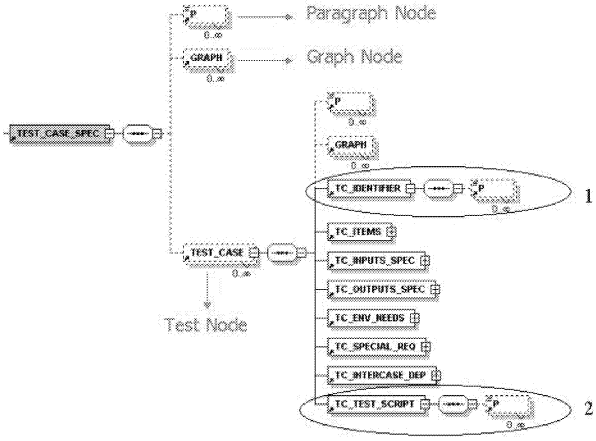


Figure 6. Software Validation Test Specification

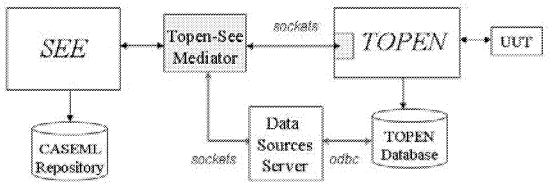


Figure 7. DSS Operational View

DSS server is designed and constructed to work with different DBMS, XML files or another type of files. The server may be located in a machine with an open communication port to communicate with DSS clients. TSM supports the configuration of the values of the machine and port to communicate with DSS server. Figure 7 shows a DSS graphical interface view, having got access to the information stored within database both for test specifications and executions, as long as the DBMS provides an access interface such as ODBC.

4. CONCLUSIONS AND FUTURE WORK

This paper has presented an approach deploy systems tests, as an important element they are of the system process development. The automation approach described within this paper plays an essential role, both for the integration of requirements and tests as for the facilities provided to manage tests. TOPEN concept was validated in the EU IST METSES project (Lazaro et al, 2002) and the integration with SEE was defined within the ESA/ESTEC DOBERTSEE project (Bollaín et al. 2003). The integration, together with some other issues, has been implemented after the end of this project and is being evaluated at present. In combination with simulation, the current approach provides a good support for incremental component-based development processes, useful for large systems.

In the future it is planned to integrate complementary approaches into the environment. These approaches include automatic test generation such as that achieved through the use of model checking techniques. Other line for the future will be to exploit the MIB database. This will be useful to implement alternative testing techniques such as statistical testing, and other issues such as system monitoring, diagnosis.

REFERENCES

ISO/IEC, ISO/IEC 15288:2002 Systems engineering -- System life cycle processes.

Alarcón P.P., Garbajosa J., Yagüe A., García, C., "Data Sources Server: An Approach to Heterogeneous Data Integration" In Proceedings of the ICEIS Conference, 2002.

Bollaín M., Alarcón P. P., Garbajosa J., Amador J., "A low-cost document-centric Software/System Engineering Environment" in Proceedings of the ICSSEA Conference, 2003.

ECSS, Space Engineering: Software (ECSS-E40B), ESA Publication, 2003.

Lázaro J.M., Peñafiel F., Mahillo M.A. and Gonzalo A. "Testing of Interconnected Machines". SV02 Workshop in ICSSEA Conference, 2002.

Garbajosa J., Alandes M., Mahillo M.A., Piattini M., "Assisting the Definition and Execution of Test Suites for Complex Systems", in Proceedings of the 7th ECBS Conference, 2000.

ACKNOWLEDGEMENTS

This work has been partially sponsored by project AGMOD TIC2003-08503. Ministerio de Educación, Ciencia y Cultura de España, and by the DOBERTSEE project, European Space Agency ESA/ESTEC Contract No. 15133/01/NL/ND

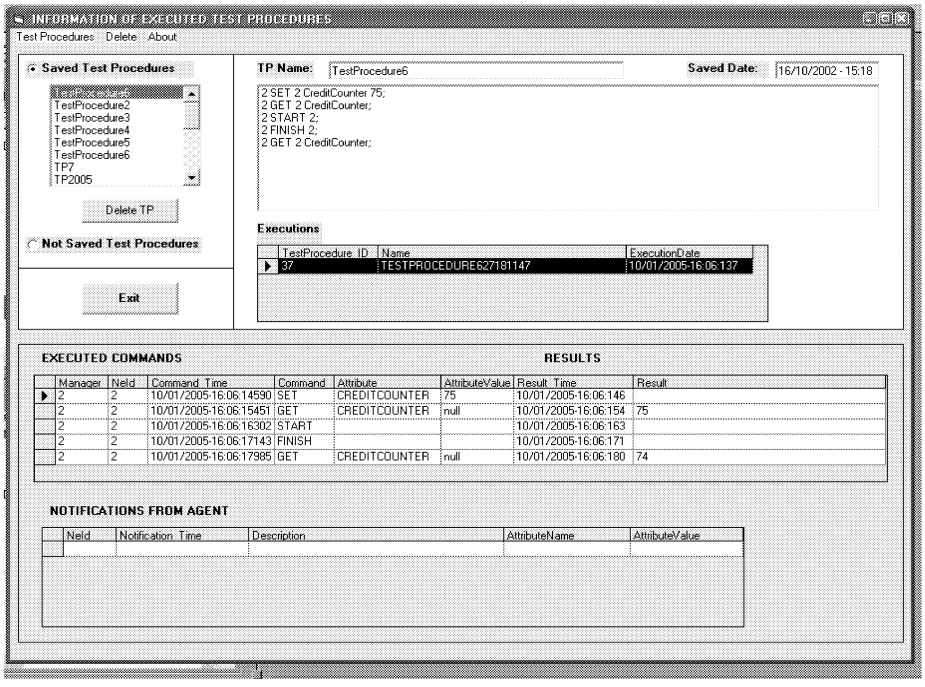


Fig. 7. View of the Test Procedure Executions

SOM VALIDATION METHOD FOR WASTEWATER TREATMENT

Hilario López and Iván Machón
Universidad de Oviedo.
Escuela Politécnica Superior de Ingeniería. Departamento de Ingeniería Eléctrica, Electrónica de Computadores y Sistemas.
Edificio Departamental 2. Zona Oeste. Campus de Viesques s/n. 33204 Gijón (Asturias). Spain
E-mail: hilario@isa.uniovi.es

KEYWORDS

Computer software, environmental science, estimation, validation, neural network.

ABSTRACT

Neural networks are widely used to supervise physical processes. These models must be validated to achieve a general approach. Therefore the validation of the model is an important task. In this paper a validation method for SOM training is outlined and a predefined criterion to determine the SOM map size is tested. After validation a software tool is developed to obtain the aerobic end-point detection of a sequencing batch reactor in a coke wastewater treatment plant, achieving the plant improvement.

PURPOSE OF THE PAPER

This work is part of the KNOWATER II project which is sponsored by ECSC. The contractors are Centro Sviluppo Materiali S.p.A., Corus RT&D, Betrieb Forschung Institut (BFI) and Universidad de Oviedo. The main objective of this project was the development of plant supervision techniques for implementation in wastewater treatment plants.
The present work was focused on verifying the method to determine the SOM map size and the optimum training data set for SOM validation. Once the model has been validated, a software tool was developed to supervise a biological wastewater treatment at a coke wastewater treatment plant (CWTP) of Arcelor in Avilés (Spain). The estimation of the current process state is calculated. Thus, important on-line knowledge is obtained.

COKE WASTEWATER TREATMENT PLANT

The coke wastewater treatment plant (CWTP) consists of three zones: ammonia stripping towers, homogenization tank and biological reactor. The instrumentation of the CWTP was selected to control some key process variables (pHs and dissolved oxygen concentration) and to achieve the process monitoring by means of the designed software tool and its built-in AI technique. Firstly, the influent stream is treated applying vapor in order to remove the ammonia in the stripping towers. A pHmeter is installed and a PID, which controls a caustic soda dosing pump, controls the pH. The set-point is fixed at 12.
The second stage is the homogenization tank. At this point, a second pHmeter is established for the purpose of controlling pH by means of a PID controller with a set-point equal to 7. Sulphuric acid is added in order to neutralize the pH.
Finally, the wastewater is treated biologically in a Sequencing Batch Reactor (SBR). The dissolved oxygen concentration is controlled by a third PID. Air is pumped into the reactor and a valve is regulated. The set-point is 3-5 mg/l.
An initial off-line study of the process was done and the PID controller output of the oxygen closed loop was connected and registered as one of the process variables to train the SOM network (López and Machón 2004a).

SOM

Self-Organizing Map (SOM) is a useful tool for process supervision and was used to construct a model that can be used as a pattern (Kohonen 2001). The SOM consists of a regular lattice typically defined in a two dimensional space composed of several neurons placed in the nodes of the lattice. SOM training implies assigning a set of coordinates in the input data space (prototype vector) to each neuron. Thus, the input data set is mapped into the output space.
The present study was carried out using the SOM toolbox (Vesanto et al. 1999) developed at the HUT (Helsinki University of Technology). The steps taken to analyze the data are outlined in a previous work (López et al. 2003). Firstly, the most significant process variables are selected. These variables are described in table 1. Secondly, the data were normalized to make SOM treat them in the same way. After normalizing a batch training algorithm is carried out. The results are visualized in several ways: Interneuron distance matrix (Umatrix) that shows in gray or color levels the mean distance of each unit to its closest neighbors; the component planes that display the value of a given input variable throughout the whole data set in the 2D lattice; the best clustering structure that allows the main working zones of the process to be visualized (López and Machón 2004b).

Table 1. Training Variables

Name	Description
OXYGEN	Dissolved oxygen concentration (mgO_2/litre)
CONTROLLER OUTPUT	Output of the PID controller OUTPUT of the oxygen closed loop (0-100)
TEMPERATURE_SBR	Temperature in the SBR (C)

VALIDATION METHOD

The SOM network must achieve the topology preservation of the data. The topographic error check this property and can be calculated (Kiviluoto 1996) by equation (1) as the proportion of sample vectors for which two best matching neurons are not adjacent. N is the number of samples, x_k is the kth sample of the data set and $u(x_k)$ is equal to 1 if the first and second best matching neurons of x_k are not adjacent neurons, otherwise zero.

$$e_t = \frac{1}{N}\sum_{k=1}^{N} u(x_k) \qquad (1)$$

Moreover, the prototype vectors approximate to the data set trying to substitute a data vector for a prototype vector of the SOM. A consequence of this approach is the quantization error. Equation (2) is usually used to calculate the average quantization error over the whole data set. N is the number of samples, x_i is the ith data sample and m_b is the prototype vector of the best matching neuron for x_i.

$$e_q = \frac{1}{N}\sum_{i=1}^{N} \| x_i - m_b \| \qquad (2)$$

The SOM toolbox uses equations (3) and (4) to determine the output space size. An objective of this paper is the checking of equations (3) and (4) to validate the SOM training. The number of neurons of the output space is determined by equation (3). M is the number of neurons and N is the number of samples of the training data. On the other hand, the ratio between the number of rows n_1 and the number of columns n_2 of the output space is calculated according to equation (4), e_1 is the highest eigenvalue and e_2 is the second one.

$$M = 5 \cdot \sqrt{N} \qquad (3)$$

$$\frac{n_1}{n_2} = \sqrt{\frac{e_1}{e_2}} \qquad (4)$$

Five data sets showed in figure 1 and figure 2, which correspond to the aerobic phase of the SBR, are available to carry out the validation of the model. Each sample is the mean values of the process variables for 8 minutes and 20 seconds. The objective is to find out the model that minimizes the quantization and topographic errors. A specific data set and an optimum map size must be selected. The method of validation can be summarized in the following steps (Machón and López 2004):

1) A data set or pattern p_i is chosen to train the network. The data are normalized to a distribution with zero mean value and unitary variance.

2) Batch training is carried out on the SOM map whose sidelengths are calculated by means of equations (3) and (4) using pattern p_i as training data.

3) Once the trained model is obtained, the topographic and quantization errors are calculated for the remaining patterns p_j which have not been used during the training. These patterns must also be previously normalized.

4) The size of this trained map is increased and reduced respecting the proportionality of its sidelengths (width and length). Once the size has been modified, the neural network is again trained using pattern p_i.

5) The third and fourth steps are repeated for different map sizes.

6) Steps 1 through 5 are repeated for the remaining patterns p_j, assuming each of these the role of pattern p_i.

Several map sizes have been trained using the five patterns. The mean values of the errors over the available patterns in function of the map number are shown in figure 3 and figure 4 using each pattern as training data. The sidelengths of the trained maps for each utilized pattern are showed in table 2. It can be seen that the larger the map size the lower the quantization error but the higher the topographic error because the neural network folds to reduce the quantization error. Moreover, the larger the map size the higher the computational cost. Therefore, there is compromise between the increase of the topographic error and the reduction of the quantization error. The model whose sidelenghts have been calculated by means of equations (3) and (4) correspond to a horizontal axis value equal to 6 (map number equal to 6). The quantization error has been reduced and the topographic error has been incremented not very much, as can be seen in figure 3 and figure 4. The results verify that the criterion to determine the map sidelengths is a reasonable optimum solution. The selected model, which is integrated into the software tool, was trained using pattern 1 and the demonstrated criterion, equations (3) and (4), to determine the output space size.

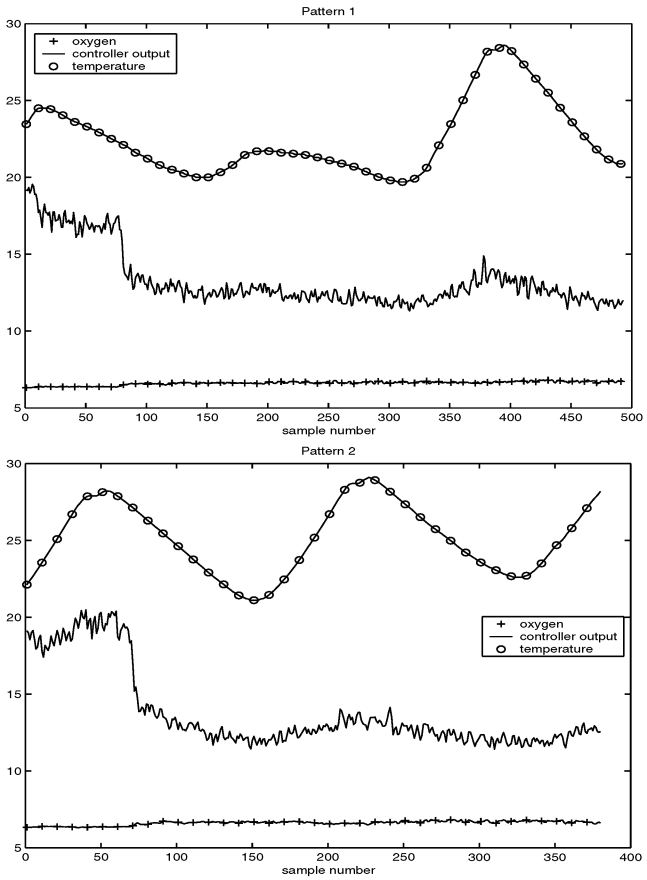


Figure 1. Pattern 1 and 2

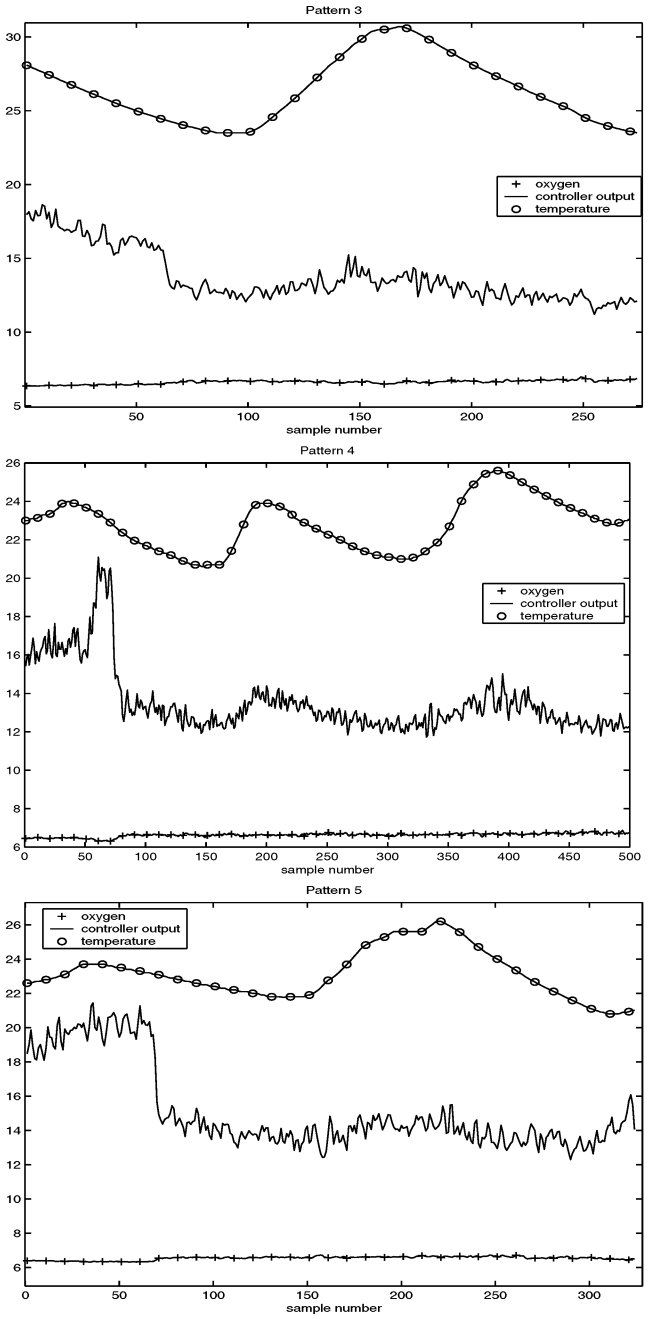


Figure 2. Pattern 3, 4 and 5.

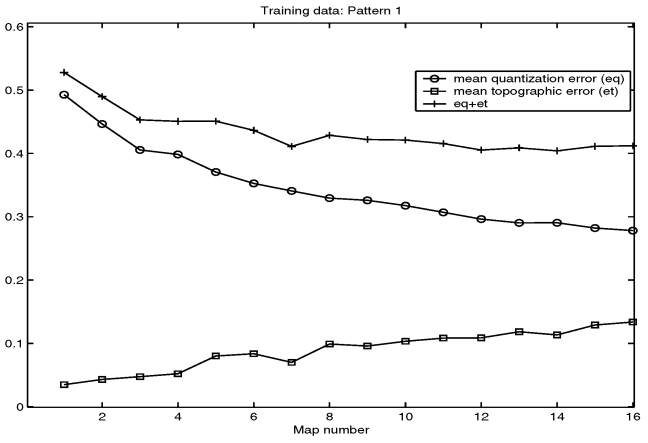


Figure 3. Mean of Errors using Pattern 1 as Training Data

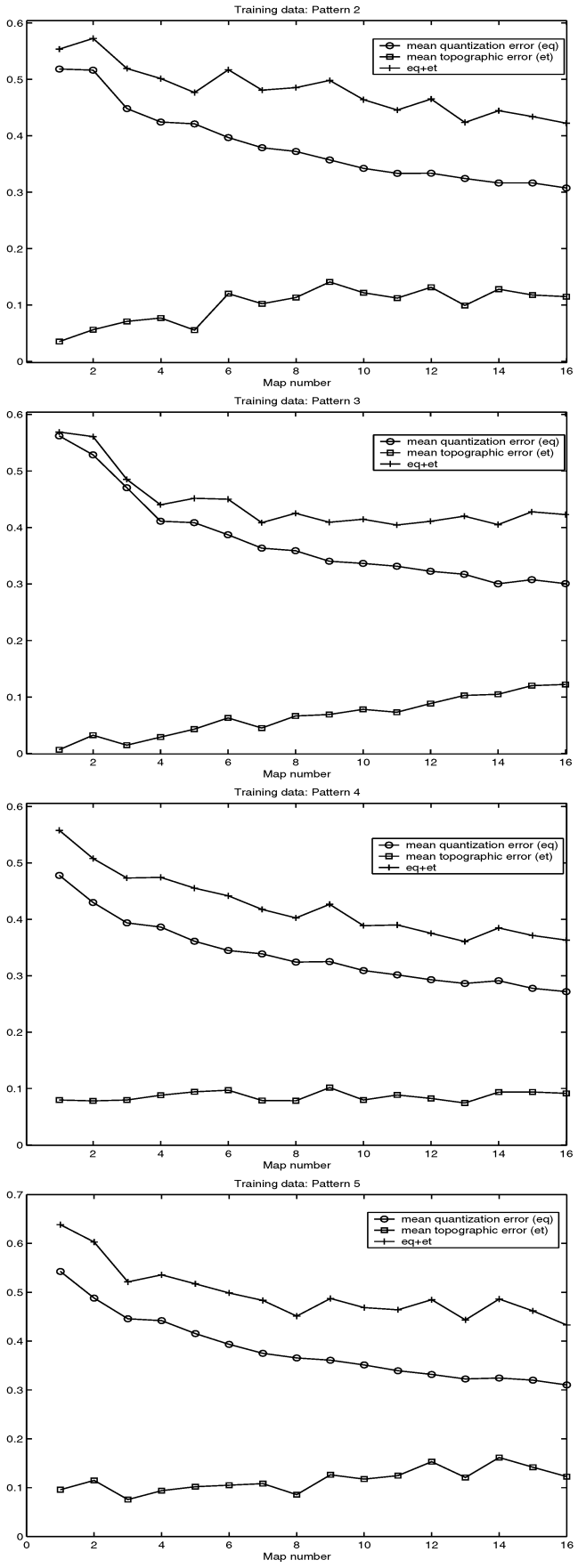


Figure 4. Mean of Errors using Pattern 3, 4 and 5 as Training Data

Table 2. Trained Map Sizes

Map number	Map size (training data: Pattern 1)	Map size (training data: Pattern 2)	Map size (training data: Pattern 3)	Map size (training data: Pattern 4)	Map size (training data: Pattern 5)
No 1	7 x 4	7 x 4	6 x 4	7 x 4	6 x 4
No 2	8 x 4	8 x 5	7 x 4	8 x 5	7 x 5
No 3	10 x 5	10 x 6	8 x 5	10 x 6	8 x 6
No 4	11 x 6	11 x 6	10 x 6	11 x 6	9 x 6
No 5	13 x 6	13 x 7	11 x 6	13 x 7	10 x 7
No 6	14 x 7	14 x 8	12 x 7	14 x 8	11 x 8
No 7	15 x 8	15 x 9	13 x 8	15 x 9	12 x 9
No 8	17 x 8	17 x 10	14 x 8	17 x 10	13 x 10
No 9	18 x 9	18 x 10	16 x 9	18 x 10	14 x 10
No 10	20 x 10	20 x 11	17 x 10	20 x 11	15 x 11
No 11	21 x 11	21 x 12	18 x 11	21 x 12	17 x 12
No 12	22 x 11	22 x 13	19 x 11	22 x 13	18 x 13
No 13	24 x 12	24 x 14	20 x 12	24 x 14	19 x 14
No 14	25 x 13	25 x 14	22 x 13	25 x 14	20 x 14
No 15	27 x 13	27 x 15	23 x 13	27 x 15	21 x 15
No 16	28 x 14	28 x 16	24 x 14	28 x 16	22 x 16

PROPOSED AI TECHNIQUE

The developed software tool carries out a clustering process which consists of a two-stage procedure (Vesanto and Alhoniemi 2000). Firstly, the prototype vectors are obtained training the data of the aerobic phase using a SOM algorithm and then clustering them using a K-means algorithm, see (McQueen 1967; Dubes and Jain 1976). Ten clustering structures were obtained varying the predefined number of clusters.

Finally, the best clustering structure between the ten structures, which have been obtained from the K-means algorithm, is selected using the Davies-Bouldin index (Davies and Bouldin 1979). This index searches the model that minimizes the withincluster distance and maximizes the between-clusters distance. The Davies-Bouldin index is suitable for evaluation of K-means partitioning, because it gives low values indicating good clustering results for spherical clusters. Figure 5 shows the Davies-Bouldin index after being applied to the data from the aerobic stage of the treatment. The best clustering corresponds to a number of two clusters and can be seen in figure 5. Once the best clustering has been selected, the clustering structure must be visualized in order to label the input data space onto the 2D-lattice. This allows the visualization of the different features of the process.

The best clustering has been projected onto the SOM and a new component plane has been obtained that is displayed in figure 6. Cluster 1 corresponds to the first hours of the aerobic treatment where the values of the controller output are high due to the high chemical oxygen demand (COD). During this period the biological activity is high and the toxic substances are eliminated by means of the cellular metabolism, whereas cluster 2 represents the data collected after this high biological activity where the values of the controller output are lower because the COD has decreased. If this is the state of the treatment plant, the biological treatment of the aerobic stages can be finished improving the capacity of the plant.

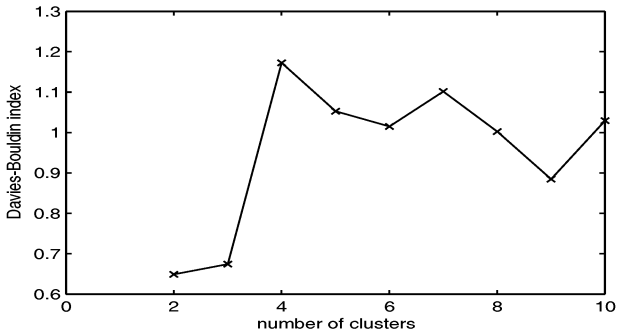


Figure 5. Results of Davies-Bouldin Index

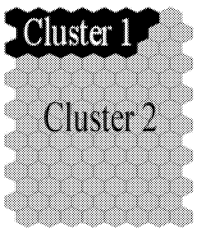


Figure 6. Best Clustering Structure

SOFTWARE TOOL

The developed software is running in a PC station that is connected to a data acquisition interface by means of Ethernet connection and TCP/IP protocol (MODBUS protocol in particular). The proposed AI techniques are integrated into this application to achieve the process monitoring and the process state estimation. Data from the aerobic stage is collected on-line automatically to train a SOM network. The plant operator can visualize the latest SOM network that corresponds to the latest aerobic treatment cycle of the plant, viewing the correlation between the process variables and the data classification

can be obtained. The estimation of the current process state is also calculated by the validated pattern (pattern 1).

The program improves the plant operation with the estimation of a key variable (oxygen uptake rate) by means of the controller output dynamic from the dissolved oxygen control (Cavalcanti et al. 1999). The oxygen uptake rate is one of the most important variables in the wastewater biological treatments and is the associated demand which is necessary to oxidize the organic substrate by the heterotrophic biomass, see (Jeppson 1993; López et al. 2004). The estimation of the time of the main activity of the treatment (aerobic phase end-point) achieves operating cost savings and increases the plant performance; see (Yu et al. 2001; Paul et al. 1998; Cho et al. 2001; Andreottola et al. 2001).

The training data set must only contain the samples of the aerobic stage and is determined by the mean value of the controller output because this signal can be considered as a key variable to estimate the states of the treatment, see (López and Machón 2004a; López and Machón 2004c).

The results of the latest SOM network can be visualized in figure 7 and they correspond to the latest aerobic stage of the wastewater treatment. They are the U-matrix, the component planes and the best clustering structure. The U-matrix shows the distances between prototype vectors of neighboring neurons. Thus, a group of neurons, which is too concentrated, implies a cluster. Distanced zones of neurons must separate several clusters. Each component plane shows the value of each neuron to estimate the data variable of the input space. It is useful to determine the several zones where the variable value is high or low and to observe any correlation or relationship between process variables. The correlations among variables can be observed, for example, between the controller output and the oxygen concentration. Also there is a correlation between the controller output and the temperature in the reactor due to the higher the temperature the lower the dissolved oxygen concentration and the controller must compensate this effect. As mentioned above, the best clustering structure is composed of 2 clusters and is calculated by means of the Davies-Bouldin index. A cluster corresponds to HIGH COD and the other is the LOW COD.

The cycles of the biological treatment at the sequencing batch reactor can be clearly observed in figure 9 and figure 10. The higher values correspond to the anoxic stage when the controller output is saturated and equal to 100%. The rest of the data corresponds to the aerobic stage (including sedimentation).

An important aspect appears: the end-point of the aerobic reaction. This end-point detection can be used to finalize the aerobic stage and in this way the duration of the cycle is shorter increasing the operating capacity of the plant. Moreover, the oxygen consumption is high, influenced by the temperature in the reactor as could have been expected. The duration of the cycle was initially 48-72 hours one year ago, see figure 9, and it has been reduced to 24 hours as is showed in figure 10. In this way the operating capacity of the plant has been increased by reducing the retention time.

In figure 8 the process state is estimated projecting the current values onto a SOM network by means of standing out the best matching neuron from the rest of the neurons. This SOM network is the validated pattern (pattern 1). The projection is carried out onto the component planes and the best clustering structure. Thus, important on-line knowledge is obtained and the end of the main biological activity can be identified.

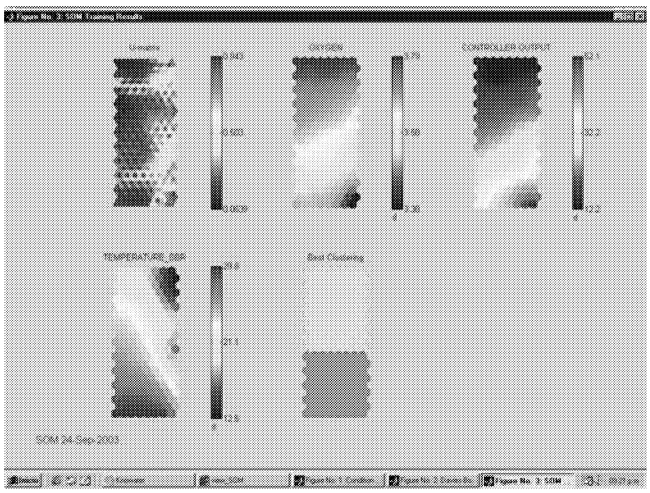


Figure 7. SOM Training Results

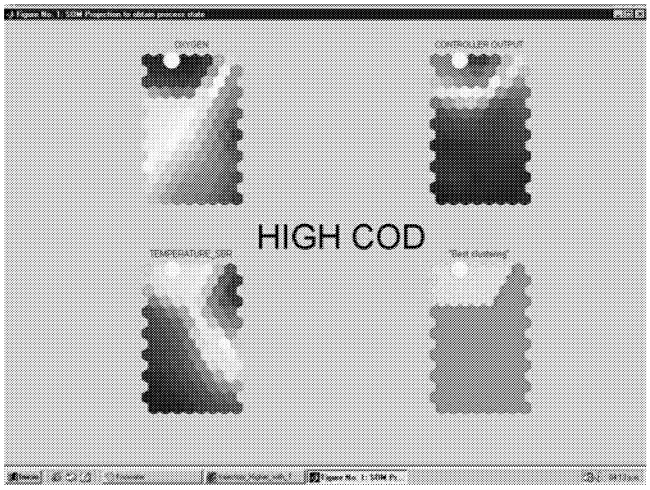


Figure 8. Process State Estimation

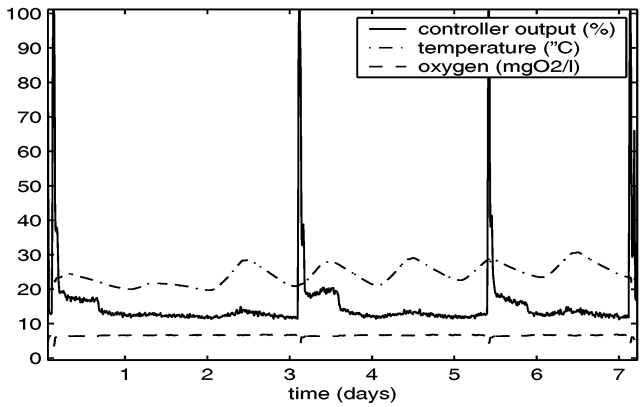


Figure 9. Process Values of one year ago

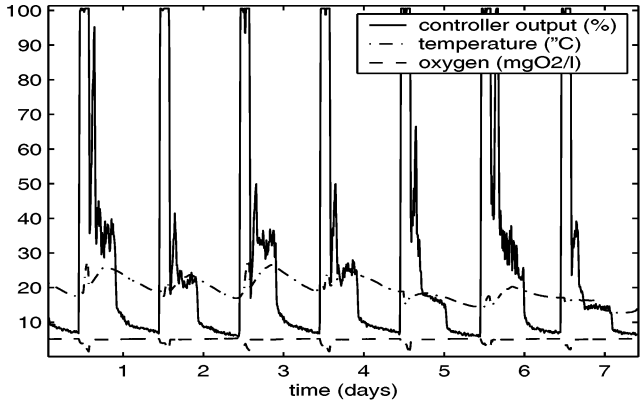


Figure 10. Current Process Values

CONCLUSIONS

The results verify that the criterion, which is used by the SOM toolbox to determine the SOM map size, is a reasonable optimum solution for SOM training. Also a procedure is established to determine the training data set to be used for SOM training. After validation a software tool was developed to supervise a coke wastewater treatment plant (CWTP) and is a stand-alone application which is composed of the data acquisition system from the CWTP and the proposed AI technique. The data set of the aerobic stage is collected to train automatically a SOM network. The data classification is obtained using K-means algorithm as partitive clustering algorithm and Davies-Bouldin index for clustering validation. The estimation of the current process state can be assigned calculating the best matching neuron that corresponds to the current process values. The end-point of the aerobic reaction can be detected by this AI technique. So, operating cost savings are achieved and the plant performance is increased reducing the total retention time from 48-72 hours to 24 hours.

ACKNOWLEDGMENTS

Warm thanks are expressed to the financial support from the Commission of the European Communities, European Coal and Steel (ECSC), supporting KNOWATER II project "Implementation of a Knowledge Based System for Control of Steelworks Waste Water Treatment Plant" with 7210-PR- 234 as agreement number.

REFERENCES

Andreottola G.; P. Foladori and M. Ragazzi. 2001. "On-line control of a sbr system for nitrogen removal from industrial wastewater," *Water Science and Technology*, vol. 43, no. 3, pp. 93–100.

Cavalcanti S.; G. Singh; A. Cornelius and R. Silvrio. 1999. "Feedback control method for estimating the oxygen uptake rate in activated sludge systems," *IEEE Trans. Instrum. Meas.*, vol. 48, no. 4, pp. 864–869.

Cho B.; S. Liaw; C. Chang; R. Yu; S. Yang and B. Chiou. 2001. "Development of a real-time control strategy with artificial neural network for automatic control of a continuous-flow sequencing batch reactor," *Water Science and Technology*, vol. 44, no. 1, pp. 95–104.

Davies D. and D. Bouldin. 1979. "A cluster separation measure," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 1, no. 2, pp. 224–227.

Dubes R. and A. Jain. 1976. "Clustering techniques: The user's dilemma," *Pattern Recognition*, vol. 8, pp. 247–260.

Jeppson U. 1993. "On the verifiability of the activated sludge system dynamics," Licentiate thesis, Sweden.

Kiviluoto K. 1996. "Topology preservation in self-organizing maps," in *IEEE International Conference on Neural Networks*, vol. 1, pp. 294–299.

Kohonen T. 2001. *Self-Organizing Maps*. New York: Springer-Verlag.

López H.; I. Machón and S. Roces. 2003. "Waste treatment monitoring using self-organizing map and condition achievement maps," in *Proc. IFAC 5th Symposium on Intelligent Components and Instruments for Control Applications*, Aveiro.

López H. and I. Machón. 2004a. "Biological wastewater treatment analysis using som and clustering algorithms," in *Proc. 12th Mediterranean Conference on Control and Automation*, Kusadasi.

López H. and I. Machón. 2004b. "Self-organizing map and clustering for wastewater treatment monitoring," *Engineering Applications of Artificial Intelligence*, vol. 17, no. 3, pp. 215–225.

López H. and I. Machón. 2004c. "An introduction to biological wastewater treatment explained by som and clustering algorithms," in *Proc. IEEE International Symposium on Industrial Electronics*, Ajaccio.

López H.; E. Marañón; S. Roces; I. Machón; J. Rodríguez and I. Vázquez. 2004. "Modelling of an activated sludge plant using simulink," in *Proc. 16th International Congress of Chemical and Process Engineering*, Prague.

Machón I. and H. López. 2004. "An application for on-line control of a sequencing batch reactor," in *Proc. IFAC Workshop on Modelling and Control for Participatory Planning and Managing Water Systems*, Venice.

McQueen J. 1967. "Some methods for classification and analysis of multivariate observations," in *5-th Berkeley Symposium on mathematics, Statistics and Probability*, no. 1, pp. 281–298.

Paul E.; S. Plisson-Saune; M. Mauret and J. Cantet. 1998. "Process state evaluation of alternating oxic-anoxic activated sludge using orp, ph and do," *Water Science and Technology*, vol. 38, no. 3, pp. 299–306.

Vesanto J.; E. Alhoniemi; J. Himberg; K. Kiviluoto and J. Parviainen. 1999. "Self-organizing map for data mining in matlab: the som toolbox," *Simulation News Europe*, pp. 25–54.

Vesanto J. and E. Alhoniemi. 2000 "Clustering of the self-organizing map," *IEEE Trans. Neural Networks*, vol. 11, no. 3, pp. 586–600.

Yu R.; S. Liaw; B. Cho and S. Yang. 2001. "Dinamic control of a continuous-inflow sbr with time-varying influent loading," *Water Science and Technology*, vol. 43, no. 3, pp. 107–114.

SIMULATION-BASED FRAMEWORK FOR MAINTENANCE OPTIMIZATION

Lust Thibaut
Roux Olivier
Riane Fouad
Production and Operations Management, CREGI - FUCaM
151, chaussée de Binche
B-7000 Mons, Belgium
E-mail: {lust|roux|riane}@fucam.ac.be

Dehombreux Pierre

Mechanical Service of Engineering, FPMs
53, Rue du Joncquois
B-7000 Mons, Belgium
E-mail: pierre.dehombreux@fpms.ac.be

KEYWORDS
Discrete simulation, Maintenance strategies, Stochastic optimization.

ABSTRACT

Maintenance managers have to optimize the availability of their production machines to ensure production throughputs at lowest expenditure. They aim at developing policies for their complex production systems in order to maximize production availability and/or to minimize maintenance average costs. In this article we develop a modelling approach that combines optimization and simulation and makes it possible to capture the dynamic behaviour as well as the reliability of the multi-component systems. The use of simulation provides an easy evaluation for the system's availability as well as the average cost per unit of time with regard to different maintenance policies. The integration of optimization techniques helps to find the optimal various parameters of these strategies. The paper is devoted to the presentation of the architecture of the decision-making framework.

INTRODUCTION

Maintenance activities influence the entire production process, from product quality to on-time delivery. Poor maintenance procedures can cost millions of Euros in repairs, end up with very poor quality and substantial lost of production. Meanwhile good maintenance practices can cut production costs immensely. That is why the maintenance function should no longer be considered as a source of cost but as a strategic asset and thus a profit centre.
Maintenance managers deal with manufacturing systems that are subject to deterioration and failure. They are often confronted to a rethink of the way they should deal with maintenance policies and maintenance organization issues. One of their major concerns is the way to allocate the right budget to the appropriate equipment or component while minimizing the total expenditure and maximizing the effective availability of production assets.
A maintenance policy can be defined as a set of maintenance actions to be executed according to various schemes and structures. These actions can be divided into different categories: Failure Based Maintenance, Use Based Maintenance, Detection Based Maintenance, Condition Based Maintenance and Design-Out Maintenance (Naert and Van Mol 2002).

Current maintenance policies are time oriented and are still based on expected cost or availability optimization rather than being optimized with risk-cost considerations.
In real life situations, managers face a complex decision making problem when they consider the availability aspect as well as the economic issue of their maintenance's activities. In fact they continuously look for the way to improve the availability of their production machines in order to assure given production throughputs. But they want to achieve this goal at the lowest expenditure.

The aim of this research is to provide such decision makers with a framework that makes it easy the design, modelling and optimization of maintenance strategies.

MAINTENANCE STRATEGIES

A lot of work has been done in the field of maintenance policies of deteriorating equipment. A maintenance strategy may be defined as a decision rule which establishes the sequence of maintenance actions to be undertaken according to the degradation level of the system and with regard to the acceptable exploitation thresholds. Each maintenance action consists of maintaining or restoring the system in a specified state using the appropriate resources. A cost and duration are incurred to execute each maintenance action (Jamali et al. 2000).
In a previous contribution (Jamali et al. 2002), we have gone through different maintenance strategies (Barlow and Proschan 1965, Cho and Parlar 1991, Nakagawa 1979, Pierskalla and Voelker 1976, Sherif and Smith 1981). They concern the replacement of systems subject to random failures and whose states are known at all time. All the studied policies are governed by analytical models that make it possible to evaluate over an infinite horizon the associated performances under a series of hypothesis (Ait-Kadi et al. 2002). These strategies are distinguished by the nature and the action sequel that they suggest, by the selected performance criteria, by the deterministic or stochastic character of the parameters that they take into account and by the fact that the system is considered as a sole entity or as a system constituted of many components which state may be known at all time or after inspection.
Using an analytical formulation, one can model the considered maintenance strategy using its characteristic parameters and decision variables to describe the technical as well as the economical objectives to optimize. If one succeeds to solve to optimality such analytical models, he can establish the existence and the uniqueness conditions of an optimal strategy. He can also derive sensibility analysis.

The major inconvenient of such approaches is that it becomes difficult to develop models and very complex to solve them if one wants to consider other factors that have significant impact on the system's behaviour and performance. This fact has led us to explore simulation's possibilities in order to handle the situation and efficiently evaluate maintenance strategies.

SIMULATION-BASED APPROCH FOR MAINTENANCE OPTIMIZATION

In the manufacturing and service industries context, many problems are of such complexity that managers must built their decisions on aid support systems that are adequate to provide satisfactory answers to their strategic, tactical or operational questions. Most of these operational problems can be modelled as complex discrete systems, whose related decision making requirements are approached using simulation techniques.

It is the objective of this research to develop a simulation-based decision support approach that can operate for designing maintenance strategies for complex production systems. In particular, it is concerned with the development of a unified, graphical framework that makes it possible to the decision maker to, first, understand and then model the dynamics of the system under consideration and, second, to design and optimize the appropriate maintenance policy. A key element of such a framework is the development of a graphical language that enables the automatic code generation for simulation purposes and optimization analysis.

To handle the elaboration of maintenance strategies for complex system one has to specify the structure of the system, its logical organization and its decision parameters (preventive and corrective costs, preventive periodicity, etc.). The models of the system as well as the data will be developed following a unified methodology that leads to the specifications of the real system. Then, detailed level analysis that integrates different scenarios' comparison is derived. The methodology shall use a systemic and hierarchical approach.

The design and the development of an open environment that includes concepts, methodologies, languages, and solving engines are presented here. The overall objective of our research is the integration of several tools in a unified manner (cp. Figure 1). The unification aspect of the framework relies on the application of the same set of concepts at different levels of the modelling and solving process, from the modelling methodology to the solving engines.

The simulation language used in the framework is RAO, based on the RAO (resources - actions - operations) method (Artiba et al. 1998), which uses modified production rules for describing complex discrete systems and processes.

In the RAO method (cp. Figure 2), production equipment can be modelled as a complex discrete system using resources and performing operations. The resources are depicted in a database. The set of operations (actions) fulfilled by the resources are described by the modified production rules in a knowledge base. Unlike traditional production rules, the modified ones make it possible to describe the dynamics of the system thanks to the temporal specifications and dependence relations between activities.

To reproduce the functioning process of a Complex Discrete System (CDS) in the RAO simulator, we need to describe the concurrency of the irregular events and the way they influence the realization of the different actions.

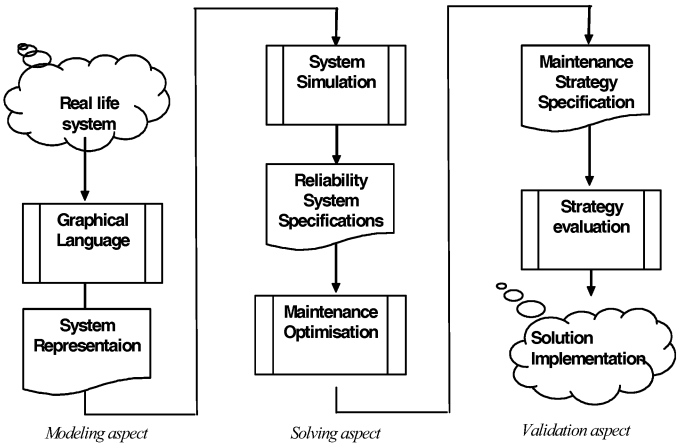


Figure 1: A unified approach for solving decision making problems

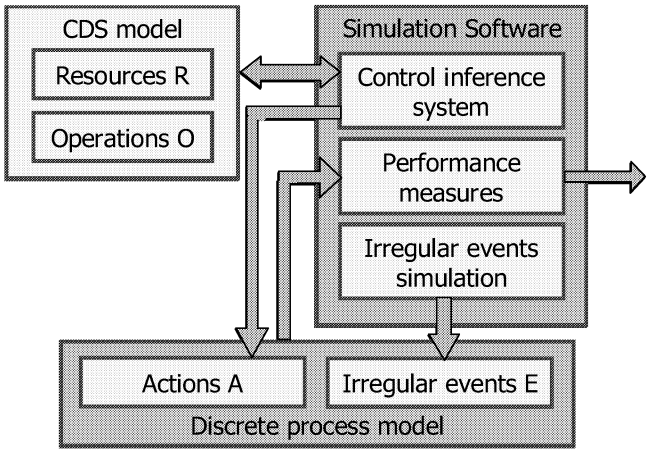


Figure 2: principles of RAO running

MODELING MAINTENANCE POLICIES

We can develop simulation models for different maintenance strategies. In order to illustrate the approach we detail hereafter the implementation of the Modified Bloc Replacement Policy (MBRP) using RAO simulation language (Jamali et al. 2004).

Let us consider a single component system that is subject to random failures. Each time the system breaks down we replace it by a new one. On the other hand, when performing preventive maintenance, we replace the system only if its age is greater than a given threshold b. These preventive actions are scheduled at given dates kT (k = 1,2,3,…). The above described scheme is the one called the Modified Bloc Replacement Policy. It was suggested in order to improve the Bloc Replacement Strategy. The Achilles heel of this later lies in the fact that components are changed preventively even if they are almost new.

Maintenance manager who chooses to implement this policy looks for the replacement period T and the age threshold b that maximize the steady-state availability of the system or

that minimize the expected cost per unit of time. A detailed simulation model of the periodic and conditional maintenance strategies has been developed and implemented using RAO simulator (Jamali et al. 2004).

We can improve this strategy in order to integrate risk analysis as a condition for maintenance policy elaboration (Bardey et al. 2003, Riane et al. 2004). This strategy follows the same scheme as MBRP. The difference lies on the preconditions to be satisfied in order to fulfil a preventive action. The component life cycle is punctuated by a sequence of failure events and related preventive actions. The single component system swings between a stopped state (broken down or to maintain) and a ready state. The state transition is dictated by the moments of breakdown and preventive maintenance. Failure occurs if the generated moment of breakdown is less or equal to the remaining time-to preventive action. Otherwise, the system is still working but can be maintained preventively.

In this later case, our improved policy suggests to realized the preventive action only if the value of risk, denoted by P_D, is lower than the level accepted by the decision maker (captured in the model by accepted_risk_level). A mathematical analysis is then necessary to evaluate the value risk based on the computing of the failure's probability between k and $k+T$, knowing that component has survived until k. This involves evaluating integrals of the probability density function f from k to $k+T$. We use a numerical approach when density functions are not suitable for integral computation which is the case of the normal distribution.

The model's diagram is depicted in Figure 3 thanks to ALIX modelling formalism that is suitable for use with RAO simulator (Pichel et al. 2003).

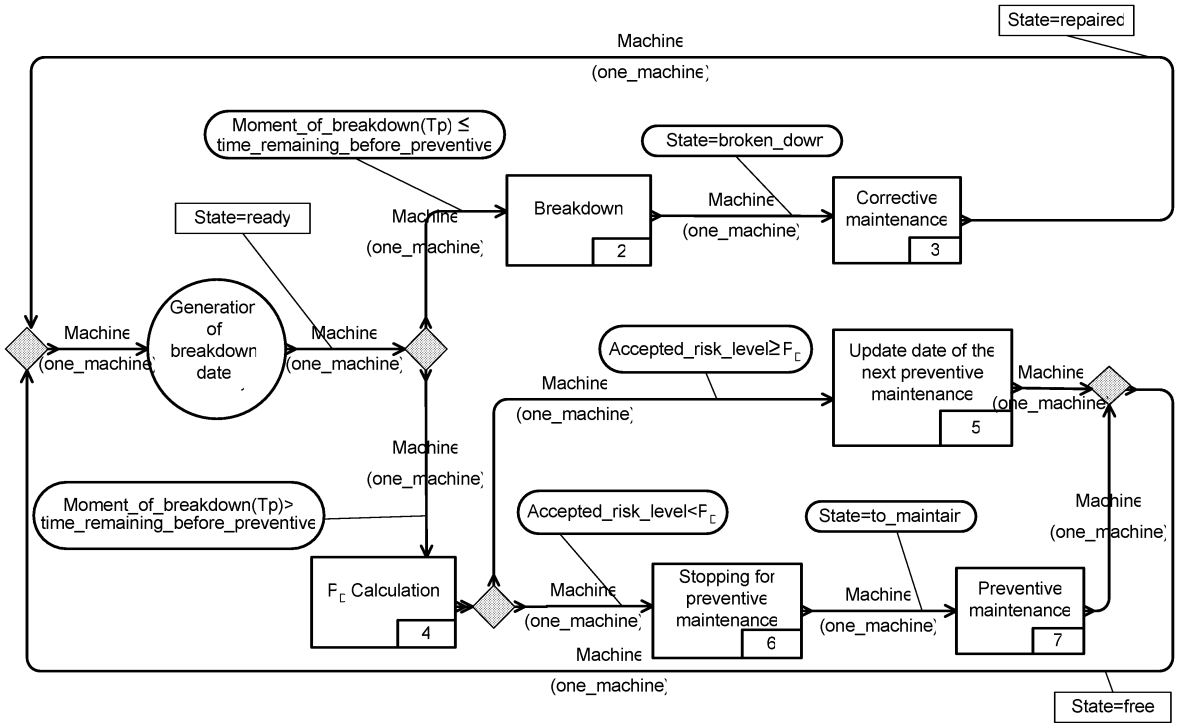


Figure 3: Simulation model's flow chart of the Modified Bloc Replacement Policy

SIMULATION-BASED FRAMEWORK

The policy described above can easily be modelled and implemented thanks to the developed user-friendly interface that supports our framework (cp. Figure 4). It allows the user to easily model its system using block diagram formalism. For each system's component, the user specifies the reliability function, the cost and the time of the preventive and corrective maintenance, the quality of the maintenance actions and the mode of detection of the component's failure. He can also choose the strategy of maintenance to apply to each component. All the system specifications and the appropriate data are saved in a XML format.

The simulation of the dynamics of the system is then possible thanks to the use of RAO simulator. All the necessary models are generated automatically by the framework, so that the user should not write any code of simulation.

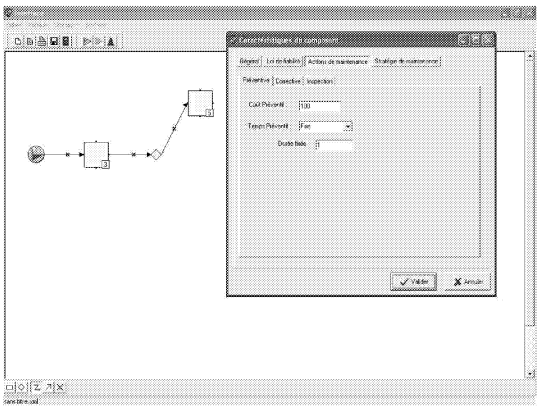


Figure 4: Simulation Maintenance Interface

Simulation makes it possible to derive the reliability distribution and the dynamic behaviour of the multi-component system over a long period.
The reliability of the system is represented thanks to the diagram of distribution of failure's times of the system (cp. Figure 5).
The dynamic behaviour of the system appears in the form of Gantt diagrams illustrating the actions of maintenance undertaken on each component. One can also see the trend of average maintenance cost per unit of time and the availability of the system along the simulation period (cp. Figure 6). Other performance indicators are also available as the number of preventive or corrective maintenance, the total cost of preventive or corrective maintenance, the period of unavailability of the system or components, etc.

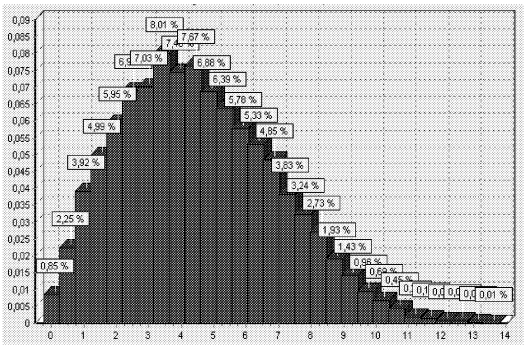


Figure 5: Distribution of failure's times

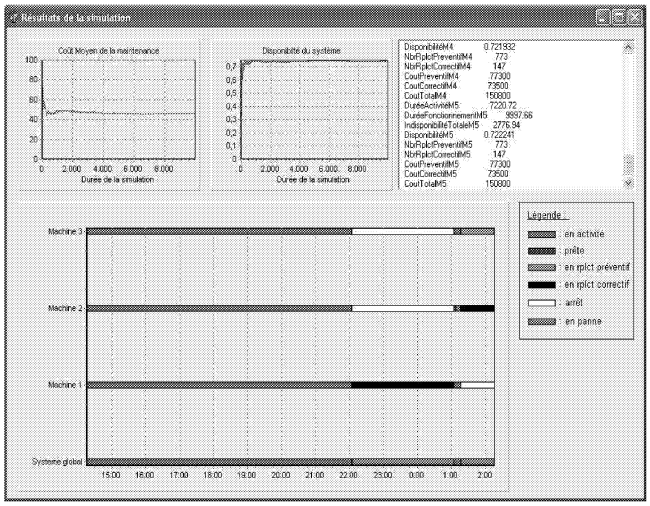


Figure 6: Simulation results presentation

The user has also the possibility to obtain the optimal parameters of the selected maintenance strategies applied to the components of the system. For example, he can compute the parameters T and b for the strategy MBRP presented in this article so as to minimize the cost of maintenance (cp. Figure 7) or to maximize the availability of the system (cp. Figure 8).
To determine the optimal values of the parameters, we use a coupling between the model of simulation and a module of optimization. The coupling, described in Figure 9, allows to evaluate thanks to simulation each solution provided by the model of optimization until a satisfactory threshold is achieved. One can then fix the near optimal values of the parameters according to the selected criterion (maintenance average cost or equipment availability).

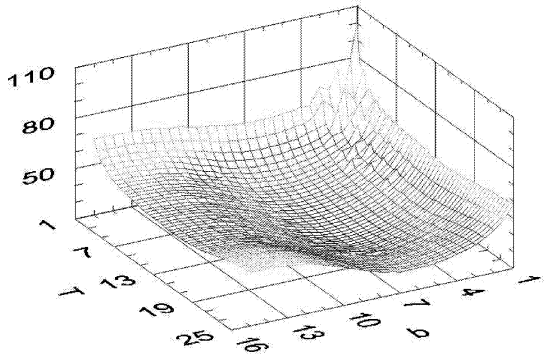


Figure 7: Minimization of the cost per unit of time

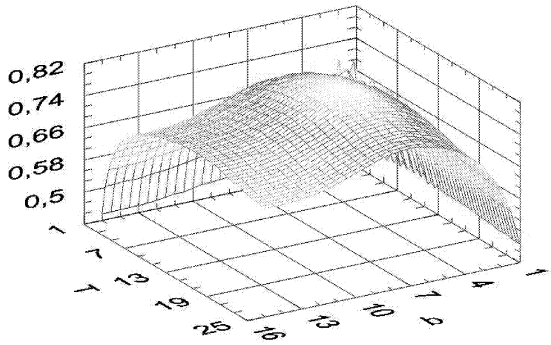


Figure 8: Maximization of the availability

The module of optimization uses the method of the simplex of Nelder-Mead developed by Nelder and Mead in 1965 (Nelder and Mead 1965). It is a local optimization method, using directly the value of the function objective, without use of derived (Chelouah 1999). This powerful and simple to program method lends itself well to a function objective coming from a simulation, since it is tolerant with the noises in the values of the function objective. The algorithm was adapted according to Luersen et al. (Luersen et al. 2003) to hold account of the bounds and the constraints on the parameters.

.

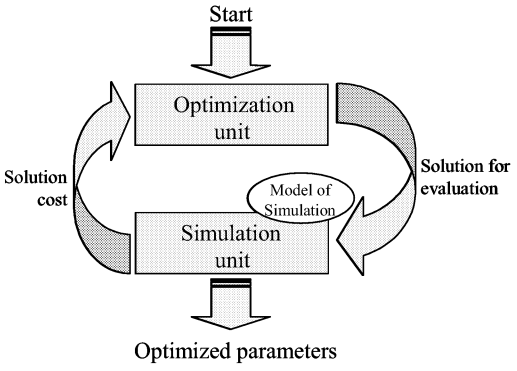


Figure 9: Coupling simulation and optimization

The decision maker can thus apply the module of optimization to obtain the optimal values of the parameters of the various maintenance strategies implemented in the framework. The comparison of the strategies' results in term of average cost per unit of time and system's availability is then possible. In general, no strategy is better on the two criteria, and the manager will have to choose the strategy corresponding best to his preferences, according to whether he privileges the cost of maintenance or the availability of the system.

CONCLUSION

We have developed a simulation-based framework for maintenance strategies elaboration.
In order to optimize the availability of their production machines to ensure production throughputs at lowest expenditure, maintenance manager can use this framework to specify the structure of the system, its logical organization, and its maintenance strategies. The modelling approach combines optimization and simulation and makes it possible to capture the dynamic behaviour as well as the reliability of multi-component complex systems. The use of simulation provides an easy evaluation for the system's availability as well as the average cost per unit of time with regard to different maintenance policies. The integration of optimization techniques helps to find the optimal various parameters of these strategies.
This framework must still be improved in order to take into account other maintenance strategies using the concept of replacement by used components, inspections, human and material resources limitations, and to widen the module of optimization to integrate other methods like genetic algorithms. We also hope to offer a real multicriteria decision-making aid for the choice of the maintenance strategy.

ACKNOWLEDGMENTS

This research is done in the context of OPTIMAIN project supported by the funding of the Walloon Region and DGTRE 215206 (FUCaM) and 215362 (FPMs).

REFERENCES

Ait-Kadi, D.; Jamali, M.A. and Artiba, A. 2002. "Optimal periodic replacement strategies using new and used items*". Journal of Decision Systems.*

Artiba, A.; Emelyanov, V.V. and Iassinovski, S.I. 1998. *Introduction to intelligent simulation: The RAO language*. Kluwer Academic Publishers.

Bardey, D.; Riane F. and Artiba, A. "To maintain or not to maintain: What should a risk-averse decision maker do?". In *Proceedings of the 2003 International Conference on Industrial Engineering and Production Management* (Porto, May 26-28).

Barlow, R. and Proschan, F. 1965. *Mathematical Theory of Reliability.* John Wiley & Sons, New York.

Chelouah, R. 1999. "Adaptation aux problèmes à variables continues de plusieurs métaheuristiques d'optimisation combinatoire : la méthode tabou, les algorithmes génétiques et les méthodes hybrides. Application en contrôle non destructif". Phd thesis of the Cergy-Pontoise university.

Cho, D.I. and Parlar, M. 1991. "A survey of maintenance models for multiunit systems". *European Journal of Operational Research*, No.51, 1-23.

Jamali, M.A.; Ait-Kadi, D. and Artiba, A. 2000. "Outils d'aide à la mise en œuvre d'un système de la gestion de la maintenance". *Journal Européen des Systèmes Automatisés*. Vol. 34, No.2-3(Apr), 391-407.

Jamali, M.A. ; Ait-Kadi D. and Artiba A. "Stratégies optimales de remplacements périodiques". In *Proceedings of the JTEA 2002* (Sousse Nord, Mars). 242-252.

Jamali, M.A.; Roux, O.; Ait-Kadi, D. and Artiba, A. "An integrated approach using optimization methods and simulation in maintenance strategies analysis". In *Proceedings of the 2004 CIRP Design Seminar.*

Luersen, M.A; Le Riche, R. and Guyon, F. 2003. "A constrained, globalized, and bounded Nelder–Mead method for engineering optimization". Springer-Verlag.

Naert, A. and Van Mol, T. 2002. "Development and Implementation of A Total Productivity Improvement Concept". Centrum voor Industrieel Beleid, KULeuven.

Nakagawa, T. 1979. "Optimum policies when preventive maintenance is imperfect". *IEEE Transactions on Reliability*, No. 4, 331-332.

Nelder, J.A and Mead, R. 1965. "A simplex method for function minimization". *The Computer Journal,* Vol. 7, 308-313.

Pichel, D. ; Dontaine, A. ; Iassinovski, S. ; Artiba, A. and Fagnart, C. "ALIX : une méthode de modélisation des flux d'un système de production". In *Proceedings of the 2003 CPI Conference* (Maroc).

Pierskalla, W.P. and Voelker, J. A. 1976. "A survey of maintenance models: The control and surveillance of deteriorating systems". Naval Research Logistics Quarterly, 23, 3, 353-388.

Riane, F.; Roux, O.; Dehombreux, P. and Bardey, D. "Maintenance Strategies development from a risk aversion point of view". In *Proceedings of the 2004 Intelligent Maintenance Systems International Conference* (Arles, France, Jul.14-17).

Sherif, Y.S. and Smith, M.L. 1981. "Optimal maintenance models for systems subject to failure – a review". Naval research Logistics Quarterly, 28, 47-74, 1981.

AUTHOR BIOGRAPHY

THIBAUT LUST was graduated from Mons (Belgium). He is an industrial engineer specialized in electro mechanics. He currently a researcher at FUCaM University and works on the development of Optimain project, funded by the Walloon region. His main research interest is the integration of optimization in the management of maintenance activities. He continues studying to be graduated as a civil engineer in computer science and management.

COMPUTING SIMULATION AND HEURISTIC OPTIMIZATION OF THE MARINE TWIN SHAFT GAS TURBINE

Josko Dvornik
Ante Munitic
University of Split
Faculty of Maritime Studies
Zrinsko-Frankopanska 38, 21000 Split, Croatia
e-mail: josko@pfst.hr

Enco Tireli
University of Rijeka
Faculty of Maritime Studies
Studentska 2, 51000 Rijeka, Croatia
e-mail: tireli@pfri.hr

KEYWORDS
System dynamics, modelling, twin shaft gas turbine, continuous and discrete simulation and heuristics optimization

ABSTRACT

Simulation Modelling, together with System Dynamics and intensive use of modern digital computer, which mean massive application, today very inexpensive and in the same time very powerful personal computer (PC-a), is one of the most suitable and effective scientific way for investigation of the dynamics behaviour of non-linear and complex: natural, technical and organization systems.
The aim of this paper is to show the efficiency of the application of the System Dynamics Simulation Modelling in investigation of behaviour dynamics, one of the complex marine system and process i.e. "gas–turbine". Twin shaft gas turbine shall be presented with mental-verbal, structural and mathematical-computing modules, and will simulate working process of turbine.

SYSTEM DYNAMICS SIMULATING MODELLING OF TWIN SHAFT GAS TURBINE

Basic equations of twin shaft gas turbine:

Equations of shaft of turbo compressor with low pressure – consumer:

$$T_{a1}\varphi_{\omega 1} + k_1\varphi_{\omega 1} = \mu_G + k_{\omega 2}\varphi_{\omega 2} \qquad (1)$$

$$\frac{d\varphi_{\omega 1}}{dt} = \frac{k_1}{T_{a1}}\left[\frac{\mu_G}{k_1} + \frac{k_{\omega 2}\varphi_{\omega 2}}{k_1} - \frac{f(t)}{k_1} - \varphi_{\omega 1}\right] \qquad (2)$$

$$\frac{d\varphi_{\omega 1}}{dt} = \frac{1}{T_{a1}}\left(\mu_G + k_{\omega 2}\varphi_{\omega 2} - f(t) - k_1\varphi_{\omega 1}\right) \qquad (3)$$

Equations of shaft of turbo compressor with high pressure:

$$T_{a2}\varphi_{\omega 2} + k_2\varphi_{\omega 2} = \mu_G + k_{\omega 1}\varphi_{\omega 1} \qquad (4)$$

$$\frac{d\varphi_{\omega 2}}{dt} = \frac{k_2}{T_{a2}}\left[\frac{\mu_G}{k_2} + \frac{k_{\omega 1}\varphi_{\omega 1}}{k_2} - \varphi_{\omega 2}\right] \qquad (5)$$

$$\frac{d\varphi_{\omega 2}}{dt} = \frac{1}{T_{a2}}\left(\mu_G + k_{\omega 1}\varphi_{\omega 1} - \varphi_{\omega 2}\right) \qquad (6)$$

where are:

T_{a1} and T_{a1} -Time of shaft 1.and 2. running
k_1 and k_2 Self regulating coefficient of shaft 1

Mental- verbal model for low pressure – consumer:

When relative variation of fuel consumption μ_G and product $k_2*\varphi_{\omega 2}$ are increasing, relative angular speed variation is increasing also, resulting in positive cause- consequence relation UPV (+).
When relative variation of possible external act of cargo and product $k_1*\varphi_{\omega 1}$ are increasing, relative angular speed variation is decreasing and observed UPV(-) is negative.
Further, when time of shaft running T_{a1} is increasing, relative angular speed variation is decreasing, resulting in negative sign UPV (-).

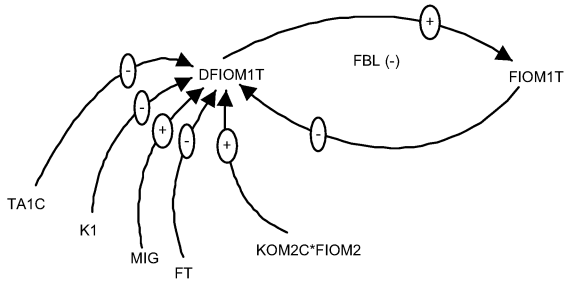


Figure 1. Structural diagram of turbo compressor with low pressure

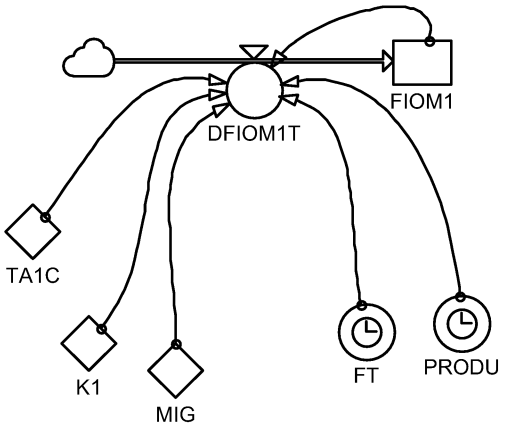


Figure 2. Flow diagram of turbo compressor with low pressure

In observed system there is one feed back loop (FBL):
FBL1(-): DFIOM1T=>(+) FIOM1T=>(+)
FIOM1T=>(-)DFIOM1T; with self regulating dynamic character (-), because the addition of negative sign is odd number.
Mental- verbal model, structural and flow diagram for high pressure – consumer are presented at ESS 2003 (12).

COMPUTER SIMULATION MODEL OF THE MARINE GAS TURBINE

Scenario:

Run of the turbine is triple-stage, which mean that in TIME=1 second we accelerate turbine by bringing the fuel. On 10% of nominal number of revolution in TIME=5 seconds we increase fuel consumption to the 35% of nominal number of revolution, and in TIME=10 we increase fuel supply, and in that way obtain “uniform” heating of turbine i.e. nominal number of revolution. In TIME=15 PID regulator is under the shock loading from gas turbine, in amount of 50% from nominal load, meaning that F (t)=0.5. In TIME=30 stochastically load occurs.

Graphics results of the simulation:

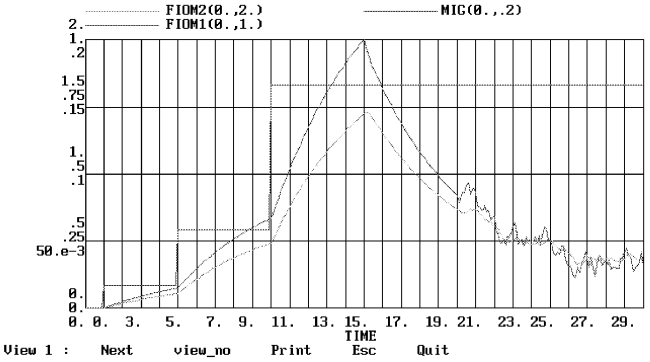


Figure 1. Relative angular speed variation, relative angular speed variation of second shaft, relative variation of fuel consumption

CONCLUSION

The application of System Dynamics Simulation Modelling Approach of the complex marine dynamic processes, which the authors, together with their graduate students, carried out at the Maritime Faculty University of Split - Croatia eleven years ago, revealed the following facts:
l. The System Dynamics Modelling Approach is a very suitable software education tool for marine students and engineers.
2. System Dynamics Computer Simulation Models of marine systems or processes are very effective and successfully implemented in simulation and training courses as part of the marine education process.

REFERENCES

Forrester, Jay W. 1973/1971. "Principles of Systems", MIT Press, Cambridge Massachusetts, USA,

Munitic, A. 1989. "Computer Simulation with Help of System Dynamics", BIS, Croatia,

Munitic, A., 1989, "Application Possibilities of System Dynamics Modelling", *Proceeding of SCS Western Multiconference*, San Diego, California, USA,

He:eB4ha, P. A., 1975, "Abtomat4⊥a∗4β cP∗ob∴ix ⊥hepŋet4Peck4x Pctahobok", Moskva, Rusia,

Munitic, A.,Kuzmanic, I.,Krcum, M.,1998., "System Dynamic Simulation Modelling of the Marine Synchronous Generator", *Proceeding of Modelling and Simulation Conference, IASTED*, Pittsburgh, 372-375.

Munitic A., Antonic R., Dvornik J., 2003., "System dynamics simulation modeling of ship-gas turbine generator", *Proceeding of ICCC'03,* SLOVAK, 357-360

Munitic A., Orsulic M., Dvornik J., 2003, "Computer Simulation of Complex Ship System "Gas turbine.Sysnhronous Generator" *Proceeding of ISC 2003,* Valencia, Spain, 192-197

Munitic A., Orsulic M., Dvornik J., 2003, "Continuous Computer Simulation Model of the Marine Gas turbine", *Proceeding of ESS 2003,* Delft, The Netherlands, Delft, The Netherlands, October 26-29, 2003.

NEURAL NETWORK AND GENETIC ALGORITHM SIMULATION

CORTI'S ORGAN PHYSIOLOGIST-BASED COCHLEAR MODEL: NEURAL LATERAL INHIBITION

Francisco Ríos
Raquel Fernández-Ramos
Jorge Romero-Sánchez
José Francisco Martín
A. Ignacio Molina-Conde
Francisco Javier Marín
Universidad de Málaga
Campus Teatinos s/n
29017-Málaga
Spain
Email: fjmarin@uma.es

KEYWORDS
Neural Networks, Speech síntesis, Bio-engineering

ABSTRACT

Corti's Organ is an Electro-Mechanical transducer that allows the energy coupling between acoustical stimuli and auditory nerve. Although the structure and functionality of this organ are complex, state of the art models have been currently developed and tested. Cochlea model presented in this paper is based on Bekesy's theories and others and concerns on the behaviour of auditory system on spatio-frequency domain and mechanisms of lateral inhibition. This paper applies the intensity adaptation and lateral inhibition with artificial neural networks.

INTRODUCTION

Corti's Organ is composed of more than 15.000 cilia excited by means of travelling waves. These waves produce relative pressures distributed along the cochlea, exciting a specific number of cilia in a local way. Nonlinear mechanisms of local adaptation to the intensity (external cilia cells) and lateral inhibition (internal cilia cells) allow the selection of very few excited elements. These transmit a very precise intensity and frequency information. These signals are the only ones coupled to the auditory nerve and only a signal without redundancies reaches the brain.
Distribution of pressure waves matches a quasi-logarithmic law due to Cochlea morphology. The cochlear model presented takes Bark's law as an approximation to this behaviour consisting on grouped arbitrary elements composed of a set of selective coupled exciters (bank of filters according to Patterson's model). These sets apply the intensity adaptation principles and lateral inhibition (our contribution in this paper) implemented with artificial neural networks. Elements excited during the process generate a bioelectric signal in the same way than cilia cell. The spikes signals of excited cilia are constant both in amplitude and frequency related to acoustic pressure.
This paper is organized in four parts, as follows: First, we study the cochlear sound codification remarking the following hypotheses: a) precise definition of frequency, b) few excited elements at the same time, c) temporal masking, and d) critical band. Second, we show a high level architecture that executes them. Third, we present the algorithms proposed to verify these principles: bank of filters, lateral inhibition, masking, compression, quantification and codification. And, finally, we show the efficiency of our model, technologically feasible, for cochlear implants.

CORTI'S ORGAN SOUND PROCESSING

Cochlea is a rigid spiral shape tube (34 mm length in humans), longitudinally divided by both, basilar and reissner membranes. Corti's organ transforms pressure waves to nervous impulses (Duphar 1986). It exists around 3500 internal cilia cells and over 15000 external cilia cells, which present connections with afferent and efferent nervous fibres (towards and from the brain), conforming the auditory nerve (figure. 1).

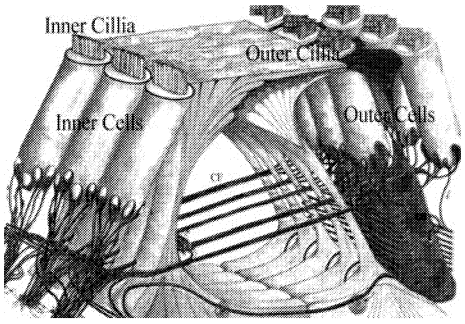


Figure. 1: Corti's organ three-dimensional.

The system behaves like a mechanical energy regenerative system (positive feedback) or an automatic sensitivity control system (Nobili et al. 1998). However, there is a question, how it is possible that ear frequency perception is so precise when simple filters of low quality factor can be used to model pressure waves. We think that the travelling wave presses on a lot of cilia cells, but these cells like

nervous cells of the retina or other sensorial nervous centres, have the property of lateral inhibition. The lateral inhibition in a set of excited nervous cells is a mechanism of cellular competition that produces only one cell or few of them send nervous impulse towards crust zones.
By using psychoacoustic methods, it is possible to determine psychophysics parameters such as: frequency response, sound pressure thresholds for the excitation, precision in the frequency acquisition, masking thresholds and critical band. Two first parameters try to delimit the system operation range with a quasi-logarithmic response. Lateral inhibition mechanism selector, is provided by a very small group of cilia cells that maintains their excitation until the superior nervous centres. Masking phenomenon is a process producing a sound hearing threshold rising due to presence of another sound (Zwicker and Tilmann 1991). Pre-masking periods are short: 2-3 ms, simultaneous masking periods are in the range of 150 ms and post-masking ones can be extended to 200 ms. Critical Band (CB) is the frequency interval that represents the maximum frequency resolution of auditory system. It can be expressed as follows:

$$\Delta f_{CB}(f)=25+75[1+1.4(f/1000)^2]^{0.69} \qquad (1)$$

This experimental formula let us divide the audio range frequencies in adjacent non-overlapping CB intervals, and representing a first problem approach to model the frequency selectivity of the Corti's organ. We have chosen the Patterson model and the highly efficient algorithm developed by Slaney (Slaney 1993).

COCHLEAR MODEL: BLOCKS FUNCTIONAL DIAGRAM

Figure 2 shows the blocks functional diagram and processing stages flow of our cochlear model. The signal is amplification and band limited to adapt the dynamic range of the converter and the sampling frequency to avoid aliasing. 14 bits A/D converter is connected to a bank of 32, 64 or 128 ERB band-pass filters, whose central frequencies are inside other CB. The set constitutes a quasi-sine signals array (with modulated amplitude) which excites the nonlinear filter of lateral inhibition. The neural lateral inhibition is applied to the amplitude mean value of the quasi-sine signals provided by each one of the previous stage filters. Channels selector choose five signal maximum of the processed signals, per frame, producing a 5x2 matrix, if it exists, that contains the precise frequency and amplitude of each selected maxima. Timing masking is reproduced matching the current 5x2 matrix with a stored sample of previous results covering a temporal rank around 150 milliseconds.

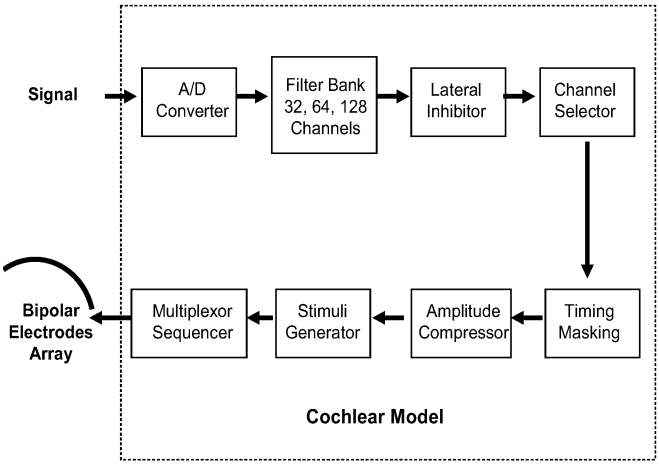


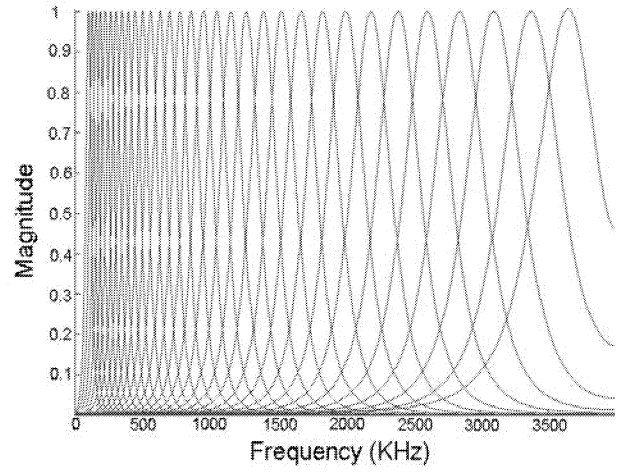
Figure. 2: Cochlear Model. Sound transformation for the creation of stimuli signal.

After masking process, electrical stimulus is synthesized. The system has a look-up table to generate wave forms from the stimuli. This look-up table stores a quantized logarithmic compression law (i.e. μ law). The generated impulses have zero mean to avoid polarizations and ionic migration. Finally, last operation consists of mapping the stimuli in a pair of electrodes whose position is identified with the corresponding cilia cells frequency. There is a clear relationship among the frequency precision, the number of electrodes, and the cochlear location. On the other hand, to avoid stimuli interferences, it is necessary time multiplexing of the excitations. Sampling time in the multiplexing process has to be smaller than persistence due to masking phenomena. We propose a sampling time in the range of 2-3 milliseconds.

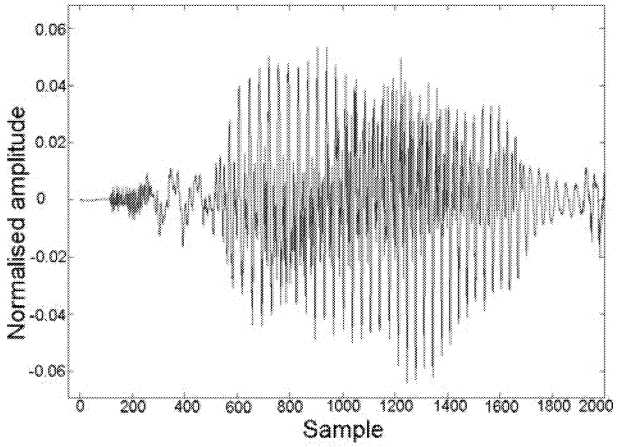
COCHLEAR MODEL: IMPLEMENTATION

Cochlear Filter Bank

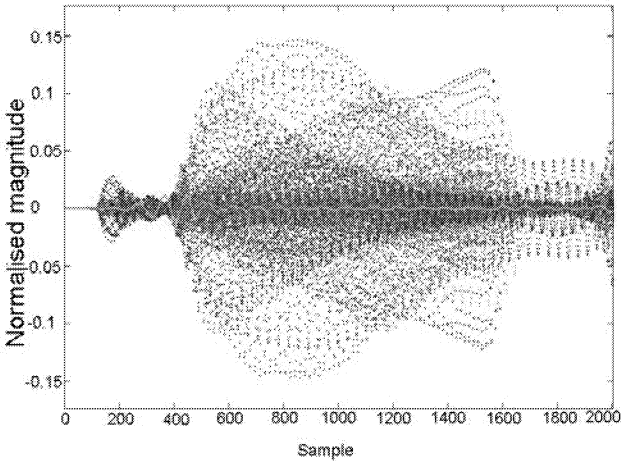
Patterson's cochlear model is based on an independent vector band-pass filters basis. The bandwidth of each filter is described by its equivalent rectangular bandwidth (ERB: Rectangular Equivalent Bandwidth).This measurement is the CB bandwidth. Figure 3 shows the magnitude shape of a 32 units filter bank distributed in a 4KHz band, starting at 100Hz, with a minimum bandwidth of 24 Hertz and quality factor of 9.3 per filter. Even though few band pass filters are used (in our experiments we have used banks of 32, 64 and 128 filters), frequency precision does not depend on filters selectivity, but on the lateral inhibition principle application.



(a)



(b)



(c)

Figure 3. Magnitude of filter bank (a), probe signal (b), and signal after filter bank processing (c).

Neural Lateral Inhibition

We think that the travelling wave presses on a lot of cilia cells, which have the property of lateral inhibition as a mechanism of cellular competition that produces only one cell or few of them send nervous impulse towards crust zones. Thus, we propose the self-organizing map (Kohonen 1990) which neighbouring neurons compete in their activities by means of mutual lateral interactions, and developed adaptively into specific detectors of different signal patterns. On the other hand, in the auditory cortex there are the so-called "tonotopic maps", which represent pitches of tones in terms of the cortical distance (Suga and O'Neill 1979). Thus, we propose an internal representation of the signal (organized spatially) by means of a self-organizing map. The problem is reduced to interpreter the ultrametric topological order that is not easily interpretable. We want to emphasize the signal maxima and minima and their spatial distribution by means of the pattern's picture, which must be time invariable. This pattern's picture is the weight matrix of the neural network. Only a signal without redundancies reaches the brain, the pattern's picture.

Let us consider an aligned array of internal cilia cells (neurons) and suppose that an arbitrary signal generates an excitation pattern in all of them. Our neural model postulates that only one cell at a times gives the active response to the current input. Let **x(t)** be the input vector that is connected in parallel to all the neurons i in the network (i=1..64, one per frequency channels), and $\mathbf{w_i}$ the weight vector of neuron i. The matching criterion is based on the Euclidean distances between **x** and **w.** The updating process in discrete-time notation for the best-matching is as follows:

$$w_i(t+1) = w_i(t) + \alpha(t)\left[x(t) - w_i(t)\right] \quad (2)$$

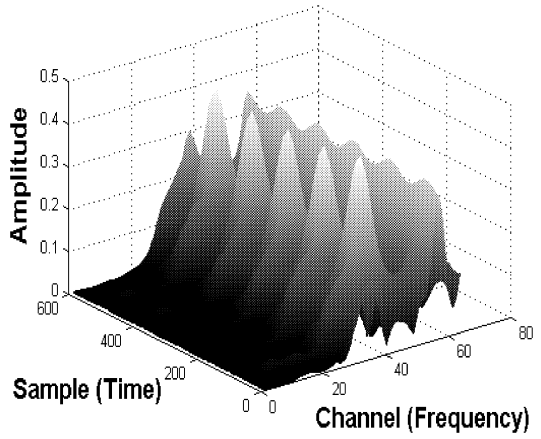
where α(t) is a suitable, monotonically decreasing sequence of scalar-valued gain coefficients, $0 < \alpha(t) < 1$. α(t) is so-called dynamic rate learning parameter, which evolves according to equation (4):

$$\alpha(t) = \frac{\alpha_0}{(1 + \frac{c.t}{nn})} \quad (3)$$

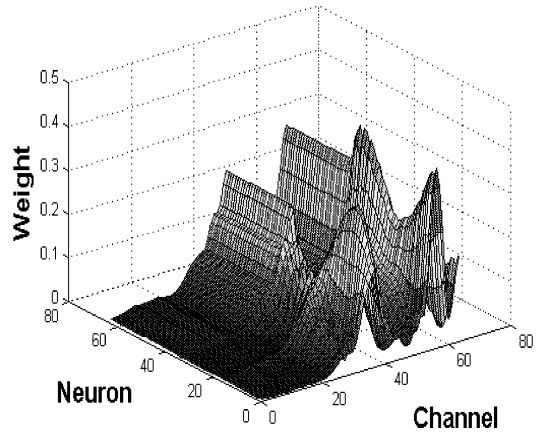
where α_0 is the initial rate learning, 0.3, *c* is constant, 0.2, and *nn* is the number of neurons, 64. When the number of neurons is increased, then the generated map is more precise in frequency. Figure 4 shows inhibition algorithm pattern, considering a bank of 64 filters. It can be noticed the sharp envelope of the response, with it maximum located at any frequency delimited by CB.

With this scheme, the number of selected channels to be excited will be reduced to a smaller group, in our case from 64 to 36. Besides, we will only excite with the values of the 5 first relative maximums in the precise positions indicated by the inhibition algorithm. In order to verify that the information codified in this small number of maximums is complete, we have developed an inverse algorithm of

reconstruction capable of reproducing the original signal from this small set of excitations. The reconstruction algorithm reproduces the sound adding a set of sinusoidal signals. These signals have frequencies provided by the algorithm, random phases and amplitudes defined by the relative maxima.



(a)



(b)

Figure 4. Time-frequency filter-bank response before (a) and after (b) neural lateral inhibition.

Temporal Masking

We postulate the selection from the excitation matrix, those values that are maxima and their time distance in the same channel are given by measured times in psychoacoustic experiences (Wilson et al. 1995). The minimum interval is equal to the frame size associated to the sampling frequency. Thus, the temporal masking algorithm is:

- If a maximum amplitude signal located in the same frequency or same channel exists, then it will be taken as a mark to measure the time interval (of 100 to 150 milliseconds).
- If a greater maximum is found, then the mark will be updated. In other case, it will not change.
- If the generated signal and time distance between them is greater than the time interval, then an excitation will be generated in any channel.
- In alternate channels the algorithm is not applied.

Compression, Quantification and Construction of Stimulation Signal

External cilia cells compress the acoustic signal amplitude. This amplitude compression characteristic is typical of some audio compression systems that use logarithmic laws and work correctly in voice communication standard systems. Each electrode must provide a signal stimulus that consists of an impulses train of zero mean (to avoid ionic migration), whose frequency is related with signal amplitude (Loizou 1998). The impulses train will be activated in a sequential way with a suitable frequency, in those channels with position indicated by the lateral inhibition algorithm.

CONCLUSIONS AND FUTURE WORKS

We have presented a cochlear model based on the Corti's organ physiology functions. Each of the psychoacoustic experiences has been justified and several algorithms that execute them have been proposed. Special emphasis has been placed on the proposal of the neural lateral inhibition, which allows very precise frequency selection. Also, a sound processor has been developed, which adds a novel element based on the stimuli of greater energy codification. Consequently, this solution brings the possibility of integrating high density electrodes array. This work constitutes a first approach for the development of cochlear implant prosthesis. Future works must be supported by a multidisciplinary team composed by a group of surgery medical specialists, physicians and engineers.

REFERENCES

Duphar Farmaceutica S.A., *El oido interno*. Duphar. Spain. 1986.

Nobili, R.; F. Mammano and J. Ashmore. *How well do understand the cochlea. Frequency filtering by the cochlea.* 1998.

Zwicker, E. and U.Tilmann "Audio Engineering and Psychoacoustics: Matching Signals to the final Receiver, the Human Auditory System". Audio Engineer Society, 39, no. 3. pp. 115-126 1991.

Slaney, M.. "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank". *Apple Computer Report* 35. *Perception Group-Advanced Technology Group*. Apple Computer, Inc. 1993.

Kohonen, T. "The Self-Organizing Map", *Proc. IEEE,* vol. 78, no. 9, pp. 1464-1480. 1990.

Suga, N. and W.E.O'Neill, "Neural axis representing target range in the auditory cortex of the moustache bat", *Science, vol. 206*, pp. 351-353. 1979.

Wilson, B.; D. Lawson and M Zerbi. "Advances in coding strategies for cochlear implants". *Advances in Otolaryngology-Head and Neck Surgery*, 9, pp. 105-129. 1995

Loizou, P.C. "Mimicking the human ear. An overview of Signal-Processing Strategies for converting sound into electrical signals in Cochlear implants". *IEEE Signal Processing Magazine.* pp. 101-130. 1998.

RULE REPRESENTATION AND INITIAL POPULATION IN GENETIC LEARNING

I. Bruha
Department of Computing & Software, McMaster University
Hamilton, Ont., Canada, L8S4K1
E-mail: bruha@mcmaster.ca

KEYWORDS
Genetic algorithm, rule-based systems, initial population, decision rule representation.

ABSTRACT

Genetic algorithms (GAs), relatively new powerful optimization tool, have been applied in various simulation projects, particularly for optimization as well as search simulation. In Machine Learning (ML), genetic algorithms have been utilized for search through a space of all possible concept descriptions. We designed and implemented a system that merges a domain-independent GA into the covering learning algorithm CN4, a large extension of the well-known CN2.

In this paper, we present some recent improvements of the above learning algorithm (which we call *GA-CN4*). The paper particularly discusses the representation of decision rules as chromosomes (objects, individuals) in GAs, and the ways of initialization of a population for a GA.

We conclude by description and analysis of a few experiments that were carried out in order to compare various approaches to initialization of a population for GAs.

INTRODUCTION

Genetic algorithms represent a powerful tool for simulating optimization and search procedure. A genetic algorithm (GA) encompasses a long process of evolution of a large population of chromosomes (individuals) before selecting optimal values that have a better chance of being globally optimal compared to the traditional methods. On the other hand, the selection operation gives a 'chance' to weaker, worse objects to be involved in the evolutionary process.

The fundamental idea is simple: chromosomes selected according to an 'evaluation' are allowed to crossover so as to produce a 'slightly different' new one – the offspring. It is clear that the algorithm performs according to how 'slightly different' and 'evaluation' are defined.

Details of the theory of genetic algorithms may be found in several books, e.g. (Goldberg 1989; Holland 1975). There are many papers and projects concerning genetic algorithms and their incorporation into data mining (Bala et al. 1995; De Jong et al. 1993; Giordana and Saitta 1993; Janikow 1993; Turney 1995).

A designer of a GA has to provide an evaluation function, called *fitness*, that evaluates any individual (chromosome). The fitter individual has the greater chance in forming a new population. Given an initial population of individuals, a genetic algorithm proceeds by choosing individuals to become parents and then replacing members of the current population by the new individuals (offsprings) that are modified copies of their parents. The designer must also select types of *genetic operators*, namely *selection, crossover*, and *mutation.*

They are two important characteristics of the search performed by a GA: the search is usually global and is parallel in nature since a GA processes not just a single individual (chromosome) but a large set (population) of individuals.

Note that GAs exhibit various parameters that must be selected by a designer; there are, however, two important parameters whose selection and/or definition is the most important, namely the initialization of the population of individuals and their representation. Last-but-not-least, GAs can process individuals represented by discrete attributes only. Therefore, another important issue is the discretization of numerical attributes; see e.g. (Bruha and Berka 1996; Bruha et al. 2000).

Let us describe our learning algorithm GA-CN4 in more details. Originally, a few years ago, we designed the rule-inducing CN4 learning algorithm (Bruha and Kockova 1994), a large extension of the well-known algorithm CN2 (Clark and Boswell 1991) by many

efficient routines. In our new genetic learner GA-CN4, we have replaced the original search strategy by a domain-independent GA.

A single rule is treated as an individual (chromosome). After a stopping condition is satisfied, the best individual (decision rule) is selected. GA-CN4 generates decision rules from a set of K training examples, each accompanied by its *desired class* C_r , r = 1,...,R . Examples are represented by N discrete (symbolic) *attributes*. An attribute A_n , n=1,...,N, comprises J(n) distinct values $V_1,...,V_{J(n)}$. The algorithm yields an *ordered* list of decision rules of the form

`Rule: if Cond then class is` C_r

The condition *Cond* is represented by a conjunction of selectors (attribute-value pairs).

Our domain-independent GA processes the fixed-length chromosomes by means of the following procedure (here T is a training set):

procedure GA(T)
1. Initialize a new population
2. **Until** stopping condition is satisfied **do**
   - 2.1 Select individuals by the tournament selection operator
   - 2.2 Generate offsprings by the two-point crossover operator
   - 2.3 Perform the bit mutation
   - 2.4 Execute the hill-climbing local-optimization
   - 2.5 Check whether each new individual has the correct attribute and class values (within their ranges); if not the individual's fitness is set to 0 (the worst value)

   enddo
3. Select the fittest individual (i.e. a decision rule)
4. **If** this object is statistically significant
   then return this object
   else return nil

The above algorithm mentions some particular operations used in our genetic learner. Their detailed description can be found e.g. in (Bruha et al. 2000; Goldberg 1989; Holland 1975). More specifically, the fitness function is calculated by the Laplacian criterion for expected accuracy (for the class C_r)

$$Lapl(C_r, Cond) = \frac{K_r(Cond)+1}{K(Cond)+R}$$

where $K_r(Cond)$ is the number of training examples of the class C_r covered by the condition *Cond*, $K(Cond)$ the number of examples covered by the condition, R the number of classes.

The default parameter values in our genetic algorithm: size of population is 30, probability of mutation *Pmut* = 0.002 . The genetic algorithm stops the search when the Laplacian criterion does not improve after 10000 generations.

Our GA also includes a check for statistical significance of the fittest individual. It has to comply with the statistical characteristics of a database which is used for training; the P^2-statistics is used for this test of conformity. If no fittest individual can be found, or it does not comply with the P^2-statistic, then *nil* is returned in order to stop further search; the details can be found in (Bruha and Kockova 1994).

The next two sections of this paper describe both the issue of rules representation and that of initial population. The last two sections introduce experiments that exhibit the characteristics of various scenarios in our genetic learner, and analysis including future work.

REPRESENTATION OF RULES

Each decision rule generated by the genetic learner is portrayed by a chromosome (individual). Since the number N of attributes for a task (problem) is given, each chromosome (decision rule) is represented by a fixed-length string of the length of N fields.

There exist various ways of actual depiction of these fields that form the entire string (i.e., chromosome's representation). We now survey the most interesting ones (Kralik 2001).

(a) Byte Representation

Each attribute field is placed in one byte (8 bits); i.e., maximum range of values of any attribute is from 0 to 255.

Let us illustrate this representation on the well-known 'weather' problem (Quinlan 1986) with these attributes:

```
windy:          false, true
humidity:       normal, high
temperature:    cool, mild, hot
outlook:        rain, overcast, sunny
```

Each rule is represented by a string of 4 bytes, each attribute value by an integer from the range 0 to J(n) . If a byte is within this range, then it is interpreted as the corresponding attribute value; otherwise, the given selector is excluded from the rule. For instance, the string

```
00000001 10111101 01011111 11001011
```

corresponds to values 1, 189, 95, 203; since only the 1st value is permissible, i.e. it represents

`if windy=true then class is C`

where C is the majority class of the training examples covered by this rule.

Evidently, this representation requires

$Lchrom = 8 * N$

bits. The probability of inserting a selector A_n into a rule is thus

$Pins(A_n) = \frac{J(n)}{256}$

i.e. usually very small. In our example, the probability $Pins(\texttt{windy}) = 0.0078$ (0.78%).

However, this representation allows to easily introduce a *negation* of a selector. The left-most bit in any byte can serve as a flag for negation (equal 1 if the selector is to be negated). The maximum range of attribute values is consequently half, i.e. from 0 to 127.

(b) Reduced Mask Representation

In this representation, each attribute A_n in a chromosome representation is depicted by $Lmask(n)$ bits which is the minimum integer greater than or equal to $J(n)$. The probability of inserting a selector A_n into a rule looks thus better:

$Pins(A_n) = \frac{J(n)}{Lmask(n)+1}$

In our example, $Pins(\texttt{windy}) = 0.75$.

Also, by adding an extra bit for each rule allows to introduce a *negation* of the corresponding selector.

(c) Bit representation

Each attribute A_n is represented by $J(n)+1$ bits. Each attribute field begins with one bit (toggle) for negation. Each remaining bit exhibits one value of this attribute; it is set up to 1 if the corresponding value is present in the selector. E.g., the string

```
0 01 1 00 1 101 1 001
```

in our task represents this rule:

```
if windy=true &&
   temperature\=[cool or hot] &&
   outlook\=sunny
then class is C
```

Hence, this representation allows *negation, internal disjunction*, and *negation of the internal disjunction*. The length of a chromosome is

$Lchrom = N + \sum_{n=1}^{N} J(n)$

and the probability of inserting a selector A_n into a rule is always 0.5 .

(d) Representation of numerical attributes

A numerical attribute A_n is represented by two real numbers: the lower bound $Lbound_n$ of an interval and its upper bound $Ubound_n$, i.e. this field depicts the interval

$Lbound_n < A_n \# Ubound_n$

If a bound does not fit to the entire range of the numerical attribute, then it is eliminated from the above formula; if both bounds do not fit, or if the lower bound is greater than the upper one, then the entire selector is eliminated from the corresponding rule.

INITIALIZATION OF A POPULATION

Another important issue of defining or selecting parameters for a genetic algorithm is the initialization of the initial population.

First, the initialization of a population could be done either

(Rand) randomly (using a pseudo-random number generator), or

(KB) by utilizing an initial domain-specific ('external') knowledge base (KB).

In the latter case, KB is available from an 'external' agent; consequently, our genetic learner then serves as a refinement agent. The initial population is in this case formed by chromosomes that correspond to the conditions of 'external' rules and their subsets until the entire size of the population is generated (Bruha and Kralik 2002):

procedure Initialize

1. Convert all the external rules to chromosomes of the initial population
2. **Until** the required number of initial chromosomes (size of population) is generated **do**
   - 2.1 **If** there is no external (generalized) rule left
     then generate randomly the rest of initial chromosomes and **exit**
   - 2.2 Generalize the external rules by removing last selector from their conditions
   - 2.3 Convert all these new rules to initial chromosomes

   enddo

Next, we have to realize that a population is initialized before a new rule is generated (step 1 of the procedure GA above). There exist two possibilities:

(A) to initialize a new population for each new rule, or

(B) to initialize a new population for the first rule only, thus using the final population from the last call (loop) as the new one for the current call of the procedure GA.

In our genetic learner, we thus provide four possible routines of initialization: *(A-Rand), (A-KB), (B-Rand),* and *(B-KB).*

EXPERIMENTS

To compare and analyze the efficiency and power of the two important routines/parameters (rule representation and population initialization) in our genetic learner GA-CN4, we run various tests above several datasets from the Machine Learning Repository and other sources. Particularly, the following databases have been exploited:

! *Japanese Credit*, *Iris*, *Pima Indian Diabetes*, *and Australian Credit*:
Tab. 1 exhibits the characteristics of these databases taken from the Machine Learning Repository.

! *Quality of Car Engines*:
This robust task of predicting the over-all quality of car engines has been provided by the Mercedes-Benz Research Institute, Germany (see Tab. 1).

! *Smog Prediction*:
These smog prediction data were used for ecological testing in the Czech Academy of Sciences, Prague, Czechland; the entire problem involves two subsets, see Tab. 1.

Each of the above databases was randomly separated to two subsets (70% learning, 30% testing) and this 'splitting' procedure has been executed 15 times for each database. Comparison of the various routines was done by the paired t-test at the significant level 2.5%.

The experiments carried out for various representations of rules, however, did have not bring any significant difference in the classification accuracy. This result, as a matter of fact, has been expected because various representations do not bring any relevant contribution to prediction accuracy.

Performance of various routines for population initializations is in Table 2. The C4.5 served as the external domain-specific agent that formed the initial population for the case *(KB)*. Its pruning factor was 50% so rather inconsistent knowledge bases were generated; consequently, our genetic learner with a population initialized by this external knowledge base had enough 'space' to show its power by modifying this initial population.

ANALYSIS AND FUTURE RESEARCH

Two newer issues designed and implemented in our genetic algorithm GA-CN4, namely (a) representation of chromosomes in the genetic algorithm, (b) the ways of initialization of a population for the genetic algorithm, have been discusses in this paper.

First, the experiments with various chromosome representation did not bring anything substantial as for the classification accuracy; the t-test exhibited no significant difference.

Nevertheless, the tests performed for various initialization procedures revealed the following:

- there is no significant difference between the cases *(A)* and *(B)*;
- initialization by an external KB became significantly better than the random one.

Generally speaking, we found out that the genetic learner has better classification performance than the traditional learning algorithms. We explain this better behaviour by the fact that the traditional algorithms explore a small number of hypotheses at a time, whereas the genetic algorithm carries out a search within a robust population, in parallel in nature. The only disadvantage of GA is time consumption.

As for our future work, we plan to add and explore the following directions:

- to build in an internal discretizing routine in our GA to improve its performance;
- to exploit and test various rule representations in GA from the viewpoint of the complexity and comprehensibility of the knowledge base induced;
- to incorporate the genuine global search into the genetic learner by considering the entire set of decision rules as a single chromosome in a population.

REFERENCES

Bala, J. et al. 1995. "Hybrid learning using genetic algorithms and decision trees for pattern classification", *Proc. IJCAI-95*, 719-724.

Bruha, I. and Berka, P. 1996. "Numerical attributes in symbolic learning algorithms: Discretization and fuzzification", *2nd International Conference Neural Networks & Expert Systems in Medicine*, Plymouth, UK, 131-138.

Bruha, I. and Kockova, S. 1994. "A support for decision making: Cost-sensitive learning system", *Artificial Intelligence in Medicine*, 6, 67-82.

Bruha, I., Kralik, P., and Berka, P. 2000. "Genetic learner: Discretization and fuzzification of

numerical attributes", *Intelligent Data Analysis J.*, 4, 445-460.

Bruha, I. and Kralik, P. 2002. "Genetic learner GA-CN4: The ways of initializations of new populations", Dept Computing & Software, McMaster Univ, Techn. Rept.

Clark, P. and Boswell, R. 1991. "Rule induction with CN2: Some recent improvements", *EWSL-91*, Porto, Springer-Verlag, 151-163.

De Jong, K.A., Spears, W.M., and Gordon, D.F. 1993. "Using genetic algorithms for concept learning", *Machine Learning*, 13, 161-188.

Giordana, A. and Saitta, L. 1993. "REGAL: An integrated system for learning relations using genetic algorithms", *Proc. 2nd International Workshop Multistrategy Learning*, 234-249.

Goldberg, D.E. 1989. *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley.

Holland, J. 1975. *Adaptation in natural and artificial systems*, Univ. of Michigan Press, Ann Arbor.

Janikow C.Z. 1993. "A knowledge-intensive genetic algorithm for supervised learning", *Machine Learning*, 5, 189-228.

Kralik, P. 2001. "Application of genetic algorithms in machine learning", PhD thesis, Czech Techn. Univ., Brno.

Quinlan, J.R. 1986. "Induction of decision trees", *Machine Learning*, 1, 81-106.

Turney, P.D. 1995. "Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm", *J. Artificial Intelligence Research.*

database	# classes	frequency of majority class	# examples	# attributes	numerical attributes	average # of values per symbolic attribute
Japanese Credit	2	68%	125	10	5	3
Iris	3	33%	147	4	4	-
Pima Indian Diabetes	2	65%	768	8	8	-
Australian Credit	2	55%	690	14	6	4.6
Quality	2	72%	269	21	5	5.4
Smog-1	2	73%	207	4	3	4.0
Smog-2	4	47%	139	5	4	4.0

Table 1. Characteristics of the seven databases used for our experiments.

database	*(Rand) (A)*	*(KB) (A)*	*(Rand) (B)*	*(KB) (B)*
Japanese Credit	97	98	97	98
Iris	98	98	98	97
Pima Indian Diabetes	86	89	87	89
Australian Credit	92	94	93	94
Quality	97	98	96	98
Smog-1	71	79	72	78
Smog-2	69	71	68	70

Table 2. Classification accuracy for various initializations of populations for GA-CN4. Here:
(Rand) a population is initialized randomly
(KB) a population is initialized by utilizing an external knowledge base
(A) initialize a new population for each new rule
(B) initialize a new population by using the final population from the last call

VISUAL SIMULATIONS

A Simplified Aerodynamic Model for Flight Simulation

H. Chen, T.R. Wan and R.A. Earnshaw
School of Informatics
University of Bradford,
Bradford, UK, BD7 1DP
Email :{ h.chen3, t.wan}@bradford.ac.uk

Abstract

With the revolution in computer technology in recent years, the capability of simulating aircraft's performance with high-fidelity in real-time situations has arrived on desktop computers. In this paper, we present a simplified nonlinear six degree-of-freedom aircraft model for flight simulations. The proposed nonlinear aircraft model reduces the need for large amount of flying data and calculation while keeping the nonlinearities of the general aerodynamic models. Therefore, a practical balance between the simulation speed and the simulation fidelity can be achieved. Such a model is useful for flight simulation applications on desktop computers.

I Introduction

Simulation is a useful tool for aviation researches. The current state of the arts in simulation technology has provided today's aerospace industry with much valuable assistance that could not be obtained elsewhere. Real time flight simulation is one of the key techniques in the development and innovation of future aviation concepts and systems [1]. Flight simulation has involved in many aspects of the aviation research which includes: flying quality evaluation, flight controls, flight test support, human factor investigation and crew training, etc [2]. Flight simulation is conducted on a broad front, both at research establishments and in industry. Flight simulation ranges from comprehensive representations of the operational equipments and environments in support of full-mission performance to modest simulations which involving a single item of equipment or part task. Flight simulation allows the researchers to explore the implications of different design options without having to incur the expense and delay arising from building and testing a range of prototypes. In addition, flight simulation research has provided a means of evaluating the likely behavior and consequences arising from abnormal operating configurations without jeopardizing the safety of the flight crews.

In the past, mathematical calculations performed during real-time flight simulation required the high-performance floating point processing of supercomputers. With the recent advances in microprocessor technology, the desktop computer can provide enough computational power to properly operate a real-time flight simulation. The essence of real time flight simulator is using the aerodynamic model to create an illusion of the real flight in virtual environments. The aerodynamic model in real time flight simulation systems is expected to offer high fidelity simulation quality while keeping the simulation speed within an acceptable scope.

A complete aerodynamic model includes fully articulated control surfaces and airflow divergence patterns over the aircraft. Such kind of models is normally relies on large amount of experiment data and divided into numerous sub-models to guarantee the accuracy, which would seriously affect the real time performance of the simulation system and does not suitable for desktop computer simulation applications [10]. On the other hand an over simplified aircraft model such as linear model, which is often used in the research of trim condition flight control [13], or performance-based models, which focus on specific motion behavior (longitudinal or lateral). Such models involve only two or three degrees of freedoms and can not meet the requirements of full envelope flight simulation. These models either only reproduce the response of the aircraft within specific conditions (linear model) or can not fulfill the task to evaluate the complex maneuver behavior of the aircraft (performance-based model) such as the combination of both longitudinal and lateral movement.

In this paper, we introduce a simplified nonlinear 6 DOF aircraft model for flight simulation. This nonlinear aircraft model reduces the need for large amount of flying data and calculation while keeping the nonlinearities of the general aerodynamic models. Therefore, this model is expected to be a good choice for general flight simulation applications on desktop computers. Our paper is organized as following: the background knowledge about flying theory is introduced in the next section; in section III we discuss the concept of coordinate systems; in section IV we discuss and compare the most often used orientation models in flight simulation and in section V we present the full 6 DOF nonlinear aerodynamic model; in section VI we show some simulation results and at the final section we draw the conclusion.

II Background Knowledge about Flying Theory

The word "aerodynamic" comes from two Greek words: *aerios*, concerning the air, and *dynamic*, meaning powerful. Aerodynamics is the study of forces and the resulting motion of objects through the air [4]. A flying aircraft is subject to a pressure depending on the airspeed and the air density. This pressure increases exponentially with increasing of the airspeed. The aircraft's resistance to the

airflow (drag) depends on the shape of the fuselage and flying surfaces.

The density of the atmosphere has three primary effects on the aircraft. Two are aerodynamic effects while the other relates to propulsion. The wing's profile of an aircraft is usually asymmetric, which causes the airspeed on the wings' upper side to be higher than the underside and in turn makes the pressure on the upper side to be lower than the underside. Therefore, a lift force is created. The lift force of a symmetric profile is based on the airspeed and on a positive angle of attack to on-coming flow. A discussion about the design of wing's profile is beyond the scope of this paper. We do not go to further detail here.

In theory when the aircraft climbs to higher altitudes, the density of the atmosphere decreases, which results in the decrease in both drag and lift. Less drag means the aircraft can fly faster, assuming the aircraft engine delivers the same amount of thrust. Indeed, the aircraft must fly faster in order to generate enough lift. Otherwise, the aircraft will descend to a lower altitude to keep the balance between the lift and the weight of the aircraft. However, this theory does not mean that the aircraft can fly faster and faster when the altitude increases. The question arises by the issue of the effect of atmosphere density on propulsion. The aircraft's engines operate by compressing incoming air, mixing the compressed air with fuel, burning the mixture, and exhausting it to generate thrust. This process becomes more difficult as the air density decreases because the compression is less efficient. The air density eventually decreases so much that there is not sufficient compression to support combustion. In summation, the decrease in air density that occurs as an aircraft climbs to higher altitudes has three effects: 1) reduces lift, 2) reduces drag and 3) reduces thrust. The propulsion effect is the most significant. As a result, the maximum altitude an aircraft can reach is always bounded.

III Coordinates Systems

Coordinate systems in which the aircraft model is described can vary to a great extent depending on the application and the preference of the user. In aircraft simulation, coordinate systems fall into two broad classes, "body" coordinates and "earth" or "inertial" coordinates [5]. Body coordinates have their origin based at the centre of gravity of the flight and continually move with the aircraft. Inertial coordinates, on the other hand are defined with respect to the earth and have the origin positioned at some suitable location of the simulated world. Other coordinate systems exist such as wind coordinate and stability coordinates, etc, these coordinate systems are defined by the flight's states [6]. The flight model presented in this paper is based on body and inertial coordinate systems. In general, all aerodynamic forces, accelerations and velocities are calculated in the body coordinate system, and then convert to the world coordinate system before updating the aircraft's position and attitude.

The body coordinate system is one of the most often mentioned coordinate system when discussing the aerodynamic model. The x-axis of the body coordinate system coincides with the fuselage of the aircraft and points forward; the z-axis lies in the aircraft's symmetry plane and points downward, the y-axis is perpendicular to x and z axes forming a right-handed coordinate system. An illustration of the body coordinate system is shown in figure.1.

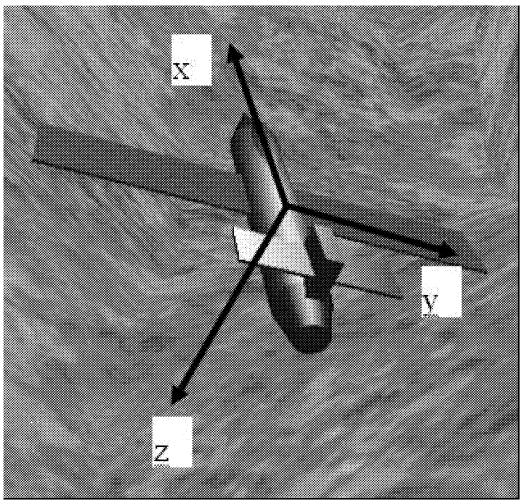


Figure.1. Body coordinates system.

World (or inertial) coordinate system is defined with respect to the earth (or the virtual world) and has the origin positioned at some suitable location. In the real world the origin is the center of the earth, the x-axis points to north, y-axis points to east and z-axis points to downward.

IV Orientation Model

In flight simulation application, the linear and angular velocities are normally calculated in the body coordinates system. However, when the aircraft updates its position and orientation conversion of the velocities from the body coordinate system to the world coordinate system is needed. Three methods exist for defining the conversion of angular velocities to orientations in world coordinate system. They are Euler angle, Direction Cosines and Quaternion Method. We discuss these three methods in this section.

Euler angles: The most common method of defining an aircraft's orientation in world space is by Euler angles. According to Euler's Rotation Theorem an arbitrary rotation may be described by only three parameters [7]. Using a sequence of three angles, the Euler method provides an intuitive description of aircraft attitude in world space [5]. These angles are the azimuth angle ψ, the pitch angle θ and the roll angle φ. The orientation of one Cartesian coordinate system with respect to another can always be described by three successive rotations, and the angles of rotation are called Euler angles. Euler angles have an immediate intuitive appeal and are perfectly adequate for the simulation of civil aircraft. However, the Euler angles are not unique for a given orientation, which means that the order of these rotations is important. Euler chose the convention of rotating first about the z-axis, then about the new y-axis and finally about the new x-axis. This convention is most often used in graphics and has been proposed as the standard for aerospace engineer applications [8]. A further restriction has been done on the amplitude of Euler angles. The pitch angle has been restricted within the scope of $[-\pi/2, \pi/2]$. When the pitch angle reaches $\pm\pi/2$, the roll and yaw angles change abruptly by π radians. This is a more mathematically convenient choice. There is a direct relationship between Euler angles

and the angular velocity of the aircraft within body coordinate system [9], which is shown in below equations.

$$\dot{\phi} = p + q\sin\phi\tan\theta + r\cos\phi\tan\theta \quad (1)$$

$$\dot{\theta} = q\cos\phi - r\sin\phi \quad (2)$$

$$\dot{\psi} = q\sin\phi\sec\theta + r\cos\phi\sec\theta \quad (3)$$

Direction Cosines Matrix: The direction cosine matrix is so-called because its element can be determined from dot product that involves the direction cosines between corresponding axes of the new and old coordinate systems. In the flight simulation, transforming between body coordinates and world coordinates is done frequently. The Direction Cosines Matrix is a convenient way to represent the transformation between two coordinate systems. Using matrix notation and the direction cosines (a, b, c), the transformation between the body coordinate system and the world coordinate system are expressed by the following equations [3]:

$$\begin{bmatrix} Xw \\ Yw \\ Zw \end{bmatrix} = \begin{bmatrix} a1 & b1 & c1 \\ a2 & b2 & c2 \\ a3 & b3 & c3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (4)$$

The relationships between the Euler angles and the Direction Cosines Matrix can be expressed by below equations:

$$\begin{cases} a1 = \cos\theta\cos\psi \\ a2 = \cos\theta\sin\psi \\ a3 = -\sin\theta \end{cases} \quad (5)$$

$$\begin{cases} b1 = \sin\phi\sin\theta\cos\psi - \cos\phi\sin\psi \\ b2 = \sin\phi\sin\theta\sin\psi + \cos\phi\cos\psi \\ b3 = \sin\phi\cos\theta \end{cases} \quad (6)$$

$$\begin{cases} c1 = \cos\phi\sin\theta\cos\psi + \sin\phi\sin\psi \\ c2 = \cos\phi\sin\theta\sin\psi - \sin\phi\cos\psi \\ c3 = \cos\phi\cos\theta \end{cases} \quad (7)$$

Also, when the Direction Cosines Matrix is fixed, Euler Angles can be determined by the following equations:

$$\theta = a\sin(-a_3) \quad (8)$$

$$\psi = a\cos(a_1/\cos\theta)(sign[a_2]) \quad (9)$$

$$\phi = a\cos(c_3/\cos\theta)(sign[b_3]) \quad (10)$$

Quaternion Method: An alternative method has gained popularity in presentation orientation is using the unit quaternion. Quaternion has been introduced because of their "all-attitude" capability and numerical advantages in simulations and controls. W. R. Hamilton introduced the quaternion form, which provides an efficient means for updating orientations [7]. The basic idea of quaternion method is to find a way of describing the general rotation of one 3D coordinate system to another by using a general rotation matrix T, which can be written in terms of four quaternion parameters q0, q1, q2 and q3. The matrix has the following expression [6]:

$$T = \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1q_2 - q_0q_3) & 2(q_0q_2 + q_1q_3) \\ 2(q_1q_2 + q_0q_3) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_2q_3 - q_0q_1) \\ 2(q_1q_3 - q_0q_2) & 2(q_2q_3 + q_0q_1) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix} \quad (11)$$

One coordinate system (body coordinate system) can be brought into coincidence with a reference coordinate system (inertial coordinate system) by a single rotation D (here, D is rotational angle) about a fixed axis making angles A, B and C with the reference coordinate system. The four parameters A, B, C and D define the orientation of the aircraft in world coordinates system [11]. q0, q1, q2 and q3 can be determined by A, B, C and D using below equations.

$$\begin{cases} q_0 = \cos(D/2) \\ q_1 = \cos A\sin(D/2) \\ q_2 = \cos B\sin(D/2) \\ q_3 = \cos C\sin(D/2) \end{cases} \quad (12)$$

The quaternion parameters may be related to the Euler angles through the following equations:

$$\begin{cases} q_0 = \cos(\phi/2)\cos(\theta/2)\cos(\psi/2) + \\ \qquad \sin(\phi/2)\sin(\theta/2)\sin(\psi/2) \\ q_1 = \sin(\phi/2)\cos(\theta/2)\cos(\psi/2) - \\ \qquad \cos(\phi/2)\sin(\theta/2)\sin(\psi/2) \\ q_2 = \cos(\phi/2)\sin(\theta/2)\cos(\psi/2) + \\ \qquad \sin(\phi/2)\cos(\theta/2)\sin(\psi/2) \\ q_3 = \cos(\phi/2)\cos(\theta/2)\sin(\psi/2) - \\ \qquad \sin(\phi/2)\sin(\theta/2)\cos(\psi/2) \end{cases} \quad (13)$$

Quaternion and Euler angles each have their own advantages and disadvantages. Euler angles use only three components instead of four to represent orientation; it is the most intuitive method. Furthermore, comparing with quaternion method Euler angle is more computationally efficient [10]. The most significant advantage of quaternion is that no singularity exists when the pitch angle passes through ±π/2. In the Euler method $\dot{p}$ and $\dot{r}$ both become undefined in this situation due to division by zero. This drawback can be solved by using Truncating values as ±π/2 is approached. If the pitch angle is truncated at values of 89.99 and 90.01 then a 0.02 degree rotation skip results [11]. However, in higher accuracy flight simulation, such a strategy will hurt the simulation fidelity.

V Aerodynamic Model

In this section we discuss the aerodynamic model. The nomenclatures used in this section are listed in the appendix. To develop the aircraft equation of motion, six equations are usually derived for the three applied forces along each of the three axes and for the three applied torques around each of the three axes [12]. In vector form we can write:

$$F = m\dot{V} + \Omega \times mV \quad (14)$$

$$T = I\dot{\Omega} + \Omega \times I\Omega \quad (15)$$

Where V= [u, v, w], Ω= [p, q, r], F= [F_X, F_Y, F_Z] and T= [L, M, N]. m is the mass of the aircraft and I is the inertia of the aircraft. The symbol "×" represents the cross product. F_X, F_Y and F_Z are the forces along the aircraft's body axes, which are the resultant of the external forces on the aircraft. These forces include lift, drag, side-force and engine thrust. In order to simplify the aerodynamic model, engine thrust is

limited to the X-axis of body coordinate system; the gravity force is considered as a constant, which is listed separately in the below equations in order to simplify the expression. The linear accelerations of the aircraft can be calculated by the below equations:

$$\dot{u} = vr - wq - g\sin\theta + \frac{F_X}{m} \tag{16}$$

$$\dot{v} = wp - ur + g\sin\phi\cos\theta + \frac{F_Y}{m} \tag{17}$$

$$\dot{w} = uq - vp + g\cos\phi\cos\theta + \frac{F_Z}{m} \tag{18}$$

The effect of the gyroscopic on angular acceleration has not been considered here since the relevant literature indicates that it only has second order effect on the aircraft's performance [11]. The moments around the aircraft's body axes can be expressed as following:

$$L = I_{XX}\dot{p} - I_{XZ}\dot{r} - I_{xz}pq + (I_{ZZ} - I_{YY})rq \tag{19}$$

$$M = I_{YY}\dot{q} + (I_{XX} - I_{ZZ})pr + I_{XZ}(p^2 - r^2) \tag{20}$$

$$N = I_{ZZ}\dot{r} - I_{XZ}\dot{p} + (I_{YY} - I_{XX})pq + I_{XZ}qr \tag{21}$$

The aerodynamic forces and moments are assumed can be expressed as a function of the instantaneous value of the perturbation variables [9]. The perturbation variables also known as non-dimensional coefficients in aerodynamic engineer, which are the derivatives of forces (lift, drag, side-force) or moments respect to aircraft's states. Normally the aircraft's states include the linear velocity, angular velocity, angle of attack, sideslip angle, control deflections and their derivatives. Non-dimensional coefficients are generated in actual aircraft testing, which are available for most aircrafts. Using the non-dimensional coefficients, lift, drag and side-force can be calculated by below equations [10]:

$$Lift = 0.5*[C_{L0} + C_{L\alpha}\alpha + C_{Lq}q\frac{c}{2V} + C_{L\dot{\alpha}}\dot{\alpha}\frac{c}{2V} + C_{L\delta e}\delta e]\rho V^2 S \tag{22}$$

$$Drag = 0.5*[C_{D0} + C_{D\alpha}\alpha + C_{D\delta e}\delta e]\rho V^2 S \tag{23}$$

$$S_F = 0.5*[C_{Y\beta}\beta + C_{Y\delta r}\delta r]\rho V^2 S \tag{24}$$

The moment around the body axes are determined by the following equations:

$$L = 0.5*[C_{Ll\beta}\beta + C_{Lp}p\frac{b}{2V} + C_{Lr}r\frac{b}{2V} + C_{L\delta a}\delta a + C_{L\delta r}\delta r]\rho V^2 S \tag{25}$$

$$M = 0.5*[C_{M0} + C_{M\alpha}\alpha + C_{Mq}q\frac{c}{2V} + C_{M\dot{\alpha}}\dot{\alpha}\frac{c}{2V} + C_{M\delta e}\delta e]\rho V^2 S \tag{26}$$

$$N = 0.5*[C_{N\beta}\beta + C_{Np}p\frac{b}{2V} + C_{Nr}r\frac{b}{2V} + C_{N\delta a}\delta a + C_{N\delta r}\delta r]\rho V^2 S \tag{27}$$

Once Lift, drag and side-force are calculated, they are translated into forces along the aircraft body coordinate axes. F_X, F_Y and F_Z represent the resultant aerodynamic forces.

$$F_X = Lift\sin\alpha - Drag\cos\alpha - S_F\sin\beta + F_{Thrust} \tag{28}$$

$$F_Y = S_F\cos\beta \tag{29}$$

$$F_Z = -Lift\cos\alpha - Drag\sin\alpha \tag{30}$$

In order to simplify the expression of angular acceleration, two terms are introduced here:

$$L' = L + I_{XZ}pq - (I_{ZZ} - I_{YY})rq \tag{31}$$

$$N' = N - (I_{YY} - I_{XX})pq - I_{XZ}rq \tag{32}$$

The angular velocity can be expressed as following:

$$\dot{p} = (L'I_{ZZ} - N'I_{XZ})/(I_{XX}I_{ZZ} - I_{XZ}^2) \tag{33}$$

$$\dot{q} = (M - (I_{XX} - I_{ZZ})pr - I_{XZ}(p^2 - r^2))/I_{YY} \tag{34}$$

$$\dot{r} = (N'I_{XX} + L'I_{XZ})/(I_{XX}I_{ZZ} - I_{XZ}^2) \tag{35}$$

VI Simulation Results and Discussion

This 6 DOF aircraft model has been successfully implemented in our flights simulation experiment. The simulation code is written in C++ and graphics interface OpenGL. The virtual scene currently includes three flights and a 3D environment with texture mapping and lighting. The simulation program is tested on a desktop computer, the CPU is Pentium 4 (3.0 GHz) and the video card is Quadro FX 500 with 128MB memory. The average frame rate is beyond 80fps. One thing worth mention here is that the aerodynamic model exhibits a tendency to "blow up" if the time step between aerodynamic calculations becomes too great. Figure.2 shows an example of the "blow up" tendency in coordinate-turn flying conditions.

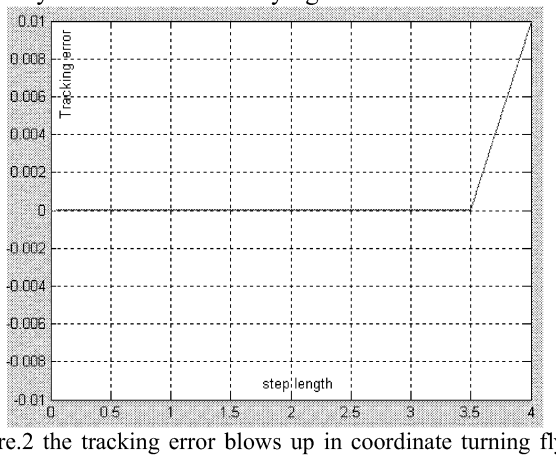


Figure.2 the tracking error blows up in coordinate turning flying condition when the time step increase.

When the time step is bigger the numeric simulation results becomes instable. Such kind of "blow up" will happen when the aircraft possesses big rotation rates. A simple method is used in our simulation to solve this problem. The time step is adjusted according to the aircraft's angular velocity, which can be formulized by below equation.

$$step = scale_cof/(rotation + Maxrate) \tag{36}$$

As the aircraft's angular velocity increasing, the time step decreases. The "Maxrate" is a plus constant, which preserves the maximum time step in level flying conditions. Figure.3 shows a simulation case using this aircraft model combine with a simple feedback PID controller. In this simulation the aircraft is controlled to track a pre-designed trajectory, which is marked by dash-curve in figure.3, the actual flying trajectory is marked by solid curve. The average position tracking error is 10.98 ft in this figure.

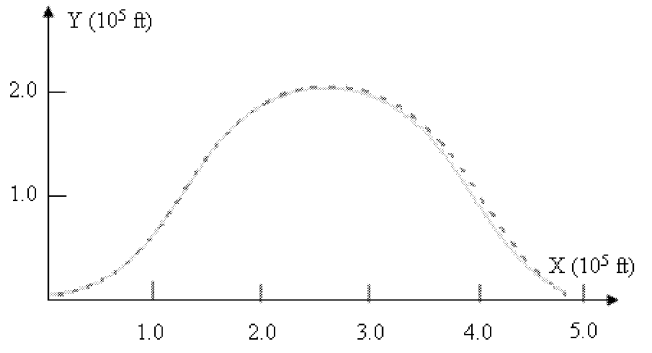


Figure.3. One case of flight simulation: tracking the pre-designed trajectory.

VII Conclusion

In this paper, we present a nonlinear 6 DOF aircraft model for general flight simulation on desktop computers. This nonlinear aircraft model reduces the need for large amount of flying data and calculation while keeping the proper nonlinearities of the general aerodynamic models. Therefore, a practical balance between the simulation speed and the simulation fidelity can be achieved. This model is suitable for general flight simulation applications on desktop computers. This nonlinear aircraft model has been successfully implemented in our 3D flight simulation platform, the performance is satisfactory.

References

[1] Mario S.V. Valenti Clari, Rob C.J. Ruigrok, Bart W.M. Heesbeen and Jaap Groeneweg, "Research Flight Simulation of Future Autonomous Aircraft Operations", Proceedings of the 2002 Winter Simulation Conference.

[2] E. Bruce Jackson, "Results of a Flight Simulation Software Methods Survey", AIAA Flight Simulation Technologies Conference August 7-9, 1995.

[3] Albert J. Rehmann, Michael C. Reynolds, Robets D. Mitman "A Handbook of Flight Simulation Fidelity Requirements for Human Factors Research". Report No: DOT/FAA/CT-TN95/46, Crew System Ergonomics Information Analysis Centre.

[4] Tom Benson, "Beginner's Guide to Aerodynamics", NASA Glenn Research Center. http://www.grc.nasa.gov/WWW/K-12/airplane/bga.html.

[5] Rolfe. J. M., Staples. K. J.,"Flight Simulation", Cambridge University Press, New York, NY 10022, 1986.

[6] Ilan Kroo, "Aerodynamic Modeling for Simulation and Control". http://adg.stanford.edu/aa208/amsc.html

[7] Eric W. Weisstein, "Euler's Rotation Theorem", From MathWorld http://mathworld.wolfram.com/EulersRotationTheorem.html.

[8] UCF/IST, "Military Standard (DRAFT) for Protocol Data Units for Distributed Interactive Simulation", University of Central Florida Institute for Simulation and Training, June, 1990.

[9] Nelson Robert C, "Flight Stability and Automatic Control", McGraw-Hill Book Company, New York, 1989.

[10] Joseph M. Cooke, Michael J. Zyda, David R. Pratt and Robert B. McGhee, "NPSNET: Flight Simulation Dynamic Modeling Using Quaternions", Department of Computer Science, Monterey, Califrnia, USA.

[11] Goldiez, Brian and Lin, Kuo-Chi, "The Orientation Representation in the Draft Military Standard for Distributed Interactive Simulation", AAI, University of Central Florida, Orlando, Florida, 1991.

[12] Raymond T. Stefani, Bahram Shahian, Clement J. Savant, Jr and Gene H. Hostetter, "Design of Feedback Control Systems", OXFORD UNIVERSITY PRESS, 2002.

[13] Istvan Szaszi and Balazs Kulcsar, "Robust Control and Fault Detection Filter Design for Aircraft Pitch Axis", Periodica Polytechnica Ser Transp. ENG. VOL. 29, NO.1-2, PP 83-100.

Appendix

Nomenclatures used in this paper:

Φ, θ, ψ Roll, Pitch and Azimuth angle.

u, v, w Linear velocity along X, Y and Z body axes.

p, q, r Angular velocity around X, Y and Z body axes

L, M, N Moments about the X, Y and Z body axes

F_x, F_y, F_z Forces on aircraft within body axes

α Angle of attack

β Sideslip angle

S Surface area of wing

b Wing span

c Chord length

I_{xx}, I_{yy}, I_{zz} Roll, Pitch and Yaw inertia

$C_{L0}, C_{L\alpha}, C_{Lq}, C_{L\dot{\alpha}}, C_{L\delta e}$ Non-dimensional coefficients of Lift

$C_{D0}, C_{D\alpha}, C_{D\delta e}$ Non-dimensional coefficients of Drag

$C_{Y\beta}, C_{Y\delta r}$ Non-dimensional coefficients of sideforce

$C_{M0}, C_{M\alpha}, C_{Mq}, C_{M\dot{\alpha}}, C_{M\delta e}$ Non-dimensional coefficients of M-moments

$C_{L\beta}, C_{Lp}, C_{Lr}, C_{L\alpha a}, C_{L\delta r}$ Non-dimensional coefficients of L-moments

$C_{N\beta}, C_{Np}, C_{Nr}, C_{N\delta a}, C_{N\delta r}$ Non-dimensional coefficients of N-moments

3D VISUALISATION OF MUSICAL CONTENT

Jacek Grekow
Faculty of Computer Science
Technical University of Bialystok,
Wiejska 45A, Bialystok 15-351, Poland
E-mail: grekowj@ii.pb.bialystok.pl

KEYWORD
Visualisation, Musical content analyse, Sound simulation, 3D presentation, Acoustic information.

ABSTRACT

Audio information we receive is a complicated phenomenon. Consequently, an extensive analysis of a piece of music is a complex task due to the sheer number of data attributes such as tempo, rhythm, pitch, consonance, harmony, dynamics. That is the reason why all the research conducted to date in this field has concentrated solely on selected parameters, and that fact determined the aim of the experiments as well as their results. This paper presents a totally new approach to creating 3-D images to the analysis and simulation of musical content. As the main criterion for the synthesis of the 3-D images we have chosen harmonic content, which has close relation to the musical content of a piece. Additionally, the graphic display of its main components in shape of 3-D figures is a crucial aspect in the presented approach. The figures, which were named AKWETs, are applicable in simulation and analyse musical content. The paper also states that the the order of AKWETs identified in a given phrase forms a meaningful visual sequence, connected with its musical original.

INTRODUCTION

The following questions have been asked for a long time: why do people admire a particular piece of music, whilst are not attracted to another? What criteria need to be fulfilled for the listener to accept or reject a given performance? By what means do people's hearing and brain analyse the information carried by sound? Only recently academic papers have emerged that study the musical content with application of computer techniques.

Audio information we receive is a complicated phenomenon. Consequently, an extensive analysis of a piece of music is a complex task due to the sheer number of data attributes such as tempo, rhythm, pitch, consonance, harmony, dynamics. That is the reason why all the research conducted to date in this field has concentrated solely on selected parameters and that fact determined the aim of the experiments as well as their results. Some authors analyse different performances of one piece of music, restricting their attention to two expressive dimensions: tempo and loudness (Dixon et al. 2002). Typical example, where musical and spatial parameters are related one to another and form an interactive visual tool are also served (Graves et al. 1997).

This paper presents a an entirely new approach to creating 3-D images to the analysis and simulation of musical content. As the main criterion for the synthesis of the 3-D images we have chosen harmonic content, which has close relation to the musical content of a piece. Additionally, the graphic display of its main components in shape of solid figures is a crucial aspect in the presented approach. The figures, which were named AKWETs, are used to simulation and analyse musical content. The paper also states that the the order of AKWETs identified in a given phrase forms a meaningful visual sequence, connected with its musical original.

Visualisation is often used in academic research to illustrate relation between various parameters. Applied to analysis of musical content, visualisation reveals its aesthetical value, as it combines two independent human senses. What is more, it helps obtain information that is not always clearly perceived when considerable amount of data is given. We can assume that we obtain information of a different level and our cognitive potential extends. Consequently, the presented AKWETs enable us to:

- demonstrate correlation between 3-D images and music
- obtain additional information about chords - their type, internal consonance and dissonance etc.
- draw conclusions about how chords follow in succession
- get a new look at the overall structure of a piece of music

This area is still open for further research.

METHOD OF CREATING 3-D IMAGES

Harmonic content of a piece of music consists of chords that are related to its consecutive segments, and that relation depends on the sounds of a chord as well as other principles of musical theory (Sikorski 1965).

The main idea of the method presented is based on chords. As mentioned above, chords divide the analysed musical content into segments, the length of which is not equal. It is determined by the mentioned principles (Sikorski 1965). The segments may include a part of a bar or several bars, so that the difference in length between particular segments may be considerable. Our method of audio analysis is completely different in comparison with the traditional ones (Czyżewski 2001).

The fact that most chords contain three notes or it is possible to choose three crucial components from a multi-note chord was taken into account (Sikorski 1965).

Every musical note can be represented in a simplified form as a sine curve of frequency f. Consequently, each

component of chord A is described by means of sine function $S_i(t)$ of frequency f_i.

Each function obtained is associated with one of the axes (X, Y or Z) of Cartesian coordinates {U}, and a 3-D figure Φ_A assigned to chord A is formed.

$$A \rightarrow \Phi_A \qquad (1)$$

We name figure "**AKWET**" = **AK**ord**W**izualny**ET**alon.

The forming of each AKWET may be described as the result of the movement of point P, spatial position of which is determined by function $S_i(t)$ values put on X, Y and Z axes respectively (Figure.1).

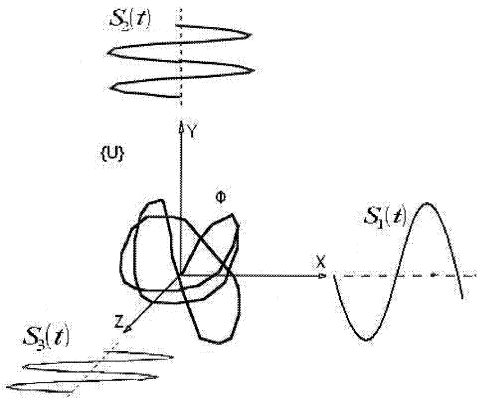


Figure 1: Forming of an AKWET

Each axis of the {U} coordinates is related to a sinusoidal signal $S_i(t)$ of frequency f_i where i = 1,2,3. The signals illustrate consecutive components of the AKWET. When digitised with frequency F >> f_i, the signals $S_i(t)$ form a sequence of samples s_{ji}. Each element of the sequence is a three-dimensional vector that determines position of point P_j.

The coordinates of a single point P_j (P_{jx}, P_{jy}, P_{jz}) , which are calculated on the basis on common time t, and the signal function $S_i(t)$ may be recorded as follows:

$$\begin{aligned} P_j &= (P_{jx}, P_{jy}, P_{jz}) \\ P_{jx} &= S_1(t_j) = A\sin\omega_1 t_j \\ P_{jy} &= S_2(t_j) = A\sin\omega_2 t_j \\ P_{jz} &= S_3(t_j) = A\sin\omega_3 t_j \end{aligned} \qquad (2)$$

Since all the components taken into account are equally important in an AKWET's structure, we investigate three functions of identical amplitude A.

AKWET Φ is a spatial and periodic figure (Figure 1) and its period depends on the values of component signals.

ACHIEVEMENTS TO DATE AND FUTURE WORK

AKWET's structure is similar to Lissajous figures (Szczeniowski 1972). Significant difference lies, however, in the fact that we achieve 3-D figures, which are then used to visualise and analyse sophisticated musical content.

Correlation between AKWET's shape and the sound of chords

Classical music theory uses two basic keys – minor and major, to describe the scale of a particular piece. The keys are represented by minor and major chords, the sound of which substantially differs; simplifying the matter we can state that a minor key is "sad", while a major – "cheerful" (Sikorski 1965). We can consider keys basic AKWETs.

It results that the figures representing major and minor chords respectively differ significantly (see Figure.2). The major chord's shape is simpler and clearer. The minor chord, on the other hand, is denoted with a more complex, multidimensional figure, which cannot, despite its regular shape, be associated with peace.

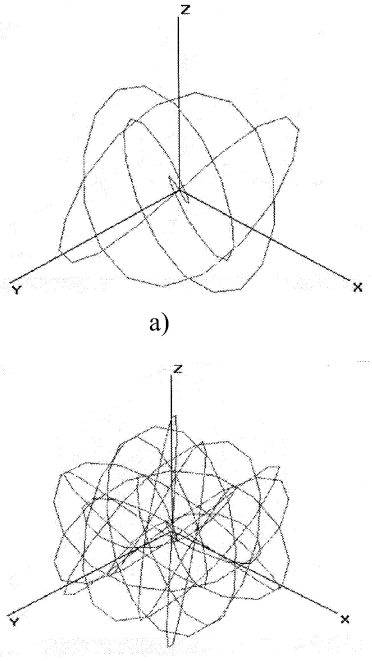


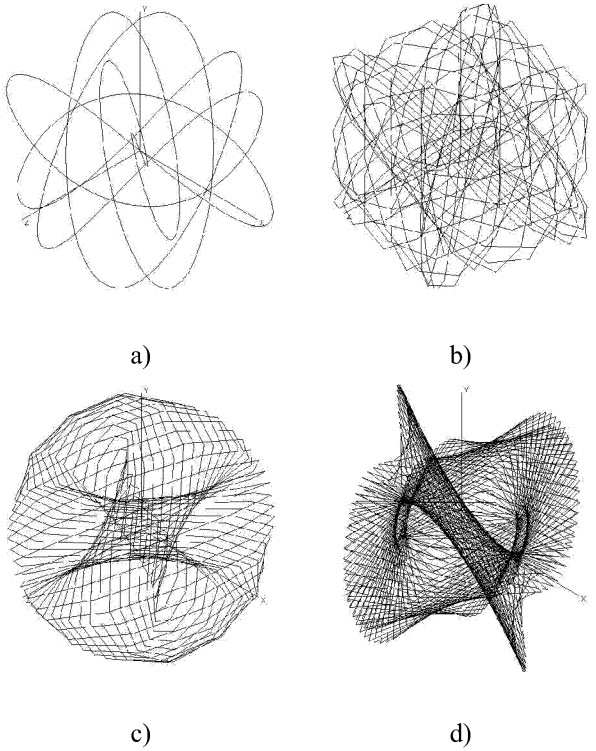
Figure 2: Model patterns a) major AKWET b) minor AKWET

This example is the first evidence that the AKWET is based on the right criterion.

Further development of AKWETs might make use of another criterion – aesthetic qualities of a 3-D image.

Additional information about chords that might be extracted from the figures' shape

An AKWET simulator has been designed and built for further experiments and all the basic, according to the musical laws (Sikorski 1965), chords have been tested. Furthermore, simulator-generated graphics originated ideas of new AKWETs. Figure 3 shows several model patterns depicting internal consonance (a) and dissonance (b) as well as two additional AKWETs (c) and (d), not subjected to musical laws, but interesting for further research.



a) b)

c) d)

Figure 3: Several model AKWETs: a) consonance, b) dissonance, c) new 1, d) new 2.

Simulation of musical content

The sequence of appearance of static 3-D images is an equally interesting field of research. One of the basic patterns of development of a musical piece is called cadence (Sikorski 1965). Figure 4 enables us to visually track the development of a major cadence K_d (T S D^7 T). Consecutive AKWETs form a visual sequence that is temporally connected with the musical original.

Each chord from the sequence K_d has meaningful content reflected in a visual sequence: introduction, development, climax and denouement. The order of AKWETs identified in a given phrase forms a meaningful visual sequence, connected with its musical original.

CONCLUSIONS

The method presented in this paper is an innovative approach to simulation and analyse pieces of music. It is based on combining two human senses: hearing and vision, which results in mutually linked abstract figures that illustrate music and to analyse musical content (AKWETs).

The AKWET method may be applied in:

- visualising music;
- studies and prediction of the impact of music on listeners;
- supporting composer's work in evaluating combination of chords in a composition by visual simulation

We are certain that continuation of this research may bring, especially when the computer programmes are developed, new discoveries in the area that combines music with technology.

REFERENCES

Czyżewski A. 2001. "Dźwiek cyfrowy". EXIT, Warszawa, 3-40.

Dixon S.; Goebl W. and Widmer G. 2002. "Real time tracking and visualisation of musical expression." *Technical Report 2002-04,* Austian Research Institute for Artifical Intelligence.

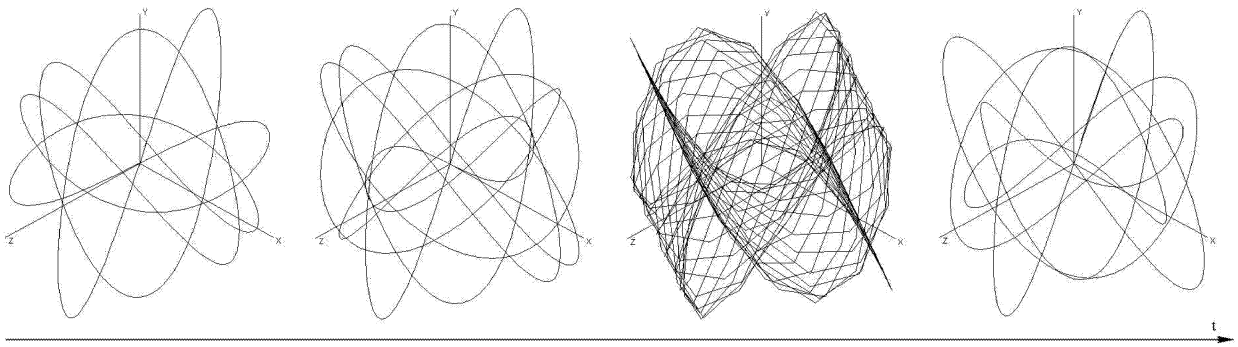
Graves A.; Hand C. and Hugill A. 1997. "Interactive Visualisation of Musical Form using VRML" *The Fourth UK VR-SIG Conference*, Brunel University.

Grekow J 2004. "The method of conversion an audio signals." *Digital Signal Processing and its Applications*, Moskwa.

Sikorski K. 1965. "Harmonia". Polskie Wydawnictwo Muzyczne, Kraków, 25-99.

Szczeniowski S. 1972. "Fizyka doswiadczalna". PWN, Warszawa, 120-129.

JACEK GREKOW was born in Bialystok, Poland. In 1994 graduated from the Technical University of Informatics Sofia, Bulgaria. Graduated also of the Vienna Conservatory and after obtaining his musical degree in 1997, he worked for a year in IBM Vienna, as computer programmer. Since 1998 he worked for a couple of years as computer programmer in T-matic Company before moving to the Technical University in Bialystok, where he is now doing his researches in music visualisation and analysis. His e-mail address is: grekowj@ii.pb.bialystok.pl.



a) b) c) d)

Figure 4: Visualisation of a model cadence K_d (T S D^7 T): a) T (I), b) Subdominant (IV), c) $Dominant^7$ (V^7), d) T (I_1).

A KNOWLEDGE REPRESENTATION SYSTEM TO SUPPORT MEDICAL DIAGNOSTIC IMAGING

Edoardo Ardizzone
Daniele Peri
Dipartimento di Ingegneria Informatica, Università degli studi di Palermo
Viale delle Scienze, Ed. 6
90128 Palermo, Italy
E-mail: peri@csai.unipa.it

KEYWORDS
Health Care, Pattern recognition, AI-supported simulation, Expert Systems, Decision

ABSTRACT

An experimental system to support medical diagnostic imaging is presented. The system is based on an anatomical knowledge representation paradigm able to capture relationships among region of interest in images produced by medical imaging modalities such as MRI ad CT. A description language based on first order logic and sets is used to represent regions of interest and relationships among themselves. Fuzzy logic expressions permit to encode human expert's rules characterized by uncertainty. Applications of the system to encoding of anatomical knowledge describing Carotid arteries in the neck and Multiple Sclerosis lesions are also presented.

INTRODUCTION

The availability of imaging modalities capable of exporting digital data makes medical imaging an interesting and promising field to either test already well known Intelligent Data Analysis techniques or to develop new ones. The target scenario of studies in this field is populated by automatic tools ranging from measurements and estimations aimed to support medical diagnosis and track evolution of diseases, to feature extraction for monitoring patients conditions or easing surgical planning. The automation level achievable in such tools is massively conditioned by the nature of data itself. The great variability which characterizes such data, makes in fact easier to focus mainly over properties of image data which are independent from anatomical knowledge such as, for instance, global measurements, intensity peaks, shape of areas. More robust approaches to automatic imaging data analysis should rather embed some anatomical knowledge (Brinkley et al. 1999). Various examples of applications of methodologies and tools of Artificial Intelligence for representing and manipulating medical knowledge can be found in literature (Wilson et al. 1995). In (Alonge et al. 2001) a software architecture is presented that is able to perform classification of Multiple Sclerosis Disease focal lesions in MR images of the brain depending on a set of measurements related to lesions shape, position and brightness once trained. In that architecture, the clinical knowledge about the disease under investigation was learned by supervised training. Alternatively, the employment of deformable models, showed itself useful in interactive 3D segmentation of medical images (Hinshaw and Brinkley 1997). In this work a methodology to represent knowledge of the anatomical domain in order to perform automatic processing tasks over from medical imaging modalities data is presented. In the followings we describe at first the knowledge description paradigm and the description language structure and main features. Then we show two different applications of the proposed paradigm to medical diagnosis support.

KNOWLEDGE MODELING

Knowledge Modeling Our work is focused on studying the realization of automatic processing tools in the medical imaging domain. In particular we were interested in investigating whether sets of rules, kept close as possible to natural language, could encode enough knowledge to extract autonomously extract clinical information from a set of imaging data. In order to address this problem, we formulated a paradigm based on sets, relationships among them and dynamic classification, we implemented it as a software environment and we tested it in automatic processing of imaging modalities data sets. Our paradigm relies on a knowledge base, initialized with input data sets and rules, which is monotonically and dynamically enriched by applying rules to sets. Sets can be either initialized with data or described by rules encoding relationships with other sets. Rules are expressed by means of first order logic and, in order to encode relationships suitable for the target domain, can contain either crisp or fuzzy expressions (Zadeh 1965) though the paradigm is flexible enough to be extended to the probabilistic realm. In such schema models of anatomic regions can be implemented at different levels of detail and tested over real data, maximization of truth value of sets of rules and crossed relationships among sets can reinforce models. Moreover, the declarative approach makes simpler to the human expert to encode knowledge.

THE SOFTWARE SYSTEM

In order to test our methodology, a software system was designed and implemented. The main goal in designing the system was to realise a general purpose yet powerful tool to perform automatic data processing. Another goal of the design was to let the user interact with processing results through real-time 3D data visualization. The core of the system is constituted by a Prolog environment

(Diaz GNUprolog) enriched with image processing tools and a 3D data visualization module implemented using 3D graphics industrial standards API (OpenGL and Mesa). A description language was designed to encode the knowledge in the system and to perform queries. As aforementioned, our language is closely related to first order logic and Prolog and its syntax is based on predicates and processing scheme is performed by unification and backtracking. Like in Prolog, terms can be numbers, strings, written as series of characters enclosed in single quotes, and variables, whose first letter is capitalized. Programs are constituted by a sequence of linguistic variables and sets declarations.
In general, a declaration is composed of a type identifier (lvar or set), a label, the assignment operator (":="), the body of declaration enclosed in curly brackets and a semicolon as the termination character. Linguistic variables are declared as a series of term each one associated to a membership function of the piecewise linear kind described by couples of real numbers respectively representing the element in universe of the discourse and its fuzzy truth value, which belongs to the closed interval [0, 1]. Linear interpolation allows to evaluate the membership function over the entire real numbers domain. Outside the discrete support of the series of couples the membership function is considered equal to zero.
A linguistic variable declaration is in the form:

```
lvar name: ={
  term term1:=[[x11,y11],[x12,y12],...]];
  term term2:=[[x21,y21],[x22,y22],...]];
  ...
};
```

Fuzzy expressions are evaluated through the predicate "fuzzy_truth(Expression, Value)". A generic fuzzy expression is composed of linguistic variable evaluators, whose syntactic template is "lvar LVar(Term, CrispValue, FuzzyValue), and fuzzy operators (and, or, not).

Sets can be declared in two ways. In the former they are declared as list of elements:

```
set name :={
 [element1, element2, ..., elementn]
};
```

in the latter they are declared through expressions describing relationships with other sets. A declaration is composed of a "constructor" and a logic expression the constructor must satisfy separated by a colon:

```
set name :={
 constructor : logic_expression
};
```

The constructor can be either a single term or a list of terms, the logic expression is constituted by a series of clauses separated by commas and terminated by a dot.
Clauses can be predicates, value assignment expressions and relational expressions composed of algebraic expressions and the relational operators "equal" ("="), "different" ("\="), "greater than" (">"), "less than" ("<"), "greater or equal to" (">=") and "less or equal to" ("=<").
Like in Prolog, clause's variables have a capital letter as first character and underscore is used to specify predicates' parameters which no unification is requested with.
Backtracking is used to find out dynamically constructor values satisfying all the clauses from the leftmost to the rightmost.
An example of set declaration is:

```
set set3 :={
 A :
  member_of_set([A, B], set set1),
  member_of_set(B, set set2).
};
```

which defines set3 as the set of all "A" elements which are the first element of couples belonging to set1 whose second element "B" belongs to set2.
Set can also have parameters to give more generality to relationships among themselves:

```
set section(Z):={
 A :
  member_of_set([A, _, _, Z], set objects).
};
```

defines set section(Z) as the set of 4-uples belonging to set objects having Z as last element.
Once inferred, sets become facts of the knowledge base in order to avoid redundant calculations. For the current applications knowledge base monotonicity is not a concern though a more sophisticated set assertion and retraction mechanism is currently under investigation. Several pre-defined predicates implement operations of image processing and three dimensional representation suitable to operate on CT and MR scan images and to visualize geometric data. Other predicates implement objects and operators of fuzzy logic. Graphics predicates allows definition of 2-dimensional curves, voxel objects, bitmap images, extrusion of curves and generation of solids. The underlying Prolog implementation makes really simple to add other predicates even interfacing with other software subsystems.

APPLICATION TO CT IMAGES: RECONSTRUCTION OF THE CAROTID ARTERIES IN THE NECK

The system was employed to recognize and reconstruct carotid arteries in the neck from the set of CT scan images taken from the Visible Human Dataset. We coded a set of rules stating the conditions to be satisfied from the image areas belonging to Carotid and Vertebral arteries in the neck. These conditions are expressed as possibility distribution of fuzzy linguistic variables in terms of distance and angle values among centroids of the segmented areas. Our declarative approach does not require encoding any algorithm description; nevertheless processing can be seen as a three steps task. In the first step low level image segmentation is performed. Images of the input set are processed one by one and each one is

segmented by thresholding and region growing. As each image represent a slice of a volume, image pixels are actually 2D projections of elementary volumes or voxels. Segmentation extracts from each slice clusters of 8-connected voxels whose intensity, which is proportional, in CT scan images, to volume density, and total volume makes them possible sections of the target anatomical components. In the second step the extracted clusters of voxels are classified applying rules encoding spatial relationships among sections of the target anatomical components in order to find, in each slice, the configuration of clusters maximizing the truth value of the given set of rules. In the declarative approach finding the winning configuration equals labeling each of its clusters as belonging to a specific target section. The third step performs 3-D reconstruction and visualization of the target anatomical components. Voxels belonging to the classified clusters of all the slices are grouped in the four sets associated respectively to right and left Vertebral Arteries and right and left Carotid Arteries. The four sets are then rendered into the 3D interactive environment. In detail, classification relies on a model of geometrical relationships among Vertebral and Carotid arteries, the former being added to the model to improve classification robustness. For each Carotid one or two sections can be present in each slice according to whether the processed section lays above the artery's bifurcation or below.
We then encoded into a set of rules stating the conditions to be satisfied from the image areas belonging to Carotid and Vertebral arteries in the neck. These conditions are expressed as possibility distribution of fuzzy linguistic variables in terms of distance and angle values among centroids of the segmented areas. Angles were represented as positive values in 0-360° range, in the counter-clockwise way. The set "images" contains file name and position, expressed in millimetres, along the normal to scans of every CT scan image arranged in couples. The set "sections(Images, Z)" is declared to contain all the clusters of 8-connected pixel whose intensity and area values lie in the thresholding interval. Its items are extracted from the image of slice whose position in "Z" by the segment predicate.
Classification relationships are based on distance and angle among centroids of pixel regions contained in the set "centroids(Images, Z)" which creates a correspondence among centroids and region's groups of pixels.
Considering the symmetry of the problem and for the sake of simplicity, just the code regarding the right arteries will be showed, where possible. The set:

```
set vcandidates(Images, Z):= {
  [A, B, Truth] :
    member_of_set(A,
     set centroids(Images, Z)
    ),
    member_of_set(B,
     set centroids(Images, Z)
    ),
    distance(A, B, Dis),
    angledeg(A, B, Ang),
    fuzzy_truth(
     lvar distance(v_right_v_left, Dis)
      and lvar(angle, v_right_v_left, Ang),
      Truth
    ),
    Truth > 0.5.
 };
```

is declared to contain all the couples of centroids whose evaluation in fuzzy expression "v_right_v_left" is greater than 0.5. The predicate "member_of_set(Element, set SetName)" unifies Element with an element of SetName set.

Set:

```
set rvertebral(Images, Z):= {
  A :
    max_of_set(
     set vcandidates(Images, Z),
     [A,_, Tval],
     Tval
    ).
};
```

is declared to contain respectively the first element of "vcandidates" set element whose third element, the membership value of the first two in the fuzzy expression vcandidates, is maximum. The predicate "max_of_set(set SetName, Element, Var, Max)": unifies Var with Element and then unifies Max with the maximum value of variable Var in SetName set.
Sets containing sections of right is declared with respect to the Vertebral arteries:

```
set rcarotid(Images, Z):= {
  B :
   member_of_set(A,
    set rvertebral(Images, Z)
   ),
   subtract_set(
    set centroids(Images, Z),
     set rvertebral(Images, Z),
    One
   ),
   subtract_set(
    One,
    set lvertebral(Images, Z),
    Two
   ),
   member_of_set(B, Two),
   distance(A, B, Dis),
   angledeg(A, B, Ang),
   fuzzy_truth(
        lvar distance(v_right_car_right, Dis),
        and
        lvar angle(v_right_car_right, Ang), Truth
   ),
   Truth >0.5.
};
```

in which "subtract_set(set Set1, set Set2, set Set3)" unifies Set3 with the set obtained subtracting Set1 to Set2. Similarily the set lcarotid(Images, Z) is defined.
Sets rvertebral(Images) is declared as the group of all the sections of Vertebral arteries:

```
set rvertebral(Images):=(
  A :
   member_of_set(
    [A,Z],
    set rvertebral(Images, Z)
   ).
};
```

A similar scheme applies to carotid arteries:

```
set rcarotid(Images):={
  A :
   member_of_set(
    A,
    set rcarotid(Images, Z)
   ).
};
```

In figure 1 an interactive model rendered by the system of the set of voxels of the target anatomical structures is showed.

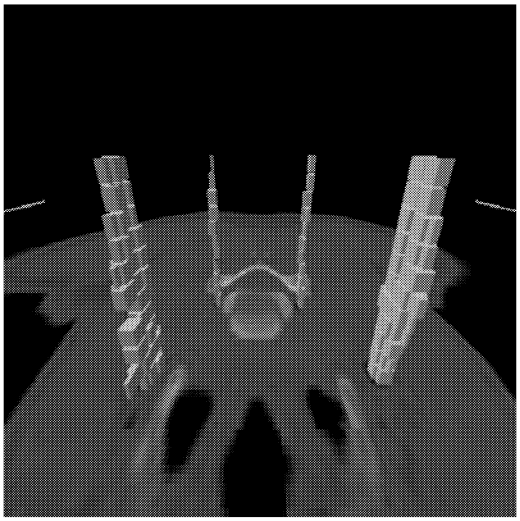


Figure 1: Reconstructed voxel models of Carotid and Vertebral arteries in comparison with one of the CT slices

APPLICATION TO MR IMAGES: VALIDATION OF SEGMENTATION OF MULTIPLE SCLEROSIS LESIONS

Segmentation aimed to extract multiple sclerosis lesions in MR images requires some specific anatomical knowledge (Alonge et al. 2001). This is necessary also to explain the results obtained by segmentation algorithms such as the FCM described in (Alonge et al. 2001) and more effective ones as the two-channels FCM that was used in this study (Ardizzone et al. 2002). Starting from this considerations, we tried to encode such knowledge into our system in order to validate results of segmentations algorithms by comparison with the human experts' diagnosis. In figure 2 PD and T2 images of a brain slice are showed with the images obtained at two different steps of the segmentation process. In the first one, each pixel represents the membership value of the corresponding voxel of the original slice to the cluster of lesions. Thresholding this image to the 50% of its dynamics yielded the second one. Doing this we focused on the knowledge of the human expert without enclosing in our system of rules too many details of the low-level segmentation. Briefly, lesions tend to appear, in brain MR images, as characterized by very high values of either PD and T2 signals and round shape.
Segmentation can also be confused by ventriculi and other cavities containing fluid.

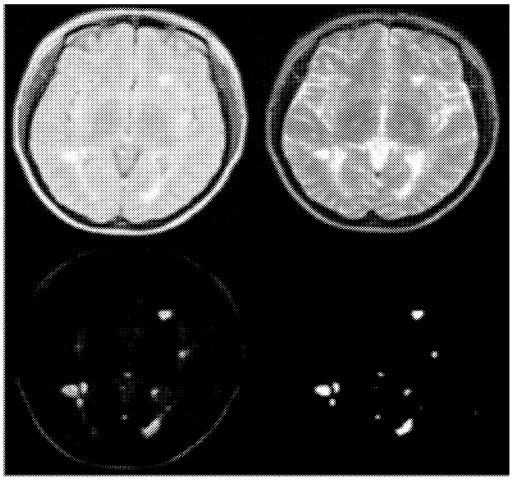


Figure 2: From top to bottom, left to right: PD image, T2 image, segmented fuzzy cluster 7 of a MRI slice of the brain

With respect to the data set of brain MR images we conducted the experimentation on, composed of couples of PD and T2 images for 18 slices of brains, we formulated the fuzzy rules a priori. We also added a rule to consider the distance from the centroid of the image to exclude outer objects. On the basis of these considerations we encoded a rule stating that lesions have very high PD signal values, very high T2 signal values, round shape and not very large distance from centroid of the image:

```
set fval(Slice):={
  Truth :
      member_of_set(
       [AvgPDVal,AvgT2Val,Dist,Area,AspRat],
       set values(Slice)
      ),
      fuzzy_truth(
       lvar signal_pd(very_high, AvgPDVal)
       and
       lvar signal_t2(very_high, AvgT2Val)
       and (
       (lvar aspect_ratio(round, AspRat) and
        not lvar area(very_small, Area) or
        lvar area(very_small, Area)) and not
        lvar distance_from_centroid(
         very_large, Distance
        )
      ),
      Truth
      )
};
```



Figure 3: From top to bottom, areas extracted by the segmentation process, on the left, and their respective validation, on the right for five slices of MR scan. Higher degrees of confidence correspond to darker objects.

Aspect ratio was defined through a rectangular membership function in order to avoid nonlinear effects. Sets "images(Slice)", "clusters(Slice)" were defined as containing respectively input image names and segmented clusters for the slice "Slice". Set "values (slice)" was defined as containing sets of the five values expressing the features for each of the cluster segmented from the slice. Results contained in the "fval(Slice)" set were then used to obtain the images showed in the right column of figure 3 in which higher degrees of confidence correspond to darker objects. These preliminary results encouraged us to close the loop of validation with further steps involving supervised rule induction.

CONCLUSIONS

Conclusions The methodology presented in this work aims to provide tools to express anatomical knowledge by means of a formal description language to be employed in automatic feature extraction from medical imaging data. The shown paradigm proved itself a simple yet powerful mean to express relationships and even hierarchies among anatomical components that can be used in knowledge base driven automatic analysis tools. The implemented software system, whose sets of predicates can be easily extended to implement new data analysis techniques, proved itself to be a flexible environment to implement and test such tools. Wider experimentation is necessary to validate the methodology with respect to specific applications of the type of the presented ones and extend it to other domains. Further investigations are in fact being carried over larger data sets obtained by MRI scans in several parts of the body. The preliminary results seem very interesting.

REFERENCES

Alonge, F., Ardizzone, E. and R. Pirrone. 2001. "An architecture for the recognition and classification of Multiple Sclerosis lesions in MR images". *International Journal of Knowledge-Based Intelligent Engineering Systems* 5, No. 4, 228–233.

Ardizzone, E., Gambino, O., Peri, D. and R. Pirrone. 2002. "Two channels fuzzy C-Means detection of Multiple Sclerosis lesions in multispectral MR images". *In Proceedings of IEEE 2002 International Conference on Image Processing,* (Rochester, NY, September 22-25), .

Brinkley, J.F., Wlong, B.A., Hinshaw, K.P. and C. Rosse. 1999. "Design of an anatomy information system". *IEEE Computer Graphics and Applications* 19, No. 3, 38–48.

Dellepiane, S.,Venturi, G. and G. Vernazza. 1991. "Model generation and model matching of real images by a fuzzy approach". *Pattern Recognition 25,* No.2, 115–137.

D. Diaz. Gnuprolog. http://gnu-prolog.inria.fr.

Hinshaw, K.P. and J.F. Brinkley. 1997. "Shape-based interactive three-dimensional medical image segmentation". *SPIE Medical Imaging: Image Processing, pages* 30, No. 34, 236–242.

NIH. Visible human. http://www.nlm.nih.gov/research/visible/ visible human.html.

Wilson, L.S., Brown, M.S., Talhami, H. , Gill, R.W. , Sun, C. and Doust B. 1995. "Medical image understanding using anatomical models: application to chest X-rays". In *Information Processing in Medical imaging*, Yves Bizais (Ed.), Kluwer:38–48.

Zadeh, L.A. 1965. "Fuzzy sets". *Information and control* 8, 338–353.

Dynamic Generation of Hypermedia Presentations from Distributed Digital Archives using CIDOC CRM

Chia-Hung Lin, Chia-Hung Yeh, Bai-Hsuen Chen, Jen-Shin Hong,
Dept. of Computer Science and Information Engineering, National Chi-Nan University, Taiwan
Email: lch@ncnu.edu.tw

KEYWORDS
Hypermedia Authoring, CIDOC CRM, Digital Museum Exhibitions

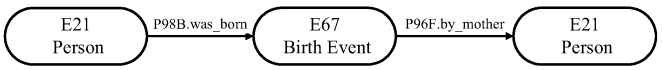
ABSTRACT
This paper proposes a framework for dynamic generation of digital museum exhibitions based the architecture of Semantic Web. In the digital archives, media objects are described based on a metadata model using the CIDOC CRM as the underlying ontology. By pre-mapping the daily concepts to their associate CRM paths, the retrieval interface for the media objects is more user-friendly to exhibition editors. The proposed framework considers various technical aspects in the lifecycle of the exhibition implementations, including the data storage of digital archive, the distributed data access, the query interface design, and the hypermedia styling process.

INTRODUCTION
Over the last decade, aiming to preserve and distribute precious cultural objects, numerous digital archives have been developed worldwide. Based on appropriate designed metadata for annotating the cultural objects, content experts can now search for digital surrogates of cultural artifacts though Internet. Yet, for the general public (who usually have limited domain knowledge) to explore and apprehend the significance of the cultural objects, online hypermedia exhibitions serve a better role than archives. However, to author interactive hypermedia presentations from scratch is complicated. The authoring process requires a sequence of tasks including, topic selection, composition, relevant media objects retrieval, hypermedia composing, and aesthetic style designs. It requires tightly cooperation between content experts, hypermedia engineers, and designers, and is inevitable time-consuming and labor-intensive. How to alleviate the efforts for transforming the digital artifacts into hypermedia exhibition poses a significant challenge.
In the literature, [10] proposed a framework for dynamic generation of hypermedia presentations through semantic inferencing. In the framework, media objects in digital archives are annotated using Dublin Core (DC) metadata model. The media objects relevant to an exhibition can be dynamic retrieved through certain primitive semantic inferences based on the metadata. Various DC elements are associated with certain presentation style parameters such that the style could be assigned automatically to a certain extent by an intelligent styling engine. Yet, since DC Element Set defines only basic access points for digital resources by a minimal set of semantic fields [7], the precision and robustness of the inferences based on DC should be limited.

Other than using field-based metadata model to annotate the digital resources, in the Semantic Web architecture [11], the metadata models are based on domain specific ontologies that specify the underlying knowledge structure pertaining to real life occurrences. This provides a possibility for constructing a "semantic network" of cultural resources with which more extensive inferences could be applied to access the information. In the cultural heritage domain, CIDOC Conceptual Reference Model (CRM) is considered as a carefully designed ontology [6]. For dynamic linking media objects from digital archives to generate exhibitions, we believe CRM could serve a better role than Dublin Core.

However, the complexity associated with CRM poses a challenge for the design of a user-friendly query interface to access digital resources. How to develop a system with which a content expert can easily access digital objects annotate based on CRM model need to be explored. This paper describes our approach of a framework for dynamic hypermedia generations out of CRM-based digital archives. Particular emphasis is put on the design for a user-friendly interface for retrieving CRM-based data. A prototype system will be presented.

FRAMEWORK
This project intends to develop a system with which content authors can easily produce exhibitions without dealing with complex query processes for accessing the media objects annotated by CRM-based metadata. CIDOC CRM provides an extensible vocabulary and knowledge structure about possible states of cultural heritage related events in order to drive queries. In CRM, concepts are interlinked to other concepts through properties. For example, "the mother of a person" is mapped to a "CRM path" as shown in Figure 1. In the Semantic Web framework, assertions could be encoded using RDF (Resource Description Framework) syntax [9]. RDF is a framework for describing and interchanging metadata for the semantic web.



Figures 1: the CRM path for "a person's mother".

There are currently 84 classes (entities) and 141 properties (relationships) defined in CRM. Naturally, it is not practical to expect end users to explore such complex knowledge structure to be able to formulate the queries into the digital archives. A user-friendly interface for assessing the RDF-based digital archives is a must. We present our approach to attack this issue. In the proposed framework, the basic functional module includes (Figure 2):

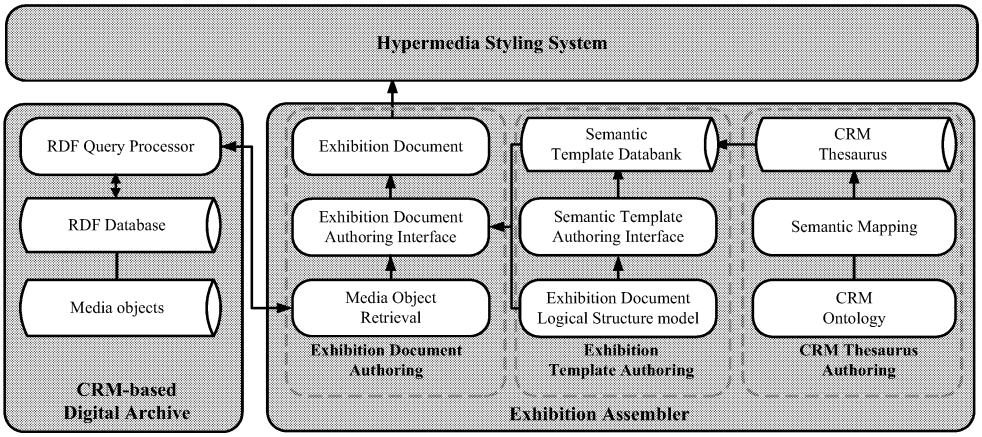


Figure 2: the overall framework

1. CRM-based Digital Archives: serves as a repository of digital objects annotated based on CRM model.
2. Exhibition Assembler: for assembling automatically the relevant media objects into exhibition documents.
3. Hypermedia Styling System: for fine tuning and assigning the presentation styles of the exhibition documents.

1: CRM-based Digital Archive

All the media objects should be properly annotated based CRM information model. The assertions should be described by a sequence of RDF triplets (subject/ predicate / object) to construct a semantic network that represents the knowledge of the media objects. The *RDF Query Processor* processes the query by traversing and exploring the semantic network.

2: Exhibition Assembler

The system provides a user-friendly interface for the exhibition editors to easily access the media objects deposited in digital archives. Formulating the queries in RDF-based framework is not intuitive as conventional keyword-based queries. In general, most current design of the interfaces to query RDF database are implemented on top of certain textual-based RDF Query Languages, such as RQL, SquishQL, etc [1, 2, 4, 5, 8]. These approaches provide SQL-like syntax with which the end users formulate queries using commands like "select", "where", etc. The users need to understand the underlying domain ontology as well as the query language syntax to formulate the query. In our application, since we can't expect the exhibition editors to delve in the complicated CRM architecture, we have to design a more intuitive interface with which they can access the digital archives with ease.

The *Exhibition Assembler* includes three major parts which manage the authoring tasks of *CRM thesaurus, exhibition templates and exhibition documents, r*espectively. The CRM thesaurus is a databank of the mapping between numerous daily concepts and their associate CRM paths. This needs to be done by CRM experts. Appropriate constraints should also be specified in the CRM thesaurus to provide an interface for further refining the potential query patterns through the daily concepts. For example, if the user want to find a person's children who was born in USA, the constrain should be defined as "E21 Person → P98B was born → E67 Birth Event → P7 took place at → E53 Place= USA").

In the *Exhibition Template Authoring* module, an exhibition editor can design exhibition templates by organizing a number of concatenated daily concepts with which the CRM paths has been pre-defined in the *CRM Thesaurus*. This facilitates the mass production of exhibitions in which the underlying semantic architecture is identical. (e.g., to present the person's name in the title of the exhibition document, the general information in first paragraph, a sequence of images in the second paragraph). In this module, an exhibition editor does not have to understand the CRM model to author the exhibition template.

In the *Exhibition Document Authoring* module, an exhibition editor can first select a specific exhibition template and then input the subject to automatically generating the exhibition document on the fly. The *Media Object Retrieval* module first looks in the CRM thesaurus for the CRM paths of the daily concepts specified in the exhibition document. The CRM paths then are passed to the digital archives to retrieve relevant data to be assembled into an exhibition document. The exhibition document can now be send to the *Hypermedia Styling System* for further fine tuning and styling. The exhibition document is based on our design of a hypermedia document logical structure model [3].

3: Hypermedia Styling System

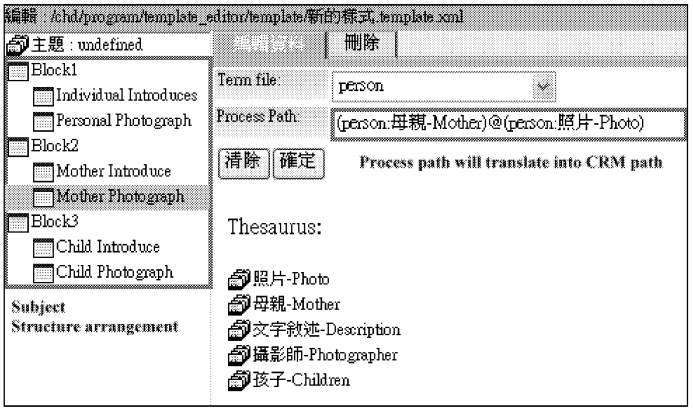
The *Hypermedia Styling System* provides the exhibition editors an interface to revise and specify the presentation styles of the composed exhibition documents. The more detailed discussion of this issue was elaborated in [3].

SYSTEM IMPLEMENTATION

In the prototype system, to facilitate the realization of the proposed framework, all the CRM-based RDF statements are stored in a relation database [1, 2, 4, 5]. The *Exhibition Assembler* communicates with the *CRM-based Digital Archives* using Web Service technology [12]. Web Service architecture is an interoperability framework which provides a standard mean of communications between different software applications in a distributed computing environment. The *Exhibition Assembler* uses SOAP (Simple Object Access Protocol) to dispatch the query commands to RDF database for retrieving media objects. The *RDF Query Processor* then rewrites the CRM-based query into SQL syntax and sends to the relational database. The prototype system was implemented using ASP.NET.

Example Scenario

In this section, we present a demonstrative scenario to illustrate the exhibition authoring process using the prototype system. In this example, an exhibition editor intends to produce a sequence of hypermedia exhibitions to introduce various fundamental personal information of certain historical celebrities. Each exhibition contains three blocks (paragraphs). The first block contains the textual information of the person, and then the images of the person's photos. The second block contains the textual information of the person's mother, and then the images of the person's mother's photo. The third block contains the textual information of the person's children, and then the images of the person's children's photos.



Figures 3: A snapshot of the authoring interface for the *Exhibition Template Authoring* module.

To create a linkage of the "images of the photos of a person's mother" into an exhibition template, the exhibition editor can simply click on the icon of mother, and then on the icon of photo ("person @ mother @ photo", Figure 3). This concatenates the CRM paths of the daily concepts-mother and photo. The other blocks can be done in the same manner. In the next step, the exhibition editor then selects the exhibition template and input the subject (e.g., Napoleon) in the *Exhibition Document Authoring* module. The system then retrieves and assembles all relevant data into an XML-based exhibition document and sends to the *Hypermedia Styling System* for subsequent styling. Figure 4 shows a snapshot of a final presentation of an exhibition document.



Figures 4: A snapshot of an example hypermedia presentation

CONCLUSION and FUTURE WORK

This paper proposes a framework for dynamic generation of hypermedia exhibitions out of RDF-based digital archives. We particular emphase the pre-mapping of CRM-path to daily concepts to simplify the retrieval interfaces of digital resources. Our system provides a graphical interface for the content editors to easily compose an exhibition document without any knowledge of the query syntax associated with the CRM and the data storage schema within the digital archives. Based on a primitive evaluation on the usability of the prototype system, our approach appears to be promising.

In the future, we plan to conduct more in-depth research on systematic approaches for mapping the daily concepts to CRM paths. We also intend to convert a number of existed digital archives to CRM compatible one. Only then, we can know the exact efficacy of the proposed approach in real life applications.

ACKNOWLEDGEMENT

This research was supported in part by a grant by National Science Council of Taiwan, grants No. 93-2422-H-260-002. The authors thank Dr. Martin Doerr for kindly providing us the instruction in CIDOC CRM.

REFERENCES

1. Aimilia Magkanaraki et al. 2002. "Ontology Storage and Querying." ICS-FORTH, Technical Report, No. 308.
2. Gregory Karvounarakis et al. 2002. "RQL: A Declarative query language for RDF." The Eleventh International World Wide Web Conference, Honolulu, Hawaii, USA, pp. 592-603.
3. Jen-Shin Hong et al. 2004. "Toward intelligent styling for digital museum exhibitions: modularization framework for aesthetic hypermedia presentations." International Journal on Digital Libraries Vol. 4, No. 1, pp. 64-68.
4. JENA, http://www.hpl.hp.com/semweb/javadoc/
5. Libby Miller et al. 2002. "Three Implementations of SquishQL, a Simple RDF Query Language." The International Semantic Web Conference, Sardinia, Italy, pp. 423-435.
6. Martin Doerr. 2002. "The CIDOC CRM - an Ontological Approach to Semantic Interoperability of Metadata." AI Magazine, Vol. 24, No. 3, http://cidoc.ics.forth.gr/.
7. Martin Doerr. 2000. "Mapping the Dublin Core Metadata Element Set to the CIDOC CRM." ICS-FORTH, Technical Report 274.
8. Peter Haase et al. 2004. "A Comparison of RDF Query Languages." The Third International Semantic Web Conference, Hiroshima, Japan, pp. 502-517.
9. RDF, http://www.w3.org/RDF/
10. Suzanne Little et al. 2002. "Dynamic Generation of Intelligent Multimedia Presentations through Semantic Inferencing." Proceedings of the Sixth European Conference on Research and Advanced Technology for Digital Libraries, pp.158-175.
11. Semantic Web, http://www.w3.org/2001/sw/
12. WebService, http://www.w3.org/2002/ws/

SIMULATION IN ROBOTICS

USING „COMPLETE KINEMATICAL STRUCTURE" IN A FASTER METHOD FOR ROBOT KINEMATIC MODELLING

Viorel Stoian
Mechatronics and Automatic Control Department
University of Craiova
Bd. Decebal, no. 107, 200440, Craiova
Romania
E-mail: stoian@robotics.ucv.ro

KEYWORDS
Kinematical model, Complete Kinematical Structure, Modelling algorithm.

ABSTRACT

This paper presents a faster method for robot modelling by direct kinematics using „*Complete Kinematical Structure* – CKS". It is a stucture with 3 revolute joints and 3 prismatic joints which can assure a complete positioning and orientation in space for an object attached to its terminal. Its homogeneous transformation operator between fix frame (R_F) and mobile frame (R_M) is determined. An algorithm for robot modelling by CKS is enunciated. Finally, two applications of some robotic structures are presented.

INTRODUCTION

Usually, in order to deal with the complex geometry of a manipulator and to obtain the kinematical equations of the robot models we will attach frames to the various parts of the mechanisms and then describe the relationship between these frames. Since the links of a robot arm may rotate and/or translate with respect to a reference coordinate frame, the total spatial displacement of the end-effector is due to the angular rotations and linear translations of the links (Schilling 1990). Denavit and Hartemberg (1964) proposed a systematic and generalized approach of using matrix algebra to describe and represent the spatial geometry of the links of a robot arm with respect to a fixed reference frame. The advantage of using the Denavit-Hartemberg representation of linkages is its algorithmic universality in deriving the kinematical equation of the robot arm (Asada 1986), (Bishop 2002), (Coiffet 1983), (Craig 1989). By Denavit-Hartemberg method results a homogeneous transformation matrix for every couple of two adjacent rigid mechanical links and all those matrices have to be multiplied. Present method is based on DH approach but decreases drastically the number of the matrices and simplifies a lot the calculations of matrix product (Stoian 1994).

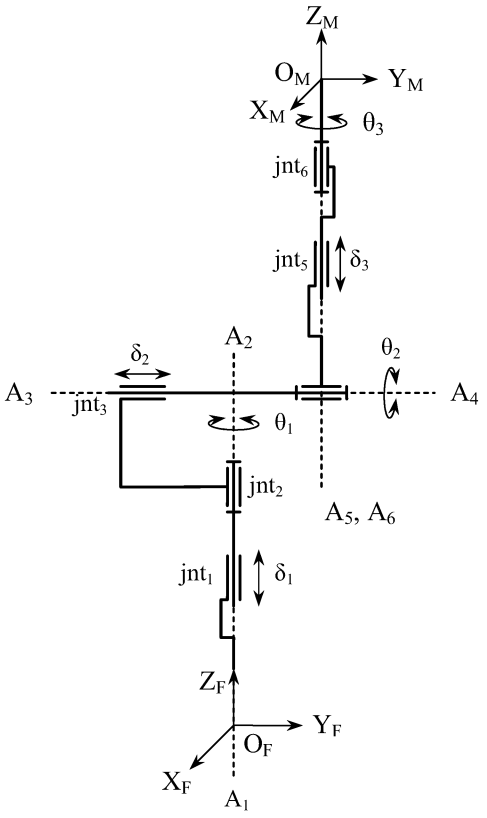


Figure 1: "Complete Kinematical Structure" (CKS)

COMPLETE KINEMATIC STRUCTURE

We name the kinematical structure from Figure 1 a "*complete kinematical structure*" (CKS) because, with 3 rotation joints and 3 translation joints, it can assure a complete positioning and orientation in space for an object attached to its terminal. The above mentioned two operations are not uncoupled. This structure is a useful tool for a faster establishing of the

kinematical model (direct problem) for the robot arms and helps to resolve the inverse kinematical problems in some situations.

THE HOMOGENEOUS TRANSFORMATION OPERATOR T_{CKS}

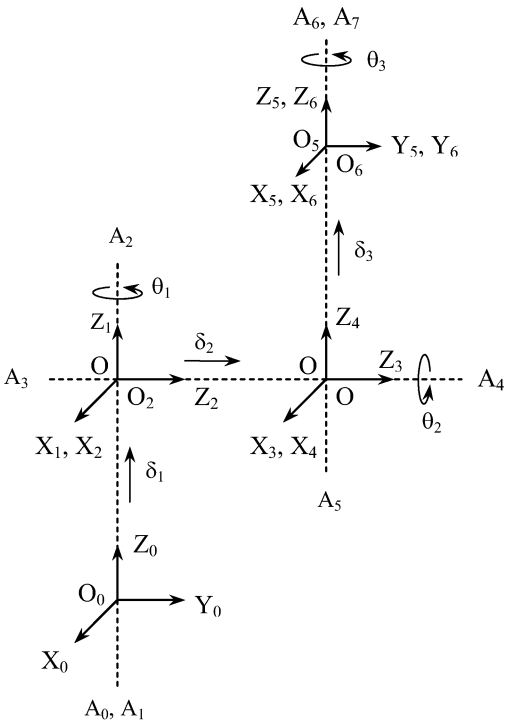


Figure 2: Maping of the Body-Attached Coordinate Frame

For determining of the transfer homogeneous operator T_{CKS} we will use the DH approach aplicated to the structure from Figure 1. In Figure 2 we presente the maping of the body-attached coordinate frame over the rigid mechanical links by first steps of the DH algorithm. The next steps establish the sets of the DH parameters for each transformation between two adjacent frames and these sets are presented in Table 1. The joint coordinate vector has the following structure:

$$\mathbf{q} = [\delta_1, \theta_1, \delta_2, \theta_2, \delta_3, \theta_3]^T \qquad (1)$$

Table 1: DH Parameters

	a_i	α_i	d_i	θ_i
T_1^0	0	0	δ_1	0
T_2^1	0	$-\pi/2$	0	θ_1
T_3^2	0	0	δ_2	0
T_4^3	0	$\pi/2$	0	θ_2
T_5^4	0	0	δ_3	0
T_6^5	0	0	0	θ_3

We obtain the general transfer operator $\mathbf{T_{CKS}}$ by product between the six intermediate homogeneous transformation operators obtained from Table 1 and by DH operator (Groover et al. 1986), (Klafter 1989):

$$\mathbf{T_{CKS}} = \mathbf{T}_6^0 = \prod_{i=1}^{6} \mathbf{T}_i^{i-1} = \mathbf{T}_1^0(\delta_1)\mathbf{T}_2^1(\theta_1)\cdot$$

$$\mathbf{T}_3^2(\delta_2)\,\mathbf{T}_4^3(\theta_2)\,\mathbf{T}_5^4(\delta_3)\,\mathbf{T}_6^5(\theta_3) =$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \delta_1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} C_1 & 0 & -S_1 & 0 \\ S_1 & 0 & C_1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \delta_2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} C_2 & 0 & S_2 & 0 \\ S_2 & 0 & -C_2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \delta_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} C_3 & -S_3 & 0 & 0 \\ S_3 & C_3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} =$$

$$= \begin{bmatrix} C_1C_2C_3 - S_1S_3 & -C_1C_2S_3 - S_1C_3 & C_1S_2 & -\delta_2S_1 + \delta_3C_1S_2 \\ S_1C_2C_3 + C_1S_3 & -S_1C_2S_3 + C_1C_3 & S_1S_2 & \delta_2C_1 + \delta_3S_1S_2 \\ -S_2C_3 & S_2S_3 & C_2 & \delta_1 + \delta_3C_2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (2)$$

ALGORITHM FOR KINEMATICAL MODEL

The next algorithm is established as follows:

- The kinematical structure of the robot is determined.
- A suitable representation with symbols of revolute/prismatic joints is made (kinematical schema).
- On this structure, a minimum number of CKSs is identified in order to cover it.
- For each identified CKS, a set of geometrical and movement parameters is appropriately established to the associated area from the kinematical schema.

- For each CKS, a homogeneous transformation operator (1) is calculated with the set of the parameters established to the previous step: $\mathbf{T^1_{CKS}}$, $\mathbf{T^2_{CKS}}$ …
- If multiple operators result, we make their product to obtain the general transformation operator T.

The homogeneous transformation operator, which makes the transformation between the coordinates of a point **P**, related to the mobile coordinate frame $\mathbf{R_M}$ and the point **P** coordinates related to the reference frame $\mathbf{R_F}$ considered fix, is presented bellow:

$$\mathbf{T} = \mathbf{T^1_{CKS}} \cdot \mathbf{T^2_{CKS}} \cdot \dots \quad (3)$$

Remark: Using this procedure, a great part of tedious and time-consumer operations of matrix multiplication (present in classical method) is eliminated.

APPLICATIONS

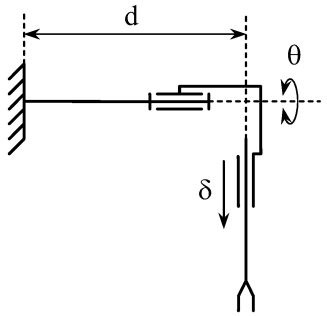


Figure 3: Robotic Structure for the Application 1

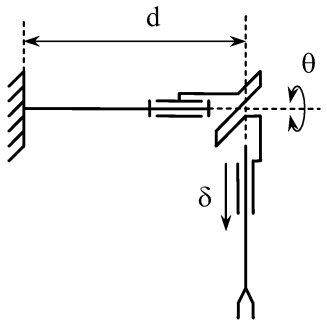


Figure 4: Modified Robotic Structure

Application 1

We want to obtain the kinematical model of the robotic structure from Figure 3. This structure is redrawing in figure 4 for an easier comparing with CKS. In Figure 5 we represented the mapping of the necessary frames. After comparing of the structure from Figure 4 with a CKS and determining of the geometrical and moving parameters from Figure 5, we obtain (4):

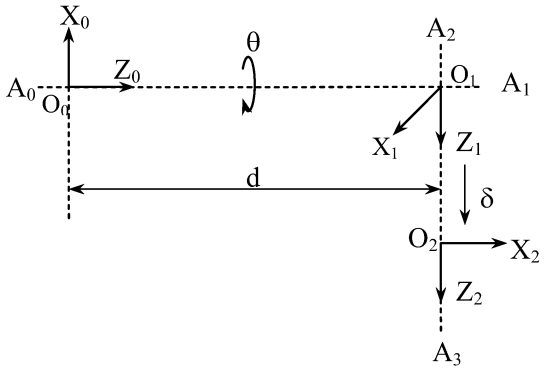


Figure 5: Mapping of the Coordinate Frame

$$\begin{array}{lll} \delta_1 \leftarrow d & \delta_2 \leftarrow 0 & \delta_3 \leftarrow \delta \\ \theta_1 \leftarrow \theta & \theta_2 \leftarrow -\pi/2 & \theta_3 \leftarrow 0 \end{array} \quad (4)$$

Those parameters represent entries for the operator $\mathbf{T}_{CKS}$ from (2). The expression of $\mathbf{T}_{CKS}$ becomes (5):

$$\mathbf{T}_{CKS} = \begin{bmatrix} 0 & -S\theta & -C\theta & -\delta C\theta \\ 0 & C\theta & -S\theta & -\delta S\theta \\ 1 & 0 & 0 & d \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (5)$$

Note: By classical methods we obtain the same operator if we have the same position and orientation for the frames $\mathbf{R_0}$ (fix) and $\mathbf{R_2}$ (mobile).

Application 2

In this second case we want to obtain the kinematical model of the robotic structure from Figure 6. This structure is a more complex structure and we need 3 CKSs for cover it because there are 3 joints which have the parallel axis.
The first CKS covers the first two joints, the second CKS covers the third joint and the third CKS covers the last two joint of the robot. In this case we need to atach 2 intermediary frames: first (**R'**) on link index 2 and the second (**R''**) on link index 3.
For the first substructure (Figure 7) we identify the parameters (6) of the CKS_1 and the operator $\mathbf{T^1}_{CKS}$ (7) results applying (2), for the second structure (Figure 8) we identify the parameters (8) of the CKS_2 and the operator $\mathbf{T^2}_{CKS}$ (9) results applying (2) and for the last structure (Figure 9) we identify the parameters (10) of the CKS_3 and the operator $\mathbf{T^3}_{CKS}$ (11) results applying (2).

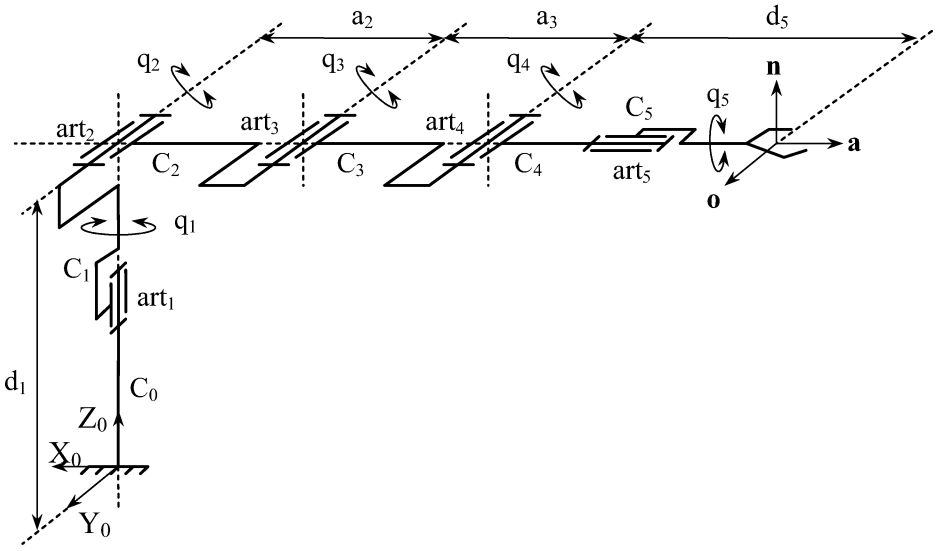


Figure 6: Robotic Structure for the Application 2

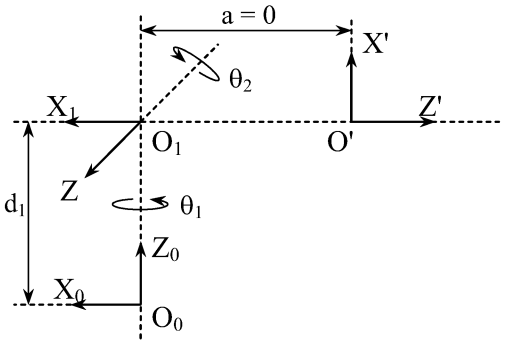


Figure 7: First Robot Substructure

$$\delta_1 \leftarrow d_1 \quad \delta_2 \leftarrow 0 \quad \delta_3 \leftarrow 0$$
$$\theta_1 \leftarrow \theta_1 \quad \theta_2 \leftarrow -\pi/2 + \theta_2 \quad \theta_3 \leftarrow 0 \qquad (6)$$

$$\mathbf{T}^1_{CKS} = \begin{bmatrix} C_1S_2 & -S_1 & -C_1C_2 & 0 \\ S_1S_2 & C_1 & -S_1C_2 & 0 \\ C_2 & 0 & S_2 & d_1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (7)$$

$$\delta_1 \leftarrow a_2, \quad \delta_2 \leftarrow 0 \quad \delta_3 \leftarrow 0$$
$$\theta_1 \leftarrow 0 \quad \theta_2 \leftarrow \theta_3 \quad \theta_3 \leftarrow 0 \qquad (8)$$

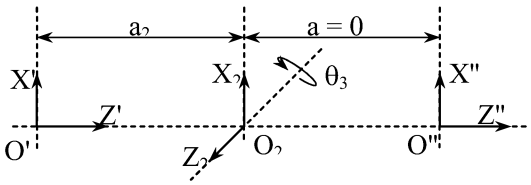


Figure 8: Second Robot Substructure

$$\mathbf{T}^2_{CKS} = \begin{bmatrix} C_3 & 0 & S_3 & 0 \\ 0 & 1 & 0 & 0 \\ -S_3 & 0 & C_3 & a_2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (9)$$

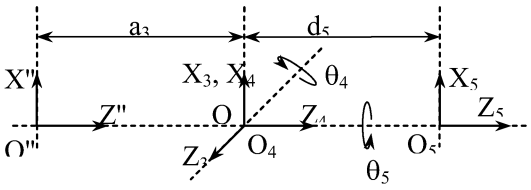


Figure 9: Third Robot Substructure

$$\delta_1 \leftarrow a_3 \quad \delta_2 \leftarrow 0 \quad \delta_3 \leftarrow d_5$$
$$\theta_1 \leftarrow 0 \quad \theta_2 \leftarrow \theta_4 \quad \theta_3 \leftarrow \theta_5 \qquad (10)$$

$$\mathbf{T}^3_{CKS} = \begin{bmatrix} C_4C_5 & -C_4S_5 & S_4 & d_5S_4 \\ S_5 & C_5 & 0 & 0 \\ -S_4C_5 & S_4S_5 & C_4 & a_3 + d_5C_4 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (11)$$

The operator between the end-effector coordinate frame and the reference coordinate frame is:

$$\mathbf{T}(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) =$$

$$\mathbf{T}^1_{CKS}(\theta_1, \theta_2) \cdot \mathbf{T}^2_{CKS}(\theta_3) \cdot \mathbf{T}^3_{CKS}(\theta_4, \theta_5) =$$

$$\mathbf{T}_{ARM}(\theta_1, \theta_2, \theta_3) \cdot \mathbf{T}_{TERM}(\theta_4, \theta_5) =$$

$$= \begin{bmatrix} C_1S_{234}C_5 - S_1S_5 & -C_1S_{234}S_5 - S_1C_5 & -C_1C_{234} & -C_1(a_2C_2 + a_3C_{23} + d_5C_{234}) \\ S_1S_{234}C_5 + C_1S_5 & -S_1S_{234}S_5 + C_1C_5 & -S_1C_{234} & -S_1(a_2C_2 + a_3C_{23} + d_5C_{234}) \\ C_{234}C_5 & -C_{234}S_5 & S_{234} & d_1 + a_2S_2 + a_3S_{23} + d_5S_{234} \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (12)$$

Notations: $S_2 = \sin\theta_2$; $C_2 = \cos\theta_2$; $S_{23} = \sin(\theta_2 + \theta_3)$; $C_{23} = \cos(\theta_2 + \theta_3)$; $S_{234} = \sin(\theta_2 + \theta_3 + \theta_4)$; $C_{234} = \cos(\theta_2 + \theta_3 + \theta_4)$

CONCLUSIONS

The direct kinematics problem is reduced to find a transformation matrix that relates the body-attached coordinate frame to the reference coordinate frame using vector and matrix algebra. By Denavit-Hartemberg method a homogeneous transformation matrix results for every couple of two adjacent rigid mechanical links and all those matrices have to be multiplied.

This paper presents a faster method for robot modelling by direct kinematics using „Complete Kinematical Structure" which is a stucture with 3 revolute joints and 3 prismatic joints which can assure a complete positioning and orientation in space for a object attached to its terminal. Its homogeneous transformation operator between fix frame and mobile frame is determined and used in an algorithm for robot modelling. Present method is based on DH approach but decreases drastically the number of the matrices and simplifies a lot of the calculations.

REFERENCES

Asada, H. and J.-J.E. Slotine. 1986. *Robot Analysis and Control*, John Wiley & Sons, Inc., New York.

Bishop, R.H. (Editor-in-Chief). 2002. *The Mechatronics Handbook*. CRC Press, New York.

Coiffet, P. 1983. *Modelling and Control*. Rob. Tech., Hermes Publishing.

Craig, J.J. 1989. *Introduction to Robotics. Mechanics and Control*, Addison-Wesley Publishing Company, New York.

Denavit, J. and R.S. Hartemberg. 1964. *Kinematic Synthesis of Linkages*. McGraw-Hill, New York.

Groover, M., et al. 1986. *Industrial Robotics. Technology, Programming and Applications*. McGraw-Hill, Inc., New York.

Klafter, R.D. 1989. *Robotic engineering: An Integrated Approach*. Prentice-Hall, New Jersey, USA.

Renaud, M. 1980. *Contribution a la modelisation et a la commande dynamique des robots manipulateurs*. These de Docteur d' Etat, Univ. Paul Sabatier, Toulouse, France.

Schilling, J.R. 1990. *Fundamentals of Robotics, Analysis and Control*. Prentice Hall, New York, USA.

Stoian, V. 1994. "Extension of the DH method for robotic structure kinematical model". *Proceedings of the SINTES 7 Symposium*, Craiova, Romania, may 20-21, vol. 1, pp. 252-260 (in Romanian).

AUTHOR BIOGRAPHY

Prof. **VIOREL STOIAN** was born in Segarcea, Romania, and went to the University of Craiova, Romania where he studied Automatic Control and obtained his degrees in 1983. After doing his military service, he worked for a couple of years for the Research Institute on Control of Automatic Equipments before moving 1990 to University of Craiova, Automatic Control Department. From 2004 he is working to Mechatronics and Automatic Control Department.

Simulation of mobile robot applications with *VirtualRobot*

Martin Mellado, Carlos Correcher*, Rosa Mª de Molina, José Camacho, Ginés Benet
Instituto de Automática e Informática Industrial (ai2)
Universidad Politécnica de Valencia (UPV)
Camino de Vera, s/n, 46022 – Valencia, Spain
Phone: +34-963879815 robotica@ai2.upv.es http://www.ai2.upv.es

* Currently working in *Institute for the Protection and Security of the Citizen* (IPSC), Joint Research Centre – Ispra.

KEYWORDS

Robotics, mobile robots, robot simulation

ABSTRACT

This paper introduces the basis of a graphical software application, designed for low cost hardware, as a flexible platform to work on mobile robots. *VirtualRobot* has proven to be a useful tool in mobile robot and sensor simulation. In this paper, some simulations of mobile robot applications will be shown with the help of *VirtualRobot Simulator*.

INTRODUCTION

In robotics, simulation goes further than a graphical software, becoming a proper engineering method. Nowadays, graphic simulators for robotic systems are indispensable in most of the robot design, learning and exploitation steps. Since this kind of applications were implanted during the eighties (Stauffer 1984; Howie 1984; Parker 1985), technological advances and improvements in computing engineering allow them to be applied on any field in robotics: industrial robotics (Freund et al. 1999), mobile robotics (Yang et al. 1997), sub-aquatic robotics (Chen et al. 1997) or aero-spatial robotics (Yoshida 1995). In addition, they have become significantly more powerful and flexible.

Some simulators go further than usual 3D graphical representation, applying virtual reality techniques (Matsumoto et al. 1999; Pérez and Schilling 2001). Sensor data inclusion allows more advanced simulators (Kimoto and Yuta 1995; Hale and Gosine 1999) to represent, in a dynamic way, robot environment as detected by sensors. With these methods, simulation software is not only dedicated to simulate robot behaviour, but also is useful for design, analysis and validation of techniques as collision detection, motion planning, sensor modelling, evaluation of control architectures and so on.

In most of the cases, they are very complex simulation systems (usually with expensive hardware requirements) designed to study and analyse a punctual characteristic within a specific branch of robotics. Therefore, there is still an open subject for future simulation software with general features and open software architecture, less specific and more flexible, to be adapted for most topics in robotics education, research and industrial implantation.

In this paper, the basis of a graphical software application as platform to work on robotics is explained, concentrating the details for mobile robotics and mainly its use for the simulation of different applications in research projects. On the next section, *VirtualRobot* use for mobile robot modelling is introduced. The following sections present practical simulation cases from different application examples carried out in the context of research projects.

MOBILE ROBOT MODELLING WITH *VirtualRobot*

VirtualRobot, a freeware software suite[1] created originally for remote articulated-robot monitoring, programming and simulation of the specific robot control system GENERIS (Generalised Software Control System for Industrial Robots) (Ruiz 1999) has become also a useful general tool in mobile robot simulation, including sensor simulation. *VirtualRobot* main features and architecture can be found in (Mellado et al. 2003).

In the *VirtualRobot* framework, a craft is some kind of vehicle designed for navigation in water (rarely in other liquids), in the air or through space, which can move freely in its surrounding constrained to its kinematics and dynamics characteristics (Mellado 2003). Note that fixed articulated robots and walking robots are not included in the concept of craft.

When using *VRS*, the simulator software of *VirtualRobot*, crafts are modelled at two levels: a low level, to model the craft shape and kinematics, and a high level, to emulate the craft behaviour. In order to shape a craft, a basic geometric modeller is used for each of the robot parts. To model craft kinematics, one in a set of predefined Dynamic Link Libraries (dlls) or a user defined dll is used.

The emulation of the craft behaviour is carried out by means of an external user application, which uses an external access library called *VRCraft*, an extension of the *VRS External Access Library VReal*.

[1] *VirtualRobot* can be downloaded in http://www.isa.upv.es/~vrs

Geometric Model

A craft is made up of two main parts: the craft base and the craft drives. The craft base is the main body of the robot. It has to be graphically modelled as an object in an ASCII file using the *VirtualRobot* modeller or imported from other file formats. The model must contain all the fixed parts of the robot.

The craft drives are the devices which produce the robot motion. Each of them has to be modelled as an object in an ASCII file, too. There are three types of drive components: ahead-drive, steer-drive and steer-ahead-drive. An ahead-drive produces movements in a linear direction. The ahead-drive which rotation produces craft motion is called active whereas the ahead-drive which rotation is produced by the movement of the craft is called passive. A steer-drive is a rotating device for the craft to describe turns. Also, applying different velocities to the ahead-drives it is possible to steer the craft. A steer-ahead-drive is an ahead-drive coupled to a steer-drive. Additionally, it's possible to define sensors and auxiliary coordinate systems into the final craft model.

Craft Kinematics Component

Craft Kinematics Component is used to solve the kinematics of the different robots in *VRS*. Thus, the movement of crafts in *VRS* is obtained by the kinematics module, which re-locates the craft every certain time interval, depending on the drives velocities. This time period is dynamically adjusted by *VRS* in order to achieve the maximum human real-time effect.

To emulate craft kinematics, in addition to the kinematics dll, it is necessary to create an ASCII file, which contains the craft definition data in a fixed format. This file must comprise the craft configuration (craft type, number of drives, number of sensors, number of frames, craft kinematics file and transformations), the craft base configuration (name, geometric file, transformations and global kinematics parameters), every craft drive configuration (name, drive type, geometric file, transformations and kinematics parameters), every craft sensor configuration (name, sensor type, transformations and craft sensor file) and every craft frame configuration (transformations).

VirtualRobot includes a Dynamic Link Library (dll) for four predefined types of *VirtualRobot* crafts: bicycle, tricycle, car and differential vehicle. Any user application can choose one of those libraries to model the robot kinematics. Anyway, the user can develop a new kinematics dll just following a fixed interface and using an empty template provided for this purpose.

The predefined kinematics interface must include an initialisation function (*ckInitKinematics*) and an update function (*ckKinematics*). *VRS* will call the corresponding initialisation function once, when a craft is loaded. The update function will be called to update the craft localization every time interval. The initialisation function parameters correspond to the parameters defined in the craft definition file (craft type, global kinematics parameters, numbers of drives and a matrix which contains each of the kinematics parameters of all the drives). The update function parameters are time interval (since the update function was called for the last time), current craft localization and current velocity and position of every drive. Note that the craft geometric model should be consistent with the craft kinematics model.

MONITORING AND SIMULATION OF A REAL MOBILE ROBOT

The robot YAIR is an autonomous robot prototype that is currently being developed as a part of a research project (Benet et al. 2002). The goal of this project is to define and implement a multi-layer architecture that allows the development of several autonomous mobile agent prototypes with sufficient intelligence, and to develop a prototype of this kind of robot. The prototype must be able to be remotely planned and react to unexpected events. This kind of system requires low and medium perceptual capabilities in order to extract, understand, and manipulate spatio-temporal information about the environment and for the cooperation of those capabilities with internal plans of actions. Therefore, the development of new techniques and methods for intelligent sensors, data fusion, local navigation, and real-time operating systems kernels for the construction of mobile robots is especially interesting.

In Figure 1, a photograph of the robot YAIR prototype is shown. Its periphery has an octagonal shape, having two active ahead-drives to produce the robot movements, and two passive free-steering wheels. The approximate size is 57 cm wide per 62 cm high.



Figure 1. YAIR prototype aspect

The robot has enough sensors to be used in partially structured or unknown environments, and can make indoor movements with two independent DC motors driving the robot. Three kind of distance sensors: ultrasonic, infrared, and bumper, give the robot the ability to cover a wide range of distances to detect obstacles, build maps (ultrasonic), avoid imminent collisions (infrared), make close approximations to objects (infrared and bumpers), or detect collisions.

For map-building and object location tasks, two sensor types are used: an ultrasonic rotary sensor placed on top, and a ring of 16 infrared sensors distributed in eight pairs around the octagonal perimeter of the robot. Each pair of sensors is centred at each edge of the octagon, with 15cm of separation between them. The maximum detection range is approximately one metre.

The distributed architecture of the robot has several modules that are interconnected using the CAN bus, a bus that has been chosen for its real time features. Each module performs its part in the distributed architecture. The actuator and the sensory modules are intelligent and perform basic algorithms of control in the case of actuators, or basic data processing algorithms in the case of sensors (Benet et al. 2005).

A *VRS* simulation application for YAIR robot has been very useful for system design and application developments even when the robot was not available. For example, as shown in figure 2, some planning modules for maze navigation according to vision control can be implemented in simulation before the full system is available in the real robot.

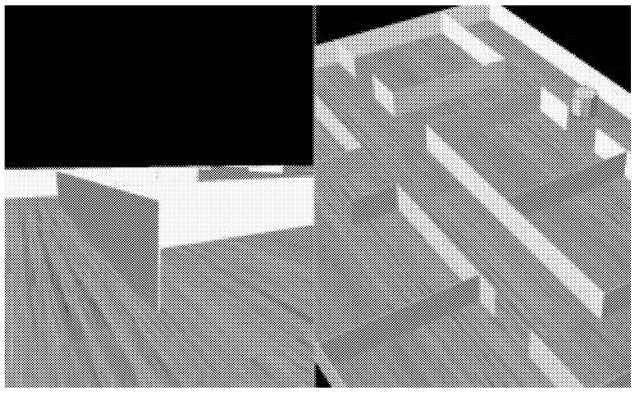


Figure 2. YAIR simulation in *VRS*

SIMULATION OF MULTI-ROBOT SYSTEMS

When simulating the behaviour of multiple robots, communication among them has to be achieved solely by means of their sensors and actuators. In *VRS*, the main thread acts as the real environment and each robot will communicate directly with it. This communication has to be constrained to simulate sensors in a proper manner.

VRS includes a library (*VRSens*) developed to simulate distance sensors. Anyway, the user is able to design his or her own functions to constrain the information that the robot takes from the environment, so that specific sensors can be simulated. In the application shown in the next section, communication sensors were developed Ad-hoc whereas obstacles detection sensors used this library.

There are two ways of programming multiple robot systems. On one hand, a program with the whole robot functionality and behaviour can de developed. *VRS* supports the communications establishment between its main thread and several applications at the same time. To investigate robots interactions is as simple as running the program several times. This is the best choice when simulating completely autonomous robots, which do not need any command from user.

On the other hand, when robots are not autonomous or they show different operation modes, a visual interface by which the user is able to control the parameters of the simulation has to be developed. If a program is thrown for any robot in the simulation and there is one different interface for each robot, the user will probably get confused. Robots can be represented by an independent thread of the user's application and thus only one interface is needed. In that case, the user is who has to avoid communication between threads. This was our choice to develop both application examples shown next.

ROBOTS FOLLOWING SIMULATION

As an example of multiple mobile robots simulation in VRS, we have designed a following task program. Robots following is a flocking task in which robots navigate arranged in a queue and tracking the first one movement. This task has many applications, including Automatic Highway Systems (Laski and Ravani 1993).

The main thread of the program throws as many robot threads as the user specifies in the user interface. Once the robots are properly arranged in a queue, the user can drive the first robot using the cursors in the interface.

Robots are driven by two controllers: the longitudinal and the lateral controller. The longitudinal controller is a Fuzzy PID (Naranjo et al. 2003). Its inputs are the reading of the obstacle detection sensors of the robots, which obtain the distance of each robot with the predecessor. The lateral controller simulates a sound based controller.

Results observed in the simulation in *VRS* by modelling one of the robots were the same as those obtained in *Simulink* (*Matlab*). Nevertheless, *VRS* also shows the interactions among multiple robots in the queue. This feature made possible for us to observe an error in the turn radius, as it can be seen in Figure 3.

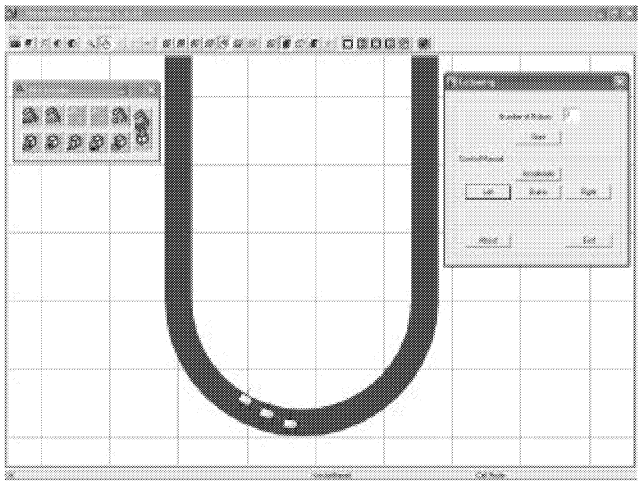


Figure 3: Following Task.

SIMULATION OF CRICKET-ROBOTS

In this example, *VRS* was used to test the design of a cooperative framework among mobile robots inspired in phonotaxis behaviours (tracking the source of a specific sound) observed in cricket mating. This framework was made up of the five basis cooperative behaviours proposed by (Mataric 1995). A pair of real robots were designed and developed.

The simulation showed all potential interactions while using any of the individual and cooperative behaviours implemented in the robots. An interface was developed to control the robots with the same possibilities as with the real robots (figure 4).

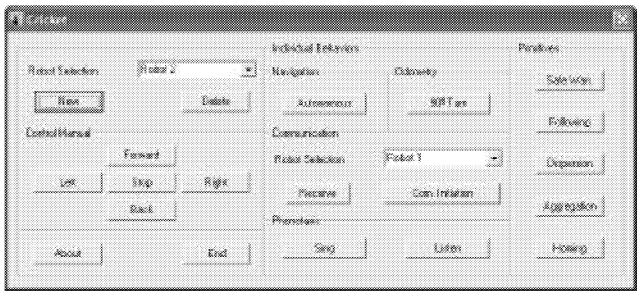


Figure 4: Cricket robots Interface.

In the experience in figure 5, a combination of aggregation and following tasks (group homing) was simulated. Three steps of the simulation can be seen: starting point, where all robots are far from each other, aggregation around top robot and queue formation.

Results in simulation were of special interest to identify problems in the upgrading of the real framework from 2 to n robots.

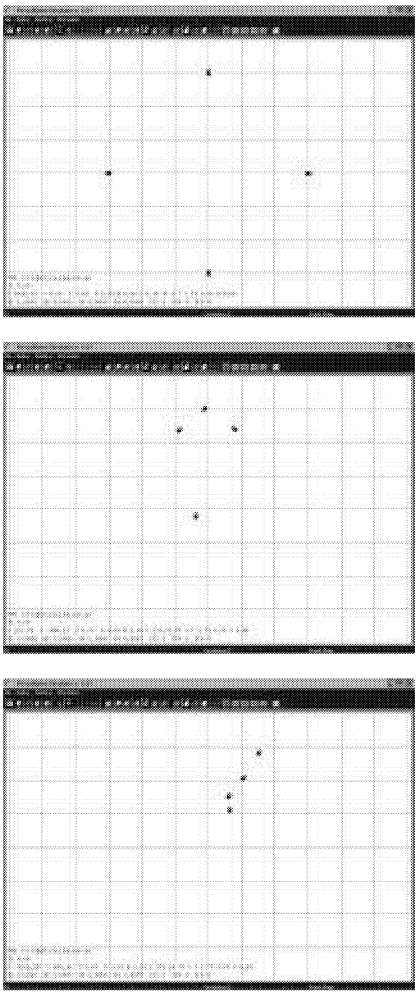


Figure 5: Group Homing Task. a) starting point, b) aggregation, c) following leader to home.

CONCLUSIONS

The possibilities of *VirtualRobot* software for mobile robot simulation have been presented in this paper through different applications. The main implementation considerations for mobile robot simulation have been exposed to show its advantages. Three application examples have been shown, one for monitoring a real mobile robot and two for multi-robot systems simulation.

ACKNOWLEDGEMENTS

This work has been partially funded by FEDER-CICYT projects with references DPI2002-04434-C04-03 and DPI2002-04434-C04-04 and by the FPU grants program, Secretaría de Estado de Educación y Universidades (Ministry of Education and Science).

REFERENCES

Benet, G.; Blanes, F.; Simo, J.E.; Pérez, P. 2002. "Using Infrared sensors for distance measurement in mobile robots" *Robotics and autonomous systems*, ISSN: 0921-8890, Elsevier Science, vol. 40/4; pp: 255-266

Benet, G.; Martínez, M.; Blanes, F.; Pérez, P.; Simo, J.E. 2005. "Differentiating Between Walls and Corners Using the Amplitude of Ultrasonic Echoes". *Robotics and autonomous systems*, ISSN: 0921-8890, Elsevier Science, vol. 50/1; pp: 13-25

Chen, X; Marco, D; Smith, S; An, E; Ganesan, K.; Healey, T. 1997. "6 DOF Nonlinear AUV Simulation Toolbox". *IEEE*, Oceans 97, Halifax.

Freund, E.; Rokossa, D.; Rossman, J. 1999. "Intuitive Off-line Programming of Industrial Robots Using VR-Techniques", *15th ISPE/IEE Int. Conf. on CAD/CAM, Robotics and Factories of the Future*, Águas de Lindóia, SP, Brazil.

Hale R.D.; Rokonuzzaman M.; Gosine R.G. 1999. "Control of Mobile Robots in Unstructured Environments using Discrete Event Modeling". *The Int. Society for Optical Engineering*.

Howie, P. 1984. "Graphic Simulation for Off-Line Robot Programming". *Robotics Today*. February. ISSN: 0193-6913.

Kimoto, K.; Yuta S.I. 1995. "Autonomous Mobile Robot Simulator: a Programming Tool for Sensor-based Behavior". *Autonomous Robots*. IS: 0929-5593.

Laski, T. A. and Ravani, B. 1993. "A review of Research Related to Automated Highway Systems (AHS)". *Research Report UCD-ARR-93-10-25-01,* Advanced Highway Maintenance and Construction Technology Research Center (AHMCT), Dept. of Mecanical and Aeronautical Engineering, University of California, Davis (Oct).

Mataric, M.J. 1995. "Designing and Understanding Adaptive Group Behavior". *Adaptive Behavior*, vol. 4, Issue 1 (Dec), pp. 51-80.

Matsumoto, Y.; Miyazaki, T.; Inaba, M.; Inoue, H. 1999. "View Simulation System: A Mobile Robot Simulator using VR Technology". *IEEE Int. Conf. on Intelligent Robots and Systems*. Piscataway, NJ, USA.

Mellado, M. 2003. "Simulación en robótica mediante *VirtualRobot*". *Editorial Universidad Politécnica de Valencia.* Ref: 2003.443.

Mellado, M.; Correcher, C. ; Catret, J.V.; Puig, D. 2003. "*VirtualRobot:* An open general-purpose simulation tool for robotics". *European Simulation and Modelling Conference, ESM'2003*, Naples, Italy, October 2003

Naranjo, J.E.; González, C.; Reviejo, J.; García, R.; and de Pedro T. 2003. "Adaptive Fuzzy Control for Inter-Vehicle Gap Keeping". *IEEE Transactions on Intelligent Transportation Systems*, vol. 4, Issue 3 (Sep), pp. 132-142.

Parker, J. 1985. "Simulating a Robot Arm using Graphics and Animation". *AI, Graphics and Simulation* Published by Soc for Computer Simulation, La Jolla, CA, USA.

Pérez, M. and Schilling, K. 2001."Virtual Reality for Tele-Education Experiments with Remote Mobile Robot Hardware". *Workshop on Internet Based Control Education*.

Ruiz, E. 1999. "GENERIS: The EC- JRC Generalised Software Control System for Industrial Robots". *Int. Journal of Industrial Robot*, Vol. 26, No. 1.

Stauffer, R.1984. "Robot System Simulation". *Robotics Today*. June. ISSN: 0193-6913

Yoshida K. 1995. "Experimental Study on the Dynamics and Control of a Space Robot with Experimental Free-Floating Robot Satellite (EFFORTS) Simulators". *Advanced Robotics*. ISSN: 0169-1864.

Yang, L.; Yang, X.; He, K.; Guo, M.; Zhang, B. 1997. "Research on Mobile Robot Simulation & Visualization". *IEEE Int. Conf. on Systems, Man and Cybernetics*. Piscataway, NJ, USA.

PATHFINDING WITH KOHONEN-MAPS

Matthias Reuter
Department Modelling & Simulation
CUTEC Institut GmbH
Clausthal-Zellerfeld, Germany
reuter@in.tu-clausthal.de

Steffen Harneit
Department Modelling & Simulation
CUTEC Institut GmbH
Clausthal-Zellerfeld, Germany
steffen.harneit@cutec.de

KEYWORDS: Robots, Pathfinder, Kohonen-Maps

ABSTRACT –Subject of our work to be reported here was to develop a compact algorithm with which a robot is able to move from a starting position to a target position without colliding with obstacles in a priori unknown area. To do so, we use a modification of Vleugel's kind of Kohonen-Map to cover the explored area with a growing grid of neurons, which results in a mathematical graph G=(V,E). The set of used neurons was subdivided in safe and unsafe neurons. While safe neurons represent parts of the area not covered with obstacles, unsafe neurons represent parts with obstacles, respectively. Furthermore we used a modification of Dijkstra's algorithm to determine a shortest path from the actual position of the robot.

INTRODUCTION

One of the most important problems building autonomous robot-systems is to solve the so called "motion planning problem" which can be formulated as follows: If a robot and an environment with a set of (immovable) obstacles are given, find a path from a starting position of the robot to a defined target location such that the robot is travelling along an optimal path without colliding with any obstacle. Many approaches to this problem have been proposed [3,4,5], but most of them require a complete knowledge about the environment (position and shape of any obstacle). Since these prior information is often not available in practical applications, we propose an approach, where only locally acquired sensorial information is used to construct a desired path in an environment which is explored step by step. The only type of sensing we use is the determination of the distances to the next obstacles in the robots viewing cone. In our simulations we assumed furthermore that the angle of this viewing cone has 180 degrees and the maximum viewing distance is freely configurable. For constructing the path we use a simplified kind of Vleugels coloured Kohonen network [4], which will lead to a graph on which a shortest path to the next temporary desired target position can be determined by the application of a modified version of Dijkstra's algorithm. Hereby we use the constraint that the robot only is able to move forward and additionally is able to rotate 180 degrees without moving.

PRODUCING THE GRAPH G

For the internal representation of the environment (the neural mental representation of the environment) we subdivide the "real world" in a number of square fields. A good starting point to do so is to allocate memory for

$$\frac{2\left|p_{target,x}-p_{start,x}\right|}{b}\times\frac{2\left|p_{target,y}-p_{start,y}\right|}{b} \qquad (1.1)$$

squares, whereby b denotes the width of one single square. Furthermore it should hold

$$|p_{target,x}-p_{start,x}|>q\ ,\ |p_{target,y}-p_{start,y}|>q \qquad (1.2)$$

whereby q denotes a minimum size of the area to be explored.

Hereby the best value for b has to be determined experimentally, because on the one hand a too large value of b often will show the effect that the algorithm cannot find a way through small gateways, while on the other hand a too small value of b leads to an unacceptable amount of required memory. Our experiments showed that good results can by obtained with $b=r_{robot}$, whereby r_{robot} announces the radius of the outer bounding sphere of the used robot.

If $S_{x,y}\in S$ is the square in column x and row y of the created grid, a function t assigns to every square one of the three properties "unknown", "free" or "occupied":

$$t:S\rightarrow\{\text{unknown, free, occupied}\} \qquad (1.3)$$

By this definition a so far unknown square has not yet been explored and may contain an obstacle or may not, a free square has already been explored and is known to contain no obstacle while an occupied square at least partly is covered with an obstacle.

Initially only the starting point $p_{start}\in R^2$ of the robot and the desired target position $p_{target}\in R^2$ are known. Since the robot does not know its environment, to every square is assigned the property "unknown". Initially no neurons exist, so it holds: $N(0)=N_{safe}(0)=N_{unsafe}(0)=0$. The grid of neurons that will be created next is used to generate a graph G=(V,E), on which shortest paths of the robot can be calculated. For this purpose we use a modification of Dijkstra's algorithm. The generation of the grid is done

dynamically and depends on the newly explored area the robot has scanned. The set $N(t)$ of neurons $n_i(t)$, $i=1,2,...,|N(t)|$, $t \in N_0$, at an actual iteration step t, is separated into two disjoint sets $N_{safe}(t)$ and $N_{unsafe}(t)$. In addition to every neuron the property either "frontier" or "no frontier" is attached. This frontier flag will be used later on to determine the next target position the robot has to move to, as described further below.

Now to each neuron n_i a two-dimensional vector $w_i \in R^2$ (movement in the third dimension will be straight-forward) is assigned - in the context of artificial neural nets called weight vector. Because every weight vector describes a point in the input space (the real environment), we call them *position vectors*.

After the robot has sent out a sensor scan in the environment, which results either in the distances of the next obstacles within the viewing cone of the sensor or in the maximum viewing distance, the grid of neurons, inspired by Vleugels's model of the coloured Kohonen map, will be placed over the (newly) explored squares. In the following adaption phase some as occupied known squares (resp. their coordinates) will be presented to the net and a winner neuron $n_{bmu} \in N_{safe}$ (bmu, best matching unit) will be determined. This winner is that safe neuron, whose Euclidian distance to the input vector is the smallest of all the distances between the input vector z and the position vectors of safe neurons, for which holds

$$\|w_{bmu} - z\| \leq \|w_i - z\| \forall i : n_i \in N_{safe} \qquad (1.4)$$

The winner neuron and his nearest neighbours now will be pushed away from the input vector. In this way the safe neurons are moved away from the obstacles and the shortest path from the actual position of the robot to the next desired position can be determined.

The Pathfinder Algorithm

At the start of every iteration the robot makes a sensor scan as described above. As a result the distances to the next obstacles within its viewing radius can be calculated. Every new explored square (any square previously marked with the property "unknown") that is not covered by any obstacle will be marked as "free". Next in any new free square a new free neuron is placed. Squares that are recognized to be at least covered with an obstacle are marked as occupied and a new unsafe neuron is placed in whereby its position vector is directly at the point of impact of the sensor ray and the obstacle (see Figure 1). Now to every square at least one neuron is attached, either free or not free. A list of all neurons lying in any square will be updated to save time as well for determining the winner neuron as for determining its direct neighbours.

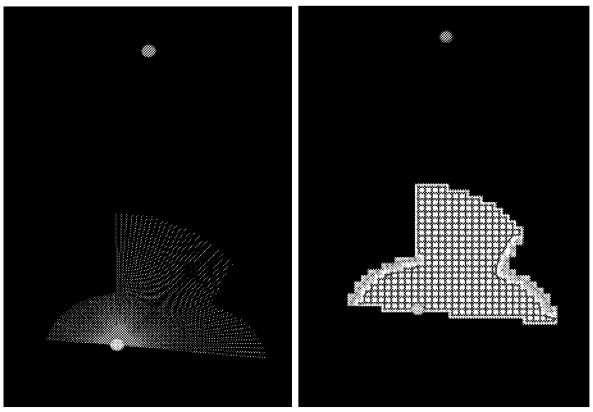


Fig. 1: Scanning and scaling a so far unknown area

Next the neighbourhood of the neurons will be updated, whereby we use a simple type of neighbourhood by attaching a maximum of four neighboured neurons to any neuron, as the upper, lower, left and right neighbour, whereby they always lie either in neighboured squares or the same square. In addition any neuron to which no four neighbours had been found is marked to be a frontier neuron. This means that this neuron is positioned at the frontier of a so far unknown area.

After that calculation step all explored squares are defined as occupied or free, they are at least covered with one neuron whereby the neighbourhood relationship defines a graph G. In our algorithm this graph G is only defined on the set of free neurons N_{free}. Next the following adaption phase pushes the free neurons away from the not free neurons. For considering the size of the robot, a corresponding parameter is included in the adaption formula.

At first a position vector z(t) will be chosen randomly from the set

$$z(t)=w_i(t) : i \in \{1,2,\ldots|N(t)|\} \text{ and } n_i(t) \in N_{unsafe}(t) \qquad (1.5)$$

Next the best matching unit amongst the free neurons will be determined. It is important to mention that by our procedure the input vector and hence the winner neuron belongs to the neurons of the actual observed square or one of its neighboured squares.

If $d_{bmu}:=\|z-w_{bmu}\|_2$ defines the Euclidian distance between the input vector and the position vector of the winner, the adaption of the winning neuron can now be defined as

$$w_{bmu}(t+1) = w_{bmu}(t) + \varepsilon \cdot \rho \cdot (w_{bmu}(t) - z) + \varepsilon \cdot \sigma \cdot \sum_{j \in J(bmu)} (w_j(t) - z) \qquad (1.6)$$

whereby J(bmu) represents the set of direct neighbours of the best matching unit, ε a constant learning factor,

$\sigma << \varepsilon$ a constant which defines the strength the winning unit will be pushed away from its direct neighbours. Furthermore ρ is connected to the influence of the radius of the robot as well as to the distance d_{bmu} of the best matching unit to the input vector. That's in that way, that neurons that lie close to an obstacle are pushed away farther than neurons that are already at sufficient range to obstacles. As our experiments showed it is advantageous to calculate ρ by

$$\rho := \frac{1}{100} + \frac{4}{1000}\left(\frac{5}{4} r_{robot} - d_{bmu}\right) \qquad (1.7)$$

Please note, that the weight adaption only takes place, if $w_{bmu}(t+1)$ does not lie in an occupied square. Otherwise no adaption will be applied. This guarantees that no free neuron is placed in a square that contains an obstacle.

In addition also the direct neighbours $j \in J(bmu)$ of the best matching unit are adapted, if they are marked as free neurons, by

$$w_j(t+1) = w_j(t) + \delta_1 \cdot \varepsilon \cdot \rho \cdot (w_j(t) - z) + \delta_2 \cdot \varepsilon \cdot \sigma \cdot \sum_{j \in J(j)} (w_j(t) - z) \qquad (1.8)$$

Here δ_1 and δ_2 correspond to constant factors for implementing a lower adaption rate of the neighboured neurons. It is important to mention, that by this procedure no free neuron will be pushed into an occupied square and no not free neurons will be adapted (moved), as they exactly describe points on the outer surface of obstacles.

The procedure described will be repeated, depending of the number of neurons, whereby we found that calculating four times the number of neurons is a good approximation for an optimal iteration.

Moving Over The Neural Grid

To lead a robot to a selected destination, at first the algorithm calculates a temporary point of interest which is placed between the robot and its destination. This point of interest is selected by a knowledge based algorithm, which accounts the map of the neural map calculated before. Next the shortest distance between the robot and the temporary destination is selected by calculating the minimum number of the neurons between these two positions. To optimize the pathway we integrated a cost function in the algorithm which benchmarks the path by a 1 if a free neuron is momentary reached on the path, which neighbours are also all free, a 6 if a neuron is reached, which neighbours are not all free and a 25 if a frontier neuron is reached. In that way we collect an optimal velocity, resp. stopover, from which the next scan of the environment shall be done.

If the cost function reaches a value larger than a given value the robot pauses anyway and scans once again.

When G=(V,E), $V=N_{free}(t)$ [V=vertexes, E=edges] is calculated, the next best target position has to be determined out of the set

$$p_{target}(t) \in \{w_i(t) \mid i=1,2,...,|N(t)|\} \text{ and } n_i \in V \qquad (1.9)$$

For that the shortest paths from the actual position of the robot $p_{robot}(t)$ - which is identical to one of the weight vectors of the free neurons (since the robot only travels "on the free neurons") - to all the other weight vectors of free neurons will be calculated by Dijkstra's algorithm [1,2]. Let $c_i(p_{robot}(t))$ denote the cost value to travel from the robot's position to the weight vector of the free neuron i. Amongst the free neurons satisfying the condition $c_i < \infty$ one is chosen to represent the next best target position $p_{target}(t+1)$ in the following manner: With D>0 an upper limit for the maximum distance temporarily allowed, $p_{target}(t+1) := w_i$ is that neuron with the shortest Euclidian distance of all free neurons to the global target position. That holds under the condition, that the distance is less or equal D. If the algorithm cannot find a neuron for which this conditions holds, that of the free neurons labeled as frontier neurons is chosen, of which Euclidian distance is the smallest one amongst all frontier neurons. If also now no such frontier neuron can be found, the whole environment had been explored. Because of the limitation to consider neurons under the constraint $||w_i - p_{target}|| \leq D$ only, there may exist a path to the target position within the already explored area. For that reason D has to be increased sometimes and a new effort for determining the best target position has to be started. If D exceeds a particular maximum value, the algorithm can be aborted with the result that no suitable path has been found.

If a new temporary target position has been determined successfully, the robot starts its movement on the shortest path to this position. The range the robot travels depends on the difficulty of the ground walked on. For that reason every step (a movement from one neuron to a neighboured one) costs a certain number of penalty points. After a certain number of penalty points are accumulated, the robot stops its movement and again scans its environment. While a walk onto a neuron, labeled as frontier neuron, costs the maximum number of penalty points because entering the border between the explored and the unexplored environment has to result in a new sensor scan, a walk onto a free neuron with only free neighbours costs only one penalty point. The third case is a walk on a free neuron with at least one not free neighbour. Here the robot has to move slower and scan this area frequently, because of as well improving the information about this difficult environment and decreasing the chance of colliding as of discovering previously unspotted gateways between different obstacles. The cost for this type of movement has to lie between one penalty point and the maximum number of penalty points.

Summarising, the algorithm for generating a path on which the robot can travel without colliding with any obstacle is as follows:

1. Let the robot examine his environment by sending out some rays, for example with an ultrasonic scan, within a predefined viewing cone, defined by the angle α and with a maximum viewing distance d_{max}. The result for any emitted ray i is the distance $d_{i,obstacle} \in [0,d_{max}]$ to the first obstacle that is hit by the ray.
2. Put in all the new discovered free squares a safe neuron and in all new discovered occupied squares an unsafe neuron.
3. Correct/Update the neighbourhood-connections between the neurons. Neurons, to which are assigned less than four direct neighbours, will be labelled as frontier neurons.
4. Learning phase: A number of input vectors (position vectors), belonging to occupied squares, will be presented and the existing free neurons will be adapted by a winner takes most-rule.
5. Apply the modified Dijkstra's algorithm to determine the shortest paths from $p_{robo}(t)$ to all free neurons. If no such path exists, terminate with the result, that there exists no way to the target position.
6. Determine the next desired target position $p_{target}(t)$
7. Move the robot in the direction of $p_{target}(t)$. The distance the robot travels depends on the cost of the movement.
8. If the new position of the robot is near enough the target's position, terminate with success. Else continue with step 1.

EXPERIMENTAL EXAMPLES

The next Figures give some examples how our algorithm works. In Figure 2 an example environment (room) is shown which has to be explored by a robot.

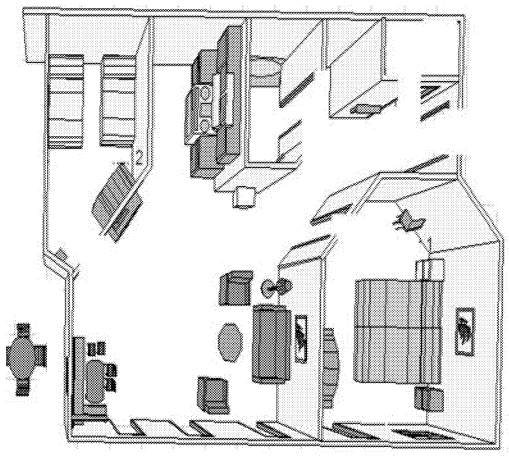


Fig. 2: Example environment to be scanned

The different states of the scanning process are shown in Figure 3a-3c. Please note that at the final stage of the scan of the environment the optimal path to the destination is shown only.

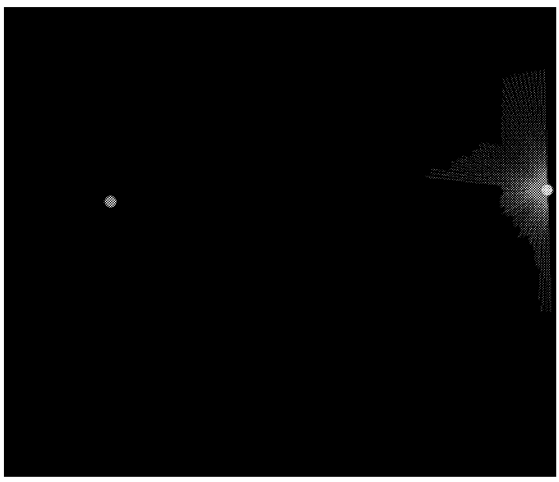


Fig. 3a: Example environment at first scan

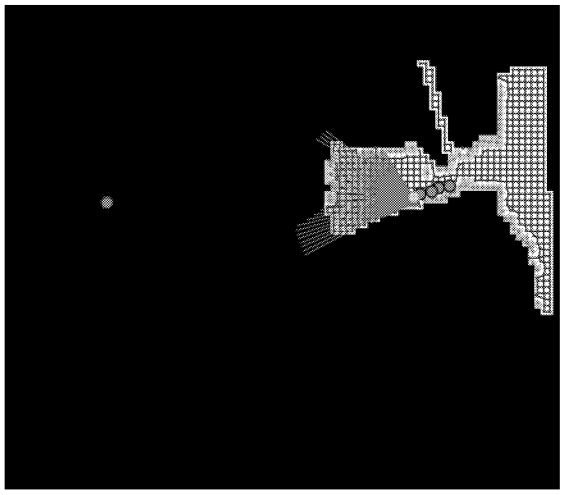


Fig. 3b: Example environment in scanning process

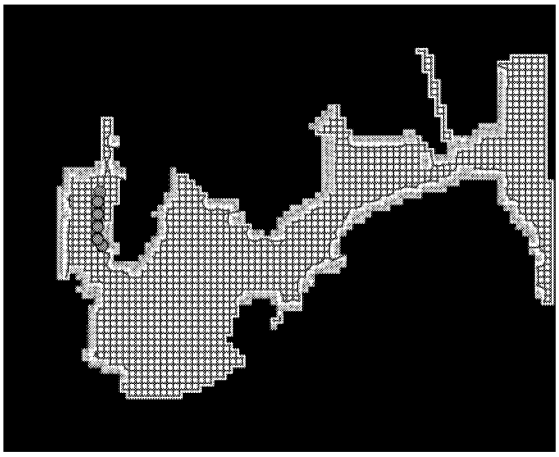


Fig. 3c: Example environment after the last scan, reaching the destination

CONCLUSIONS

We demonstrated a method with which a robot can find in a so far unknown environment an path to a defined destination by applying a neural map and the Dijkstra's algorithm. The neural map is mapping the environment in a kind of neural metal grid representation whereby this grid can be used to find an optimal path. Our kind of calculation allows it also pointing out the favourite points from which the robot has to scan the so far unknown environment.

REFERENCES

[1] T.H. Cormen, et. Al., “Introduction to algorithms”, MIT Press, Cambridge, Mass., 1980

[2] E. Dijkstra: “A note with two problems in connection with graphs”, Numerical Mathematic I, p. 265-271, 1959

[3] J. Heikkonen, et. Al: “From situations to actions: motion behaviour learning by self-organisation”, Proc. ICANN, pp. 262-267, 1993

[4] J. Vleugels, et. al.: “A self-organising neural network for robot motion planning”, proc. ICCAN, pp. 281-284, 1993

[5] J. Tani, et. al.: “learning goal-directed sensory-based navigation of a mobile robot”, Neural networks, Vol. 7, No. 3, pp. 553-563, 1994

MATTHIAS REUTER was born in Dresden, Germany, and went to the Clausthal University of Technology, where he studied physics and geophysics. He obtained his doctor in 1987 and his PD 2002. He worked as scientist at the University of the German Army in Hamburg, the University of Dortmund and the Clausthal University of Technology. He developed a couple of new neural nets structures and learning strategies. His major investigations areas are neuronal net dynamics and robotics.

STEFFEN HARNEIT was born in Lüneburg, Germany, and went to the Clausthal University of Technology, where he studied computer science. His favourite subjects of research are spatiotemporal dynamics of neural nets and multiagent systems.

Extraction of System Parameters and Their Importance for Controlling

Matthias Reuter
Department Modelling & Simulation
CUTEC Institut GmbH
Leibnizstr. 21+23
Clausthal-Zellerfeld, Germany
matthias.reuter@cutec.de

Sabine Bostelmann
Institute of Computer Science
Clausthal University of Technology
Julius-Albert-Str. 4
Clausthal-Zellerfeld, Germany
sabine.bostelmann@in.tu-clausthal.de

KEYWORDS
Controlling, System parameters, Quality management.

ABSTRACT

Even if the quality of given measured data of a system is high, often it is very hard to extract the relevant facts, trends and resulting managing guidelines because an adequate problem oriented representation is missing. Regarding this fact in this paper we discuss a new method on how to obtain a problem oriented representation for regression curves with involved time dependencies, by illustrating the influence of management decisions on the system behaviour. So we will show how management decisions can influence a system with totally different explanatory variables so that a good decision may lead to a steady improvement of the quality data while a bad decision might result in a heavy drop of the quality curve. But that exactly is the point that is really valuable for controlling because in case of bad decisions the controlling interval needs to be much smaller to enable fast intervening and thus to prevent a sizably loss of quality.

INTRODUCTION

Each company that uses quality management needs to solve the task of analysing the gathered data. First of all the data needs to be freed of outliers and system-immanent noise as well as of measurement errors. As described in detail in (Bostelmann and Reuter 2004) the right pre-processing can transform even the scattered data shown in Figure 1 into a curve that is much easier to analyse (see Figure 3), by eliminating uninteresting information and features of the data set. One of several pre-processing steps to get to such a smooth curve shown here is the double variable floating average method described in (Reuter 1993), briefly called DLS (see Figure 2). This procedure is based on the evaluation of mean values of the given data regarding time slots of different sizes. The three curves on the upper half of the figure represent the courses of the different mean values while the black filled curve at the bottom corresponds to the scattering of the data around the general data tendency.

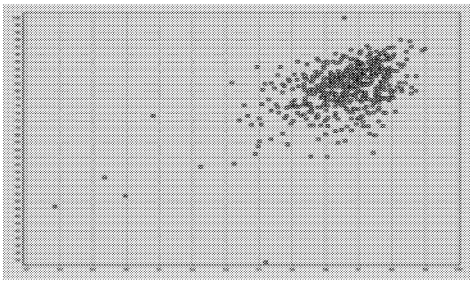


Figure 1: Originally Measured Data

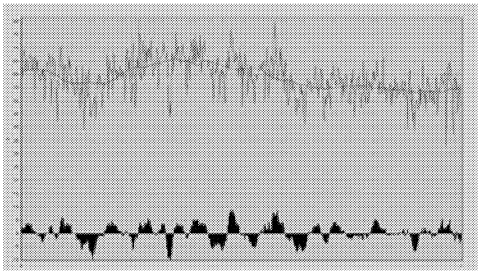


Figure 2: Double Variable Floating Average Method (DLS)

Applying the DLS on both axes the graph of Figure 3 will result from the given (pre-processed) data if the trend of the mean value is used only. Nevertheless the resulting curve still needs to be processed in order to obtain convincing system parameters for it seems that there is an overlapping of the curve, resp. of several independent trends the system follows.

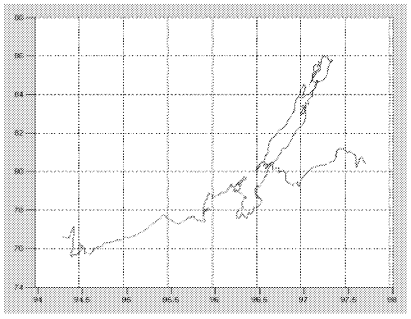


Figure 3: Resulting Pre-Processed Data

For a better understanding it is important to mention that the data shown in Figures 1 and 3 consists of two different system values: while value A, shown on the x-axis, represents the punctuality of the process to be supervised, value B, shown on the y-axis, represents the disposability of the system components which have to be used for the process. Furthermore it is necessary to verify that the eliminated "noise" actually does not contain any information. Therefore the class oriented distribution of the filtered noise is analysed. Only if it follows a Gaussian distribution it can be eliminated, otherwise it contains real information and must not be neglected. The task in this example was to analyse the interdependency of the two values and to extract the system parameters out of the given data. The processes that need to be done in order to obtain the desired parameters are explained in the following.

ANALYSIS OF SYSTEM PARAMETERS

As mentioned above, the pre-processed data still needs some processing before the final analysis. First the overlapping of the curve needs to be examined.

Decomposition of pre-processed data

Taking a closer look at the time course of the data - shown in Figure 3 - it becomes obvious that the imaginary overlapping of the curve is just a result of the two-dimensional plotting of the data. Displaying the data in a three-dimensional plot, adding a time-axis to the two-dimensional diagram, the intersections disappear (see Figure 4). The arrows in Figure 4 indicate the course of time while the ovals mark the turning points of the curve.

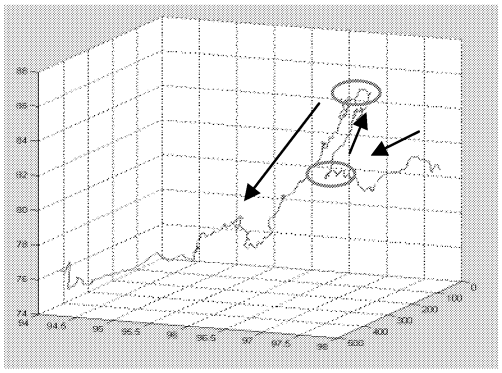


Figure 4: Pre-Processed Data Plotted with Time-Axis. Arrows Indicate Flow of Time, Ovals Mark Turning Points

Regarding Figure 4 it seems apparent that the inter-dependency of the system values develops quite smoothly – apart from the two turning points of the curve. These turning points indicate decisions made by the management whereupon the system leaves its current state and needs to orientate itself towards another steady state. Thus it is obvious to divide the pre-processed curve into three parts before beginning the actual analysis. The results of this separation are shown in Figures 5, 6 and 7. Please note that – regarding the time course of the data indicated by the arrows below the figures – only the curve in Figure 6 is rising.

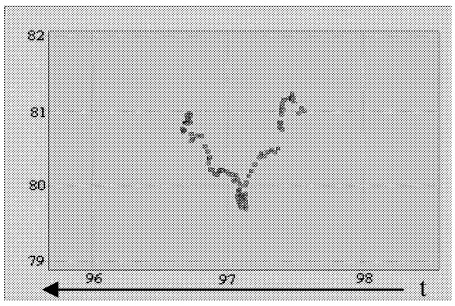


Figure 5: First Part of the Curve (No Obvious Trend)

Having separated these distinguishable system states belonging to different management trends, an independent analysis can be done for each single part of the curve. In the following these sections will be regarded one by one.

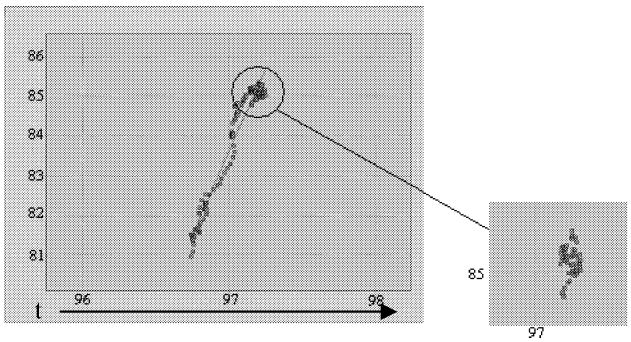


Figure 6: Second Part of the Curve (Rising with Time)

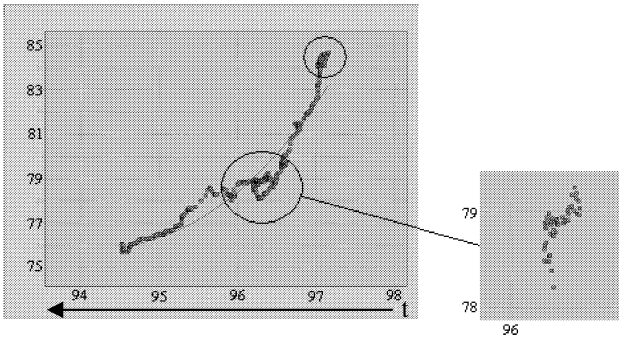


Figure 7: Third Part of the Curve (Dropping with Time)

Extraction of system parameters

Taking a closer look at Figure 5, it becomes apparent that the first part of the curve cannot be analysed as one entity because of the two different tendencies shown. It would be necessary to divide this part of the initial curve into another two sections but because of the small amount of affected data points a statistical interpretation of these two parts will not be meaningful. Therefore this first part of the data will not be discussed any further.

In the following the data of Figures 6 and 7 will be discussed. Because of the smooth course of the data a regression analysis can easily be done in both cases. There is a variety of possible regression functions, e.g. linear regression, exponential regression, power regression and logarithmic regression. The quality of the regressions was measured by a correlation coefficient. The direct usage of the correlation coefficient is only possible in the linear case because the correlation coefficient is a measurement of the linear interdependency of two values. For reasons of comparability in case of the other regression functions the axis were transformed so that the interdependency of the values was linear again according to the transformed axis. Thus the usage of the correlation coefficient is appropriate. Due to the reason that the system acts independent of the part of the imposed coordinate system in which its actual state is located, the regression is done in an own coordinate system for each part to be analysed.

According to this procedure, the data of Figure 6 follows a logarithmic regression:

$$y = 0.67 + 10.144 * \log x \qquad (1)$$

whereby the quality of the regression is very high, indicated by a correlation coefficient of 0.979. Likewise the power regression of the data of Figure 7 is determined as follows

$$y = (0.554 + 0.616\, x)^{2.6} \qquad (2)$$

also with a high resulting correlation coefficient of 0.968. These high correlations are even more astonishing since due to the system dynamics the regarded parts of the curve still show a little chaotic behaviour (circled in Figures 6 and 7). Tough even in these states the system is moving towards an attractor point it is not in equilibrium. After a short time the steady state is retrieved.

Having come so far the system parameters of the particular management phases can be directly read off formulas (1) and (2). Again, for interpretation it has to be taken into account that the second part of the initial curve is rising while the last part is dropping regarding the course of time.

Importance of system parameters

Evaluating the system parameters is a very important task of observing and controlling the system. Knowing the values it is possible to:

- detect inconsistencies of the system,
- verify management decisions,
- model the system.

In this specific example the evaluated parameters and interdependencies show that even if the curve rises steadily in the second part, the dropping after the management decision is even stronger. Thus the decision does not just result in a turning of the system, the deterioration is accelerated. This knowledge leads to smaller controlling intervals for the loss should be minimized. Even if it is not possible to extract the system parameters of a running system right away, the trend can be observed very fast so that a correction of the course of the process can be pushed. After a short time the process values can be determined and – if wanted – e.g. be used for modelling the system. If the controlling intervals cannot be read out off the system parameters, they can be detected by simulating the system and testing its behaviour in relation to disturbing influences and/or decisions.

CONCLUSIONS

This paper demonstrates a method with which the extraction of system parameters out of a running system becomes very easy. The knowledge of these parameters is a useful controlling method for it allows to define controlling intervals and to model the system. Simulating the system is yet another important tool to get intrinsic information of the system and to predict its behaviour in the future.

REFERENCES

Bostelmann, S. and M. Reuter. 2004. “Analysing trends of quality data by means of new pre-processing methods.” Proc. WAC 2004 (June), Seville, Spain, pp. 385-390.

Doty, L.A. 1995. “Statistical Process Control.” Industrial Press Inc., ISBN 0-8311-3069-5, New York.

Kazmierski, T.J. 1995. “Statistical Problem Solving in Quality Engineering.” Mcgraw-Hill, ISBN 007034048X.

Reuter, M. 1993. “Frequency Difference Spectra and Their Use as a New Pre-processing Step for Acoustic Classifiers/Identifiers.” EUFIT 93 Proceedings Vol.1, pp. 436-442.

Ryan, T.P. 2000. “Statistical Methods for Quality Improvement.” Whiley, ISBN 0-471-19775-0, New York.

Schönwiese, C.D. 1985. “Praktische Statistik für Meteorologen und Geowissenschaftler.” Gebrüder Bornträger.

MATTHIAS REUTER was born in Dresden, Germany and went to the Clausthal University of Technology, where he studied physics and geophysics. He obtained his doctor in 1987 and his PD 2002. He worked as scientist at the University of the German Army in Hamburg, the University of Dortmund and the Clausthal University of Technology. He developed a couple of new neural nets structures and learning strategies. His major investigations areas are neuronal net dynamics and robotics.

SABINE BOSTELMANN was born in Hannover, Germany and went to the University of Hildesheim and to the Clausthal University of Technology, where she studied business mathematics. She worked at the University of Dortmund and at the Clausthal University of Technology. Her favourite subject of research are neural dynamics and the evaluation of new learning strategies of neural nets.

ELECTRONICS PERFORMANCE ANALYSIS AND OPTIMIZATION

CLUSTERING PROCEDURE FOR PERFORMANCE ANALYSIS OF MOBILE COMMUNICATION NETWORKS

Dimitar Radev
University of Rousse
8 Studentska Street
7017 Rousse, Bulgaria
E-mail: dradev@abv.bg

Izabella Lokshina
SUNY Oneonta
224 Netzer Administration Bldg.
Oneonta, NY 13820, USA
E-mail: lokshiiv@oneonta.edu

Svetla Radeva
University of Architecture,
Civil Engineering and Geodesy
1 Hr. Smirnensky Blvd.
1046 Sofia, Bulgaria
E-mail: svetla_fce@abv.bg

KEYWORDS
Mobile communication networks, Markov reward models, neural networks.

ABSTRACT

Mobile communication networks with calls which belong to one of four service classes and arrive randomly, are considered in the paper. Arriving calls are granted service based on the definite service class, required maximum and minimum bandwidth, and available network resources. Clustering procedure for performance analysis of mobile communication networks and their measures for blocking probability and quality of service are introduced as Markov reward model enhanced with use of vector quantification and neural modeling. The optimal link occupancy probability distribution is determined using neural network which was trained on the base of Kohonen rules. Simulation and numerical results are shown.

INTRODUCTION

We can identify a feature of the mobile communication network that the access part is totally separated from the core network as one of the key architectural aspects of the third generation mobile communication networks. This separation becomes a reason for generating different services and developing different techniques to evaluate parameters of their performance. All multi-service networks, providing guaranteed blocking probability and guaranteed throughput level, as for example, core of third generation UMTS/IMT (short for universal mobile telecommunication system) networks, could be classified based on four service classes: conversational, streaming, interactive and background. The conversational class provides high quality access to a range of different services including high bit rate services. This class is suitable for the demanding user who wishes to receive bandwidth assurance similar to that of the constant bit rate class in an ATM. The streaming class is designed to carry high bandwidth with variable bit rate services, such as medium- or high- quality video or teleconferencing service. The interactive class supports less demanding services typically supported by today's best effort IP networks, including file transfer, web browsing, or telnet applications. The holding time of interactive class calls typically depends on the throughput. The background class is of the best effort type, meaning that background calls receive whatever bandwidth is "left over" by the calls of the higher priority service classes. Examples of this include e-mail and low quality file transfers. With respect to their holding time, this class is similar to the interactive class. The fact, that it is relatively simple to obtain the mean and variance of the distribution, but still difficult to obtain the distribution itself, becomes a reason for approximations applied to determine a discrete version of a normal density function for link occupancy distribution (Racz et al. 2001). Use of clustering and neural modeling of a single link transmission is considered in this paper.

BANDWIDTH SHARING MODEL WITH MARKOV CHAIN

For multi-service environment modeling, a continuous time Markov chain (CTMC) is used. Since the mobile core network exercises admission control on a link-by-link basis, a single transmission link is defined as Markov reward model. Based on description of the service classes it is clear that we need in modeling the service classes with only throughput guarantee (i.e., conversational, streaming and interactive), and in computing the bandwidth left over for the background class. Then we present the stochastic stationary process as a network of three queues in the state space. The input parameters of the system are the set of arrival rates (λ_1, λ_2, λ_3) and departure rates (μ_1, μ_2, μ_3), the bandwidths (b_1, b_2, b_3), and throughput constraints $(\widetilde{\theta}^{\min}, \hat{\theta}^{\min})$. The model state is uniquely characterized by the triple (n_1, n_2, n_3), where n_1 is the number of conversational flows, n_2 is the number of streaming flows, and n_3 is the number of interactive flows.
In order to obtain the performance measure, the CTMC's generator matrix Q and the bandwidth sharing policy (BSP)

are defined so that the link capacity C is divided into two parts. A common part and a part reserved only for the streaming and interactive flows are shown in Fig.1. Let assume that C - δ denotes the portion of the link capacity that is dedicated to the first three service classes, and δ denotes the portion of the link capacity which is dedicated to the background class. There are only two possibilities forδ. For δ = 0 we have Policy 1 which is a complete sharing. For $\delta > 0$ we have Policy 2 which is a complete portioning.

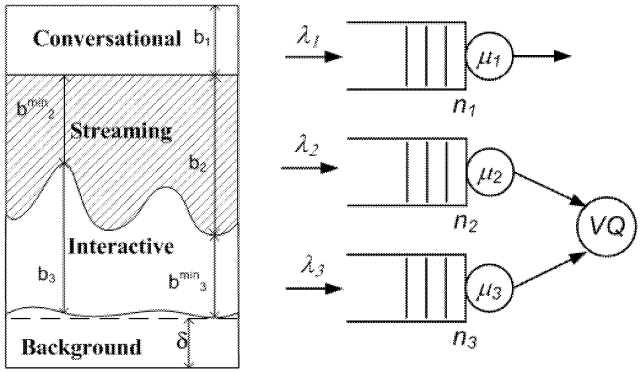


Figure 1 Model of a single transmission link

There are the following possibilities which could be true for both policies. If there is enough bandwidth for all flows to get their respective peak bandwidth demands, then class-2 and class-3 flows would occupy b_2 and b_3 bandwidth units respectively. If there is a need in bandwidth compression, i.e. $n_1 \cdot b_1 + n_2 \cdot b_2 + n_3 \cdot b_3 > C - \delta$, then the bandwidth compression of the flows would be divided in equal parts between both classes as long as minimum rate constraint is met for both classes. If there is a need for further bandwidth compression, but either one or both classes does not tolerate further bandwidth decrease at the time of a new flow arrival, then the service class which tolerates further compression would decrease the bandwidth occupied by its flows equally, as long as the minimum bandwidth constraint would be kept for this traffic class. The basic assumptions of the above rules demonstrate that both the streaming and interactive flows are greedy. It means that they always occupy the maximum possible bandwidth on the link, which is the smaller of their peak bandwidth requirements (b_2 and b_3, respectively), and the equal share of the bandwidth left for them by the conversational flows. All streaming and interactive in-progress flows share proportionally the available bandwidth among themselves, i.e. the newly arrived flow and the in-progress flows will be squeezed to the same compression values. After that, if a newly arriving flow decreases the flow bandwidth below minimal accepted value, and is not admitted to the system, then it is blocked and lost. By this reason, it is important to develop bandwidth sharing models for streaming and interactive flows. We propose the models, based on Markov chain reward model, in which only transitions between neighboring states are allowed. Non-zero transition rates for the feasible states are described according to (1), where the first three equations represent the state transitions due to all arrival, while the second three equations represent the transitions due to call departures.

$$
\begin{aligned}
&q(n_1,n_2,n_3 \to n_1+1,n_2,n_3) = \lambda_1 \\
&q(n_1,n_2,n_3 \to n_1,n_2+1,n_3) = \lambda_2 \\
&q(n_1,n_2,n_3 \to n_1,n_2,n_3+1) = \lambda_3 \\
&q(n_1,n_2,n_3 \to n_1-1,n_2,n_3) = n_1 \cdot \mu_1 \\
&q(n_1,n_2,n_3 \to n_1,n_2-1,n_3) = n_2 \cdot \mu_2 \\
&q(n_1,n_2,n_3 \to n_1,n_2,n_3-1) = n_3 \cdot \rho_3(n_1,n_2,n_3) \cdot \mu_1
\end{aligned}
\qquad (1)
$$

The $n_3 \cdot \rho_3(n_1, n_2, n_3) \cdot \mu_3$ quantity denotes the total bandwidth of the interactive flows when the system is in state (n_1, n_2, n_3), and compression of the interactive flow is denoted as ρ_3. The BSP policy is completely determined by specifying the following output parameters: the capacity of the common part, and the maximum number of streaming and interactive flows, N_2 and N_3. At this point, the set of feasible states of the CTMC model S is determined according to (2).

$$S = (N_2 + 1) \cdot (N_3 + 1) \qquad (2)$$

The stochastic system clustering presents a steady state analysis for the states j of state space S, $j \in S$; in every discrete time epoch. The states are arranged on a grid with as many dimensions as the number of queues. Each of the axes is representing the number of customers in one of the queues. Our goal is to determine the occupancy and arrival distributions, partial overlap, call blocking probabilities of feasible states of queuing system. In order to evaluate density function and performance of BSP allocation strategy, the maximal cuts and the partitions of the graph should be calculated applying Markov chain model (Kulkarni 1995). The clustering procedure is developed with use of vector quantification (VQ) and neural modeling.

CLUSTERING PROCEDURE WITH KOHONEN NEURAL NETWORKS

We can associate the clustering problem with a discrete distribution with independent margins, where the random values $X_1^{(2)},\ldots,X_k^{(2)}$ and $X_1^{(3)},\ldots,X_k^{(3)}$ are distributed into M classes $S_1,\ldots,S_M$. Then again, the number of streaming n_2 and interactive n_3 flows can be presented as a couple of time series $\{x^{(2)}\}$ and $\{x^{(3)}\}$, for which we can determine the probability function with use of VQ and Kohonen learning rules. Kohonen network algorithm (Radev et al. 2004) provides a transformation of the input space into patches with corresponding code vectors. It has an additional feature that the centers are arranged in a low dimensional structure (rectangular grid). We introduce vector quantification to the learning neural model in order to determine the probability density for each feasible steady state of Markov chain (Hofmann & Buhmann 1998). Equal width of class boundaries and equal number of target values in each class can be used as criteria for clustering (Radev & Radeva 2003). Rectangular boundaries between separated classes are orthogonal to coordinate axes. Using (3), a set of input/target couples for which the probability function of the input space is defined, the neural network can be trained according to (4), where $\mathbf{x}_j$ is the couple of two N-

dimensional input vectors, and $\mathbf{C}_j$ is the M-dimensional vector describing the conditions of target classes.

$$\{\mathbf{x}_1,\mathbf{C}_1\},\{\mathbf{x}_2,\mathbf{C}_2\},\ldots,\{\mathbf{x}_j,\mathbf{C}_j\},\ldots,\{\mathbf{x}_N,\mathbf{C}_N\} \quad (3)$$

$$\begin{aligned}\mathbf{x}_j &= \left\{X_j^{(2)},X_j^{(3)}\right\} \qquad j=1,\ldots,N\\ \mathbf{C}_j &= \left\{S_1,S_2,\ldots S_k,\ldots S_M\right\}, \quad k=1,\ldots M\end{aligned} \quad (4)$$

The hidden neurons of the first layer initialize the weight matrix $\mathbf{W}_{kj}$, and compete, and then identify the neuron-winner, which has the minimum Euclid distance d_k. It receives output value equal to 1, and defines the corresponding target class, as is shown in (5).

$$\begin{aligned}&S_k=1, \quad for \quad d_k^{\min}, \qquad d_k=\sqrt{\sum_{j=1}^{N}\left(X_j-W_{kj}\right)^2}\\ &S_k=0, \qquad otherwise\end{aligned} \quad (5)$$

At the linear layer, the neuron-winner has a negative feedback relative to the rest of neurons and a strong positive feedback relative to itself, and this is used in the learning process. The coefficients of the neurons in each consequent epoch q of the training process are altered according to Kohonen learning rule, as is introduced in (6).

$$\begin{gathered}W_{kj}(q)=W_{kj}(q-1)\pm\xi\left(X_j(q)-W_{kj}(q-1)\right)\\ 0<\xi\le 1\end{gathered} \quad (6)$$

The coefficient ξ depends on the training epoch number q, and can be in advance adjusted in interval [0, 1] (standard 0, 1). The sign of the training coefficient ξ is positive for the neuron-winner, and negative for the neighboring neurons. As a result, the area of neighboring neurons for the neuron-winner is transformed in the training process lessening Euclid distances. We can see that the neural network is determined with two hidden neurons in each cluster class at the competitive layer. At the same time, consequent adjustment of the target class on the horizontal and vertical axes produces rectangular zones at the linear layer. Each class is described via a rectangular class face, which is proportional to the approximated probability density function.

LVQ AND OPTIMAL OCCUPANCY PROBABILITY DISTRIBUTION

We suggest use of the learning vector quantification (LVQ) in a neural model in order to determine the value of the occupancy probability density function for each discrete feasible state of two-dimensional CTMC set. The probability density function of the input vector space (VQ) corresponds to the values of streaming and interactive time series and is indicated in (7).

$$f(x_1^{(2)},\ldots,x_k^{(2)},x_1^{(3)},\ldots,x_k^{(3)}) = \frac{\partial^{n\times 2}}{\partial x_1^{(2)},\ldots,x_k^{(2)},x_1^{(3)},\ldots,x_k^{(3)}}\mathbf{F}\left(x_1^{(2)},\ldots,x_k^{(2)},x_1^{(3)},\ldots,x_k^{(3)}\right) \quad (7)$$

In this paper we consider an approach to link occupancy neural modeling, which consists of the four following basic phases: preliminary vector quantification (PVQ), bipartition and weight centers, optimal class boundaries, and learning vector quantification. We introduce a preliminary separation of the input space values into different classes in the phase of preliminary vector quantification. Then we determine rectangular boundaries of the separated classes, which are orthogonal to coordinate axes. The linear layer fits the boundaries of the target classes in such a manner that we have the same number of the target values in each class. Next, we consider a small telecommunication system with a link capacity C=25, and for two-dimensional presentation n_1=1 is kept fixed, i.e., the available bandwidth for the streaming and interactive flows is 24 bandwidth units. The peak bandwidth requirements for these flows are b_2=7 and b_3=5, respectively. The flows are characterized with minimum accepted bandwidth, which we set to $b_2^{\min}$=4.2 and $b_3^{\min}$=2.8. The cut-off parameters are N_2 =4 and N_3 =2, respectively. The system has 15 feasible states, out of which there are 5 states - (1,4,0); (1,3,1); (1,4,1); (1,3,2); (1,4,2) where at least one of the flows is compressed below the peak bandwidth specified by b_2 and b_3. These are the goals of our numerical study on the link occupancy distribution, and determine the partial overlap as well as the link allocation strategy. The 500 samples were generated for positive discrete values of n_2 and n_3 with binomial distribution in interval (1,100). For better understanding of the example, the PVQ of 15 target classes with equal number of the target values in each class is offered. A lesser class face corresponds to a superior probability density link occupancy function, as is demonstrated in Fig.2. After we optimize the PVQ with use of bipartition algorithm, which is search of the new boundaries for the target classes. Separation of the input space with bipartition helps us to determine the weight centers for each class received using the PVQ. We separate the input values into M-1 couples of clusters in this phase. The bipartition algorithm separates the input space $\Re$ into two clusters $\Re_1$ and $\Re_2$, while $\Re_1$ coincides with class S_1. The weight center W_1 for cluster $\Re_1$ is defined under Kohonen neural network learning and training. Next, the space $\Re_2$ is separated into new two clusters $\Re_{21}$ and $\Re_{22}$, while $\Re_{21}$ coincides with class S_2, and the weight center W_2 is defined. Once using the described procedure, all the weight centers W_k ($W_k^{(2)}$, $W_k^{(3)}$) of all the classes are defined. We introduce the weight centers, which are received after 100 training epochs and depicted with black points, as shown in Fig.3. Then, we use the midpoint of Manhattan distances between the weight centers of the neighboring classes as criteria of optimal class boundaries. The number of the target values may be different within the new optimal class boundaries. We determine Manhattan distances between the neighboring weight centers according to (8).

$$D_{k,k+1}=\left\|W_k^{(2)}-W_{k+1}^{(2)}\right\|+\left\|W_k^{(3)}-W_{k+1}^{(3)}\right\| \quad (8)$$

The boundaries of the target classes are shifted parallel to coordinate axis for all the classes until the calculated midpoint of Manhattan distances is going to be reached. Then we recalculate the numbers of the target values within the new rectangular classes, as is also shown in Fig. 3.

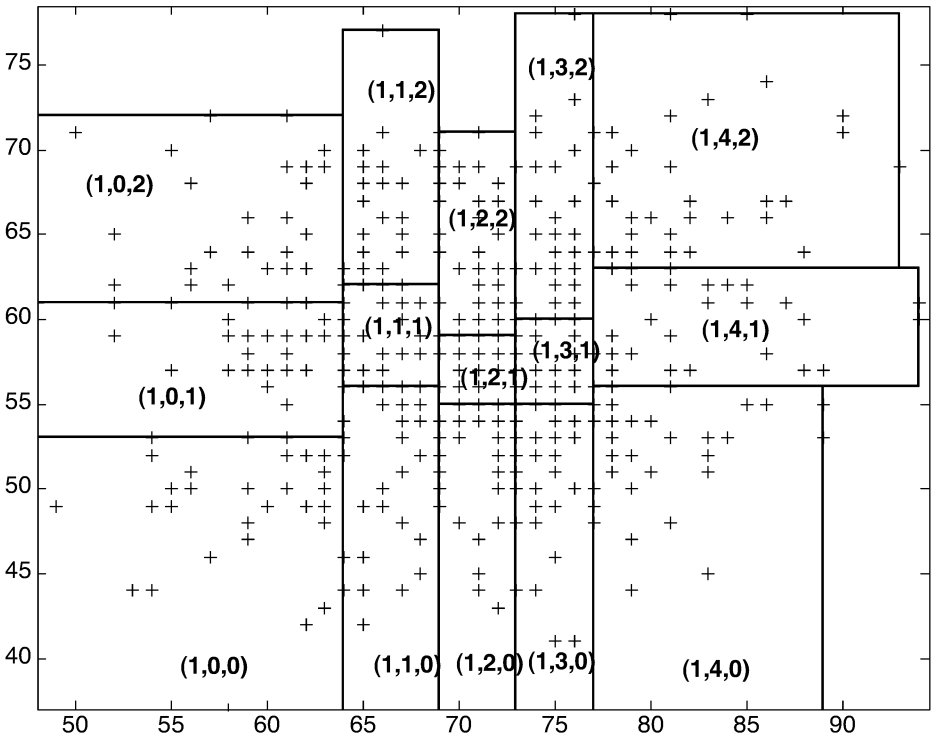


Figure 2 Preliminary VQ with equal number of target values in each class

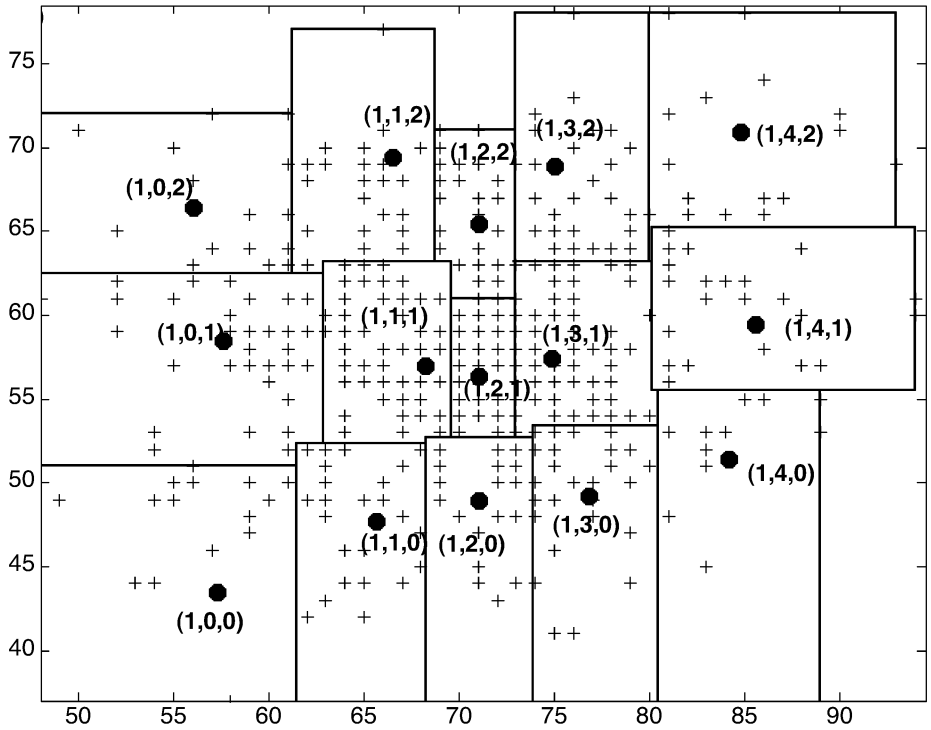


Figure 3 Optimal boundaries between classes

We use learning vector quantification to receive the approximated values of occupancy probability distribution in the target classes for input time series in this phase. The accuracy of results which we receive depends on the number of training epochs, and it increases with growth of the number of the learned target values. The efficiency of suggested clustering procedure is illustrated in Table 1, in which the occupancy probability density function is calculated for 15 feasible states. The influence of the PVQ in shifting the class boundaries is also demonstrated.

Table 1 Density Distribution %

Class	PVQ	LVQ
(1,0,0)	1.405	1.189
(1,1,0)	4.818	5.241
(1,2,0)	7.026	6.674
(1,3,0)	8.029	5.435
(1,4,0)	2.342	1.702
(1,0,1)	3.513	3.16
(1,1,1)	11.242	14.202
(1,2,1)	28.104	23.74
(1,3,1)	16.863	14.943
(1,4,1)	3.011	2.759
(1,0,2)	1.916	2.327
(1,1,2)	4.497	3.998
(1,2,2)	1.044	9.011
(1,3,2)	4.684	4.375
(1,4,2)	1.506	1.248

The LVQ probability density function is decreased for the both classes with minimum and maximum values. For example, the minimum density distribution for class (1,0,0) is decreased from 1.406 to 1.189 and the maximum density distribution for class (1,2,1) is decreased from 28.104 to 23.74, respectively. Cutbacks of LVQ density function and low values of density distribution in the classes with partial overlap are results of corrections of the link occupancy allocated strategy.

CONCLUSIONS

A clustering procedure with Kohonen neural network, described with two-dimensional Markov reward model is developed. Suggested approach includes four phases of neural modeling: preliminary vector quantification, weight centers determined with use of bipartition, class boundaries optimization, and learning vector quantification. The model provides the opportunity to determine optimal probability occupancy density function for feasible steady states of Markov chain. The application of probability occupancy density function for quick and easy performance analysis of the partial overlap allocation scheme is demonstrated using the numerical example.

REFERENCES

Hofmann, T., Buhmann, J. 1998. Competitive Learning Algorithms for Robust Vector Quantization, *IEEE Transaction on Signal Processing*, 46(6), pp. 1665-1675.

Kulkarni,V. G., 1995. *Modeling and Analysis of Stochastic Systems*, Chapman Hall.

Racz, S. Telek, M., Fodor, G., 2001. Call level performance analysis of 3rd generation mobile core network, *Proceedings of the IEEE International Conference on Communications, ICC*, Helsinki, Finland, vol. 2. pp. 456–461.

Radev, D., Radeva, S., 2003. Artificial Intelligence Modelling of Stochastic Processes in Digital Communication Networks, *Journal of Electrical Engineering,* vol. 54, N:9-10, pp. 255-259.

Radev, D., Lokshina, I., Radeva, S., Insinga, R., 2004. Neural Modelling of Link Occupancy Distribution for Broadband Telecommunication Transmission, *Proc. of the DSI 2004: 35-th Annual Meeting of the Decision Sciences Institute*, Boston, USA, pp. 3001-3009.

BIOGRAPHY

DIMITAR RADEV, PhD is from University of Rousse, Rousse, Bulgaria. In 2001 he became Associate Professor at Department of Communication Technique and Technologies at University of Rousse. His mainly research interests are connected with Teletraffic Theory, Simulation and Modeling of Communication Networks, Performance Analysis of Queuing Systems.

IZABELLA LOKSHINA, PhD is an Assistant Professor of Management Information Systems at Division of Economics and Business at SUNY Oneonta. Her positions included Senior Scientific Researcher at the Moscow Central Research Institute of Complex Automation and Associate Professor of Automated Control Systems at Moscow State Mining University. Her main research interests are Artificial Intelligence: Fuzzy Systems and Neural Networks, and also System Modeling.

SVETLA RADEVA, PhD is from University of Architecture Civil Engineering and Geodesy, Sofia, Bulgaria. Since 2001 she is Associate Professor at Department of Computer-Aided Engineering at University of Architecture Civil Engineering and Geodesy. She is mainly interested in Artificial Intelligence, Process and Product Modeling.

A MEASUREMENT-BASED SIMULATION MODEL OF A WEB CLUSTER

Isabel Wagner Kai-Steffen Hielscher Reinhard German

Department of Computer Science 7
(Computer Networks and Communication Systems)
University of Erlangen-Nürnberg
Martensstraße 3
D-91058 Erlangen, Germany
Email: {wagner, ksjh, german}@cs.fau.de

***Abstract*— To evaluate the performance of distributed web servers, we have built a cluster-based web server with eleven nodes in our laboratory to conduct measurements of fine granular performance data like individual per-packet delays.**

We implemented a precise GPS-based time synchronization solution and a software monitoring solution for high-volume measurement of one-way delays in the range of microseconds.

The obtained data is used in the input modeling phase to parameterize and calibrate our UML-based simulation models for the web cluster. Additionally, we use the measured performance figures to validate and verify the output of our simulations.

Our models help us to scale cluster-based web server systems for a given application to deliver a demanded quality of service level.

I. INTRODUCTION

The growing use of rich media in the Internet, the trend to create web content dynamically and new technologies in the access networks like DSL and cable modems have dramatic impact on the performance needs of web servers. The success of Linux and the availability of low-cost PC hardware provide cost-effective solutions to handle these challenges. Commodity server hardware can be combined with load balancing mechanisms to build high-performance cluster-based web servers.

Distributed web servers are composed by a number of web server machines, denoted here as *real servers*, where the web server software, e.g. Apache, is executed. Additionally, some method of distributing the load is needed to redirect the requests of the clients to one of the real servers. The most common approach to load-balancing is the DNS-based load-balancing mechanism where the host name of the server is resolved to different IP addresses belonging to different machines in a round-robin fashion. This method, known as *round-robin DNS* is well suited for geographic load balancing, where requests of clients are redirected to geographically near web servers which are distributed over the globe.

Other approaches for local load balancing, where all real servers are located in proximity, involve changing the routing process in some way to redirect the TCP packets to a real server chosen by a scheduling algorithm.

II. THE WEB CLUSTER

For our laboratory setup we have chosen an open-source software solution for the operating system of the cluster nodes, the web server daemon and and the load balancing solution. All software is used on standard off-the-shelf PC hardware.

A. The Linux Virtual Server System

We use the *Linux Virtual Server* [6] system for load balancing in our web cluster. This Open Source project implements a layer-4-switching system that supports routing-based load-balancing of various IP-based services.

The whole distributed web server carries one single IP address called *Virtual IP Address (VIP)*. Requests sent to this address are balanced among the real servers carrying the *Real IP Addresses (RIP$_i$)*.

Three mechanisms for load-balancing are available: *Network Address Translation*, *IP Tunneling* and *Direct Routing*.

We did most of our measurements with NAT, where the packets arrive at the load balancer first. This node changes the destination IP address (VIP) in the TCP segments to the IP address of one of the real server nodes (RIP$_i$) as determined by the scheduling algorithm. The load balancer is used as the standard gateway for the reply packets of the real servers. When packets belonging to replies arrive, the source address is changed to the VIP and forwarded to the client via the Internet. NAT involves rewriting the packets twice and is limited to geographically close nodes since the load balancer has to be the gateway for all real servers.

Tunneling and Direct Routing cause less overhead because the packets sent by the real servers do not have to pass the load balancer. Since our load balancer does not reach saturation even with the NAT approach, we did most of our measurements with NAT. Details about the other two methods can be found in [6].

III. MEASUREMENT INFRASTRUCTURE

Our measurement approach is based on software monitoring. We instrumented the netfilter framework of the Linux TCP/IP stack to time stamp all relevant packets and write the

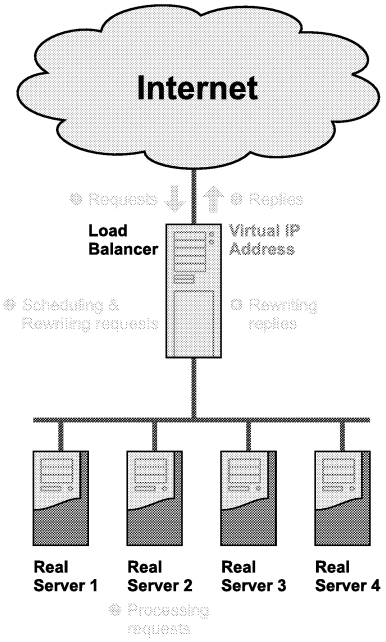


Fig. 1. Load Balancing via NAT

header information in a kernel buffer in the load generators and cluster nodes. This buffer can be read during or after the measurement and written to disk in binary form for analysis. In addition to the kernel level time stamping of TCP packets, we instrumented the Apache web server software and our own sever applications to obtain application level timestamps. Additional timing information can be obtained from the load generator software.

For a detailed model of the system, precise measurements of one-way delays are needed. For that reason, we implemented a GPS-based time synchronization solution which is based on standard time synchronization tools such as NTP and the PPS-API. We used some modifications and enhancements of the standard components to improve timekeeping accuracy.

We use the PPS-API [9] to improve the accuracy of NTP. The PPS-API provides a facility to generate high resolution time stamps for external events. Our GPS receiver has a PPS signal output that generates a pulse that marks the beginning of a second. The timestamps of the PPS pulses can be used by NTP in combination with a network timer server to discipline the kernel clock [3].

On the client side, an HTTP load generator is needed. `httperf`, a load generator developed by David Mosberger [10], is able to generate load that overloads a web server. The number of requests and the rate of the requests can be specified on the command line. Since the test client PC has certain limitations on the number of TCP connections that can held open simultaneously, `httperf` supports parallel execution on a number of client machines. This load generator is ideal for studying the behavior of a web server in extreme situations for finding the system's limits.

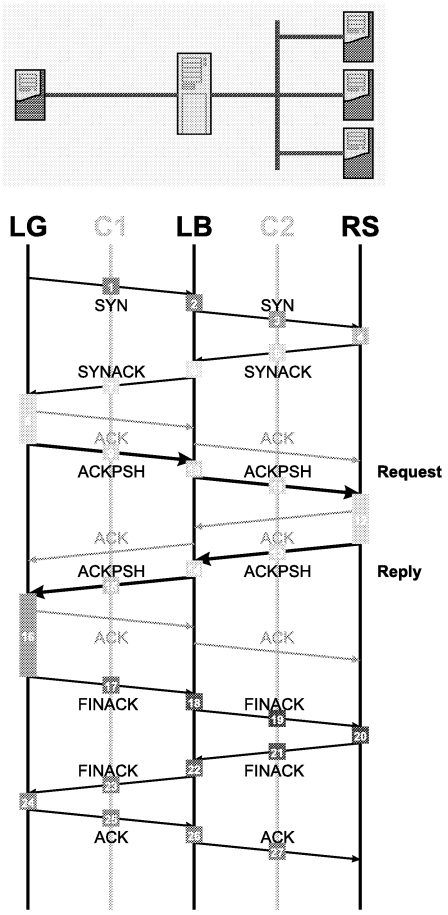


Fig. 2. Illustration of the Delays

IV. Measurements

Five real servers and one load balancer were used in a NAT environment with round-robin scheduling for the following example measurements. One test client generated HTTP requests using the `httperf` load generator.

For this example, we generated 10000 HTTP/1.0 requests for a binary file with a size of 1024 bytes. This resulted in a request size of 65 bytes. The web server added 244 bytes of header information, so the resulting replies had a size of 1268 bytes. Since this is smaller than the maximum segment size used (1500 bytes), all replies consisted of exactly one TCP segment. Figure 2 provides an illustration for the 27 individual delays in the exchange of TCP segments that contribute to the total processing time of the HTTP request drawn as colored vertical bars. Time advances along the vertical axis from top to bottom. The horizontal position shows where the delays are caused: Either by the load generator (LG), the network channel between the load generator and the load balancer (C1), the load balancer (LB), the network between the load balancer and the real servers (C2) or by one of the real servers (RS).

The delays in the channel C1 and C2 include not only the physical propagation delay and the store-and-forward delay of the switches between the hosts, but also to time between the reception of the packet at the node of the cluster and the

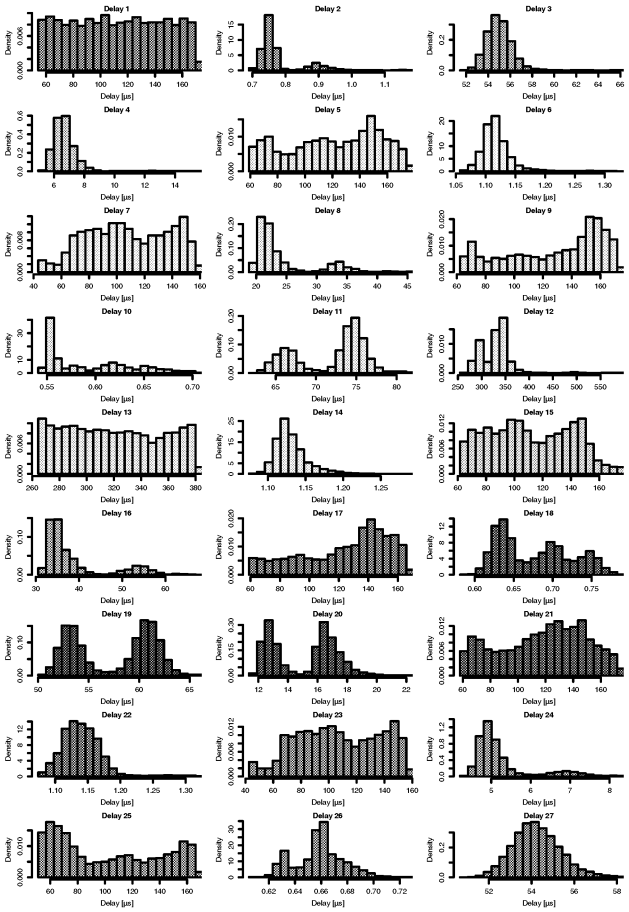


Fig. 3. Histograms of Observed Delays

beginning of the packet processing in the TCP/IP stack of the operating system.

The segments sent during delay 11, 12 and 16 are due to TCP protocol mechanisms (fast ACK) and do not mark a state change in the HTTP protocol.

Histograms of the measured delays are shown in figure 3.

All gathered data enables us to estimate theoretical distributions or to use an empirical distribution to parameterize a simulation model of our distributed web server and to calibrate and validate the model.

Care must be taken when using these probability distributions in models of the system, because most of the delays are autocorrelated.

Please note that the tails of the distributions have been cut off at the 99.5% quantile for better visualization in the histograms. The few longer delays are mainly caused by other system tasks and interrupt handlers being active during the reception of the packets.

The minimum delay that occurs as a difference between a starting point on one node and an end point on another node (delays with odd numbers) is longer than 42 μs. This suggests that a synchronization precision on the order of one microsecond is sufficient for our purposes. Some of the other delays (delays with even numbers) are much shorter, but since they are caused by single nodes of the cluster, not affected by time offsets between nodes.

V. Simulation Model

We used AnyLogic to build a simulation model of the web cluster. AnyLogic supports discrete event simulation in combination with UML state charts and enables seamless integration of arbitrary java code by being based on the java programming language.

The model imitates the real system as closely as possible. This is illustrated in figure 4 which shows the *root* object containing the topmost building blocks of our model. Network packets are generated by the *client* objects, then passed through the first network channel (*delay_c*), the *load balancer*, the second network channel and finally arrive at a *real server* where they are processed and sent back all the way to the client.

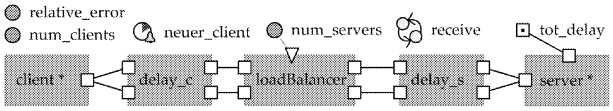


Fig. 4. Root object in AnyLogic

A. Input representation

The measured delays serve as input to our model at 27 places. Our model supports three different input types, trace-driven simulation, empirical distributions and theoretical distributions.

Complete deterministic behavior of the model can be achieved by inputting the measured traces themselves. This is mainly used to validate the correctness of the model. As the traces can only reproduce measured data, their use for estimating system behavior is rather limited.

The empirical distributions were implemented as a first step towards somewhat random simulation outputs. Their main problem is that their range is fixed, thus none of the outlier values measured in the real system can be reproduced. They also do not account for any of the autocorrelations found in the real system.

To solve these problems, we used ExpertFit to fit theoretical distributions to most of the delays. In some cases (like delay 6, see figure 3), a single theoretical distribution sufficed to achieve a good fit.

Other delays, like delay 11, are clearly multi-modal. In these cases, we divided the measured values in two groups and fitted a theoretical distribution for each group. Which one of the two distributions is sampled from in the model is decided randomly, based on the relative frequencies of the two groups.

The third case we have to deal with are delays like delay 18. It seems to have three peaks, but separating them is very hard as all the values between the peaks are also quite frequent. By looking at different plots of the delay, especially the delay over time plot, we found that the delay can be clearly divided

in an upper and a lower range. These two ranges are almost mutually exclusive, so that a phase with only delays from the lower range is followed by a phase with delays from the upper range, and so on. Thus we decided to model delay 18 (and the

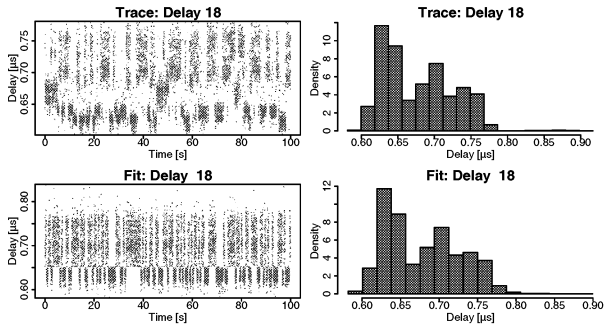


Fig. 5. Delay 18: measured values (top) and fitted distribution (bottom)

similar ones) with two bézier distributions [12]. Which phase the model is in (that is, which distribution should be sampled from) at a particular point in time is indicated by a state chart. That state chart also controls the phase changes which are based on the phase lengths observed in the measurements.

Figure 5 compares the scatter plot and histogram of the measured values to the fitted distribution. It can be seen that the fit seems to be very appropriate.

B. TCP

A core element of the model is our representation of TCP. We decided TCP to be the lowest level of the network stack that is explicitly represented in our model. The reason for this is that the lower levels (like IP or the hardware layer) offer neither very interesting procedures nor important parameters that could change system behavior significantly, while the TCP layer offers both.

Our model of TCP supports all important aspects described in the RFCs 793 (basic TCP), 1122 (requirements for Internet hosts), 2581 (congestion avoidance, slow start, fast retransmit) and 2988 (retransmission management). The model can thus deal with packets with arbitrarily large sizes and faulty network channels, just as a real system would. As our measurements were done in a loss-free environment without packet splitting, the mechanisms built into the TCP object enable us to estimate system behavior under different conditions (without repeating any measurements).

C. Client

The *client* object does not exactly replicate the behavior of the load generator in the real system (that is, one machine that generates all load), but rather how the system would behave if it were situated in a real web environment (that is, one machine generates just one request). As a consequence, the rate at which requests are sent to the server is not configurable in the client, but rather in the *root* object where the *client* objects are generated.

Each client consists of two sub-objects, *application* and *TCP*, and a queue where the client processing delays are applied to all incoming packets. The requests are generated in the application object. The length of the answer packet sent by the server can be configured here. The request is then passed on to the TCP sub-object which establishes the connection to the server and sends the request embedded in a *TCP packet* object.

D. Load balancer

The load balancer object distributes the incoming load to all available real servers (the exact number can be configured in the *root* object). The model currently supports three different load balancing schemes: *random* load distribution, *least used* and *round robin*, which is the one used in the real system.

We refrained from modeling the load balancer processor as its utilization is significantly below 1% in the real system and it therefore wouldn't have had any measurable effects on the model.

E. Servers

The *server* object is the most complex building block of our model. It consists of three main parts, the *process*, *processor*, and *system process* objects.

There's a *process* object for every unique connection to a client. These processes are responsible for processing all incoming packets and sending appropriate answer packets. No simulation time passes while packets are being processed here. Embedded are, of course, a *TCP* and a *server application* object which accurately represent the processing done in the real system.

The answer packets generated by a server *process* are handed to the *processor* object where the actual timekeeping for the server takes place. This means that the server processing delays are applied here, and there's also a queue for waiting packets in case a packet arrives while the *processor* is still busy.

The *system process* object has just the one purpose of occupying the *processor* from time to time. This is designed to reflect the actual system processes at work in a real system. The intervals between system processes and the amount of processor time taken each time have been calibrated with processor idle data measured in the real system.

F. Network channels

The network channels are modeled in several stages. First of all, we modeled the two directions in a channel separately as the physical links are full duplex. Inside each of the directions the packets are first processed by a single server queue where the transmission delay is calculated and the channel is set busy for that time. That means that we calculate the transmission delay (based upon the packet size and channel bandwidth) as the minimum delay for each packet. The other components adding to the network delay (such as propagation delay or queuing delay) are negligibly small compared to the transmission delay and have therefore been left out in order to reduce model complexity.

Afterwards, the packets are delayed by the measured delays (minus the transmission delay) by an infinite server queue.

This takes into account that parts of the measured channel delays are not caused by the network, but by the lower parts of the TCP stacks, where packets can be processed in parallel.

G. Simulation Experiments

Output data are gathered at various places throughout our model. Similar to the real system, we collect 28 timestamps for each request. These can be used to calculate and analyze the 27 simulated delays. We also record the total delay, which is used to determine the optimal number of replications in a simulation. Throughput, mean queue length and utilization are collected not only for each of the network channels, but of course also for the server processors.

The first simulation experiments we conducted were done with the measurement setup (5 real servers, 1024 byte answer data, round robin) and different request rates. Table I shows a quantile comparison of measured and simulated data for 50, 100, 200, 500 and 1000 requests per second. All numbers are in milliseconds.

	min	25%	median	mean	75%	max
50 cluster	1.750	1.934	2.039	2.247	2.146	406.000
50 model	1.785	2.114	2.194	2.197	2.272	14.020
100 cluster	1.656	1.929	2.010	2.026	2.120	2.682
100 model	1.744	2.122	2.197	2.202	2.272	16.471
200 cluster	1.664	1.952	2.060	2.590	2.152	357.000
200 model	1.801	2.120	2.200	2.218	2.281	22.736
500 cluster	1.632	1.945	2.062	24.480	2.172	902.000
500 model	1.793	2.118	2.195	2.492	2.271	63.485
1000 cluster	1.656	2.082	2.196	131.000	2.467	3002.000
1000 model	1.787	2.135	2.235	7.320	2.344	178.530

TABLE I

TOTAL DELAY QUANTILE COMPARISON: MODEL AND MEASUREMENTS

It can be seen that the model is, in most cases, quite close to the measured delay values. The excessively high maximum delays at 500 and 1000 requests per second are mostly caused by two factors. The first is the growing server load and resulting queuing effects. As described above, the model reproduces this effect to some extent although its impact seems to be still higher in the real system.

The second factor is the utilization of the load generator. As all measurements were done with only one load generator, the processor load on this machine gets close to 100% when 1000 requests per second are generated. The delays in the load generator cause about one third of the measured maximum delay, which should be ignored in the comparison between simulation and real system, as the model strives to simulate the cluster behavior in a real web environment.

Another task was to compare the model processor load with data measured in the real system. As can be seen in table II, the measured processor loads are not constantly increasing with increasing request rates. Also, no measurements are available for request rates above 250 requests per second per real server.

Even though the model processor loads are constantly increasing and do not exactly match the loads measured in the real system, the model seems to generate realistic processor loads.

request rate	system load	model load
25	1.92%	1.08%
50	0.11%	2.22%
100	0.10%	5.02%
150	1.35%	9.53%
200	18.04%	18.99%
250	13.37%	22.39%

TABLE II

CPU LOAD COMPARISON

With all that in mind, the model appears as a good representation of the real system and can thus be used to simulate similar clusters to estimate their behavior.

VI. CONCLUSIONS

The laboratory setup of the web cluster allows us to conduct fine-granular measurements of the internal behavior in a cluster-based web server.

The example measurements show how it is used to obtain a detailed view of the network delays in our distributed web server. The observed individual delays can be used in creating a detailed model of the system.

Simulation experiments with a configuration similar to the measurement setup show that the measured delays at different system loads are quite closely reproduced by our model.

This enables us to run simulation experiments with various system configurations and system loads, and do a detailed analysis of the estimated system behaviors.

REFERENCES

[1] P. Barford and M. Crovella. Generating representative Web workloads for network and server performance evaluation. *ACM SIGMETRICS Performance Evaluation Review*, 26(1):151–160, 1998.

[2] K. Egevang and P. Francis. The IP Network Address Translator (NAT). Request for Comments RFC-1631, Internet Engineering Task Force, May 1994.

[3] K. Hielscher, R. German. A Low-Cost Infrastructure for High Precision High Volume Performance Measurements of Web Clusters. In P. Kemper and W.H. Sanders, editors, *TOOLS 2003, Computer Performance Evaluations, Modelling Techniques and Tools, 13th International Conference, Proceedings*, pages 11–28, Springer, Heidelberg, 2003.

[4] R. Klar, P. Dauphin, F. Hartleb, R. Hofmann, B. Mohr, A. Quick, and M. Siegle. *Messung und Modellierung paralleler und verteilter Rechensysteme*. Teubner-Verlag, Stuttgart, 1995.

[5] J. Levine. Introduction to time and frequency metrology. *Rev. Sci. Instrum.*, 70:2567–2596, 1999.

[6] Linux Virtual Server Project. `http://www.linux-vs.org/`.

[7] D. Mills. Internet time synchronization: the Network Time Protocol. *IEEE Trans. Communications*, 39(10):1482–1493, October 1991.

[8] D. Mills. A Kernel Model for Precision Timekeeping. Request for Comments RFC-1589, Internet Engineering Task Force, March 1994.

[9] J. Mogul, D. Mills, J. Brittenson, J. Stone, and U. Windl. Pulse-per-second API for Unix-like operating systems, Version 1. Request for Comments RFC-2783, Internet Engineering Task Force, March 2000.

[10] D. Mosberger and T. Jin. httperf: A Tool for Measuring Web Server Performance. In *First Workshop on Internet Server Performance*, pages 59–67. ACM, June 1998.

[11] Netfilter/IPtables home. `http://www.netfilter.org/`.

[12] Mary Ann Flanigan Wagner and James R. Wilson. Using univariate Bézier distributions to model simulation input processes. In *25th conference on Winter simulation*, vol. 28, Los Angeles, California, USA, 1993, pp. 365–373.

NAPT: A SIMULATION TOOL FOR NETWORK ANALYSIS AND PARAMETER TUNING

Mario Barbera, Andrea G. Busà, Alessandro G. Di Nuovo, Alfio Lombardo and Giovanni Schembra
University of Catania – DIIT
Viale Andrea Doria, 6
95125 Catania,
Italy
E-mail: (mbarbera, abusa, adinuovo, schembra, lombardo)@diit.unict.it

KEYWORDS
Network Simulation, Parameter Tuning, Fluid-Flow analysis, Parallel Genetic Algorithms, Quality of Service.

ABSTRACT

In the current Internet scenario increasing interest is being shown in Quality of Service (QoS) issues, so simulation and parameter tuning are becoming increasingly important. In this paper we present a user-friendly, easily expandable and time-saving tool, which is able to perform network simulation through Fluid-Flow models and parameter tuning with Parallel Genetic Algorithms. All the network elements involved (TCP sources, links, routers and so on) are modeled with the Fluid-Flow approach as independent systems: their outputs only depend on their inputs and internal state variables. In this way each element can be implemented as a simple Java class. Network elements can be connected to each other by a powerful graphical interface: the user can simply drag and drop them in order to easily represent the network topology to be designed. Subsequently the user can select network parameters to be optimized in order to achieve prearranged QoS guarantees. During this phase the tool adopts a Parallel implementation of a Genetic Algorithm (GA) which reduces the execution time required for optimization. A case study is also presented.

1. INTRODUCTION

An increasing number of Internet users are looking for a network with Quality of Service (QoS) guarantees. Banks or medium-large industries, for example, seek basic network requirements, like minimum throughput, to improve their efficiency and to give their customers a better service. This trend has encouraged researchers to define new protocols and buffer management algorithms in order to optimize network performance. To exploit these techniques fruitfully, both users and network providers have to set a large number of protocol/algorithm parameters. RED (Floyd and Jacobson 1993), for example, is an Active Queue Management (AQM) technique implemented in the output buffers of most current routers [URL: RED]. It uses four parameters. So a network with only 20 RED buffers needs 80 interrelated parameters to be set!

In order to solve the network parameter tuning problem in a real network, researchers have developed a number of automatic methods usually classified as analytical methods, on-line methods and simulation methods.

In the literature we can find different analytical methods. Unfortunately these methods are strictly tied to the scenario being addressed, so there is no unified tool to study all the possible network topologies or technologies (Shridharan et al. 2003).

The on-line approach is an evolution of operating trial & error method in which a technical operator tunes parameters and then analyzes network changes. In (Ye and Kalyanaraman, 2002), for example, an on-line simulation of RED is used in order to achieve a high throughput while keeping a low queueing delay. On-line methods usually adopt a local optimization technique which pursues a local optimum; so, when the network size increases, on-line approaches provide a parameter tuning which might be very far from the global optimum providing the QoS requested by the user.

Finally, simulation methods could address the global optimization problem, but they need a huge computation time if a large network is to be analyzed. An example of a very popular network simulator is NS2 [URL: NS2]; it uses a packet-by-packet simulation. This approach is time consuming, and the simulation time increases exponentially when networks become larger. Better approaches are hybrid network simulators, like MAYA (Zhou et al. 2004). This simulation tool uses both packet-by-packet and Fluid-Flow approaches. MAYA results demonstrate that the simulation time decreases by about 20% in comparison with classical packet simulations, and the results are very accurate as well. However, time consumption remains exponentially tied to network size because MAYA considers a large number of packet events. For this reason MAYA does not represents a solution to the network parameters optimization problem, because a good optimization could require more than thousands of runs.

To satisfy the network provider's request for an accurate and fast network design and planning tool, in this paper we introduce a simulation methodology based on Fluid-Flow models only, and an optimization algorithm based on Parallel Genetic Algorithms. The Fluid-Flow approach results into both a simulation time linearly tied to the network size, and a good degree of results accuracy.

Parallel Genetic Algorithms ensure fast convergence towards the best parameter set, providing the requested QoS. In order to allow a simple implementation of the Fluid-Flow model solver we developed a modular approach: all the elements involved (links, routers and so on) are modeled as independent systems, that is, their outputs only depend on their inputs and internal state variables. In this way it is possible to implement each network element as an independent component, and to model the whole network by linking components to each other. The modeling modularity also gives us the possibility to expand the tool with additional components.
Finally the Fluid-Flow solver and the Parallel Genetic Algorithm have been combined into a single integrated tool, called NAPT [URL: NAPT], providing network operators with a powerful graphical interface to quickly and simply create a model of the network being addressed.
The rest of this paper is organized as follows. In Section 2, after a short description of the main characteristics of Fluid-Flow models, we show how the fluid models of two important network components are built. For the sake of clarity, Section 3 introduces some remarks on Parallel implementation of Genetic Algorithms (the skilled reader can skip this part); moreover in the same section we introduce the parametric fitness function we use. Section 4 provides an overview of the tool characteristics, while in Section 5 we present the results obtained with our tool in a network optimization problem. Finally, Section 6 concludes the paper.

2. FLUID FLOW MODELS

Fluid models have recently been applied as a scalable and accurate tool to analyze the performance of IP networks (Misra et al. 2000). Fluid models describe the network behavior through a set of ordinary differential equations that provide an abstract deterministic description of the average network dynamics. Fluid-Flow approach neglects the detailed short-term packet-by-packet behavior, which prevents the scaling of traditional approaches to the study of large IP networks. The differential equations are then numerically solved to obtain estimates of time-dependent network behavior.
The reasons why fluid flow models are very suitable to implement a network dimensioning tool can be summarized as follows:

- Fluid-Flow models are very accurate, both in transient and in steady-state analysis (Barbera et al. 2004a, Barbera et al. 2004b).
- Sources with the same average behavior can be aggregated, thus considerably reducing the number of equations needed to describe the system. This gives the model a high degree of scalability and allows large networks loaded by a huge number of traffic flows to be considered.
- The processing time required by the solver of the set of differential equations is much shorter than the packet-by-packet simulation time when network size or link capacities increase.
- The set of differential equations describing the behavior of each source or network element is independent of the network topology. The set of differential equations for the whole system (sources and network nodes) can therefore be obtained simply by considering the equations related to each source and each network element. In other words, the system model is a combination of building components.

Each component is characterized by:

- a set of Input Variables;
- a set of Output Variables;
- a set of constant Parameters;
- a set of State Variables and Internal Equations that constitute the fluid model of the component.

In order to better understand how the fluid model of the generic component of our tool is defined, we now present two meaningful examples: the TCP source group in Section 2.1 and the RED buffer in Section 2.2. For the sake of conciseness we will not discuss about how the differential equations for each component are derived; more detailed information can be found in (Barbera et al. 2004a, Barbera et al. 2005).

2.1 TCP Source Group Component

Figure 1 is a schematic representation of the TCP Source Group component.

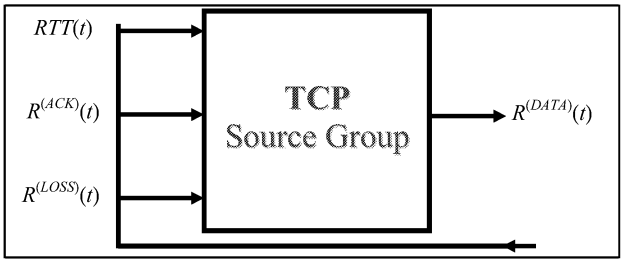


Figure 1: Schematic Representation of TCP Source Group Component

The component can be described as follows:

INPUT VARIABLES

- $RTT(t)$: round trip time at the time instant t (expressed in seconds);
- $R^{(ACK)}(t)$: arrival rate of ACK packets at the time instant t (expressed in bytes/s);
- $R^{(LOSS)}(t)$: rate of loss indications at the time instant t (expressed in bytes/s);

OUTPUT VARIABLES

- $R^{(DATA)}(t)$: total source group transmission rate at the time instant t (expressed in bytes/s);

PARAMETERS

- $s^{(P)}$: packet size (expressed in bytes);
- $s^{(A)}$: ACK size (expressed in bytes);
- $N^{(F)}$: number of flows in the group;
- $f(s)$: function that represents the number of sources in the group that have to transmit a file of a size less than or equal to s (expressed in packets);

- G: array of starting instants for each source.

STATE VARIABLES

- $W(t)$: congestion window for the generic flow in the group at the time instant t (expressed in packets);
- $T(t)$: threshold separating the Slow Start and the Congestion Avoidance ranges for the congestion window of the generic flow in the group at the time instant t (expressed in packets);
- $D(t)$: number of packets successfully sent by the generic flow from its start time to the time instant t (expressed in packets);
- $r^{(ACK)}(t)$: arrival rate of ACK packets for the generic flow in the group at the time instant t (expressed in packets/s);
- $r^{(LOSS)}(t)$: rate of loss indications for the generic flow in the group at the time instant t (expressed in packets/s);

INTERNAL EQUATIONS

$$a(t) = N^{(F)} - f\left(D(t)\right)$$

$$r^{(ACK)}(t) = \frac{R^{(ACK)}(t)}{s^{(A)} \cdot a(t)}$$

$$r^{(LOSS)}(t) = \frac{R^{(LOSS)}(t)}{s^{(P)} \cdot a(t)}$$

$$N^{(A)}(t-\tau,t) = \int_{t-\tau}^{t} r^{(ACK)}(v)dv \qquad (t > \tau)$$

$$\gamma^{(TO)}(t) = \begin{cases} r^{(LOSS)}\left(t-4RTT(t)\right) & \text{if } N^{(A)}(t-4RTT(t),t) < 3 \\ 0 & \text{elsewhere} \end{cases}$$

$$\gamma^{(TD)}(t) = \begin{cases} r^{(LOSS)}(t-\tau) \\ \text{with } \tau : N^{(A)}(t-\tau,t) = 3 & \text{if } \tau < 4RTT(t) \\ 0 & \text{elsewhere} \end{cases}$$

$$\frac{dW(t)}{dt} = -\frac{W(t)}{2}\gamma^{(TD)}(t) + \left(1-W(t)\right)\cdot\gamma^{(TO)}(t) + \begin{cases} r^{(ACK)}(t) & \text{if } W(t) < T(t) \\ \dfrac{r^{(ACK)}(t)}{W(t)} & \text{if } W(t) \geq T(t) \end{cases}$$

$$\frac{dT(t)}{dt} = \left(\frac{W(t)}{2} - T(t)\right)\cdot\left(\gamma^{(TD)}(t) + \gamma^{(TO)}(t)\right)$$

$$\frac{dD(t)}{dt} = r^{(ACK)}(t)$$

$$R^{(DATA)}(t) = a(t)\cdot\frac{s^{(P)}\cdot W(t)}{RTT(t)}$$

2.2 RED buffer component

Figure 2 is a schematic representation of a RED buffer component.

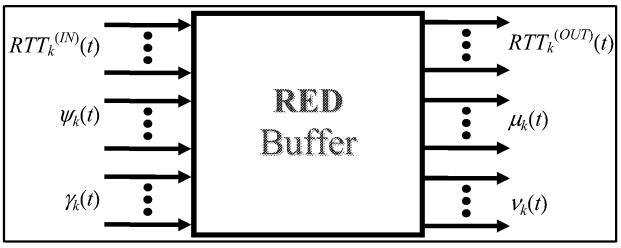


Figure 2: Schematic Representation of RED Buffer Component

The component can be described as follows:

INPUT VARIABLES

- $RTT_k^{(IN)}(t)$: delay in packet transmission from the source to the input of the buffer at the time instant t (expressed in seconds);
- $\psi_k(t)$: arrival rate of packets belonging to group k at the time instant t (expressed in bytes/s);
- $\gamma_k(t)$: loss rate at the input of the buffer for packets belonging to group k at the time instant t (expressed in bytes/s);

OUTPUT VARIABLES

- $RTT_k^{(OUT)}(t)$: delay in packet transmission from the source to the output of the buffer at the time instant t (expressed in seconds);
- $\mu_k(t)$: output rate of packets belonging to group k at the time instant t (expressed in bytes/s);
- $\gamma_k(t)$: loss rate at the input of the buffer for packets belonging to group k at the time instant t (expressed in bytes/s);

PARAMETERS

- $N^{(K)}$: number of groups that pass through the buffer;
- C: link capacity (expressed in bytes/s);
- α: weight in the EWMA filter to calculate the average queue length estimated by RED;
- T_{min}: minimum threshold (expressed in packets);
- T_{max}: maximum threshold (expressed in packets);
- p_{max}: maximum loss probability in RED discarding function;
- $s^{(P)}$: average packet size (expressed in bytes);

STATE VARIABLES

- $q(t)$: queue length of the buffer at the time instant t (expressed in bytes);
- $m(t)$: average queue length estimated by RED at the time instant t (expressed in packets);
- $p(t)$: discard probability at the time instant t;
- $\psi_k(t)$: arrival rate of packets belonging to group k at the time instant t (expressed in bytes/s);
- $\gamma_k(t)$: loss rate at the input of the buffer for packets belonging to group k at the time instant t (expressed in bytes/s);

INTERNAL EQUATIONS

$$\Psi(t) = \sum_{k=1}^{N^{(K)}} \psi_k(t)$$

$$\frac{dq(t)}{dt} = \begin{cases} -C + (1 - p(t)) \cdot \Psi(t) & \text{if } q(t) > 0 \\ \left[-C + (1 - p(t)) \cdot \Psi(t)\right]^{+} & \text{if } q(t) = 0 \end{cases}$$

$$\frac{dm(t)}{dt} = \begin{cases} \ln(1-\alpha) \cdot \left(m(t) - \frac{q(t)}{s^{(P)}} \right) \cdot \Psi(t) & \text{if } q(t) > 0 \\ \ln(1-\alpha) \cdot C \cdot m(t) & \text{if } q(t) = 0 \end{cases}$$

$$p(t) = \begin{cases} 0 & \text{if } m(t) \le t_{\min} \\ \frac{m(t) - t_{\min}}{t_{\max} - t_{\min}} p_{\max} & \text{if } t_{\min} < m(t) < t_{\max} \\ 1 & \text{if } m(t) \ge t_{\max} \end{cases}$$

$$RTT_k^{(OUT)}(t) = RTT_k^{(IN)}(t) + \frac{q(t)}{C}$$

$$\mu_k\left(t + \frac{q(t)}{C}\right) = \frac{\psi_k(t) \cdot (1 - p(t))}{1 + \frac{1}{C}\frac{dq(t)}{dt}}$$

$$\nu_k\left(t + \frac{q(t)}{C}\right) = \gamma_k(t) + \psi_k(t) \cdot p(t)$$

3. GENETIC ALGORITHMS: GENERAL REMARKS AND OUR IMPLEMENTATION

Genetic Algorithms are well-known methods used for finding solutions to optimization problems.

A Genetic Algorithm (GA) operates on a population of individuals called "chromosomes", each representing a candidate solution to a given problem (in our case each possible configuration of network parameters). An iteration of the algorithm, known as a generation, causes the current population to evolve into a new one with the aim of finding increasingly better candidates.

A GA evolves following Darwin's principle of natural selection and fittest survivor theory. Individuals are evaluated according to their *fitness value* indicating how they match the target problem solution. A fitness function has to be specifically defined for the particular problem to be solved. Accurate choice of the fitness function is the first step to obtain a good result, for this reason we will discuss our choice in Section 3.1.

The GA evolution derives from Selection, Reproduction (Copy, Crossover, Mutation) and Elimination operators.

Selection is used to determine which individuals could be used for the following Reproduction operators. According to the roulette rule, the algorithm will choose the individuals in relation to their fitness: the "fitter" a chromosome, the greater its probability of being chosen. The individuals selected for reproduction are called "parents".

The first Reproduction operator is Copy. It only chooses a parent and copies it in the new generation. This technique prevents the elimination of the best individuals which could cause a regression to worse candidates.

Crossover is the main reproduction operator and it exchanges portions of the chromosome of two parents to create two new individuals called "offspring".

The third reproduction operator is Mutation, which allows new parts of a chromosome to be inserted into a parent. It is useful to get out of local minimum or maximum values.

These operators are selected in a probabilistic way. The first way to set these probabilities is to make a balanced selection among the three operators. A second way is to set them by experience. Another way is to perform tuning while the GA is evolving. In our implementation the user can set the parameters freely and/or leave the GA to tune them itself.

A GA has other parameters to be set before it runs, such as the number of generations G and the number of individuals N in each generation. The choice of G is generally performed adopting one of three possible methods. The first method consists in choosing a number based on experience. The second one stops GA when it performs a specified fitness. The last one is an automatic technique: it stop GA when there aren't better chromosome for several generations or fitness changes are very tiny.

We implemented all previous methods. In addition, the GA can be manually stopped by the user, who could launch simulations with the current best parameters in order to see graphically how optimization evolves.

N is generally set as twice the number of parameters P to be tuned: $N = 2 \cdot P$.

In our specific field of application, network parameter optimization, the number of parameters P to be tuned could be some hundreds (as in large backbone networks, for example). So N might be close to one thousand. Since the GA has to run the solver of differential equations $G \cdot N$ times, even considering a low value of G (i.e. $G = 500$), and assuming a very fast solver like the fluid one, the computation time required to tune parameters of an entire network could be too high.

Parallel Genetic Algorithms (Tanese 1989) have been introduced into our tool to break down the computation time for real network analysis.

Nowadays there are various different implementation of this algorithms (Rivera 2001): we chose to implement coarse grained parallelization in our tool. It consists of running the GA simultaneously on some computers in a network, called "islands" or "slaves", and merging their work under the supervision of a "master". The master assigns a part of a population to a slave. Then each slave runs a simple GA exchanging one or more individuals (migration). This method not only reduces optimization time, but also improves algorithm convergence in a number of critical cases.

Coarse grained parallelization needs two other parameters to be set: the migration interval (number of generations between two successive migrations), and migration size (number of individuals to migrate). A high migration rate improves the quality of the solutions, and a high degree of connectivity between slaves increases convergence. For these reasons we choose a high migration rate (every five

generations) and full connectivity. Let us note that this does not affect the efficiency of the tool because the communication time is negligible with respect to the fitness evaluation time.

3.1 The Fitness Function

The choice of the fitness function, $\Phi(\cdot)$, is crucial to improve the convergence of the GA. Let us indicate the generic array of parameters to be tuned (the chromosome) as $\boldsymbol{C}_s$. For each $\boldsymbol{C}_s$, the fluid model solver determines the array $\boldsymbol{X}_s$ that represents the set of network performance metrics (e.g. throughput, delay, loss probability) to be achieved and/or network costs (e.g. link capacities) to be reduced. The fitness function depends on $\boldsymbol{X}_s$.
Usually, the main goal that network parameter optimization has to achieve is minimizing the cost of the equipments needed for the network realization, providing network traffic with QoS guarantees. In order to address this double target, we consider the value of the fitness function $\Phi(\cdot)$ corresponding to the generic array $\boldsymbol{X}_s$, as the sum of two weighted contributions:

$$\Phi(\boldsymbol{X}_s) = K_t \cdot F_t(\boldsymbol{X}_s) + K_c \cdot F_c(\boldsymbol{X}_s)$$

where $F_t(\cdot)$ and $F_c(\cdot)$ are two definite positive functions to be minimized. $F_t(\cdot)$ is intended to lead to the optimum, while $F_c(\cdot)$ is intended to respect the QoS constraints. K_t and K_c are weights which the user can set according to his priorities.
Given that the fitness function is the sum of two positive functions, its minimum value will be zero. So the ideal solution $\boldsymbol{X}_d$ satisfies the following equation:

$$\Phi(\boldsymbol{X}_d) = K_t \cdot F_t(\boldsymbol{X}_d) + K_c \cdot F_c(\boldsymbol{X}_d) = 0$$

We point out that $\boldsymbol{X}_d$ represents the desired solution for the optimization problem, but it may not be achievable, because the parameter set $\boldsymbol{C}_d$ corresponding to $\boldsymbol{X}_d$ may not exist.
The optimum set of parameters $\boldsymbol{C}_o$ is the one that provides an array $\boldsymbol{X}_o$ minimizing $\Phi(\boldsymbol{X}_s)$.
In general within $\boldsymbol{X}_s$ two subsets of elements can be distinguished:

- the subset $\Im^{(T)}$ of $x_i^{(S)} \in \boldsymbol{X}_s$ for which an optimal value as close as possible to the corresponding element in the optimal solution $\boldsymbol{X}_o$ is desired;
- the subset $\Im^{(C)}$ of $x_i^{(S)} \in \boldsymbol{X}_s$ representing the lower or upper thresholds for the corresponding element in the optimal solution $\boldsymbol{X}_o$.

$F_t(\cdot)$ works on the subset $\Im^{(T)}$, and for this reason we chose it as a distance metric able to guarantee that the more $x_i^{(S)} \in \Im^{(T)}$ departs from the desired values, the more the function $F_t(\cdot)$ increases:

$$F_t(\boldsymbol{X}_s) = \sum_{i \in \Im^{(T)}} \left[u_i \cdot \left(\frac{x_i^{(S)} - x_i^{(D)}}{x_i^{(D)}} \right)^2 \right]$$

where u_i are positive constants that allow some variables to be weighted more than others.
The function $F_c(\cdot)$ has the aim of imposing the constraints, so each $x_i^{(S)} \in \Im^{(C)}$ satisfying the corresponding constraint has to contribute with zero to the $F_c(\cdot)$ value, while each constraint not satisfied has to contribute positively to $F_c(\cdot)$. According to these guidelines we chose the following $F_c(\cdot)$ function:

$$F_c(\boldsymbol{X}_s) = \sum_{j \in \Im^{(C)}} \left[w_i \cdot \max\left(0, \frac{\psi\left(x_i^{(S)}, x_i^{(D)}\right)}{x_i^{(D)}} \right) \right]$$

where w_i are positive constants that allow some constraints to be weighted more than others, and

$$\psi(x, y) = \begin{cases} x - y & \text{if good values are} < y \\ y - x & \text{if good values are} \geq y \end{cases}$$

We preferred not to define $F_c(\boldsymbol{X}_s) = +\infty$ for each $\boldsymbol{X}_s$ that does not meet all the constraints, because in this way we would not have had any distinction between chromosomes that do not respect the constraints. This produces some negative effects, because when large networks are considered, the most likely situation for the first generations is that some constraints will not be met. Consequently the GA having the majority of chromosomes with the same fitness value ($+\infty$), cannot discriminate between them; it can only randomly search for good solution with a considerable waste of time. In addition, no one can assure *a priori* that a vector of parameters which can satisfy all the constraints imposed by the user exists; however, with our choice the algorithm gives at least a suboptimal solution.

4. TOOL INTERFACE

The most visible part of the tool is the graphical interface which has an important role: it makes different parts into a unified system. Users can insert various components (routers, links, sources, …) and connect them all together to represent the network topology.
Components have many properties which need to be set by the user. Specifically, for each parameter the tool needs:

- a searching range for the variables to be tuned;
- precision in order to decrease or increase GA accuracy (high precision / slower optimization – low precision / faster optimization);
- the desired values $x_i^{(D)}$;
- the possible type of constraint ("<" or ">")
- the two arrays of weights ($\underline{u}$ and $\underline{w}$)

In addition, before starting GA optimization, the tool needs the number N of individuals for each generation, the maximum number G of generations and possibly the configuration parameters of all the slaves involved in optimization.
After a simulation run or during GA optimization all the variables calculated are stored in files from which the tool makes useful charts in order to explain network performance (see Figure 3).

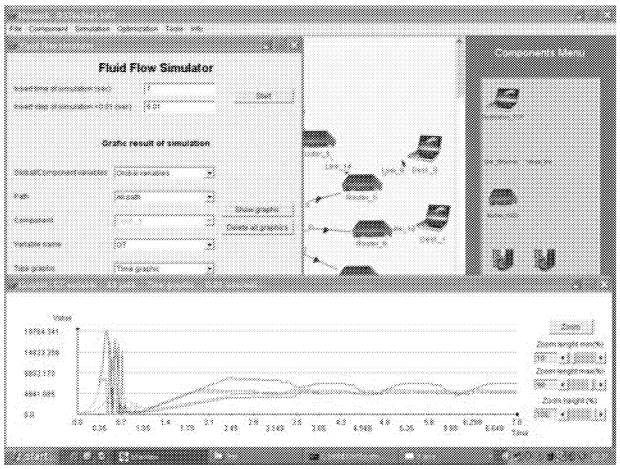


Figure 3: Tool: A Global View

For developers, the tool is realized in such a way as to permit a easier implementation of new components. They have the same characteristics as other component of the same type (e.g. routers) and so developers only add new parameters with simple functions. We also implement a parser to facilitate the implementation of equations and to improve efficiency. Otherwise developers can create new characteristics which are not implemented.

5. NUMERICAL RESULTS

In this section we will discuss tool performance and give a brief presentation of its graphical capabilities.
We first present a case study in which we use NAPT to optimize the parameters of the network shown in Figure 4.
The network is made up of 7 nodes in which 9 RED buffers are deployed, and it is loaded by 6 groups of 10 TCP flows. The propagation delay between each pair of buffers is assumed to be 20 ms, while we consider propagation delays between sources and nodes and between nodes and destinations to be negligible.
The TCP flows belonging to groups S0 and S4 are connected with destination D0, S1 and S3 with D1, and S2 and S3 with D2. TCP packets reach their destination following the path with the minimum number of hops.
Our optimization problem consists of minimizing the link capacities in the network needed to provide TCP sources with a given minimum threshold for the average goodput, that is the actual data successfully transmitted from the senders to the receivers. For the sake of simplicity we consider each TCP group of sources as requesting a 5000 packets/s minimum goodput (the packet size is 1 KB).
The vector $\boldsymbol{C_s}$ of parameters to be tuned (the chromosome) contains the output link capacity and the 4 RED parameters α, T_{min}, T_{max}, p_{max} (see (Floyd and Jacobson 1993) for references) for each RED buffer in the network. So we have 45 parameters to be optimized. The vector $\boldsymbol{X_s}$ corresponding to $\boldsymbol{C_s}$, contains the output link capacities of all the RED buffer in the subset $\Im^{(T)}$, and the goodput obtained for each group of sources in the subset $\Im^{(C)}$.
Regarding the fitness function parameters, we set all the elements of the arrays $\underline{\boldsymbol{u}}$ and $\underline{\boldsymbol{w}}$ equal to 1, and, in order to force the GA to generate solutions satisfying all the QoS constraints, we set $K_t = 1$ and $K_c = 100$.
To solve this optimization problem we ran the genetic algorithm on three common PCs connected in a classic Fast Ethernet network. Their main characteristics are in Table 1:

Table 1: Main Machine Characteristics

Number	Characteristics
1	AMD Athlon64 3000+, 512 MB RAM
2	INTEL Pentium4 2600MHz, 512MB RAM
3	AMD AthlonXP 2600+, 256 MB RAM

After about one hour, NAPT provided the results listed in Table 2.

Table 2: Buffer Tuned Parameters and Link Capacities.

Buffer	α	T_{min}	T_{max}	p_{max}	Capacity (pack/sec)	Ideal Capacity (pack/sec)	Over provisioning percentage
A	0,0043	187	386	0,1647	5016	5000	0,32 %
B	0,0087	145	495	0,2494	5004	5000	0,08 %
C	0,0020	107	228	0,0365	5023	5000	0,46 %
D	0,0062	29	386	0,0141	5002	5000	0,04 %
E	0,0087	32	412	0,2847	20533	20000	2,67 %
F	0,0043	129	490	0,0824	5000	5000	0,00 %
G1	0,0076	50	443	0,1294	10009	10000	0,09 %
G2	0,0051	179	405	0,0612	10267	10000	2,67 %
G3	0,0044	168	373	0,0400	10002	10000	0,02 %

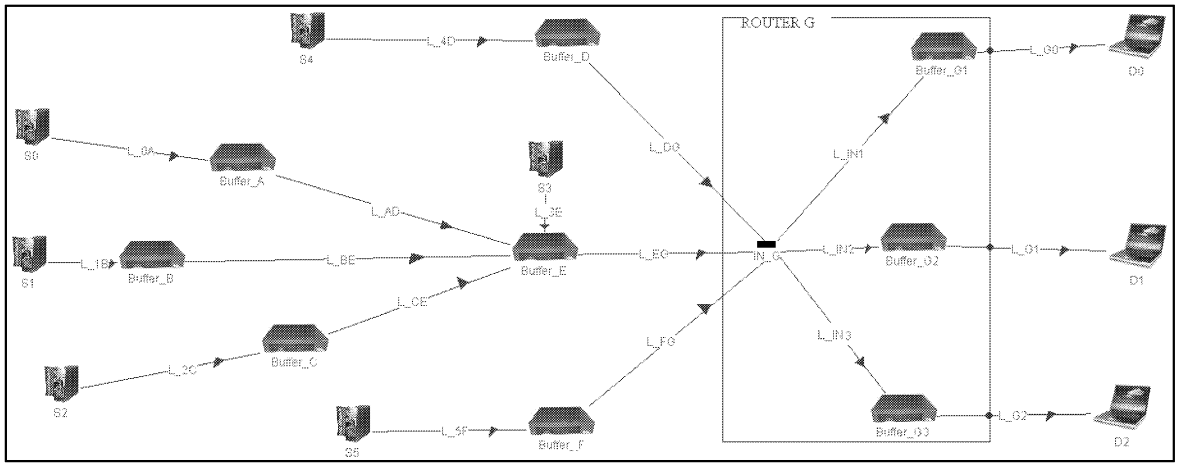


Figure 4: Optimization Problem Network Topology

We stopped execution of the genetic algorithm when all the QoS constraints were satisfied and the suggested link capacities were at most 3% greater than their ideal values. Let us note that if a 10% over-provisioning has been allowed, in less than 8 minutes we would have obtained a solution in which all the QoS constraints were met.
The tool can obtain information about suboptimal solutions because, during the evolution of the genetic algorithm, it always saves the last best solution of the optimization problem (that is, the solution with the lowest fitness function value). It is possible at any moment to analyze the temporary better solution by plotting the temporal variation or the average value of all the variables characterizing the network topology being addressed (queue length, TCP goodput, TCP window length, …).
As an example in Figure 5 we show a bar plot in which the average goodput values are represented for each group of flows. More specifically, on the right hand side of the plot there are the average values evaluated only considering the second half of the simulation interval. Since they do not take into account the initial transient behavior of the network, they are a better estimate of the steady state average goodput.
These considerations are confirmed by the temporal diagram provided in Figure 6, in which we observe a stationary goodput behavior for all the groups of flows after about 4 seconds. Figure 7 shows a zoom of the second half of the temporal evolution in Figure 6.
Figure 5 shows that all the constraints on minimum goodput values are satisfied. Only the goodput assigned to the source group S3 is significantly greater than the others. The reason for this is that flows belonging to the group S3 have a shorter round trip time than the others, so it can "grab" more bandwidth in buffer 4 than the other three groups. For this reason buffer 4 and 6 need 2,67% more bandwidth than the ideal one in order to respect the minimum requirements for the other groups.

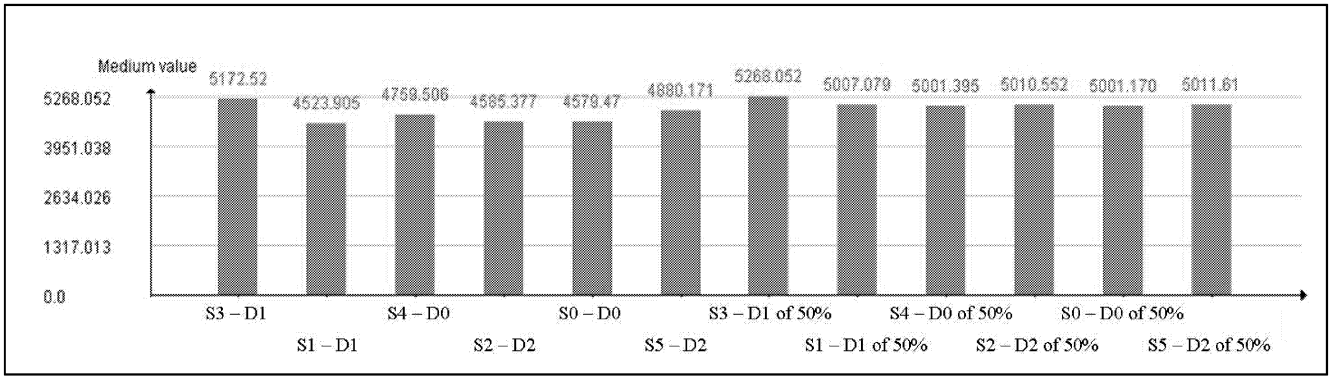


Figure 5: Bar plot: Average Values of Goodput for all Group of Sources

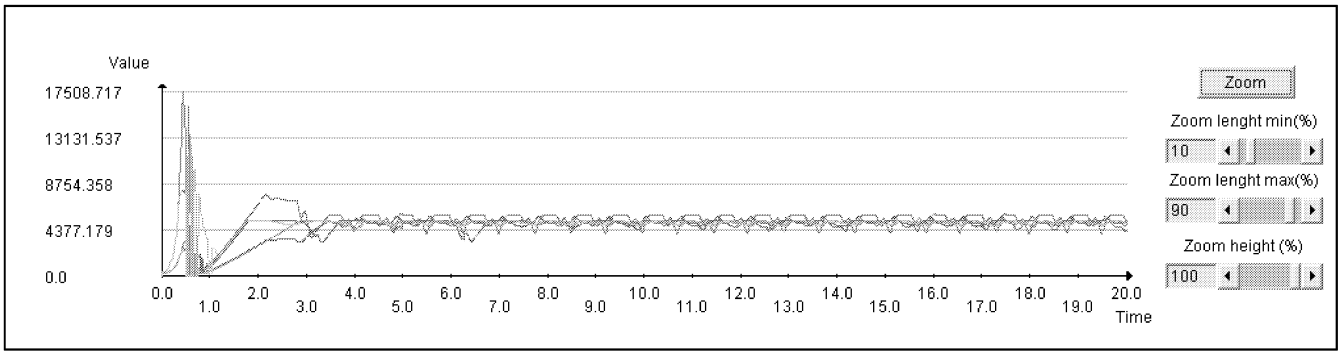


Figure 6: Temporal Diagram: Temporal Evolution of Goodput for all Group of Sources.

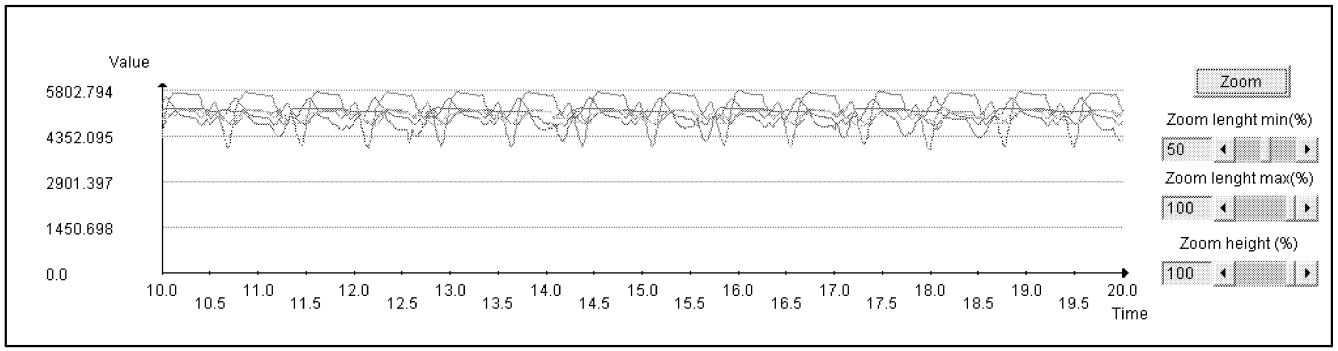


Figure 7: Temporal Diagram: Zoom of the Temporal Evolution of Goodput for all Group of Sources from 50% to 100% of Simulation Time.

NAPT is able to handle large networks. We tested it on a topology made up of over 150 elements (12 node with 44 buffers, and over 140 group of sources), but we do not present the results here for reason of space. When we used the set of PCs in Table 2, the genetic algorithm took about 15 hours to estimate the 220 optimized parameters of the network. We subsequently repeated the optimization using a larger computer network, with 12 Pentium IV processors: in this case the time required for optimization was about 4 hours.

The relatively low optimization time values were achieved thanks to the fluid flow approach to describing the temporal evolution of the network. In order to better understand why a packet-by-packet simulator cannot be used to solve optimization problems, in Table 3 we compare the processing time needed by NAPT and that required by NS2, to find network performance in three different scenarios for a single set of parameter values (chromosome). We observe that with not excessively large networks the NS2 processing time is about 10 times greater than that needed by NAPT, but for large networks the ratio between the two processing times increases considerably. This makes the use of a packet-by-packet simulator impractical for network parameter optimization.

Table 3: Processing Time: Comparison between NAPT and NS2.

System Description		Processing Time (sec.)	
# Buffers	# Group of Sources	NAPT	NS2
6	4	3,02	35
9	6	3,23	46
44	132	16,66	413

6. CONCLUSIONS

In this paper we have presented a user-friendly, easily expandable and time-saving tool, which is able to perform IP networks parameter tuning by simulating network behavior through Fluid-Flow models and adopting Parallel Genetic Algorithms as optimization methodology. The presented tool, called NAPT, has an high degree of scalability and modularity. We have assessed NAPT in different network scenarios, and we have shown its ability in finding good solutions for network optimization problem in a lower time than packet-by-packet simulators. These characteristics make NAPT a useful tool for network provider in planning and dimensioning IP networks.

REFERENCES

Barbera, M.; A. Lombardo; G. Schembra. 2004. "A Fluid-Based Model of Time-Limited TCP Flows." *Computer Networks Journal* (Elsevier), Vol. 44, No.3 (Feb.), 275-288.

Barbera, M.; A. Lombardo; G. Schembra; C. A. Trecarichi. 2004. "A Fluid-Based Model of Policed RIO Router Networks Loaded by Time-Limited TCP Flows", In *Proceedings of SPECTS 2004* (San Jose, California, U.S.A, July 25-29)

Barbera, M.; A. Lombardo; G. Schembra; C. A. Trecarichi. 2005. "An Analytical Framework to Design a DiffServ Network Supporting EF-, AF- and BE-PHBs". In *Proceedings of QoSIP 2005* (Catania, Italy, February 2-4).

Floyd, S. and V. Jacobson. 1993. "Random Early Detection Gateways for Congestion Avoidance." *IEEE/ACM Transactions on Networking*, Vol. 1, No.4 (Aug), 397-413.

Misra, V.; W.B. Gong; D. Towsley. 2000. "Fluid-based analysis of a network of AQM routers supporting TCP flows with an application to RED". In *Proceedings of SIGCOMM 2000* (Stockholm, Sweden, August 28 – September 1).

Rivera, W. 2001. "Scalable Parallel Genetic Algorithms." *Artificial Intelligence Review* 16: 153-168, Kluwer Academic Publishers.

Shridharan, M.; S. Chellappan; A. Durresi; A. Ozbay; R. Jain. 2003. "Tuning RED in Satellite Network using Control Theory". In *Proceedings of Internet Quality of Service* (Orlando, Florida, September 7-11), SPIE Vol. 5245.

Tanese, R. 1989. "Distributed genetic algorithms", In *Proceedings of 3rd Int. Conf. Genetic Algorithms* (San Monteo, USA), pp 434-439.

Ye, T. and S. Kalyanaraman. 2002. "Adaptive Tuning of RED Using On-line Simulation". In *Proceedings of IEEE GLOBECOM 2002* (Taipei, Taiwan, November 17-21).

Zhou, J.; J. Zhengrong; M. Takai; R. Bagrodia. 2004. "MAYA: Integrating Hybrid Network Modeling to the Physical World." *ACM Transactions on Modeling and Computer Simulation*, Vol. 14, No.2 (Apr.), 149-169.

[URL: RED] http://www.cisco.com/en/US/products/hw/routers/ps167/products_tech_note09186a00801e1dc1.shtml

[URL: NS2] http://www.isi.edu/nsnam/ns/

[URL: NAPT] http://www.diit.unict.it/arti/Tools/NAPT/

OPTIMIZERS WITH LEARNING DISTRIBUTIONS IN SIMULATION MODELS: AN APPLICATION TO ALL-PASS DISPERSION COMPENSATOR OF OPTIC FILTERS

Mufit Ozden[1] and Chi-Hao Cheng[2]
Department of Computer Science and Systems Analysis[1]
Department of Electrical and Computer Engineering[2]
Miami University, Oxford, OH 45056
ozdenm@muohio.edu

KEY WORDS
Meta-heuristic Optimization, Dispersion compensator in fiber optics

ABSTRCT

We apply the multiple optimizers with the learning-distributions to continuous decisions in a simulation model as integral decision-making components. In this machine learning technique, the optimizers are given performance evaluations estimated for their decisions during short simulation runs. With the feedback, the optimizers modify their decision-making distributions after the decisions with good performance judged against a threshold value and they generate new decisions for the next simulation run. With the gradually advancing threshold value, the optimizers learn the structural components of the optimal solution. Because of the probabilistic nature of their decision-making processes, the optimizers have proved to be very robust in the random environments with the noisy responses. The development of this machine learning technique and the convergence properties have been delineated in an earlier paper in detail. In this paper, we present an engineering application to a fiber optic communication problem after brief explanation of the optimizers.

INTRODUCTION

We can often consider a simulation model as a complex random system in which a number decision variables or system parameters are to be set to their "optimal" values after obtaining the performance measures through a series of simulation runs. In simulation models, the optimization of the decisions is usually based on a meta-heuristic using a search technique with a machine learning capability, such as Evolutionary Algorithms (Michalewicz 1994), Simulation Annealing (Gutjahr and Pflug, 1996), and Tabu Search (Glover et al. 1996). In this paper, we propose a new machine learning technique called "optimizers with learning distributions" as the decision-making/learning components of simulation models. These optimizers employ an adaptive probabilistic search in the continuous domain guided by the performance evaluations and they learn the pattern of good decisions in their selection distributions. This learning technique is inspired from of the learning automata theory (LA) (Narendra and Thathachar 1989). In formal mathematical terms, the decision making problem in simulated environments can be stated as follows:

$$\min_{\theta \in \Theta} J(\theta), \quad \text{where} \qquad (1)$$
$$J(\theta) = E[L(\theta,\omega)], \quad \forall \theta \in \Theta,$$

and Θ is a continuous decision interval, L is a real-valued performance function, which depends on the design $\theta \in \Theta$ and a random variable ω, and $J(\theta)$ is the expected value of this performance function. Typically, the random performance function in (1) is not known explicitly and can only be estimated with running a simulation program. Formally, the optimizer learning technique can be characterized with the triple $\{\theta, \beta, A\}$ for designs θ, performance estimation β, and a learning algorithm A. The search for the optimal decisions is conducted with the decision selection distribution P defined on the decision set θ. Initially, the distribution is usually constructed as a uniform distribution giving equal probability to every decision value. When an optimizer selects a decision θ_i randomly and implements it in a random simulation model, it receives a (normalized) performance estimation β_i from the model after a short run. On the basis of this estimation, the automaton modifies its decision selection distribution according to the learning algorithm A. We adapted a particular learning algorithm known as the "reward and inaction" (a

L_{R-I} automaton in the nomenclature of the LA theory) in conjunction with a new moving threshold level to simulation models.

In our implementation of optimizers to the continuous decision domains, every continuous decision component is first scaled over [0,1) interval and each digit after the decimal-point is represented by a discrete optimizer that has a selection distribution defined on 10 discrete values (0,1,...,9) to the desired level of accuracy. Hence, several optimizers represent a continuous decision component. The simulation model provides a binary feedback (success or failure) for the randomly selected decisions after comparing to a current threshold value. When found successful, the optimizers modify their respective selection-distributions as follows:
At the simulation run n,

if β_i is *1*,

$$p_i(n+1) = p_i(n) + a\,[1 - p_i(n)]$$

$$p_j(n+1) = (1 - a)\,p_j(n) \qquad \forall j \text{ except } i$$

otherwise, $p_j(n+1) = p_j(n)$, $\qquad \forall j \qquad (2)$

where a is small positive learning constant and $P(n)$ is a distribution vector whose i^{th} component is the selection-distribution $p_i(n)$ for the optimizer i. Formally, the process $\{P(n) : n \geq 0\}$ describes a Markov process. The long-term convergence properties of the learning algorithm can be derived using this Markov property. The development and implementations to the simulation models of various forms of this machine learning technique can be found in (Ozden 1994 and 2000) for the discrete decision domain and in (Ozden 2005) for the continuous domain.

Considering a minimization problem, we define a decision set θ as successful if

$$J(\theta) < T, \quad \forall \theta \in \Theta \qquad (3)$$

The estimation of performance for a selected decision set at the end of a run may deviate from the expected value due to the random nature of a simulation model. Therefore, sometimes it is possible to reward a mediocre decision set erroneously. However, the probabilistic nature of the decision-making is expected to correct these mistakes in the long run. Hence, if the estimated performance of a decision set θ_i is less than the threshold value T, the value β is set to *1;* otherwise to *0*. Then, the optimizers update the selection distributions according to the formula (2). If the optimizers achieve a certain rate of success, with the current threshold value after a fixed number of runs, the threshold value itself is adjusted to a lower level by

$$T \leftarrow \gamma T, \text{ where } 0 < \gamma < 1. \qquad (4)$$

When the optimizers meet a certain convergence criterion, the search is terminated. For some simulation problems, it is possible to continue until some decision set achieves very high probability close to 1 as the optimal solution. However, this may require many expensive simulation runs and therefore it may not be feasible for large simulation problems. In this case, a less stringent convergence criterion may be adopted. For example, the probability of any decision set reaches a high level, such as 0.90. Alternately, the termination criteria may be based on the success rate of the decisions or the quality of the best solution discovered.

In the learning automata theory, the learning process of a L_{R-I} in determining the best single action among a group of distinct competing actions is known to be absolutely expedient. Here, we will state the extension of this property to the optimizing components in the continuous domain without going into the formal proof. The learning algorithm described above is found to be absolutely expedient as stated in the following theorem and the formal proof can be found in (Ozden 2005):

Theorem : The optimizers with the learning distributions increase the expected probability of generating a successful solution strictly in any two consecutive runs.

APPLICATIONS OF THE OPTIMIZERS WITH LEARNING DISTRIBUTIONS

We have applied the probabilistic optimizers to a number of random benchmark problems also used in the global optimization and the genetic algorithm. Table 1 summarizes the results for some of these problems in which the parameter r indicates the percentage noise uniformly distributed around the actual performance value. In this paper, we will present another application to a serious engineering problem.

All-pass Filter Dispersion Compensation Problem

During optical signal transmission in optical fiber, signal components at different wavelengths

(frequencies) experience different group delay; therefore, the transmitted optical pulse is broadened as illustrated in Figure 1. This phenomenon is referred as chromatic dispersion, and it could increase bit error rate in optical communication systems. The definition of chromatic dispersion is differentiation of group delay with respect to wavelength, and the unit commonly used is pico-seconds per nano-meter (ps/nm). (Since wavelength equals to the speed of light divided by frequency, it is also possible to represent dispersion with unit: ps/GHz.) With advances in optical communication systems and components such as low insertion loss fiber, optical fiber amplifier, and etc, dispersion compensation has become one of most important remaining issues in optical communications.

Table 1. Solutions to benchmark random functions of 20 variables (optimal objective values are 0.0.)

Function and randomness r	Objective (Avg/S.D)	Funct. Eval. (in 1000)
Sphere	0.0048	1
0.25	0.008	5
Schwefel	0.0016	50
0.25	0.0021	20
Rastrigin	0.0254	25
0.25	0.016	10
Griewank	0.012	150
0.1	0.019	10
Rosenbrock	0.076	200
0.25	0.026	20

The optical communications system tolerance for chromatic dispersion is inversely proportional to the square of system transmission rate. For example, while the 10Gbit/sec channel can tolerate dispersion of 1200 ps/nm, the 40Gbit/sec channel can only tolerate dispersion of 75 ps/nm. Therefore, the chromatic dispersion is an extremely important problem when a high transmission-rate optical communication system is designed.

To compensate chromatic dispersion, dispersion compensation fiber (a special fiber with dispersion opposite of standard fiber dispersion) is offered as the most popular solution. Dispersion compensation fiber is a reliable and mature technology. However, it has some serious disadvantages, such as large footprint, high insertion loss, causing nonlinear distortion when input signal power is high, and others (Chen et al. 2003). Alternatively, several dispersion compensation schemes have been proposed, and the one based on all-pass filter is probably the most popular alternative dispersion compensator (Chen et al. 2003, Madsen et al. 2000, Madsen et al. 1999, and Madsen and Zhao 1999). It has advantages such as compact size, low insertion, no non-linearity under high input power, and so on.

All-pass filter is a special filter with flat magnitude spectrum and non-linear phase spectrum, so it can compensate phase distortion without affecting magnitude spectrum of signals (Madsen and Zhao 1999). Phase distortion is referred as chromatic dispersion in optical communications. All-pass filter has been widely used in different areas. One of such filters used in optical communication is an etalon with partial reflective coating on the front surface and high reflective coating (R~100%) on back surface as illustrated in Figure 2. This kind of etalon is also referred as the Gires-Tournois etalon (Hsing et al.2003). The group delay spectrum of the Gires-Tournois etalon can be represented as:

$$GD(f) = \frac{(1-R)(2d/c)}{1+R+2\sqrt{R}\cos(2\pi f(2d/c))}, \qquad (5)$$

where R is the reflectance of etalon front mirror, c is the speed of light, and d is the etalon optical thickness.

By cascading several etalons, each of which has carefully designed coating and thickness, we can increase the dispersion compensation bandwidth and amount of dispersion. Fig.3 shows the simulated group delay curve of a 50GHz-channel spacing single-cavity etalon with R=52% on the front surface and R=100% on the back surface. Fig. 4 shows a desired simulated group delay of several etalons with different front surface coatings and thickness that compensates the chromatic dispersion of a transmission fiber. In Figure 3 and 4, since it is easier to design all-pass filters with frequency unit such as GHz, the group delay is shown as a function of frequency.

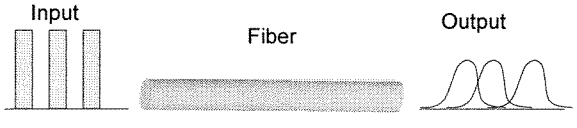


Figure 1. Dispersion effects in optical transmission fiber.

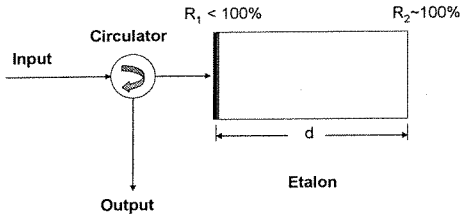


Figure 2. All-pass Etalon

Due to its periodic spectral properties, etalon-based dispersion compensator can be used for multi-channel dispersion compensation. Now, the decision problem can be stated as follows: using a set of etalons whose reflectance and cavity depths are to be determined optimally to form a group delay line with the negative slope of the fiber's over a range of frequencies in the transmission bandwidth.

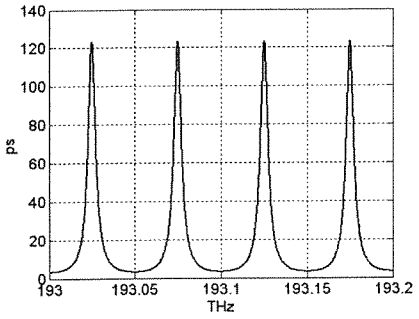


Figure 3: Simulated group delay spectrum of a single-cavity all-pass etalon

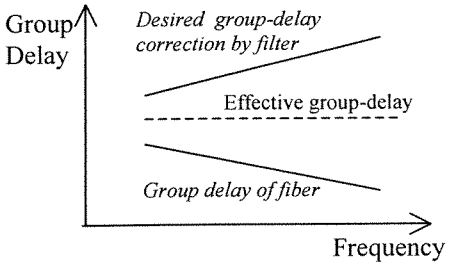


Figure 4: Group delay compensation with a All–pass filter.

The performance evaluation of a solution requires the evaluation and addition of the group delay functions (5) for the etalons at the frequencies considered in the bandwidth and determination of the slope of the fitted line through these points. Therefore, each performance evaluation amounts to a lengthy calculation. In a more realistic version of the problem, the temperature variations (a random process) also affect the group delays, but this problem is still under formulation and it will not be covered here.

Solution to All-pass filter Problem

In our problem, the objective of the problem is to construct a correction-dispersion line with the group-delay slope of 10.452 ps/GHz that cancels out the transmission fiber dispersion that is assumed to occur with same magnitude but in opposite sign over the bandwidth of 193.085-193.115 THz. For the purpose of this study, we use 10 etalons whose reflectance coefficients in the range of (0-1) and etalon depths in the interval of [0.002997536499, 0.002998312761] in meter need to be determined optimally.

This problem has been solved through our machine learning technique described above in a considerably short execution time. The objective value is taken as the minimization of the maximum deviation from the dispersion line with the desired slope and the 20 decision variables take values from their respective ranges. The final solutions of the reflectance and depths of the etalons are given in Table 2. Fig. 5 shows the group delay of the 10 etalons individually as constructed by our procedure. Fig 6 shows the combined group delay of the etalons with the linear regression attaining a slope of 10.48041 ps/GHz, which is very close to the ideal slope of 10.452 ps/GHz.

Table 2. Solutions obtained for (R,d) values.

R1=	0.449068191
d1=	0.002997987
R2=	0.16164714
d2=	0.002998149
R3=	0.035977803
d3=	0.002998291
R4=	0.110527567
d4=	0.002998234
R5=	0.235305351
d5=	0.002998044
R6=	0.139634572
d6=	0.002998048
R7=	0.038746781
d7=	0.002998292
R8=	0.153492347
d8=	0.00299809
R9=	0.084219934
d9=	0.002998308
R10=	0.233242691
d10=	0.002997665

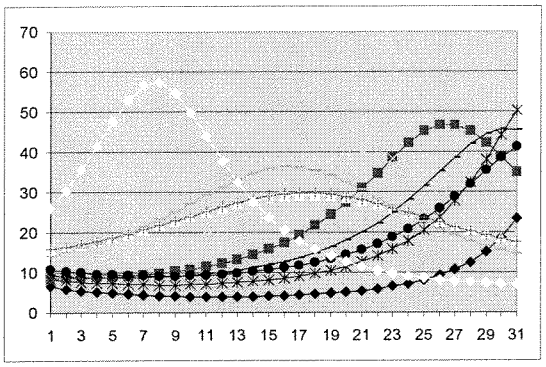


Figure 5: The group delay of the individual etalons for 31 points with 1GHz separation in the bandwidth.

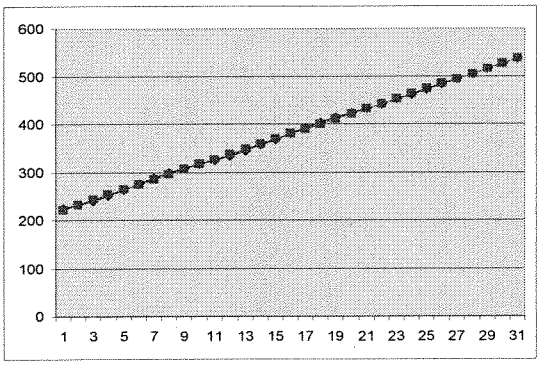


Figure 6: Combined group delay of all etalons within the selected bandwidth.

CONCLUSION

The probabilistic learning approach of this paper found very good solution to all-pass filter problem of the fiber optics as it has with the benchmark mathematical problems. For the future research, we will consider a more general version of this problem where the temperature changes are possible over a certain range. Due to the temperature changes, the thickness of the etalons will also change distorting the aggregate dispersion compensation performance of the all-pass filter device. We are planning to use the different expansion coefficients as the addition decision variables for the etalons. Hence, our goal is to design a self-tuning device for the chromatic dispersion in the fiber optics.

REFERENCES

Hsieh, Chao-Hsing, Ruibo Wang, Zhiqing James Wen, Ian McMichael, Pochi Yeh, Chao-Wei Lee, and Wood-Hi Cheng. 2003. "Flat-Top Interleavers Using Two Gires-Tournois Etalons as Phase-Dispersive Mirrors in a Michelson Interferometer," *IEEE Photonics Technology Letters* 15, No. 2 (Feb), 242-244.

Chen, Chongyang, David Z. Chen, Daniel L. Peterson, Jr., Chi-Hao Cheng, Sihan Lin, Hui Fang, Tiejun Xia, Kuangyi Wu, and Gan Zhou, 2003. "A Newly Developed Dispersion Compensation Module for ULH/Metro Applications," CIIT 2003.

Glover, F., Kelly, J. P., Laguna, M. 1996. "New advances and applications of combining simulation and optimization." *Proceedings of 28th Conference on Winter Simulation.*

Gutjahr,W.J.,Pflug,G.Ch, 1996. "Simulated annealing for noisy cost functions." *Journal of Global Optimization* 8, 1-13.

Madsen, C. K. , J. A. Walker, J. E. Ford, K. W. Goossen, T. N. Nielsen, and G. Lenz. 2000. "A tunable dispersion compensating MEMS all-pass filter," *IEEE Photonics Technology Letters* 12 , No. 6 (June) , 651 – 653.

Madsen, C. K. G. Lenz, A. J. Bruce, M. A. Cappuzzo, L. T. Gomez, R. E. Scotti. 1999. "Integrated all-pass filters for tunable dispersion and dispersion slope compensation." *IEEE Photonics Technology Letters* 11 , No. 12 (Dec), 1623-1625.

Madsen C. K. and J. Zhao. 1999. *Optical Filter Design and Analysis.* John Willey and Sons, New York.

Michalewicz, Z. 1994. *Genetic Algorithms + Data Structures = Evolution Programs*, 2nd Ed, Springer-Verlag.

Narendra, K., Thathachar, M. A. L. 1989. *Learning automata an Introduction*, Prentice Hall.

Ozden, M. 1994. "Intelligent objects in simulation." *ORSA Journal on Computing* 6, No 4, 329-342.

Ozden, M., 2000, "A Probabilistic Solution Generator For Simulation." in review with *European Journal Operational Research* 146, 35-51.

Ozden, M. 2005"A Probabilistic Search Method for Global Optimization of Continuous Functions" submitted for publication in *INFORMS Journal of Computing.*

MANAGEMENT SIMULATION TOOLS

BLENDED LEARNING APPLYING INTERACTIVE DISCRETE EVENT SIMULATION IN A WEB-BASED LEARNING MANAGEMENT SYSTEM

Wolfgang Kühn
Department of Electrical, Information and Media Engineering
University of Wuppertal
D – 42119 Wuppertal
wkuehn@uni-wuppertal.de

KEYWORDS

E-Learning, Blended learning, Discrete Event Simulation, Learning Management System

ABSTRACT

The approaches of integrating interactive discrete event simulation with E-learning technologies offer powerful features for modern blended learning structures. This type of experience based knowledge transfer processes is developed in order to improve learning processes in university und industries. The Learn Management System handles the static information, knowledge transfer and the communication regarding the teaching and learning process. The interactive discrete event simulation offers dynamic experience based learning. The technical integration of the Learn Management System with discrete event simulation is based on XML and Java technology. This approach is platform-independent. The simulation process on the simulation server can be initiated and controlled by use of XML-RPC technology from a java simulation applet.

INTRODUCTION

The use of simulation and e-learning technologies gains importance in the modern information and communication society. New types of knowledge transfer processes are developed in order to support structure changes in society and university. Electronic media offer flexible approaches for the preparation and presentation of the knowledge and to improve the knowledge transfer process in university teaching. Blended learning strategies, which combine face to face events with learning modules independent from location and time, self controlled, and learning more from experiences instead of reproduction of static knowledge will gain in future considerably importance. Even the international competitiveness of universities will force advanced teaching methods.

In industries and consulting modern simulation techniques are applied increasingly in the field of production and technology management in order to analyze and optimize production and logistic flows. The dynamic of these complex systems can be demonstrated very clear to various demands on different user levels. Due to future demands the experience with these simulation techniques is increasingly important for students and should be considered in teaching. Previous experience shows that interactive simulation is suitable to improve teaching in the field of complex system dynamics.

Information technology is going to be one of the major venues through which people will be learning in the future. Training and education that involves technology will need to be flexible as far as when and where it is taught, and the methods used. Modern university teaching is required, to develop and introduce state of the art information and communication technology in order to encourage self-dependant and self organized individual learning strategies. Net based learning approaches with individual learning speed independent of location and time can extend traditional classroom teaching.

Blended learning is learning which uses a combination of face to face and online media. This form of learning will be increasingly used in future, and the presence of both types of learning distinguish blended from wholly online and wholly classroom methods of learning. The project introduced in this paper combines face to face presence events with individual network and simulation based learning units. Blended learning strategy offers some kind of flexibility required for future demands. Therefore complex discrete event simulation running in a distributed soft and hardware environment has been integrated in a web based LMS (Learning Management System). Results of this innovative approach may be interesting for other fields also.

BLENDED LEARNING

A blend is an integrated strategy for delivering on promises about learning and performance. Blended learning involves a planned combination of approaches, such as participation in lectures and seminars, workshops, coaching by a supervisor, individual studies at home; collegial relationships and online communities.

Blended learning is a fairly new term in education, but in principle the concept has been around for a long time already. Essentially, blended learning is defined as a method of educating at a distance that can take

advantage of technology combined with traditional education or training. However due to the rapid development of computer and internet technology blended learning strategy will gain importance in modern teaching practice.

Options for blended learning go beyond the classroom. They're formal and informal, technology- and people-based, independent and convivial, and directive and discovery oriented. Students can seek the information that they need, when they need it. Therefore students are assisted through workshops, coaching sessions, online classes and performance support tools. The figure below presents some possibilities of what can constitute a blended learning approach:

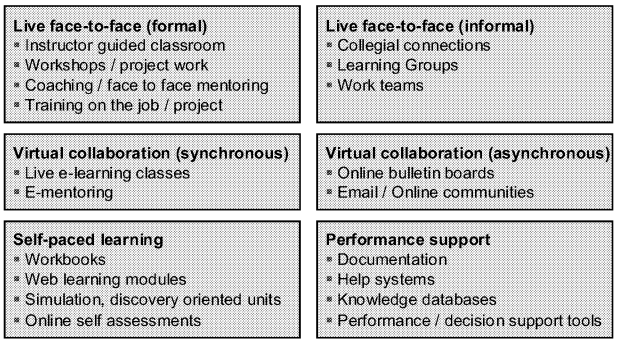


Fig. 1: Elements of Blended Learning Scenarios

Blended learning has to rivet attention on how to combine resources to achieve a strategic purpose. Therefore, staying focused on strategy is crucial to blending success. Other key factors include

Independence: Modern technology enables students to use delivered information through self study or teaching in groups. They can study on their own online while they additionally participate in online communities, phone conferences or face to face meetings.

Communication: Clear links between the various learning options, associated with a program on virtual teaming are important elements for the learning success.

Flexible options: Blending enables people to get an answer, regardless of the location, time, and learning preferences. This has positive ramifications for students who would like to access to information immediately, rather than after a scheduled event or an expert available for consultation.

Work cross functionally: Blended learning involves working with and convincing people across the organization. It is important that all those involved understand their roles and the reasons for multifaceted approaches.

Redundancy: Receiving the same and elaborated messages from several sources in various formats over time improves the learning process of the participants. Redundancy is part of any good blend.

Assets and guidance: Learners don't always know when they need additional instruction or when they're ready to test their skills. Therefore effective blends must include guidance. Direction can appear as sample paths, diagnostics, and recommendations targeted to roles, tasks, priorities and progress.

Blended learning takes advantage of the power of technology to deliver training just in time, anywhere and anytime. However, in blended learning, the technology of online education is not used as an isolated tool, but as a key part of a comprehensive workplace performance solution.

OBJECTIVE OF THE PROJECT

The introduced project is part of the initiative 'e-teaching@university – best practise' and implements a blended learning concept with simulation based experience learning in the area of production and technology management. Goal of the project is to set up advanced blended learning scenarios by use of integrating complex event oriented simulation in a web based learning management system. This encourages discovery oriented learning in the area of dynamic production systems with individual feedback und learning speed. The interactive discrete event simulation uses the potential of modern information technology, in order to improve the teaching in the field of dynamic production systems, production logistic etc.

The interactive discrete event simulation is integrated with the Learn Management System (LMS). This approach allows students to access conventional E-learning units and simulation based experience learning units from the department laboratory in the university as well as via Internet from home or an internet café anywhere. For the access a conventional Web-Browser is sufficient. All important data are stored in the learn management system or simulation database.

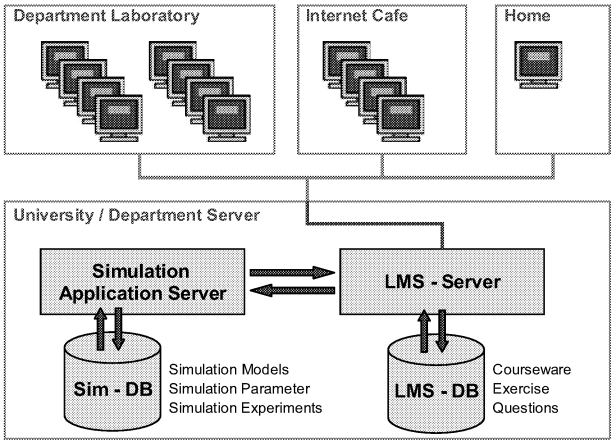


Fig. 2: Integrated LMS / Simulation Environment

The applied discrete event simulation contains event controlled simulation runs through all production and logistic processes. Therefore simulation is not the animation of dynamic problems only. Each run may contain thousands of simulation events. Numerous parameter of the production system have to be considered in real time during the run. The simulation server in this project uses Java-Sim, an advanced Java based simulation library, developed at the University of Wuppertal.

The integration of simulation in the web based learning management system enables simulation features via the internet. The user interface offers students the following possibilities:

- Parameter settings of simulation models without programming knowledge
- Simulation runs of production processes
- Interactive test of various scenarios
- Analysis of simulation results

Face to face presence events are complement reasonably with extended learning units, which are build up with integration of interactive simulation features. The interactive simulation offers students individual learning experience regarding the dynamic of complex production systems. Simulation shows interrelations in complex systems very illustrative and efficiently. The learning units are linked into a blended learning scenario, which transforms the previous presence course to a combination of presence lesson with self experienced and self controlled online learning units.

The applied concept enables students to use an interactive production simulation approach, web based and independent from location and time. It is set up for the first step in the master course 'System Optimization / Simulation Techniques' which is part of the module 'Production and Technology Management' in the master studies 'Print and Media Technology' and 'Information Technology'. In the long term the results shall be transferred even to other fields or education areas.

DIDACTIC CONCEPT

The aim of this project is an improvement of the teaching concept in the technical field of production management. The didactic concept of the introduced project combines presence events and e-learning elements reasonably to a blended learning scenario. Basically the chosen blended learning scenario contains the following components:

- **Presence events** provide the basic knowledge in classroom teaching for example by use of conventional PowerPoint presentations.
- **Self study units** uses the learning management system to deliver detailed information for self studies. These studies, individual for each student, are independent from location and time. The learning speed can be adapted to the specific demand of each particular user**.**
- **Discovery oriented study units** uses simulation technology to understand and gain experience of dynamic system behavior. These discovery oriented studies allow performing individual simulation runs for each student or student group independent from location and time.
- **Learning groups** gather knowledge in project work. Simulation experiments can be performed in small groups.
- **Self assessments components** are integrated in self study units and discovery oriented study units also.
- **Open communication**: the knowledge obtained can be exchange by use of open communication, such as email, interest groups, chats etc.

The self learning sections complement the presence events by use of interactive electronic media which support the experience learning and improve the understanding of complex system dynamics. The interactive simulation approach is supposed to offer a new media approach in the field, self learning lesson and interactive classroom lessens as well.

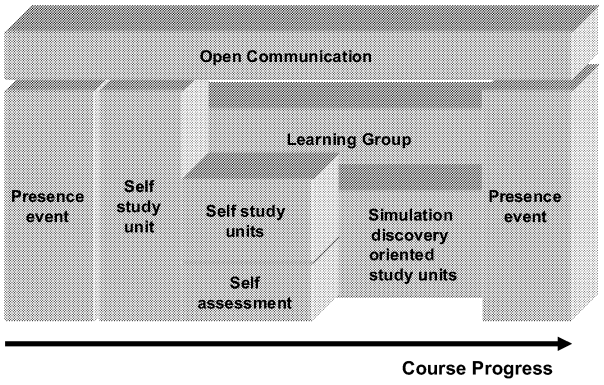


Fig. 3: Building Blocks for Blended Learning

The didactic concept shall be shown exemplary for a learning unit, 'Optimization of Queuing Systems', which is part of master course 'Systems Optimization / Simulation Techniques' in the module 'Production and Technology Management'. The blended learning scenario contains elements, such as:

- With the method of classroom teaching the basics of queuing theory are taught in presence events. These basics are explained by the faculty using classical presentation media, text, equations, drawings etc.
- In corresponding individual learning modules implemented in the LMS offer to go again through

the basic facts and processes. Interactive media containing a guided task and questions for individual progress control enable students to perform these modules individually at any time and location.

- Concerning the complex system dynamic of production systems an individual progress control is performed by use of interactive simulation models. The intention of this approach is less to simply reproduce knowledge than rather to build up empirical knowledge. For this purpose preconfigured models are prepared in advance. These do not meet a priori the requirements for the queuing problematic. Performing simulation runs with the preconfigured models the students can recognize clearly that the requirements are not met and what kind of problems occurring in the complex system dynamic.
- Next task for the students is to improve the system dynamics towards the demand. Therefore the students can change interactive some parameter of the models and test different scenarios in several simulation runs, until the actual queuing systems meets the demand. For the learning process it is essential to exchange the experiences gained in the interactive simulation in order to develop a competent empirical knowledge.
- The exchange of experiences between members of a learning group as well as with the faculty can be supported by use of communication tools integrated in a LMS or by a groupware product.

The teaching and learning approach applied in this project follows the goal not only to acquire abstract formula, but to develop an experience based expert knowledge regarding complex system dynamics of production systems.

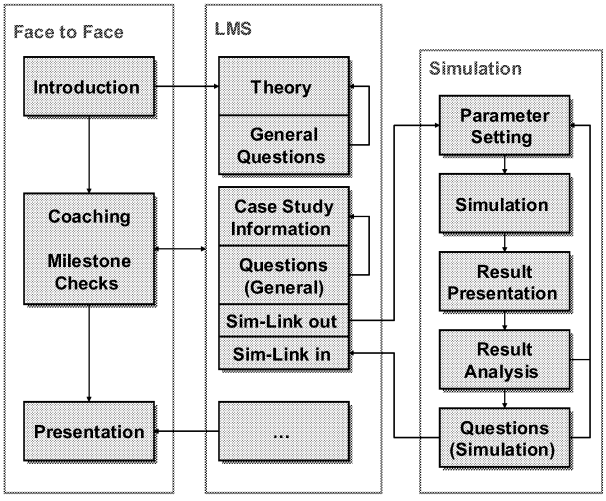


Fig. 4: Structure of a Blended Learning Unit

The simulation interface integrated in the learning management system offers the students to test the dynamics of production systems interactive in detail. Through the web based approach this can be performed in any computer lab at the university or even from the PC at home with individual learning speed. Additionally the integration of simulation in the learning environment is supposed to improve the communication between students. Blends are characterized by customization, integration, purpose, flexibility, discovery oriented studies and redundancy

SIMULATION LIBRARY

The applied Java-Sim simulation library has been developed at the University of Wuppertal for the simulation of discrete event processes. It realizes a modern simulator concept, which combines building block structures, very high flexibility and the ease of use through an advanced control concept. Further inheritance, very flexible communication and interface modules are included.

The Java-Sim library is completely implemented in native Java and offers therefore clear features for the integration on various platforms. The disadvantage of a slightly reduced processing speed compared to other programming languages is due to the increased hardware performance not an important factor any more. The Java-Sim library is not limited to one certain simulation level as many commercial simulation tools. The architecture of the library uses an object oriented approach, it is open and modular. The library offers features similar to the flexibility of a simulation language, through the level of general building blocks, such as workstations and assembly stations, up to the efficiency of simulators with special building blocks.

The Java-Sim library follows an object oriented design. It contains public simulation classes, made available for the user of the simulation library, and internal simulation classes and utility classes for internal use in the library only. The library offers general simulation objects, material flow objects, information flow objects, decision making objects and applied material flow objects.

The internal class *InfFlowObject* is used to implement information flow. Special information flow elements may be derived from this class. The class *SimEventObject* is based on the class *GeneralObject* and contains the additional functionality for creating simulation events. From this class the general classes *EventGenerator* and *MathFlowObject* and *MathFlowObjecExt* are derived. Again from these the detailed material flow classes are derived such as *Buffer*, *BufferFifo*, *BufferLifo* etc. For all simulation classes there are some general class features available, e.g. such as create object, delete object, constructor control and destructor control.

The *EventController* is the engine of each discrete event simulation. The *EventController* includes optional an

EventTracer and an *EventDebugger*. The *EventController* has to handle all events taking place during a simulation run. Therefore the *EventController* includes an *EventList*, which contains all actual simulation events. The methods of the *EventController* have to deal with incoming events, with sorting of events and calculating the event statistics. The *Eventcontroller* offers methods for parameter setting, methods for running the simulation and methods for getting information from the *EventController*.

Material flow elements are classified in the library into stationary and mobile material flow objects. Stationary objects represent mainly the available resources in a factory, mobile objects the material flow through the factory. The material flow elements are classified into active material flow elements passive material flow elements. Active material flow elements are able to pass the *ME's* actively to a next element, on the passive material flow elements the *ME's* have to be active elements or have to be moved by an external control. Active material flow objects are the *Source* and the *Drain,* the *SingleProc* and the *MultiProc*, the *Conveyor*, the *Sorter*, and the *Buffer*.

Passive material flow elements are e.g. the *Warehouse* and the *Track* which do not automatically pass ME's to a next location. Mobil material flow elements, such as the *Transporter*, the *Container* or the *Part* represent the physical or logical elements moving through a model. Mobil elements (ME's) require stationary material flow objects, such as a *SingleProc*, a *MultiProc*, a *Buffer* or a *Track* to be moved on. The Stationary Material Flow objects are classified in place oriented and length oriented elements. Capacity oriented material flow elements, such as the *SingleProc*, the *MultiProc* and the *Buffer*, may contain a certain number of elements independent from the size of each element.

Length oriented elements, such as the *Conveyor* and the *Track*, take the length of the stationary object and the mobile element into account. Extensions are Applied Material Flow Objects, which are material flow objects for specific applications. These objects are based on the basic material flow objects defined above. Some of these applied material flow objects are kept very general, such as e.g. the WorkStation, other applied material flow objects are very specific for certain applications, such as e.g. the objects for modeling transfer systems.

The Java-Sim library concept offers different user levels. On the lower level very flexible modeling and integration features are available. However on this level good Java programming knowledge is mandatory. Therefore this level is limited to face to face teaching by an expert only. The upper level focuses on users without programming knowledge. This level offers a comfortable user interface to design and to parameterize models without programming. Modeling on this level is not as flexible as on the lower level, however Java-Sim implements an advanced control concept to define flexible controls on this level.

The integration with the learning management system for the blended learning environment is set up on the upper user level of Java-Sim. A user friendly approach is required to parameterize simulation models and to perform simulation runs conveniently out of the learning management environment without any programming and system knowledge. Therefore simulation models, parameter settings and simulation results are made available to the user through Java applet with a graphical simulation user interface integrated in a learning management system. The distributed system architecture is based on the LMS server and the independent simulation server. The simulation user interface integrated in a Java applet which can be accessed out of the learning management system is strictly separated from core simulation classes. Data regarding model, parameter setting, simulation scenario, simulation results etc. have to be transferred between simulation server and LMS server.

PROJECT PLAN / REALISATION

The realization of the project contains the general concept of the blended learning scenario, the technical implementation on top of the object oriented discrete event simulation library Java-Sim, the test of the software components and the use case application towards a first prototype. The project contains the following steps:

- Definition of the entire blended learning concept
- Specification of individual learning units
- Implementation of a web based simulation user interface for the integration in the LMS
- Integration / Interfaces with the LMS
- Test of communication between simulation server and LMS
- Implementation of multimedia learning units and interaction elements in the LMS
- Implementation and test of production models
- Test of learning scenarios
- Documentation
- Prototype available for practice tests

TECHNICAL APPROACH

The implementation of the user interface and the communication classes follows a strict object oriented design in Java. For the communication between the LMS server and the simulation server XML und XML-

RPC are applied. Alternative RMI or SOAP interfaces could have been used as communication alternative.

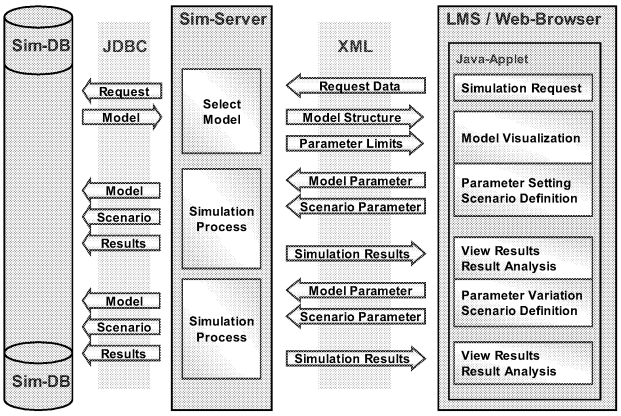


Fig. 5: Integrated LMS / Simulation Environment

The experiences gained in this project are not limited to this project but can be transferred to other applications. These are e.g. teaching experience for organizing learning scenarios as blended learning arrangements in technical university courses or experience guiding and supervising learning groups during the development of simulation strategies. Also the experience implementing asynchronous communication options with an e-Learning tool to establish empirical knowledge can be useful to many fields.

CONCLUSION

Blended learning improves learning effectiveness by addressing the changing needs in graduate and industry education, including filling knowledge gaps for learners due to educational background or time constraints. It enhances the learning experience by offering relevant material in asynchronous, learner focused formats, and reducing demands of faculty time for teaching basic material.

Blended learning may include any combination of text, graphics or audio information, streaming video, interactive learning activities, self-tests, links to related materials/tools and scenario based simulations. All enrolled students will be able to access appropriate information, regardless of the students' location. In order to improve the educational experience for learners universally and to bridge the University with industry, E-learning modules with simulation based experience units scan be made available for other universities and industry training also.

At the present approach the open source Learn Management System ILIAS has been applied. Interfacing ILIAS 3.0 to simulation gains still some technical problems. The interfacing features for dynamic processes, e.g. building questions based on parameters of an actual simulation run, are very limited respectively not available with the actual software version. Therefore for the simulation based learning experience units some features are not applied in the present approach. These features are additionally implemented in the simulation environment. This approach with partly redundant implementations is a temporary compromise, in order to enable dynamic features on base of actual existing software. On the long run the implementation of powerful interfacing features is mandatory.

REFERENCES

Kerres M., De Witt C., „A Didactical Framework for the Design of Blended Learning Arrangements“, Journal of Educational Media, 10/2003, 101 - 113

Kühn, W.; Java-Sim - An advanced Discrete Event Simulation Library, SCSC 2002, Summer Computer Simulation Conference, 2002, San Diego, USA, 393-397

Kühn W.; Standardisation of Simulation Studies by use of XML-Technology, SCSC'04, Summer Computer Simulation Conference, 2004, San Jose, USA, 334-342

Maier-Häfele K., Häfele H., „Open-Source-Werkzeuge für e-Trainings“, ManagerSeminare, 2005

Qiao, G., Riddick, F., McLean, C., 2003. “Data driven design and simulation system based on XML”, Proceedings of the 2003 Winter Simulation Conference, 1143-1148

Wiedemann, T., 2002. “Next Generation Simulation Environments founded on open source software and XML-based standard interfaces”. Proceedings of the 2002 Winter Simulation Conference, 623-628

BIOGRAPHY

WOLFGANG KUEHN studied mechanical engineering at the University of Brunswick, Germany. Afterwards he worked two years with Blaupunkt. At the University of Bremen he got his PHD in production engineering 1991, and his habilitation in the area of simulation of production systems in 1997. From 1993 to 1995 he worked as Associated Professor at the Asian Institute of Technology in Bangkok. 1996 he founded the SIPOC Simulation based Planning, Optimization and Control GmbH in Bremen. Since 1997 he has worked as professor for production planning and control in the department of Electrical, Information and Media Engineering at the University of Wuppertal.

PLANSIM, a framework to build automated planners with simulators

Ruy L. Milidiú and Frederico dos S. Liporace*
Departamento de Informática
Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro, RJ, Brazil
e-mail: {milidiu,liporace}@inf.puc-rio.br

May 16, 2005

KEYWORDS
AI Supported Simulation, Decision Support Systems, Transportation, Petroleum Engineering, Event-oriented, Software Frameworks.

ABSTRACT

In this paper we describe some issues faced when building automated planners using state-of-the-art general purpose planners, and propose PLANSIM, a software framework for building planning agents, that address those issues. PLANSIM defines a baseline architecture that integrates planning based on state space search and simulators. This strategy enables the planner developer to focus solely on aspects that are specific to the application, such as domain specific heuristics and how to model the environment through a discrete event simulator.

INTRODUCTION

Building software for automated planning of industry operational scenarios is a highly complex task. It requires expertise ranging from detailed domain knowledge, artificial intelligence, operational research techniques and software engineering.

In this paper we report the current status of our work on PLANSIM, a software framework for building automated planners that make use of discrete event simulation to model the environment.

The first critical decision when developing an automated planner is to choose the technique that will support the planner reasoning. One may use, among others, integer programming, constraint satisfaction or heuristic search (Ghallab, Nau, & Traverso 2004) techniques. In PLANSIM we use heuristic search techniques.

This decision is motivated by the fact that planners based on heuristic search were the top performers in the latter editions of the International Planning Competition (IPC) (Long & Fox 2003; Hoffmann *et al.* 2004). In this competition, the state-of-the-art general purpose automated planners developed by the artificial intelligence community are evaluated by their performance over a set of benchmarks.

The last IPC edition extended the competition benchmark to include problems inspired by real world applications (Hoffmann *et al.* 2004), as part of an effort to concentrate the research on planners that scale to handle industrial operational scenarios, and not only toy problems. Although there have been a steady improvement on the general purpose solvers performance in the last years, its currently use is still limited to a small subset of usually simplified versions of real world problems.

PLANSIM offers an alternative to the use of general purpose planning systems, but without the problems that are usually encountered when developing a domain specific automated planner.

This paper is organized as follows: first we compare the issues involved in developing an automated planner using general purpose and domain specific planners technologies. Next, we show how software frameworks may be used to obtain a compromise between easy of development and performance, and present the PLANSIM framework, with examples on how it can be instantiated. Finally, we describe how PLANSIM is being used to generate plans for multi-commodity oil pipelines.

GENERAL AND DOMAIN SPECIFIC PLANNERS

General purpose planners usually make use of PDDL(Fox & Long 2001) to describe the planning problem. PDDL defines a *domain* file that use binary predicates to represent the state of the system being modeled, and define a set of *actions* that may be applied to the system. In order to be applied, each action must satisfy its *preconditions*. The consequences of applying an action to the system are represented in the action *effects*. A *problem* file defines the system initial states and the goals that must be achieved.

The main advantage of general purpose planners is that they need only a description of the problem to start the planning process. Planners based on heuristic search, for instance, try to derive good heuristics using relaxations of the original problem, and make use of different

*Supported by CNPq

search strategies, sometimes varying them during the planning process. PDDL does not support the explicit inclusion of control rules or heuristics based on previous domain knowledge. This is a very serious drawback, since a good heuristic may make the difference between a tractable and an untractable problem. Finding a way to automatically build heuristics is interesting as a research topic for AI, but from a practitioner's point of view not being able to define heuristics can be very frustrating.
For instance, we contributed to the last IPC benchmark with Pipesworld (Milidiú, dos Santos Liporace, & de Lucena 2003), a problem inspired by the oil pipeline planning problem. Even though the original problem was simplified and the benchmark instances were small when compared to operational ones, many instances were not solved by any planner in the allowed time-frame(Hoffmann *et al.* 2004). We showed later (Milidiú & dos Santos Liporace 2004) that a dedicated planner that implements the same search algorithms used by general purpose planners, but incorporates very simple domain specific heuristics, can perform much better.
Another problem encountered when using general purpose planners lies on the limited PDDL expressiveness. For instance, numeric variables and temporal constraints have limited support in the language, and most planners still do not support them. During the Pipesworld development, for instance, we were not able to represent the tank volumes as numeric quantities because of that fact. Overall, an industrial planning problem usually must be oversimplified in order to make it fit into the PDDL model. Sometimes the changes are so significant that the resulting model is in fact a very different problem, only resembling the original one.
Domain specific planners address many of the aforementioned problems. Since they are implemented from scratch, the developer may choose the language that he or she thinks suits the problem better, not being limited by PDDL. Control rules and domain specific heuristics may also be incorporated as well. But that comes with a high price, represented by the effort to design and implement such planners. For instance, the developer must define an architecture that specifies how the search algorithms will interface with the simulators, implement the searching engines, etc.
Software frameworks may be used to drastically reduce the effort required to build such domain specific planners. From the developer's point of view, a solution based on frameworks offers a good compromise between easy of development and flexibility. This is elaborated in the next section.

PLANSIM

PLANSIM is a software framework developed in C++ for heuristic search planners that use discrete event simulators to model the environment. It makes the development of domain specific planners easier by defining an

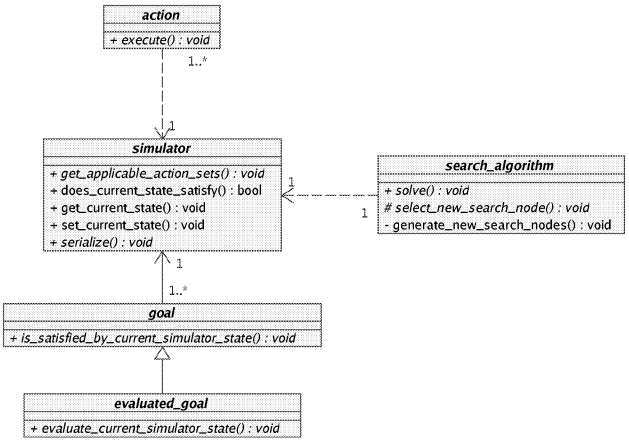


Figure 1: PLANSIM Main Classes

architecture for the application, and implementing functionalities that otherwise would have to be implemented by the planner developer. Although the architecture is fixed, the design is flexible enough to accommodate a wide range of planning applications.
The framework main classes are shown in Figure 1. Only a subset of the defined methods is shown, and the methods signatures are ommited for clarity. The complete reference may be obtained from the PLANSIM homepage on the Web. The *search algorithm* classes are responsible for performing the search in the state space until the goal is reached. It uses the *simulator* class to obtain the actions that may be applied in a given state, through the *get applicable action sets* method.
A simulator has a current state, and must be able to serialize this state, through the *get current state* and *set current state* methods. The simulator state is also updated when an *action* is applied, through the *execute* method. Every action has a cost, that may be used by the search algorithm for planning optimization.
The *goal* class represents the goals that must be achieved. The goal class must know, for a given simulator, if it is satisfied by its current state, this is performed by the *is satisfied by current simulator state* method. If the goal is an *evaluated goal*, it also must be able to evaluate, through the *evaluate current simulator state* method, the current simulator state, giving it a grade based on an estimate of the distance between the current state and the goal. This class is used to implement domain specific heuristics.
The core of the framework is composed by its searching engines, and its class hierarchy is shown in Figure 2. The *search algorithm* base class defines the basic search procedure, that is, generating the current search node neighbors at each iteration, checking if the goal is achieved and selecting the next node to be explored. Each derived class implements its own way of selecting

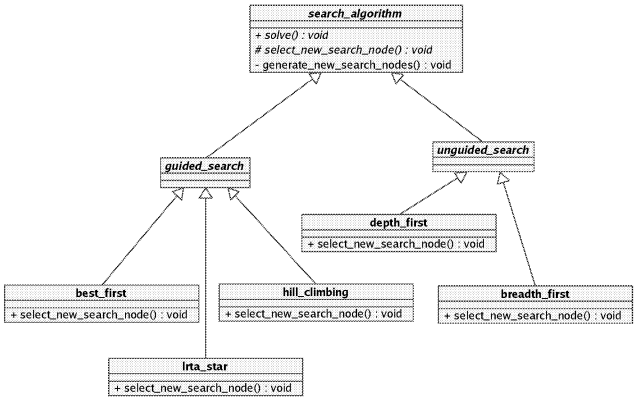


Figure 2: Search Algorithms in PLANSIM

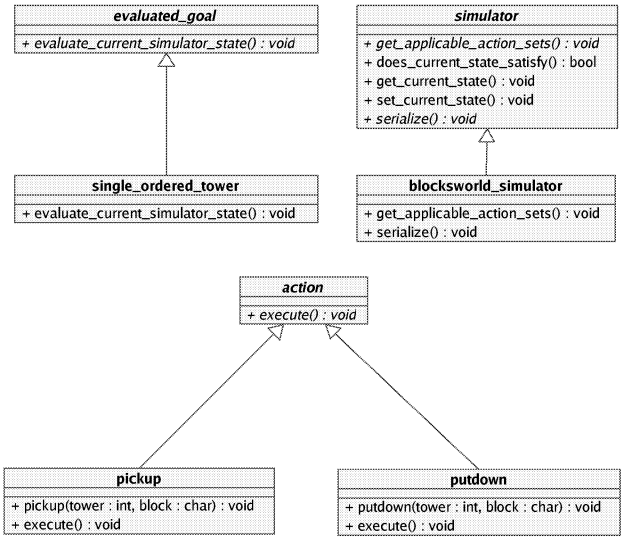


Figure 3: Blocksworld Instantiation

the next search node to be explored among the candidates. For instance, *best first* search selects the node with better estimate not yet explored, while *depth first* always explore the first node opened in the last iteration. The framework may be easily extended to incorporate new search strategies by simply implementing the *select new search node* appropriately.

The *search algorithm* classes are also concerned with the detection of duplicate states eventually found during the search procedure, that is, the generation of states that were already visited before, but have alternate paths that reach them. In that case the path with smaller cost found so far is maintained, with the other path being ignored.

PLANSIM currently implements classical algorithms such as best first search, hill climbing, learning real time A* (Korf 1990) and newly developed algorithms like enforced hill climbing (Hoffmann & Nebel 2001).

PLANSIM INSTANTIATION

In order to describe the framework usage we show in this section how it can be instantiated to solve a well known problem, Blocksworld.

In Blocksworld(Russel & Norvig 2003), we have n blocks initially ordered in m towers. Each block is assigned to a letter, and the goal is to place all blocks in a single tower, in a ordered fashion. In order to move the blocks, we have an automated arm that may hold a single block. Only blocks on the top of the towers may be moved.

Figure 3 shows how the PLANSIM framework may be instantiated to solve this problem. Only four new classes are needed: the *blocksworld simulator* class defines how the system state is represented and how the system state is serialized. We have two defined actions, namely *pickup* to pick up the top block from a tower and *putdown* to put down a block to the top of a tower. Generating which actions are applicable in a given state is a simple task for the simulator: if the arm is holding something, we have one putdown action for each tower. If the arm is not holding anything, we have a pickup action for each tower.

Finally, we have a *evaluated goal* class, *single ordered tower* that represents the goal. Being an evaluated goal, it must define a way to evaluate a simulator state. In this case, the evaluation simply counts the number of blocks correctly positioned and subtracts this number from the total number of blocks.

Although simple, this example shows how the developer may use language higher level constructs to make the modeling easier. In our case we used the C++ Standard Template Library to represent the towers.

CURRENT RESEARCH

The PLANSIM framework is currently being used to build a new version of Plumber, a tool aimed to develop plans for multi-commodity oil pipelines. An initial version of Plumber (Milidiú & dos Santos Liporace 2004) was developed to handle Pipesworld instances, the new version will work with a model that is much closer to reality, since a discrete event simulator is being developed to represent the pipeline.

PLANSIM is currently available as an opensource project under the LGPL license. It may be accessed at http://sourceforge.net/projects/plansim.

CONCLUSION

We presented PLANSIM, a software framework for building planning agents. This framework incorporates some advantages of general purpose planners, such as not requiring from the planner developer to explicitly

design and implement the searching algorithms, but in addition to that it also lets the developer use high level languages to describe the environment where the plan will be applied, and incorporate domain specific control rules and heuristics that have a great impact on the resulting planner's performance.

REFERENCES

[Fox & Long 2001] Fox, M., and Long, D. 2001. PDDL 2.1: An extension to PDDL for expressing temporal planning domains.

[Ghallab, Nau, & Traverso 2004] Ghallab, M.; Nau, D.; and Traverso, P. 2004. *Automated Planning: Theory and Practice.*

[Hoffmann & Nebel 2001] Hoffmann, J., and Nebel, B. 2001. The FF planning system: Fast plan generation through heuristic search. *Journal of Artificial Intelligence Research* 14:253–302.

[Hoffmann *et al.* 2004] Hoffmann, J.; Edelkamp, S.; Englert, R.; Liporace, F.; Thiebaux, S.; and Trg, S. 2004. Towards realistic benchmarks for planning: The domains used in the classical part of ipc-4. In *Proceedings of the International Planning Competition, International Conference on Automated Planning and Scheduling (ICAPS).*

[Korf 1990] Korf, R. E. 1990. Real-time heuristic search. *Artificial Intelligence* 42:189–211.

[Long & Fox 2003] Long, D., and Fox, M. 2003. The 3rd international planning competition: Results and analysis. *Journal of Artificial Intelligence Research* 20:1–50.

[Milidiú & dos Santos Liporace 2004] Milidiú, R. L., and dos Santos Liporace, F. 2004. Plumber, a pipeline transportation planner. In *Proceedings of the International Workshop on Harbour and Maritime Simulation, HMS*, 99–106.

[Milidiú, dos Santos Liporace, & de Lucena 2003] Milidiú, R. L.; dos Santos Liporace, F.; and de Lucena, C. J. P. 2003. Pipesworld: Planning pipeline transportation of petroleum derivatives. In *Workshop on the Competition, Trento, Italy.*

[Russel & Norvig 2003] Russel, and Norvig. 2003. *Artificial Intelligence, A Modern Approach.*

A VIRTUAL REAL-TIME MODEL FOR CONTROL SOFTWARE DEVELOPMENT – APPLIED ON A SHEET-METAL PRESS LINE

Bo Svensson and Fredrik Danielsson
Dep. of Technology, Mathematics & Computer Science
University of Trollhättan/Uddevalla
P.O. Box 957, SE-461 29 Trollhättan
Sweden
E-mail: {bo.svensson|fredrik.danielsson}@htu.se

Bengt Lennartson
Dep. of Signals and Systems
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
E-mail: bl@s2.chalmers.se

KEYWORDS

Real-time simulation, Emulation, Validation, Industrial control systems, Production, Off-line programming.

ABSTRACT

Maximizing lifetime throughput, i.e. increasing the total number of produced components in a pre-existing production line, is a continuous and never-ending process in industry. In order to attain a high lifetime throughput, it is important not only to have a high throughput rate during production but also to minimize production stops, both in number and time. In this paper a general Virtual Real-Time Model (VRTM) for simulating industrial production lines is proposed. The VRTM is intended to be used for off-line control software development and is aimed at maximizing lifetime throughput in highly automated production lines. The main features of the VRTM are that the same control software is used as in the real production line and that virtual real-time is used, implying that everything is time-synchronised and works in the same way as in the real production line. An applied VRTM simulating a sheet-metal press line is also presented together with a comparison test with the no. 53 press line at Volvo Car Manufacturing in Göteborg, Sweden, in which a satisfactory agreement was obtained. The control software from the real production line that was used includes not only the pure production control, but also all security functions. This shows that the VRTM covers all control functions in the real production line and can be used for off-line control software development.

INTRODUCTION

Usually a highly automated industrial production line consists of a number of machines and robots working together, under the control of more than one industrial control system. Each control system normally manages its part of the production line with one or more machines and/or robots. The control systems communicate with each other to get the whole production line working together as one unit. In some cases an extra control system is added to handle the supervisory control, whilst in others the supervisory control is handled by the existing control systems. Each of the industrial control systems has to be programmed, in some cases with unique control software, and in other cases with common control software with individual parameters and settings.

The initial costs of a production line have to be recovered from the revenue for the components produced on that line. For that reason it is important that the production line has a high throughput rate, not only in the short run but also throughout the entire economic life of the line. Otherwise there would be a risk that the cost per produced component would exceed the possible calculated revenue.

To attain a high lifetime throughput for an automated industrial production line it is important not only to have a high throughput rate during production, but also to avoid any unnecessary production stops. If a production stop is necessary, for instance when changing tools, it is important to make it short and to have a fast ramp-up time so that a high throughput rate can once again be achieved. To fulfil all these demands there is need for methods to develop the control software off-line, i.e. to manage programming, verification, validation, optimisation, training new operators and service personnel, etc. without using the real production line.

Many methods intended for off-line programming and verification of control software have been developed (Danielsson et al. 2003 and references herein). Some of these methods utilize simulation systems, whilst others do not. Such simulation systems usually consist of a model of an industrial control system together with a model of the plant (Hassapis 2000).

At present the control software used in most of these models is not possible to reuse in the real control systems without modifications, since a model always contains approximations and/or modelling errors. A superior method, compared to the use of a model of the control system, is to use an industrial control system emulator (Danielsson et al. 2003). The difference is that the control system emulator works in exactly the same way as the real control system, i.e. the control software used off-line is identical to that used on-line.

The model of the plant usually consists of one or more statistical model(s) and/or a selection of actions carried

out in accordance with sets of simple rules (Jönsson et al. 2003). As a consequence the model of the plant lacks the necessary sophistication for use together with the real control software. Thus there is a need for a more realistic model of the plant, a virtual plant, which can include servos, encoders, analogue values, time dependent signals, etc. Using a virtual plant and an industrial control system emulator together requires real-time simulation, due to the time dependencies.

To our knowledge, currently there are no suitable commercial simulation system tools for this type of application working in real-time.

In this paper a general Virtual Real-Time Model (VRTM), simulating automated industrial production lines, is proposed. The main features of the VRTM are that the same control software is used as in the real production line and that virtual real-time is used, implying that everything is time-synchronised and works as in the real production line.

Further, in this paper, a sheet-metal press line is described as an example of a highly automated industrial production line. Based on the proposed general VRTM, an applied VRTM simulating a sheet-metal press line is proposed and explained. In order to validate the VRTM a comparison test with the no. 53 press line at Volvo Car Manufacturing (VCM) in Göteborg, Sweden, has been performed. Finally, the comparison test results, which show a satisfactory agreement between the simulated and the real production line, are presented and future work is discussed.

THE VIRTUAL REAL-TIME MODEL

Simulating a large production system with many control systems requires an open architecture with a distributed approach that allows different modules to run on separate computers. To manage this and to handle the time synchronisation between all modules a Synchronised Distributed Simulation Protocol (SDSP) server is used. The SDSP server is not only responsible for handling the virtual real-time, but also for handling all of the common information, such as all electrical signals to/from the control systems, simulation parameters, model status, etc. (Danielsson 1999).

Based on our understanding of SDSP a general VRTM simulating automated industrial production lines is proposed and shown in Figure 1. The VRTM is intended to be used for all kinds of off-line control software development and is aimed at maximizing lifetime throughput, including programming, verification, validation, optimisation, training new operators and service personnel, etc.

The general VRTM consists of a virtual plant; *n* control system emulators, one for each controller in the real production line; and a real-time data base. All parts work together and are synchronised in time.

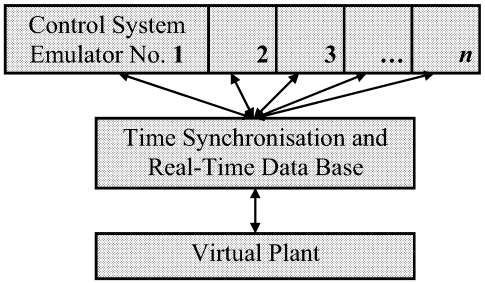


Figure 1 Proposed General Virtual Real-Time Model

The main tasks of the VRTM are to ensure that:
- all motions in the virtual plant are equal to the motions in the real production line
- all electrical signals to/from the control system emulators are equal to the signals in the real production line

The virtual plant is a vital part in the VRTM and is therefore also important for increasing the lifetime throughput. The real production line has to be analysed in detail. Geometric and kinematics, electrical analogue and digital, as well as operator interface models, etc. have to be designed and verified. All these models together form the virtual plant.

The intention is to upload the existing on-line control software from the real production line, into the proposed VRTM. Once this has been achieved the simulated production line will act as the real production line.

To summarize, the key features of the VRTM are:
- that all vital machines, robots, control systems, operator panels, etc. in the real production line are included
- that all electrical analogue and digital signals to/from the control systems in the real production line are included
- that the same control software is used in the control system emulators as is used in the control systems in the real production line
- that all vital motions of the equipment in the real production line are simulated
- that virtual real-time is used, i.e. everything is time-synchronised, albeit on a different time scale

THE SHEET-METAL PRESS LINE

A sheet-metal press line in automobile manufacture is a good example of a highly automated industrial production line. It involves a number of press control system parameters which must be tuned together with the automated material-handling equipment.

A press line normally consists of four to six press stations, feeders and extractors (often specialised robots), and intermediate stations, all working together under the control of a number of industrial control systems. The production is batch run, so that one particular component is in production at any one time. When a new component is to be produced, both the tools in all of the press stations, intermediate stations, feeders and extractors, as well as the software parameters and settings in the control systems, have to be changed. An example of a sheet-metal press line with four press stations is shown in Figure 2.

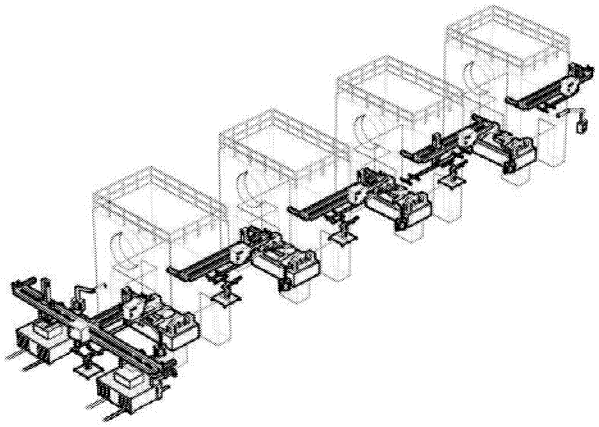


Figure 2 Sheet-metal press line with four press stations

The industrial production line simulated in this project is the no. 53 sheet-metal press line in the car manufacturing plant at VCM in Gothenburg, Sweden. This actual press line has five press stations. Between each one of the five press stations there are four combined feeder/extractors, so called Unifeeders, supplied by Binar AB (Binar), Sweden. Furthermore, there are four intermediate stations, which individually can move vertically and rotate around horizontal and vertical axes. As a feeder in the first press station and an extractor in the last press station, two specialised 2D-robots are used. Five identical industrial control systems, BiFas UHS from Binar, work together and control the entire press line. Each one of these five control systems controls one of the press station, the robot and intermediate station before the press station and all of the in- and outputs from/to sensors, push-buttons, actuators, indicators, etc. in its environment. In addition, one operator panel, OP44 from Binar, is connected to each control system, i.e. in total five operator panels are included in the press line. The control software in each of the five control systems includes all control functions in its part of the press line. This includes not only the control of the press station, robot and intermediate station, but also the robot trajectory, safety control, communication with the control systems before and after, communication with the operator, etc. The supervisory control of the complete press line is also included in the control software in the five control systems.

VRTM APPLIED ON A PRESS LINE

An applied VRTM simulating a sheet-metal press line, as described in former section, is proposed and shown in Figure 3. This applied VRTM consists of four main parts:

- five operator panels
- five control system emulators
- one time-synchronisation and a real-time data base (SDSP server)
- one virtual plant (divided into three modules)

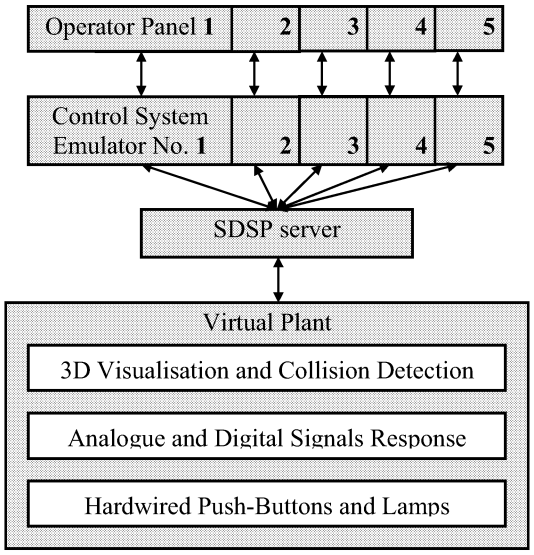


Figure 3 Proposed applied Virtual Real-Time Model simulating a sheet-metal press line

The operator panels constitute a part of the communication interface between the operator and the production line and make it possible for the operator to control the production in the simulated process in exactly the same way as in the real production line. Real operator panels, of the same type as in the real production line, are used in the simulation. The control system emulators work in exactly the same way as in real industrial control systems. Both the software in the control system emulators and in the operator panels in the simulation are the same as the software used in the real production line at VCM. This is one of the main goals of this work, i.e. that the control software used in the VRTM should be the same as in the real production. Another goal of this work is that virtual real-time is used throughout the VRTM, implying that everything is time-synchronised and works exactly as it does in the real production line. All electrical signals between the control system emulators and the virtual plant are managed by the real-time data base. The real-time data base constitutes an electrical signal interface for each control system. All these interfaces have to be modelled correctly, as otherwise the control software will not work as in the real production line. The

SDSP server manages the time synchronisation and real-time data base.

The virtual plant has been designed to simulate the operation of the real sheet-metal press line at VCM including not only all vital machines, but also all electrical analogue and digital signals to/from the control systems. It is necessary that the complete virtual plant, as well as all the VRTM constituent parts, work in virtual real-time. This is due to the fact that all parts of the virtual plant are designed with time characteristics and thus can be time-synchronised. The proposed structure partitions the virtual plant into three modules, as shown in Figure 3, each of which manages its own tasks.

The first module manages 3D visualisation, representation and collision detection of all of the vital equipment in the production line. Geometric and kinematics models have been designed for press stations, Unifeeders, intermediate stations and the feeder and extractor. All of the models are made from drawings or measurements in the real production line. The commercial software Robcad (an eMPower product from Tecnomatix) has been used for this task. Robcad has an automatic collision detection test which recognizes and reports any unwanted contacts in the simulation.

The second module calculates all electrical analogue and digital signal responses from all of the sensors, switches, transmitters, servos, encoders, etc. in the production line. A number of different software applications have been created for this task.

The third module simulates all of the hard-wired push-buttons and lamps which, together with the operator panels, constitute the communication interface between the operator and the production line. Some C++ applications have been made for this task. Electrical drawings of press line no. 53 at VCM and electrical equipment function descriptions have been used when designing the second and third modules.

VALIDATION

The main tasks of the VRTM, all of the motions in the virtual plant and all of the electrical signals to/from the control system emulators that replicate the motions and signals in the real production line, have to be validated. The virtual plant, in the VRTM, handles a rather complex system of both electrical and mechanical real-time responses to the control software activities. All of these responses have to be correct, both in value and time, otherwise the VRTM will not function in the same way as the real production line when using the same control software.

The validity of a control system emulator combined with a virtual model of a Unifeeder in the computer aided robotic tool Robcad, has already been verified. The comparison between the paths achieved on the simulated and the real Unifeeder achieved an agreement sufficient enough for this type of application (Danielsson et al. 2002).

Based on the knowledge that all of the motions in every single virtual machine are equal to the motions in the real equipment, a comparison test, using the same control software for simulation and the real production line, can be used to validate the VRTM.

A comparison test with the sheet-metal press line no. 53 at VCM in Göteborg, using the same control software, has been performed. Frame Reinforcement Sunroof P1X was used as a component in production during the comparison test. This is a component which reinforces the hatch-opening sunroof on some Volvo models. The reason for using this component is that it is a typical of the types of component produced on a sheet-metal press line and it is a component which is currently produced in the actual press line at VCM.

The physical dimensions of the component in production and all relevant tools in press stations, Unifeeders, feeder, extractor and intermediate stations in the press line were needed for the design of all geometric and kinematics models in Robcad. Most of the dimensions were obtained from drawings, but some had to be measured directly in the real production line. The accuracy is better than 5 mm in all dimensions, which is sufficient enough for this type of application.

The control software used in the five industrial control systems in the real production line was uploaded to the control system emulators in the VRTM. All parameters and settings, in the control software and on the five operator panels in the VRTM, were set to the same values as in the real production line for the actual component in production.

In the comparison test, using the same control software, the following three items were observed both in the real production line and in the VRTM:
- undesirable contacts (collisions)
- the passage of components through the production line
- throughput rate

RESULTS AND DISCUSSION

The results from the comparison test, when using the same control software, are:
- no undesirable contacts (collisions) between components, tools, presses, intermediate stations, feeders and/or extractors were reported either in the VRTM or in the real production line
- all components in the production could be visually followed through the production line in the same way in the VRTM as in the real production line

- the production throughput rate was measured to 10.4 components/minute in the VRTM and to 10.6 components/minute in the real production line

The production throughput rate measurement was in both cases done by the ordinary control software throughput rate calculator. Depending on the calculator is momentary the measurements had to be done when the production was in steady state.

The comparison test shows a satisfactory agreement between the simulated sheet-metal press line in the VRTM and the real production line at VCM.

In the VRTM some minor parts have not yet been included:
- there is no servo loop emulator; thus all servos are considered ideal, i.e. set point = actual value
- all servomotors are considered as being ideal
- no simulation or emulation of communication buses is included; thus all buses are considered ideal; i.e. received signal = send signal

CONCLUSION AND FUTURE WORK

A general VRTM for simulating industrial production lines is proposed. The VRTM is intended to be used for off-line control software development aimed at maximizing lifetime throughput in highly automated production lines. The VRTM is based on the use of the same control software in the virtual real-time simulation as in the real production line. In order to validate the VRTM an applied VRTM simulating a sheet-metal press line is proposed and a comparison test with sheet-metal press line no. 53 at VCM has been accomplished with a satisfactory agreement being obtained.

The VRTM has been validated and found to fulfil its main tasks:
- all motions in the virtual plant are equal to the motions in the real production line
- all electrical signals to/from the control system emulators are equal to the signals in the real production line

The control software, in the real production line, includes more software code handling security functions than the codes used for pure production control. The security functions are designed both to avoid damage to equipment and/or tools, but also to prevent personal injury. Using the same control software for the VRTM as is used in the real production line in the comparison test, it was shown that the virtual plant in the VRTM also handles all security signals correctly.

Use of the VRTM seems to make it possible to attain the goal of maximizing the lifetime throughput in highly automated industrial production lines. It will be possible to use the VRTM for:
- off-line programming
- verification and validation of the control software
- optimisation of the control software
- training new operators and service personnel
- testing different solutions before making changes in the real production line
- optimisation of production line processes with regard to production throughput rate
- and much more

Our experimental research with the VRTM is continues at the time of writing and further work on different approaches to maximizing lifetime throughput is planned for the future. Today there does not exist, to our knowledge, any virtual real-time simulation system used for maximizing the lifetime throughput in an industrial automated production line. The maximization procedure presently used in practice is highly empirical, based on trial and success and is largely dependent upon the experience of the operators. Furthermore, maximization in the real production line, on-line maximization, can be dangerous and costly if, for instance, a tool is damaged. In the VRTM it is possible to evaluate many different approaches without this type of risk-taking, e.g.
- optimising controller software and trajectories
- changing safety margins and rules
- investigating communication between controllers and between controller and plant
- simulating speculative 'what-if' scenarios

ACKNOWLEDGEMENTS

This work is being funded by the Swedish Foundation for Knowledge and Competence Development. The authors gratefully thank Volvo Car Manufacturing (VCM) in Göteborg, Sweden, and Binar AB (Binar) in Trollhättan, Sweden. VCM has kindly provided access to its sheet-metal press line in the car manufacturing plant and Binar has kindly provided all necessary software and development tools at our disposal.

REFERENCES

Danielsson, F. 1999. "A distributed system architecture for optimizing control logic in complex manufacturing systems". *ISCA 12th international Conference* (Atlanta, USA), 163-167.

Danielsson, F., Schüder, J., Moore, P. and Eriksson, P. 2002. "A Virtual System Using a Control System Emulator to Increase Performance of Sheet Metal Press Lines". *8th Mechatronics Forum International Conference* (Enschede, Netherlands).

Danielsson, F., Moore, P. and Eriksson, P. 2003. "Validation, off-line programming and optimisation of industrial control logic". *Mechatronics*, No.13, 571-585.

Hassapis, G. 2000. "Soft-testing of industrial control systems programmed in IEC 1131-3 languages". *ISA Transactions*, No.39, 345-355.

Jönsson, A., Bathelt, J. and Broman, G. 2003. "Interacting with real time simulations – virtual reality in industry applications". *7th Immersive Projection Technology Workshop. 9th Eurographics Workshop on Virtual Environments* (Zurich, Switzerland). 305-306.

ENTERPRISE PETRI NET BASED CAM SOFTWARE OPERA-TOPAZ

Dmitry Zaitsev
Department of Communication Networks,
Odessa National Telecommunication Academy
Kuznechnaya, 1, Odessa, 65029, Ukraine
www.geocities.com/zsoftua

KEYWORDS
Production, CAM, scheduling, Petri net

ABSTRACT

Features of enterprise Petri nets based CAM software Opera-Topaz are studied. The successful usage of Opera-Topaz during more than ten years at military plants of Ukraine in production of radar station Kolchuga and aeroplane motors confirms the practical value of ideas embodied in this tool.

INTRODUCTION

Scheduling of the production processes is one of the classical fields of Petri net application (Sleptsov and Jurasov 1986). Massive production requires stochastic nets usage whereas individual or small series production may be represented adequately with deterministic nets.

Individual or drop series production is specific and is used for creation of large-scale complex manufactured articles. For example, such military articles as aeroplanes, radar stations etc.

The process of the article production may be considered as a set of interconnected jobs on a set of resources. There are a lot of various relations on a joint set including both jobs and resources. In the simple case jobs are interconnected with a following relation and either consumable or recoverable types of resources are considered.

Complex production requires extended types of relations that cannot be represented with standard PERT models. It is the first reason for Petri net usage. Another one consists in asynchronous character of the production processes. There are several units of the same article in the production. And the same equipment is used in the production of several articles.

Petri nets allow the representation of various relations among jobs and resources required in the practice. Timed nets are adequate models of production processes (Sleptsov and Zaitsev 1988). Complete formal description of behaviour of timed Petri nets was presented in (Zaitsev and Sleptsov 1997). And dynamic priority rules usage allows the generation of timetables enough close to optimal.

Above-mentioned ideas have being successfully applied during more than ten years at the military plants of Ukraine. CAM software system Opera-Topaz (Zaitsev and Sleptsov 1990), (Sleptsov and Zaitsev 1991) was developed for Topaz Corporation wide known in the world with Kolchuga (chain mail) radar stations. Then this system was used also in Motorsich Corporation, which is a major aeroplane motors producer in Ukraine.

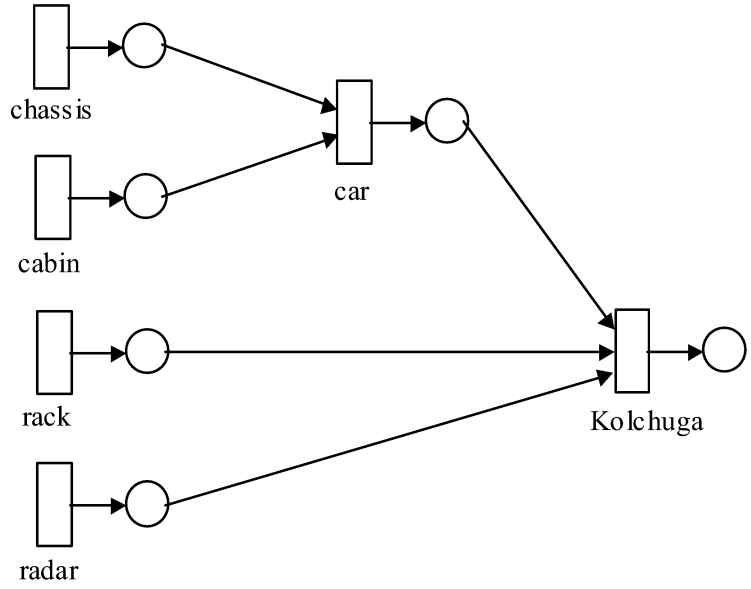
PETRI NET MODEL OF THE DISCRETE PRODUCTION

The process of the article manufacture in the discrete individual or small series production may be represented with timed Petri net (Sleptsev and Zaitsev 1988). Described in (Zaitsev and Sleptsev 1997) nets use integer values of time associated with transitions. Moreover repeated and simultaneous firing of the same active transition is allowed. So these nets was named also nets with multichannel transitions.

In (Zaitsev and Sleptsev 1997) the state equation of timed Petri nets with multichannel transitions was constructed. This equation is a complete formal description of net behaviour. On the base of this equation supplied with priority rules an ad-hoc algorithm for timetables generation was developed.

Now we consider separately a variety of processes' types and resources' relations used in the practice. Further at the figures for the convenience the places used for the description of processes have a small size whereas resources' places are lager.

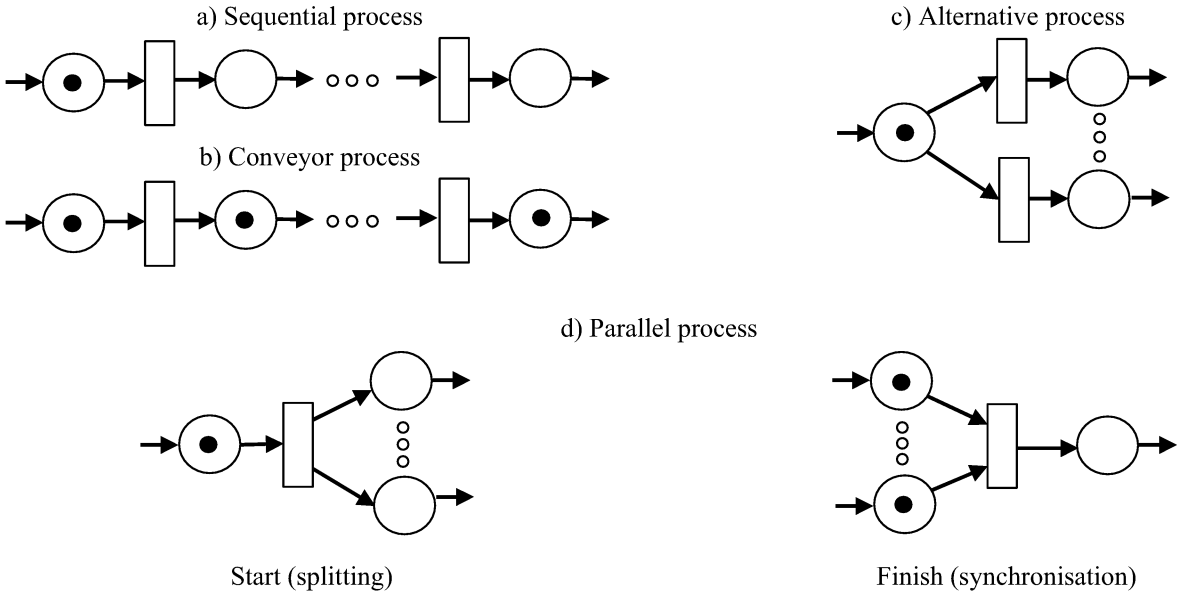
Petri net that has a tree structure describes a typical process of an article production. Article consists of units, units consist of lesser units and at least units consist of details. A tree represents this obvious hierarchy. An example of such tree is presented in Fig. 1. Only synchronisation relation is used in this model. For example, to manufacture unit "car" such units as "chassis" and "cabin" are required.



Figures 1: Technological scheme

It should be noted that even such hierarchical Petri net is more powerful than PERT net, as it represents a conveyor character of production that is reflected with marking and state of transitions. The flow of tokens moving through the net models the processes of production for a few units of the article.

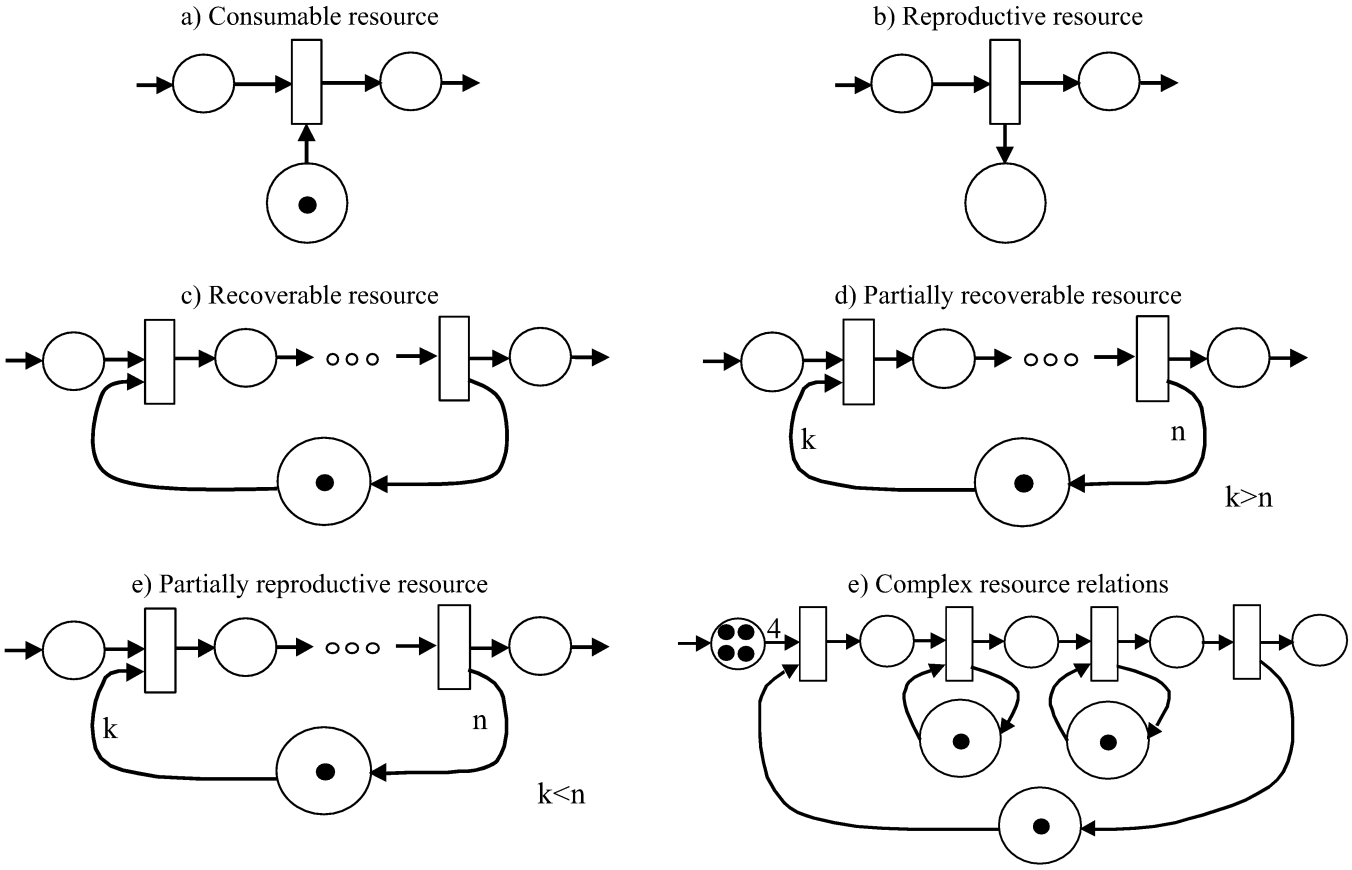
In Fig. 2 the basic types of processes represented by Petri nets are shown. The simplest is a sequential process a); a conveyor process b) is obtained from a) with marking of all places; an alternative process c) describes mutual replaceable units; a parallel process d) describes a constructive hierarchy of the manufacturing.



Figures 2: Types of processes

Parallel processes representation consists of the splitting of process, processes' descriptions and synchronisation. In most cases the splitting operation is not considered and we have a set of synchronisations of initially asynchronous branches. For example, Petri net with a tree structure (Fig. 1) is a set of synchronised jobs.

In Fig. 3 the basic types of resources' relations represented by Petri nets are shown. Consumable a), reproductive b) and recoverable c) resources are the most widespread. Consumable resources represent materials that will have been used to produce details. Reproductive resources correspond to manufactured units. Recoverable resources represent equipment, for example, machine tools.



Figures 3: Types of resources

More rarely such types of resources as partially recoverable d) or partially reproductive e) are used. In the case d) we can represent the wear of tools. The case e) is rather more sophisticated; it describes tools that are used for production of the same type tools. For example, lathe is used in the manufacturing of new lathes.

Moreover, Petri nets allow the representation of the complex multilevel resource relations also. So the net in Fig. 3 f) may be considered as the description of production with the usage of cassette with capacity of four details. Details are located in cassette for two next operations' execution: degreasing and dying.

Therefore, Petri nets are adequate models of discrete production processes.

FEATURES OF OPERA-TOPAZ

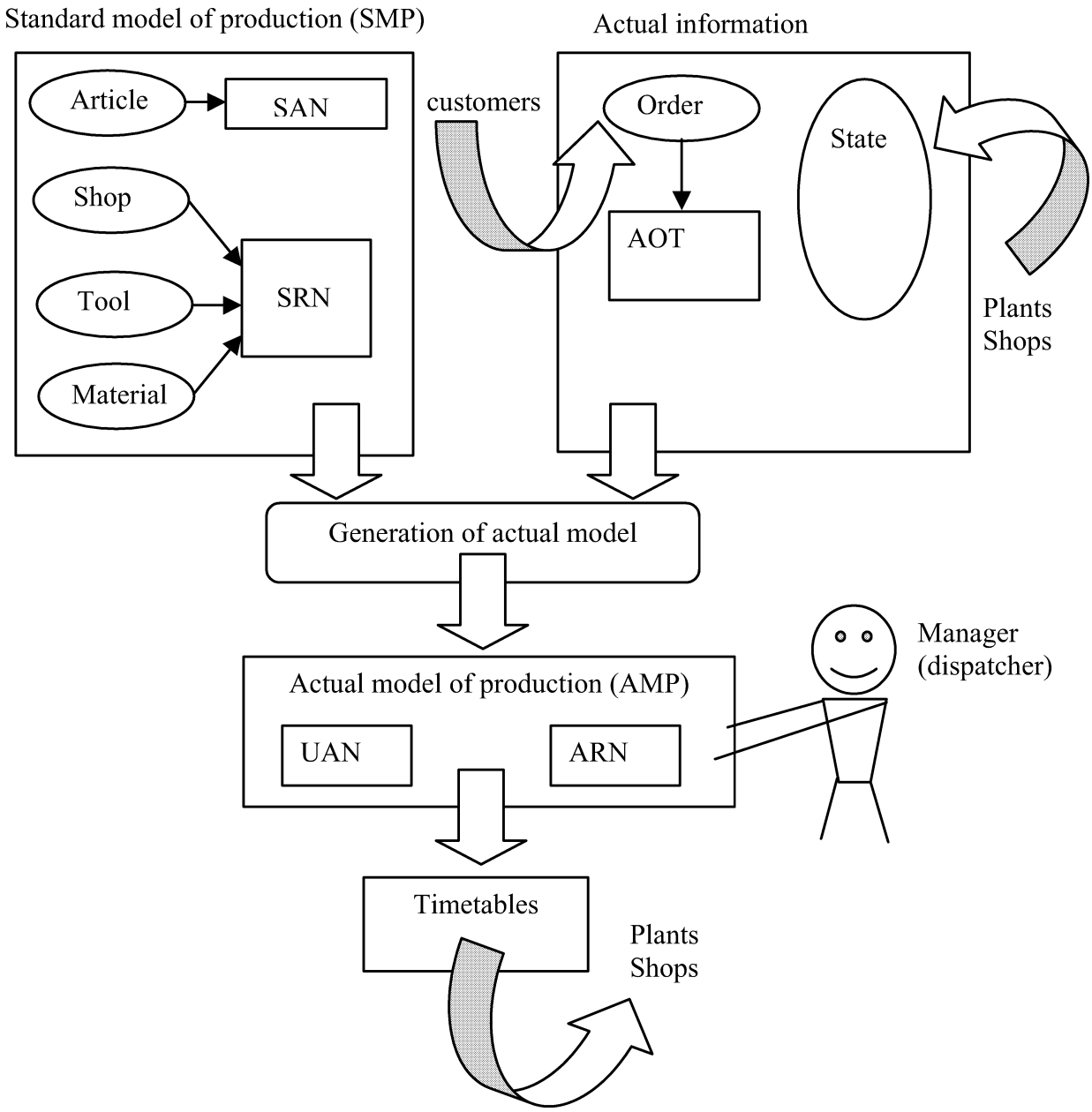
Enterprise CAM software system Opera-Topaz was described in details in (Sleptsov and Zaitsev 1991). It is based on R-technology of the modelling (Sleptsov and Zaitsev 1988) and timed Petri net (Zaitsev and Sleptsov 1997) model of production. The model is applied for a clear visual presentation of the current state of the production and for automatic and automated generation of timetables. Timetables may be represented in the standard form used in the usual process of shops' management.

The basic features of Opera-Topaz are:

- automatic and automated construction of Standard Article production Net (SAN);
- description of the resources of the plant (shops, equipment, tools, materials) with Standard Resource Net (SRN);
- maintenance of Article Orders Table (AOT) with the dates of the pass;
- automatic construction of the general actual model of production consisting of Unit of Article production Nets (UAN) connected with Actual Resource Net (ARN);
- automatic generation of quasi-optimal timetables of production;
- visual tools for presentation of current state of production and special tools for automated editing of timetables.

Now we explain features mentioned. For the scheduling purposes the Actual Model of Production (AMP) is used. This model is constructed by Opera-Topaz on the fly out of Standard Model of Production (SMP). The process of model generation is shown at Fig. 4.

SMP consists of SAN for each manufactured article and SRN that describes the structure of resources of the enterprise. There are two ways of SAN development: with the special visual graphical editor of Petri nets and via tools for automatic creation of Petri net. In the last case information about the jobs' following relation is extracted from the technological schemes.



Figures 4: Models generation and usage

In the simplest case SRN consists of the lists of shops and its resources. Moreover, it includes the information about variation of resources' amounts in time.

The major key for AMP generation is AOT. Information about actual amount of resources and actual state of processes is used also. On the base of this information the current state of the production is reflected in clear visual form.

Dispatcher disposes a set of forms for the reflection of current state of production. He may use such traditional forms as tables and Gant diagrams. Moreover, he may use Peri net on which the current state of production is represented in vivid form. Various colours reflect different states of jobs and resources. For the executing jobs the corresponding transition is filled with the part of rectangle equalling to the percentage of job's execution.

PECULIARITY OF TIMETABLES GENERATION

In the ideal case we have to generate an optimal timetable. Of course, we may describe the formal task of optimisation. Though this description is hard enough also. There are a lot of conditions that may be represented ether as criteria or as restriction.

For example, we have a list of the pass dates for units of articles. In any case we may consider the penalties for violation of the pass dates and include these penalties in the criteria that will be minimised. In other case, especially for military production, we have precise hard pass dates and have to consider them as the restrictions.

So, firstly, special tools for automated description of optimisation task is required. These tools have to process the standard list of conditions and extend this list with new required conditions. Each condition has to be represented either as restriction or be included in criteria with its own weight coefficient.

But unfortunately there are not known universal formal methods to solve described above optimisation task with the exception of the complete look over. But the complexity of the complete look over is exponential in time. It requires enormous amount of time even for small dimension tasks and, therefore, is not practically applicable.

The classical practical approach consists in the priority rules usage for the quasi-optimal timetables construction. We applied this approach to timed Petri net model. To generate a timetable we supply net with the priority rules, which allow solving conflicts among fireable transitions. From the formal point of view we extend state equation with the expressions that describe priority rules. Then we use this extended state equation to generate timetables.

It is enough nontrivial task to maintain a set of priority rules with correspondence to conditions that define optimisation task and may be included ether in criteria or in restrictions. Our goal is to extract from the practice and accumulate such a correspondence that allows generating quasi-optimal timetables enough close to optimal in a short time.

The best result was obtained in Opera-Topaz with criteria based on kan-ban technology (Sleptsov and Jurasov 1986) with the hard pass dates and soft bounds of accessible resources.

But the automatically generated plan is not a dogma because an alive dispatcher could not express in any criteria and priority rule all his skill. We give convenient tools to dispatcher for revision and correction of plans represented with Gant diagrams, tables or Petri nets. Petri net model defines the restrictions for manual adjustment of plans.

CONCLUSION

Therefore, the basic features of enterprise Petri net based CAM software Opera-Topaz were studied. The successful usage of Opera-Topaz during more than ten years at military plants of Ukraine confirms the practical value of ideas embodied in this tool.

REFERENCES

Sleptsov, A.I. and A.A. Jurasov. 1986. *Automated development of management and control systems for flexible manufactures*, Kiev, Tehnika, 160p. (in Russian).

Sleptsov, A.I. and D.A. Zaitsev. 1988. "R-technology of modeling for parallel structures and processes", Program-methodical and program-technical complexes for CAD and CAM: Proceedings of Scientific-technical conference, Izhevsk, 62-63. (In Russian).

Sleptsov, A.I. and D.A. Zaitsev. 1991. *CAM Software Opera-Topaz*, Donetsk, DPI, 68p. (in Russian).

Zaitsev, D.A. and A.I. Sleptsov. 1990. "Visualization of production processes in tool system of dispatcher in machine-building enterprise", Autometria, no. 4, 90-93.

Zaitsev, D.A. and A.I. Sleptsov. 1997. "State equations and equivalent transformations of timed Petri Nets", Cybernetics and System Analysis, vol 33, no. 5, 659-672.

Soleil: modeling a health system

Ruben González-Rubio
Département de génie électrique et de génie informatique
Université de Sherbrooke, Sherbrooke, QC, J1K 2R1 Canada
Ruben.Gonzalez-Rubio@USherbrooke.ca

Lourdes Zubieta
Williams School of Business
Bishop's University Lennoxville, QC J1M 1Z7 Canada
lzubieta@ubishops.ca

Abstract

We present a discrete-event macro model of a geographical region where a population has access to general practitioners (GP) and to a hospital emergency room (ER). We designed a queuing model to simulate a pool of primary care clinics of a region where patients interact with doctors if available; otherwise they may end in a hospital emergency room. Our simulation model can provide an insight on the expected results of different scenarios in terms of number of patients treated in cabinet, waiting to be treated, or treated in the emergency room. Each scenario can be associated with a cost for evaluation purposes. Then, the model could be used as a powerful aid for debate and discussion. The paper presents a first model as a feasibility proof of the concept and includes a critical view of the input data requirement and computing simulation costs of an more comprehensive system. This work represents a first step of a larger project, that we call Soleil. The aim of Soleil is to propose a decision aid system for health care management.

1 Introduction

Many factors are placing health care systems under increase pressure. Ageing populations, demographic changes and the increasing costs of medical care are mentioned along with unacceptable length of waiting lists. The phenomenon of "corridor beds", once considered a fire hazard, is now common in many developed countries like Canada. Examples of patient dissatisfaction and staff frustration are regularly found in the front page of newspapers [Bom04] and even recent movies [Arc03] and [Pou04] depict with not much exaggeration the current situation.

The management of health care has always been a very complex task where limited resources and expensive new technologies interact with the legal requirement of providing adequate health services to a population. Cost rationalization is present everywhere and health care systems are no exception. Decision makers must dictate a rationale for the improvement of such services (reducing waiting times, allocating resources, etc.) and provide explicit justification for the choices made. The main difficulty is to understand and quantify the interactions between the various system components; for example, assessing the consequences of adding a doctor in the emergency room instead of hiring another general doctor to work in a clinic.

This paper presents a first step in the development of the Soleil project. The innovation of this project is the modeling of a health care system at a macro level with the main purpose of building a decision aid system for health care management. The conceptual model is implemented as a simulation queuing system where human resources are pooled. The main benefit of a computer simulation tool like Soleil is the better understanding of complex processes with its inter relationships, and the ability to mimic the variation inherent to the process over time. Also, having a validated tool that simulates a current system will permit to introduce possible changes and evaluate their impact on cost and on system performance.

Computer simulation in the management of health care services is a field where researchers have studied specific areas in hospital operations like emergency rooms [BLTT04], hospital patient and resident scheduling [RWD68],[DKD+96], bed planning [Dum84] and bed allocation [Vas85],[Wri87]. Others have dealt with wider issues like health manpower planning [SPD78], evaluation of priority policies in organ transplants [HTED00],

vaccine effects [BJAK00], and mobile examination centers [VF03]. In these and other cases, the authors built a simulation model of the current system, identified inefficiencies, and once validated, explore suggestions for improvement that help to plan changes to the system under study. We took the approach of a macro model to satisfy the needs of regional health care managers instead of local administrators.

The organization of the paper is as follows. The conceptual model is presented in section 2, the questions to answer and the scenarios to analyze are described in section 3; section 4 presents the simulation results of various scenarios and the validation of the base scenario. Next, we discuss in section 5 the characteristics and issues that a project like Soleil must consider and finally section 6 summarizes our findings and sketches the road ahead.

2 The model

Following a general scheme [Wol02], we can identify on the demand side of the system, patients requesting a service on their own and patients referred by a doctor. Demand is therefore difficult to control and we can consider it as exogenous. On the supply side, we have medical and paramedical staff, material resources, budgets to pay and or acquire the resources, medical procedures approved by professional colleges and associations, collective agreements and legal obligations to provide health care services. The supply side is thus controllable, at least in principle, but highly constrained.

Usual performance measures on the quality of services are average waiting queues (the smallest the better) and service capacity utilization (the higher the better) These opposing measures have to be balanced in an effort to match supply with demand.

There are complex interactions between the different input flows and resources, output services, and the players in a health care system (budgets, x-rays, surgeries, doctors, operating rooms, patients, nurses, clinical labs). Moreover, elements on the supply side may depend on distinct government offices, making the matching of supply to demand an extremely complicated one. It is the purpose of this work to provide a tool for better understanding of the intricacies of a health system that permits to quantify the costs and benefits of different scenarios.

The model presented is a macro model that simulates the flow of patients during a period of time through a regional system of primary care. The conceptual model is illustrated in Figure 1. Our primary care model is a very simplified one and still, it can reproduce some aspects of reality, as we show later. The system resources are a given number of doctors (GP) that treat patients with appointments in their cabinets, an appointments call center (RV) having a fixed number of open lines and a fixed number of time slots available per day, and a hospital emergency room (ER) where patients are treated as they arrive. The model assumes that all the resources of the same kind work together as a pool: cabinet GP doctors, operators, and ER doctors. Patients call until they can get an appointment or until they end in the ER of a hospital. The system has a population to deserve and patients come for consultations and examinations only. We do not include in the model neither the patients coming from other regional areas, nor those that arrive to the emergency room due to accidents or other reasons. We include the emergency room *o*nly as a buffer that receives patients who either could not get an appointment with a GP or have an urgent health problem.

2.1 The calling population

We consider a clinic that provides primary care to a given population. The population requires on average, a given number of consultations per year per person. We estimated μ, the average number of visits per person per year, according to recent statistics. This parameter is relatively easy to obtain for any geographical area and if it changes over time, it is possible to include the changes in our simulation model. We assumed that the actual number of visits for each patient follows a uniform probability distribution whose mean value is μ, This is a key assumption that should be verified in a real application.

This model does not consider different groups of individuals within the population and does not distinguish among first time illnesses and recurrent illnesses, in order to keep the model simple. For all the individuals in the population it considers the aggregated number of cabinet visits to a GP doctor per year according to the same probability distribution.

The patient flow within the model is as follows (see Figure 1). When a patient calls and the clinic is open, he either gets an appointment for the same day or is asked to call again the next working day (NoGP) because no GP is available. If the clinic is closed, the patient also has to call the next working day for an appointment (NoGP). In the allocation of appointments there is a priority queue where patients already seen before have a higher priority than patients who had not yet obtained an appointment with a doctor.

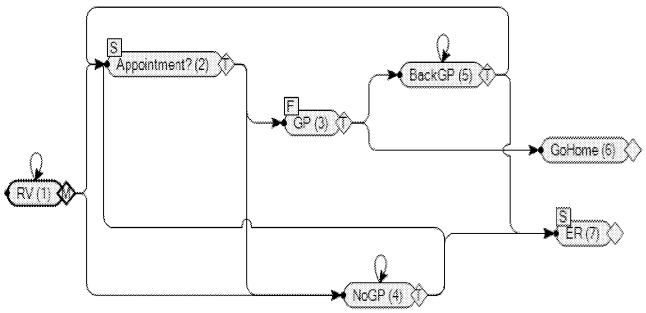


Figure 1: The conceptual model

If a time slot is available, a doctor sees the patient and then there are two possibilities: either the patient goes home (GoHome), in fact he goes out of the system for the rest of the year, or needs to come back for other appointments (BackGP). Between two visits to the doctor, a patient with a given probability, will go to the emergency room (ER).

If patient can not have an appointment, he/she will call again and again until he gets one or his health condition worsens, in which case the patient will have to go to the emergency room (ER). As explained in the next section, we assume that a visit to ER has a higher probability for patients not treated (because of the impossibility to get an appointment) than for those followed by a doctor.

2.2 System Resources and Capacity

There are three resource groups: the GP doctors, the ER doctors and the operators receiving the patients calls (RV). Depending on the number of doctors assigned to the clinic and their working hours, we calculated a number of appointments per day that we keep fixed over the simulation period of time. For example, if there is a continuous presence of 3 GPs we assume that they can give 104 appointments per day, whereas if there are 4 GPs this number increases to 117. We also assume a fixed number of 2 operators available to receive calls during a working day of 10 hours. Each appointment with a GP lasts a variable amount of time, that we assumed to follow a uniform distribution between 15 and 25 minutes. This variability in the lenght of a visit makes a day of variable length for the clinic.

The capacity of the system is thus determined by the random length of consultations, a fixed number of available doctors, a maximum number of appointments per day, and the probability distributions of the inter arrival times of patient requests and of services times (consultations and calls for appointment). Whit this base model of operation is possible to simulate different scenarios.

3 Questions and scenarios

The project Soleil should help to answer many different kind of questions and it is in the realm of the health care managers to determine the questions and the important characteristics of the system to be modeled. Here we present only a limited set as an illustration of the possible benefits of Soleil, even if we deal with an imaginary and very simplified network. The main question deals with the marginal benefits of increasing the population size with/without increasing the resources in terms of the annual number of:

- patients denied an appointment;
- patients treated, not requiring further services;
- patients in treatment, requiring further services;
- patient overflow at the emergency room.

It is obvious that if a clinic has to attend *with the same resources* the medical needs of a population that increases in size or in demands, we expect to see a higher number of denials and a higher number of people going to the emergency room because their condition requires it or even because they are tired of waiting for an appointment at the clinic. The difficulty lies in determining the magnitude of such increase and the time it may

take to see the impact in the ER. Similarly, adding a GP to a clinic will certainly reduce the number of denials, i.e., patients who cannot have an appointment, it will also reduce the number of patients waiting in the ER and will increase the number of patients whose health problem is solved; however, it is difficult to estimate in advance the magnitude of these changes and the time it may take to see an impact in the daily statistics.

3.1 Parameters and Input Distributions

The input distributions and parameters are given in Table 1. The task number indicates the place in the model network diagram (see Figure 1) where the parameter is required. All parameters are in minutes with the exception of the the number of visits required per year and the conditional probabilities to an ER visit. Note that the latter depend on whether or not the patient is being treated by a GP.

Random Variable	Task	Distribution
Requests inter arrivals	1	U[10,20]
Get appointment	2	N(2,0.5)
GP visit	3	U[15,25]
Call again	4	N(60,6)
Next GP visit	5	Normal(6000,60)
Nb of GP visits	1	U[1,3]
ER visit	7	Triang(10,30,60)
P[ER visit | seen a GP]	3	0.01
P[ER visit | not seen a GP]	4	0.10

Table 1: Base Scenario Parameters

3.2 The scenarios

We started the simulations with a base scenario, S3GP-Base, that reflects the normal operation of a clinic deserving 15,000 people that also have access to an emergency room if necessary. According to local health care administrators [Bom04] eight doctors are able to take care of a population of that size. Now, assuming that each doctor works at the clinic for an average 6.5 hours per day, during 200 days per year[1], the daily average number of doctors in the clinic should be around 3. This means a constant presence of 3 doctors, not necessarily the same, at any time during the 10 hours that the clinic is open.

At the emergency room, the actual number of doctors and patients in the waiting line are certainly higher than those provided in our results. Our aim here is not to recreate the activity at the ER but to show the impact of inadequate resources at the cabinet level on emergency rooms.

Scenario	nGp	Population
S3GP-Base	3	15,000
S3GP-37.5	3	37,500
S4GP-37.5	4	37,500

Table 2: Description of Scenarios

In the second scenario S3GO-37.5 the same resources must provide primary care to a population that boomed to 37,500 persons. The idea here is to test the impact of inadequate resources on the ER. The population increase may be the result of doctors retiring or moving to other areas. Another way to increase the demand for primary care is an ageing population, a natural disaster, a plague or an epidemics. All this events will impact the number of patients asking for an appointment and all will generate an increase in the inter arrival time of patient calls. Any increase in the demand side will certainly create a situation and will overload the emergency room.

The third scenario S4GP-37.5 will evaluate the impact of having an extra GP in the clinic, so the number of doctors present anytime in the clinic is increased to 4, with the population remaining at 37,500.

In all cases we assume that there are 2 operators receiving the patients request for appointments. The scenario characteristics and labels are summarized in Table 2:

In terms of the simulation model, to reflect the increase in demand we only need to reduce the inter arrival time of requests for appointments from a U[10,20] to a U[1,10] so that patient calls arrive on average every 5.5 min instead of 15 min as before. To increase the medical resources we just have to modify the value of a constant, the number of doctors in the clinic (nGp).

4 Results and validation

We now discuss the simulation results for each scenario. As the length of each working day at the clinic is variable (due to the variable length of appointments), we run all simulations for 300,000 minutes, so we could determine a warm up period with no recording of statistics, and enough working days to simulate one year of operation.

[1]We are not saying that a doctor works only 6.5 hours per day 200 days. Normally a doctor has other duties outside the clinic.

4.1 Base scenario

Figure 2 shows the dynamics of the clinic in terms of both the daily number of new patients entering the system (Nb_in) and the daily number of patients leaving the system after their last visit to a GP (Nb_out). After 10 days, we can observe stability in these two variables.

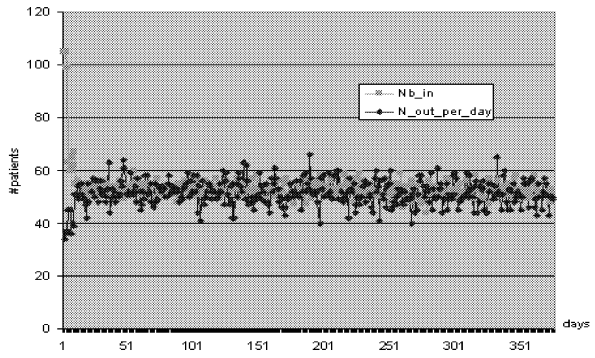


Figure 2: Normal Operation of Base Scenario

Based on the results for 365 days, we estimated daily averages and standard deviations and built 95% confidence intervals, all shown in Table 3. We found, on an average day, 52.56 patients requiring a first visit (Nb_in), 51.66 patients not requiring further visits during the year (Nb_out), 2.6 individuals unable to get an appointment (Nb_denials) and less than one person going to the emergency room.

	Base Scenario		
Variable	Average	St Dev	95% CI
Nb_in	52.56	2.52	(52.30,52.82)
Nb_out	51.66	4.50	(51.19,52.12)
Nb_denials	2.60	1.19	(2.48,2.72)
Nb_ER	0.79	0.86	(0.702,0.878)

Table 3: Daily averages

Table 4 shows the annual estimates based on daily results obtained by multiplying the daily confidence intervals of table 3 by 300, the number of working days in a year. We see that the clinic provided primary care to approximately 15,000 individuals during one year, between 742 and 816 times somebody was unable to get an appointment and that between 210 and 263 patients had to go to the emergency room. These results partially validate the hypothesis that the clinic can provide adequate primary care to a population of 15,000.

Base Scenario	
Variable	95% CI
Nb_in	(15691.26 , 15846.45)
Nb_out	(15358.23 , 15635.21)
Nb_denials	(742.8 , 816.2)
Nb_ER	(210.47 , 263.30)

Table 4: Annual volumes

4.2 Second scenario: increase population size

The second scenario S3GP-37.5 simulates how the current resources may handle a major increase in the population size. The population was increased to 37,500 people, all having access to the clinic. The model was changed to reflect this increase by modifying the parameters of the inter arrival of patient request. From an average of one patient every 15 minutes, uniformly distributed between 10 and 20 minutes, we moved to a U[1,10] with a patient request every 5.5 minutes on average. The simulation results are shown in Figure 3 where we can see a burst in the number of denials that went from less than 3 to an average of 282 per day. This created a major impact at the emergency room, where the number of patients went from less than one to 30 per day on average. The current resources were therefore appropriate for a 15,000 population, but they were operating at a high utilisation rate.

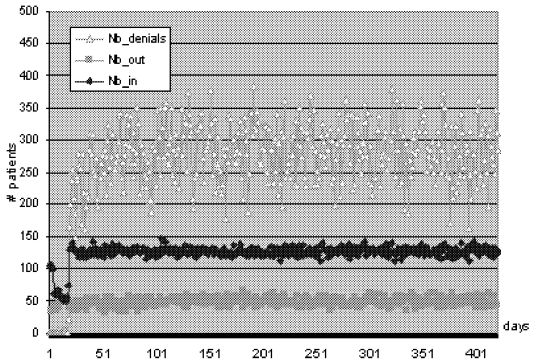


Figure 3: Impact of a population increase

In order to find out the time required to see the impact of increasing the the population size, we changed the inter arrival parameters at day 18. We can see in (Figure 4) the immediate burst in demand and the fact that it took another 10 days before the daily values start to get close to the values shown in Table 5.

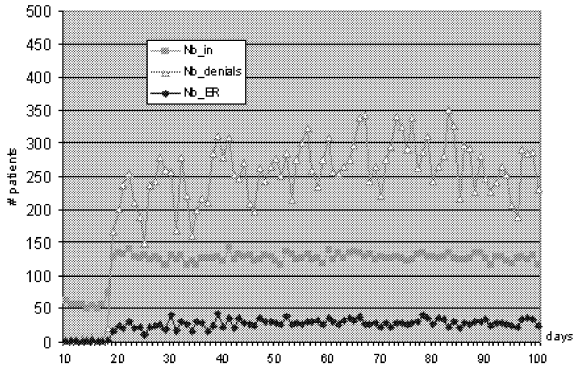


Figure 4: Detail on the impact

*S*cenario S3GP-37.5	
Variable	95% CI
Nb_in	(126.8 , 128.0)
Nb_out	(48.7 , 50.0)
Nb_denials	(279.6, 288.4)
Nb_ER	(29.7 , 30.7)

Table 5: Daily averages with 37,500

We can conclude that increasing by 2.5 times the demand of the population generates on average twice as many patient request per day but it creates an average of 30 patients per day extra in the emergency room; another consequence is that on average, 282 people per day are unable to get an appointment.

4.3 Third scenario: one extra GP

We now considered the same population as in scenario 2 and increased the number of doctors available at the clinic from 3 to 4 during the working hours. This represents going from 104 consultations per day to 117. Daily results are shown in Table 6.

*S*cenario S4GP-37.5	
Variable	95% CI
Nb_in	(107.1 , 108.2)
Nb_out	(54.1 , 55.8)
Nb_denials	(183.5, 191.9)
Nb_ER	(20.5 , 21.5)

Table 6: Daily averages with 4 GPs

Table 6 shows the impact of an extra GP in the clinic. The results show that this increase is not largely felt in the number of patients not requiring more visits (Nb_out) as it went from 51 to 55 per day, on average. However, we can see a net decline of around 100 less denials per day and a decrease of almost 30% of the patient overflow to the emergency room. The average number of patients gone daily to the ER passed from 30 to 21 with an extra doctor at the clinic.

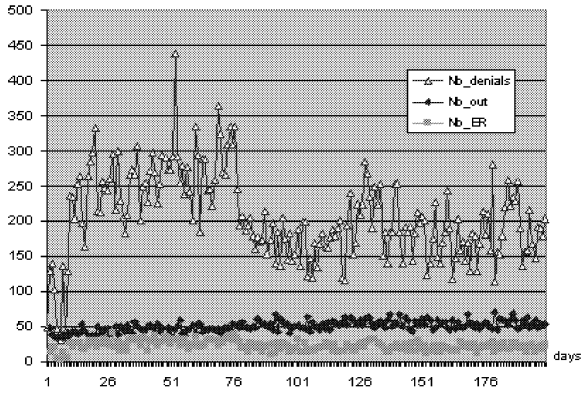


Figure 5: Impact of an extra doctor

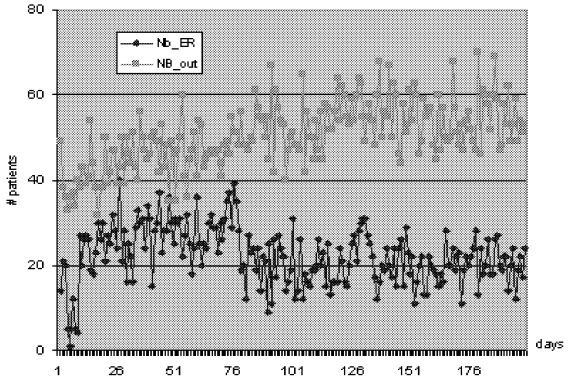


Figure 6: Detail on the impact

Again, we can see how long it takes for these results to be captured in the daily statistics by observing Figures 5 and 6. The extra doctor arrives on day 79 and in less than 2 weeks we can notice the improvement in the number of denials and in the number of patients that arrive to the emergency room. Other scenarios are also possible to implement.

5 Towards a system Soleil

The scenarios presented previously show the value of the approach, even though the conceptual model does not

represent a complete health care system as we want to develop in the Soleil project. We learn many things by building and implementing the model and scenarios and this knowledge must be used to develop a more complete system.

From a conceptual view, an obvious refinement to the model is to include a lab facility with x-rays and many types of blood and urine tests. After a visit to a GP, a patient must take an appointment for the lab, go for the tests and wait to have the lab results before going back to his/her doctor. Other refinement is to model a bigger system with specialists of a city or a large geographical region, even a province. Each refinement though will increase the model complexity and the data requirements.

Regardless of the model complexity, the system must be shared by both technical and non technical users; technical users are interested in numerical results that they can interpret whereas non technical users need to ask questions and get answers without any previous knowledge on queuing theory or probability or statistics.

Let's see what we can observe in the practical example that we built:

- We have two kind of parameters: the first are derived directly from sources like government policies and statistics; the second are those that need to be validated by an expert or by field studies.
- Simulations are programs that need to be validated and verified and the system must help here by analyzing results automatically (as much as possible).
- Even if simulation appears as a inexpensive solution compared with building a whole system, simulations take time to run, and once that the simulation program is tested and verified, it is necessary to run a program for enough (simulated) time to get results that we can trust. In our example, one run of any scenario takes about 30 to 45 minutes in a PC with an Intel Pentium(R) 1.7GHz with a memory of 2GB. A more complex model may take hours.
- Simulating scenarios takes time. A base scenario was used to simulate the present situation. Then a new scenario is proposed in which we add/remove a ressource at a particular place. Then a new run is necessary. A new scenario is developed by changing parameters to the original set and the number of runs increases very fast. If we change 2 parameters by 2 values each, for example 3 or 4 GPs and 1 or 2 specialists in the emergency room, that means that we need to run the simulation 4 times. The number of simulation is a function of the number of parameters and the number of values, so scenarios must be chosen carefully.
- The simulations were done using an off-the-self PC program, namely Micro Saint Sharp Version 2.0 [Mic04]. Programming the simulation is not a straightforward task because there are many dialogs to enter and it is necessary to test and look for bugs in all those dialogs. As many other commercial programs, it has the advantage of displaying the model in a graphical way which helps to see the model at work, but this takes processor time.

We have outlined the data requirements for the building and validation of the model as well as some of the technical characteristics that the complete system must have. In fact, we used our experience in developing software for no technical users. Following software engineering guidelines, in particular making iterative development with an active role of the system users. First it is necessary to explore multiple solutions as shown in this paper. Then, it is possible to build or make extensions by applying the knowledge gained in the previous iteration. We have already successfully developed a system to build timetables: Diamant [Gon00].

To build and adapt a health care management system, we suggest to base our developments on simulation programs built around existing libraries. This can eliminate some problems linked with a commercial program and make it more efficient an flexible.

The program could take a XML file to indicate which parameters will be used in the run. XML allows to identify the parameter name and the value in a very readable file. The system must contain assertions as a way to verify and validate results by itself. This increases the user confidence in the software.

To improve the efficiency of the system, we need to avoid redundant computations. During scenario evaluation many computations were done redundantly, as scenarios differ on small details only. For example generating the patients arrivals was repeated many times, as some scenarios have the same patient arrival process.

6 Conclusions and future work

In this paper we presented a first step towards a management aid system for regional health care planning processes. The use of a simple example shows some of the benefits that such a system can provide. There are still many unsolved problems such as the right level of abstraction to retain in the daily operations and in the

aggregated behavior of the human components. Further studies are needed to obtain the required data and specifications for a more complex system. We are persuaded that further developments of Soleil will conduct to a very successful system.

References

[Arc03] D. Arcand. The Barbarian Invations, 2003.

[BJAK00] D. C. Barth-Jones, A. L. Adams, and J. S. Koopman. Monte carlo simulation experiments for analysis of HIV vaccine effects and vaccine trial design. *WSC*, 2000.

[BLTT04] S. C. Braisford, V. A. Lattimer, P. Tarnaras, and J. C. Turnbull. Emergency and on-demand health care: Modelling a large complex system. *Journal of the Operational Research Society*, 55:34–42, 2004.

[Bom04] D. Bombardier. Aucun médecin de famille de plus. *La Tribune, Sherbrooke*, Décembre, 8 2004.

[DKD+96] R. S. Dittus, R. Klein, D. J. DeBrota, M. Dame, and J. Fitzgerald. Medical resident work schedules: design and evalaution by simulation modeling. *Management Science*, 24(6):891–906, June 1996.

[Dum84] M. B. Dumas. Simulation modeling for hospital bed planning. *Simulation*, 8:69–78, 1984.

[Gon00] R. Gonzalez-Rubio. Diamant Web Site, Groupe ϵXit, 2000. http://tictacserv.gel.usherb.ca:9606/diamantbug/index.jsp.

[HTED00] A. M. Harper, S. Taranto, E. Edwards, and O. P. Daily. An update on a successful simulation project: The UNOS liver allocation model. *WSC*, 2000.

[Mic04] Micro Analysis and Design. *Micro Saint Sharp*, 2004.

[Pou04] J-F. Pouliot. Seducing Dr. Lewis, 2004.

[RWD68] G. H. Robinson, P. Wing, and L. E. Davis. Computer simulation of hospital patient scheduling systems. *Health Services research*, 3:130–141, 1968.

[SPD78] C. R. Standridge, A. Pritsker, and H. Delcher. Issues in the development of a model for planning health manpower. *Simulation*, 31:9–13, 1978.

[Vas85] G. A. Vasilacopoulos. Simulation model for bed allocation to hospital inpatient departments. *Simulation*, 45:233–241, 1985.

[VF03] Osidach V.Z. and M. Fu. Analyzing incentives and scheduling in a major metropolitan hospital operating room through simulation. *WSC*, 2003.

[Wol02] E. Wolstenholmed. Patient flow, waiting and managerial learning - a systems thinking mapping approach. *Manuscript Cognitus Management Consultants*, 2002.

[Wri87] M. B. Wright. The application of a surgical bed simulation model. *European Journal of Operational Research*, 32:32–36, 1987.

A SIMULATION SYSTEM WITH NEUTRAL DATA INTERFACE FOR SHIPBUILDING OPERATION

Guodong Shao
Charles McLean
Swee Leong
Manufacturing Simulation and Modeling Group
National Institute of Standards and Technology (NIST)
100 Bureau Drive, MS 8260
Gaithersburg, Maryland 20899-8260, U.S.A.
E-mail: gshao@cme.nist.gov

KEYWORDS
Simulation, XML, UML, Data Model, Shipbuilding.

ABSTRACT

Simulation technology is underutilized in the shipbuilding industry. The development of neutral data interfaces between simulation and other software applications, could make information sharing within a shipyard more effective and efficient, and hence promote the utilization of simulation in the shipbuilding industry. A generic shipbuilding simulation system with such a neutral data interface has been under development at the National Institute of Standards and Technology (NIST). An interface information model represents shipyard data and facilitates data sharing among the production system; the scheduling system and the simulation system have been developed. In this paper, the objectives and requirements of a generic shipbuilding simulation and the data interface are discussed. An overview of a generic simulation system of shipbuilding operations is presented.

DISCLAIMER

No approval or endorsement of any commercial product by the National Institute of Standards and Technology is intended or implied.

INTRODUCTION

Scientists and researchers at NIST have worked on the U.S. Navy's National Shipbuilding Research Program (NSRP) project that is a collaborative effort among U.S. shipyards, government, industry and academia. The shipyard production process technology is a major NSRP initiative that addresses all production processes used to transform raw material, components, and equipment into completed products. The simulation of shipbuilding operations was carried out as part of this initiative; simulation models and data interface software were developed at NIST for the NSRP project (McLean and Shao 2001). However, that simulation model only solves one specific shipyard's problems; the shipyard's proprietary data was collected and used as input of the simulation model. Often times, the shipyard might not be willing to share the model and data with others, and other shipyards might have different methods to perform the time study for the work contents calculation. To build a generic simulation model with a neutral data interface structure will benefit shipbuilding industry in general. The mission of the generic shipbuilding simulation system is to reduce the cost and establish U.S. shipbuilding industry's international competitiveness.

A neutral interface could help to reduce the costs associated with simulation model construction and data exchange between simulation and other software applications, and thus make simulation technology more affordable and accessible to a wide range of potential shipbuilding industrial users. Currently, most of the shipyards do not take advantage of using simulation technology because of various difficulties and obstacles associated with model development, data collection and translation. Shipyards typically do not have enough resources with the appropriate technical qualifications. They cannot afford to develop customized simulations of their own operations and data translators. The data translators were needed to import their data from other production systems. Even in the case that they have customized simulation models for their shipyard, exchanging production data with other shipyards is still very difficult because of the different data format and structure from yard to yard. Furthermore, simulators are not designed to use traditional shipyard data in its native format, so models and data import routines usually must be developed for different applications from scratch. This is even true for simulation software vendors. If they were to try to develop generic simulation models, they would have the same problem, as there are no standard formats available for the models data. Custom data translators would still need to be developed at a considerable expense if someone wanted to input a specific shipyard's data into one of these simulators.

The NIST-developed neutral shipyard information model based on the NIST machine shop data specification (McLean, et al., 2005) is used to try to standardize the shipyard internal data structure. The information model provides neutral data interfaces for shipyard simulation. The model is developed using both the Unified Modeling Language (UML) (OMG 2003) and the eXtensible Markup Language (XML) (W3C 2000). The neutral shipyard information model is used to describe the important information about the shipyard production facilities and processes, to provide data for running the simulation models and to exchange data between simulation and other

manufacturing applications. Other than the traditional way to modify process parameters manually in discrete event simulation tools, this XML based simulation interface specification enables users to change the data more efficiently. The objective of the information modeling effort is to develop a computer-interpretable representation that allows for exchange of information in a shipyard environment. The information model satisfies the following needs: support data requirements for the entire shipyard production life cycle, enable data exchange between simulation and other manufacturing software for shipbuilding industry, provide input data for the shipbuilding simulations, and support testing and evaluation of the shipyard's manufacturing software applications. The information model provides mechanisms for describing data about organizations, calendars, work, resources, parts, process plans, schedules, and work orders, and layouts within a shipyard environment. A data editor will be used for entering and editing raw production data from a shipyard.

SHIPBUILDING SIMULATION SYSTEM ARCHITECTURE

In Figure 1, a system architecture of a shipbuilding simulation system with neutral data interface is presented. The intent of this system is to be able to collect and edit shipyard production data and exchange information with other production software applications within the shipyard or among shipyards.

The architecture for the generic shipbuilding simulation system is divided into the following components:

- Shipyard production data (in shipyard's format)
- Shipyard production data editor
- Neutral shipyard data files (in XML format)
- XML data processor
- A generic shipbuilding simulation model (Discrete event simulator)
- Simulation reports

Raw Shipyard Production Data: Typically four types of data files are needed by the simulation system: project management data (one for each project); schedule loading (one for each simulation run), labor configuration file, and resource configuration file. The primary sources for "raw" shipyard data normally reside in Microsoft Excel and project management applications such as Primavera or Microsoft Project, and sometimes in people's heads. By using Data Editor, these manufacturing data can be entered and interpreted into XML format. The project management data for the simulation includes project task decompositions, precedence relationships between tasks, resource requirements, timing data such as due dates and duration, other scheduling data such as project loading for the shipyard, project due dates, production area assignment options, and resource configuration data.

Shipyard Production Data Editor: The data editor is an application that provides the capability to create and modify XML documents that adhere to the shipyard data interface model. The shipyard data model has been developed using XML and an XML Schema. Any application that needs to interact with the data must be able to read in and interpret the data. The standard way for interacting with XML data is to use the World Wide Web (W3C) XML Document Object Model (DOM) (W3C 2003). DOM will read in XML data and convert it into internal data representations that can be directly manipulated by applications. The Data Editor application is able to create and modify shipyard production information. A XML data parser and data generator were developed to import and export data using the DOM library. The different functions of the data editor were implemented as a collection of Microsoft ActiveX Dynamic Link Library (ActiveX DLL) components. This allows any application that supports the Microsoft Component Object Model (COM) specification to interact with the data editor. A graphical user interface was developed for the data entry and population of the internal object structure and XML data file.

Neutral shipyard data files: This interface data file is a key element of the system; it will be the input of the simulation model. The data file provides a mechanism for configuring the shipyard data model and sharing data between simulation and other manufacturing software applications within the shipyard and among the shipyards. Other manufacturing software applications include the production scheduling systems, time study system, job-task production system, etc. XML is used to encode the data in the files. The file contains not only executable or computable data to be processed by the simulation, but also descriptive text that are intended for human interpretation. It also contains a network of cross-reference links between the various types of data required to plan and manage operations of the shipyard. It supports references to other external computer files and or paper documents that provide more appropriate mechanisms or

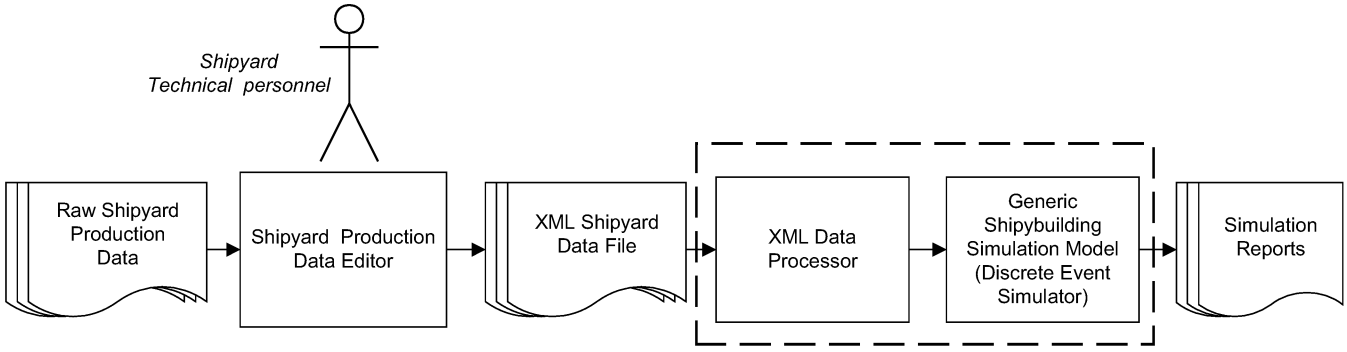


Figure 1: Shipyard Simulation System with Neutral Data Interface

standards for encoding or representing data, (e.g., part drawings). Subsets of individual data types, i.e., substructures, may be created, stored, and/or exchanged using the file.

XML data processor: This is a library of routines that handle the import and export of data in the prescribed XML format of the neutral shipyard data file. The primary function of it is to read the neutral shipyard data file in XML and translate the data into the internal object structure of the Discrete Event Simulation tool, in this prototype project, ProModel 6.0 was selected as the simulator. Therefore, ProModel needs to read in and interpret the XML data. By using the XML DOM, XML documents can be converted into a ProModel readable data structure such as integer and real. C++ and ProModel Xsub function were used to perform the task.

A generic shipbuilding simulation model (Discrete event simulator): It contains the commercial simulation software system, generic modeling elements. It is responsible for providing master clock functions, event queues and queue management, randomization functions based on embedded statistical distribution generators, and accumulating statistical data on resource state changes and work items. The ProModel software was used for simulation development, execution and analysis of simulation models. It was used for supporting analysis of shipyard layout, labor behavior, and ergonomics. The generic shipbuilding simulation system can be used as a tool to analyze the schedule impact of a new workload, evaluate production scenarios, and identify resource problems for shipbuilding in general. The simulation helps identify resource constraints and conflicts between competing jobs. The simulation model has a 2D graphical display. The simulation model is also responsible for managing the emulation of the operation of the shipyard and manipulating required data. This includes, but is not limited to:

- The production calendar and shift schedule
- The execution of orders, jobs, and tasks
- The enforcement of precedence constraints in process plan routing and operation sheets
- The sequencing of steps in the operation sheet associated with an individual task
- The availability of resources: work areas, equipment, tools, fixtures, cranes, employees, and inventory
- The allocation of resources to tasks
- The selection of statistical distributions to be applied to various processes and events.

DATA INTERFACE MODEL

To solve the different data formats currently available in the shipbuilding industry, we need a mechanism to represent the production information in a standard way. The information needed by simulation is often scattered in various sources like databases, product data management (PDM) systems, computer generated drawings and spreadsheets on different computers in the facility or people's head. More than one Commercial off-the-shelf (COTS) or custom-built applications may need to be used to create and maintain these information, so often there are situations where the needed information is incomplete, hard to find or in an indifferent, incompatible format. The problems with storing, retrieving, and exchanging information between simulations and other manufacturing applications widely exist in shipyards. The development of neutral interfaces for exchanging data between simulations and other manufacturing applications is urgently necessary. Simulation models use information of process plans, resources, and other manufacturing processes as input data. Information from different applications in different formats makes simulation modeling and analysis difficult, and exchanging information with other shipyards is almost impossible unless the information is in a standard data format. The NIST-developed information model that facilitates the exchange of manufacturing information between simulation applications, other manufacturing applications, and data sources (McLean et al. 2002) will address those issues. The XML format allows for the definition of documents that are both human and computer interpretable. XML documents can be exchanged easily between applications as text files using basic mechanisms. There are many free or low cost tools that support the definition, creation, modification, validation, and display of XML documents.

The information model describes the important elements of shipbuilding operations, the attributes of those elements, and the relationships among the elements. Both UML static structure diagrams and XML Schemas are being used to develop the information model. The UML static structure diagrams provide a graphical description of the model, while XML Schemas provide a textual description of the model that facilitates the creation of the XML instance documents. Figure 2 shows an example of a UML model of work areas. Figure 3 shows an example of an XML model for schedules.

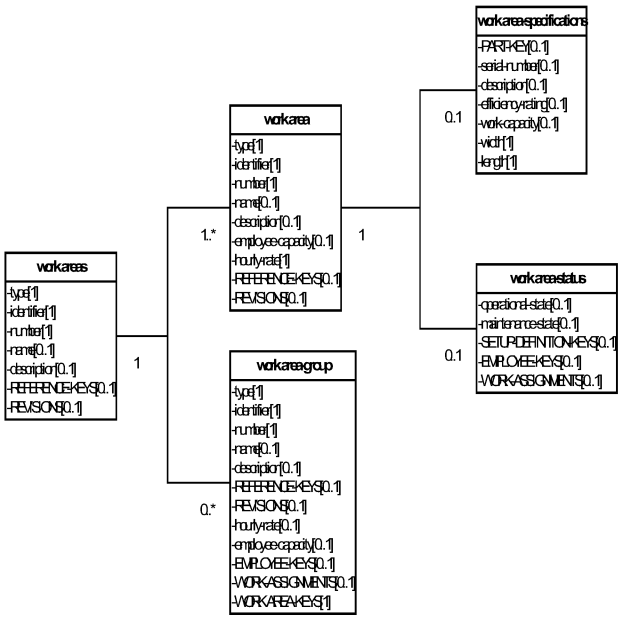


Figure 2: An Example of UML Model for Work Areas

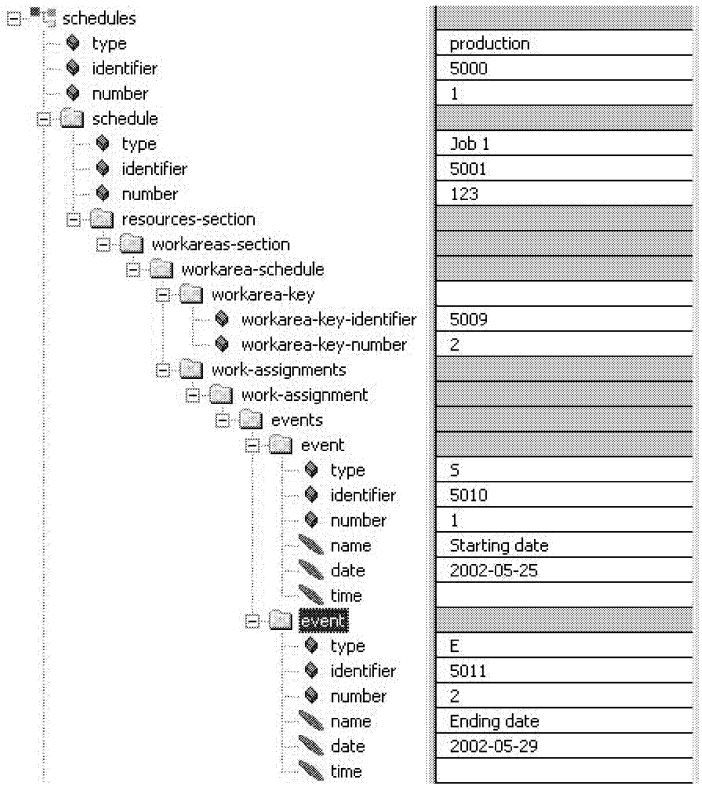


Figure 3: An Example of an XML Model for Schedules

GENERIC SHIPBUILDING SIMULATION MODEL

Generic Shipbuilding Simulation Functional Requirements

Simulations can help to identify project and resource management issues that have a major cost impact for the shipbuilding industry. Simulation was used to model and evaluate complex interactions between overall project workload and available resources. It also can be used to identify resource conflicts, "labor-hiding," and obtain a big picture of overall operations that is not available by other means. Multiple runs of models can provide sensitivity data on the potential impact of random variations in operations and break down of equipments. The simulation model enables the analysis of schedule impact due to additional projects, differing production scenarios and identification of labor resource problems, identifying resource constraints and conflicts between competing jobs (including cranes and forklifts), and demonstration of the expected results of inserting new technology or equipment into the shipyard, particularly with respect to operating costs and schedule impact. The generic simulation system is intended to be suitable and applicable for the shipbuilding industry in general.

Generic Shipbuilding Simulation System Components

Simulation system is used to refer to not only the simulation model and the simulation engine, but also the other software applications that are used to generate data for the model, a user interface system, and associated data files. The simulation system is divided into four component elements: simulator, control logic, user interface, and internal data management.

Simulator – The simulation engine used is ProModel 6.0. ProModel is a commercial simulator that allows users to build a graphical representation of an application system and test it in a variety of scenarios to provide better solution for the problems in the organization. The graphical animation and reports are useful tools for visualizing, understanding, and improving the real system. A Microsoft COM interface is included that allows the simulator to execute external code.

Control Logic – The control logic is customized code and data developed to model the production process of a shipyard. The control logic was used to manage the execution of the following modules: graphical user interface, input data files, job and task management, production area allocation and space management, labor allocation, resource allocation, calendar and clock, and report generator.

User Interface – The animated user interface allows users to interactively use the simulation model to evaluate and analyze the production process of the shipyard. The graphical display of work location, resources, production flows, and simulation results will help shipyards to make better production management and resource allocation decisions. There are nine simulation screens developed. Figure 4 shows one of them, the screen capture of simulation status.

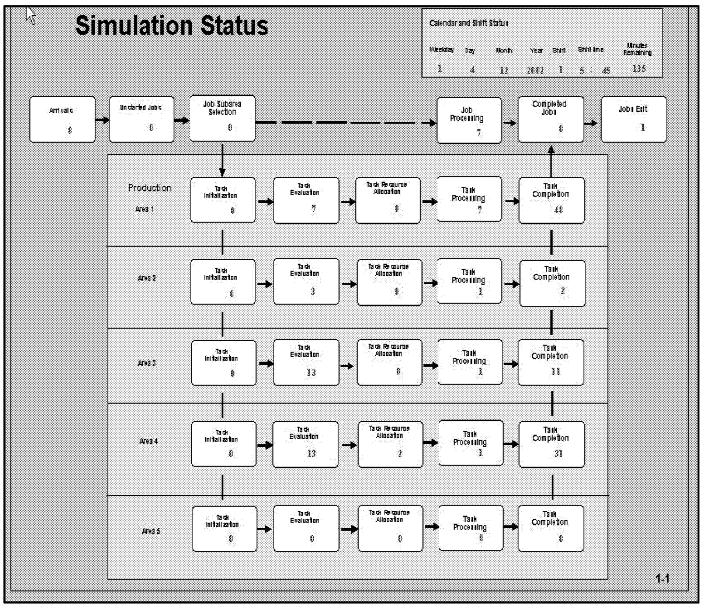


Figure 4: Screen Capture of Simulation Status

Internal Data Management – As discussed above, the primary sources for "raw" shipyard data reside in Microsoft Excel spreadsheets, Microsoft Project files, Primavera project files, Microsoft Access databases, and Comma-Separated Value (CSV) text files or even people's heads. Since the input data of the shipyard simulation system, the data types in shipyard source data files or databases are not in XML format or in the appropriate representation required for internal use by the ProModel simulator, therefore these raw data need to be converted to a standard XML structure based on the shipyard data specification. The Data Editor will be used for various shipyard production data entry and

editing in order to convert the shipyard data into XML format. In the current version of ProModel, all data must essentially be represented as an integer or real numbers. In addition, the XML data need to be read into ProModel in integer or real formats using Input data conversion utilities implemented using C++ and Xsub function. The following major sets of data are required to configure and prepared for the shipyard simulation models: schedule file, shipbuilding project plans, labor configuration data, calendar and shift information.

- *Schedule file* – The schedule is used to identify the jobs to be executed during a particular run of the simulation.
- *Shipbuilding project plans* – Project planning data for the shipyard is normally maintained in a project management file that contains task identifiers, activity descriptions, total staff hours, remaining staff hours, early start date, early finish date, craft code, staff per shift, a calendar identifier, and precedence relationships between activities. A unique identifier is assigned for each project and job (subproject or group of activities/tasks). The identifier is used to reference the input data files and track the project/job through the simulation.
- *Labor configuration data* - The shipyard configuration data file is used to identify shipyard labor, material handling, machine, and space resources that are available during a particular simulation run. The file identifies the number of each craft type and skill level that is available on each shift during a typical day. It specifies labor rates by craft type and skill level, whether new hires or contractors may augment labor pools, and simple rules for laying-off idle workers.
- *Calendar and shift information* - This data is currently updated using the development interface. It includes production calendar information, i.e., holidays, shift schedules, overtime constraints.

CONCLUSION

This paper described the work being carried out by NIST in developing a shipbuilding simulation system with a neutral interface data model and data exchange format for shipyards data. The objective of the project was to use a standardized, computer interpretable representation that allow for the efficient storage and exchange of manufacturing life cycle data in a shipbuilding environment. The information model addresses the exchange of real world data between simulation and other manufacturing software applications. It would be used to maintain data regarding the mapping of real world objects into their simulated representation.

This paper also provided a brief overview of the simulation of shipbuilding operations that is being developed by the staff of the NIST Manufacturing Interoperability Program. By using simulation in shipbuilding operation, traditional problems in a shipyard can be solved. The simulation model enables the analysis of schedule impact due to additional projects, differing production scenarios and identification of Labor resource problems, identification resource constraints and conflicts between competing jobs (including cranes and forklifts) and demonstration of the expected results of inserting new technology or equipment into the shipyard, particularly with respect to operating costs and schedule impact. The animated and graphical display of work location, resources, production flows, and simulation results will help shipyards to make better production management and resource allocation decisions. Formatted report and log files archive the simulation runs to preserve historical data for later use.

ACKNOWLEDGEMENTS

This document was prepared based on the prior work of NIST Manufacturing Simulation and Modeling Group. Work described in this paper was sponsored by NSRP Advanced Shipbuilding Enterprise, the NIST Systems Integration for Manufacturing Applications (SIMA) Program and the Software Engineering Institute's Technology Insertion Demonstration and Evaluation Program (SEI TIDE) Program.

REFERENCES

Lee T., C. McLean, and G. Shao. 2003. "A Neutral Information Model for Simulation Machine Shop Operations". In *Proceedings of the 2003 Winter Simulation Conference.* Ed. S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, 1296-1304.

McLean, C., A. Jones, T. Lee, and F. Riddick. 2002. *"*An Architecture for a Generic Data-Driven Machine Shop Simulator". In *Proceedings of the 2002 Winter Simulation Conference*, Ed. E. Yucesan, C. Chen, J. L. Snowdon, and J. M. Charnes, Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, 1108-1116.

McLean, C., T. Lee, G. Shao, and F. Riddick. 2005. *"Shop Data Model and Interface Specification."* NISTIR 7198. Gaithersburg, Maryland: National Institute of Standards and Technology.

McLean, C. and G. Shao. 2001. "Simulation of Shipbuilding Operations". In *Proceedings of the 2001 Winter Simulation Conference*. Ed. B.A. Peters, J.S. Smith, D.J. Medeiros, and M.W. Rohrer, Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, 870-876.

Object Management Group (OMG). 2003. *Unified Modeling Language.* Available online via http://www.uml.org/

ProModel Manufacturing Simulation Software Reference Guide, 1996, Orem, UT

ProModel Manufacturing Simulation Software User's Guide, 1996, Orem, UT

Richard L. Storch, Colin P. Hammon, Howard M. Bunch and Richard C. 1995. Ship Production, 2nd Edition. Cornell Maritime Press, Inc.

World Wide Web Consortium (W3C). 2000. *Extensible Markup Language (XML) 1.0 (second edition).* Available online via http://www.w3.org/TR/REC-xml.html

W3C. 2003. *XML Document Object Model (DOM).* Available online via http://www.w3.org/DOM/

ENTERPRISE AND MANAGEMENT SIMULATION

VIRTUAL ENTERPRISE RISK AND SECURITY MANAGEMENT

Tomas Feglar
International Consultant in Information Systems Research and Architecture
Vondrousova 1199, 163 00 Prague 6, Czech Republic
Phone: +420 235 313 380, Fax: +420 235 313 380
feglar@centrum.cz

KEY WORDS
Virtual Enterprise, Risk Management process, Security Management process, GERA

ABSTRACT

Global Engineering and Manufacturing in Enterprise Networks (GLOBEMEN) research project significantly influenced Virtual Enterprise (VE) theory and practice. VE Risk Management and VE Security Management processes were covered only in very limited range in this project. This paper analyzes the reasons for VE Risk Management primarily in relation to the threat of terrorism and the lack of trust. VE Security Management is closely coupled with Risk Management using risk driven countermeasure recommendation. Both management processes require appropriate automation and architecture modeling support. This paper briefly describes a new combined CRAMM/CAF approach that meets these requirements.

1. VIRTUAL ENTERPRISE (VE) CONCEPT AND METHODOLOGY

Global Engineering and Manufacturing in Enterprise Networks (GLOBEMEN) research project realized in the period of 2000 – 2003 years significantly influenced Virtual Enterprise (VE) theory and practice. Its objectives were to define reference architecture for virtual manufacturing enterprises, to implement proof of concept industrial prototypes, and to demonstrate core features of the architecture [11].
GLOBEMEN addressed three main aspects of manufacturing: sales and services, inter-enterprise delivery process management and distributed engineering. Based on industrial requirements specifications the work was integrated into Information Technology (IT) architecture for enterprise networks and virtual enterprises. Core IT architecture concepts applied three basic building blocks – Network, VE and Product (Figure 1).
Figure 1 depicts general key activities derived from Trolle and Vesterager [14]. Each basic building block is associated with eight GERA Life-cycle phases [1]. Activities necessary for a specific virtual enterprise establishment are managed by VE methodology. The first part of the methodology is addressing the life cycle of the enterprise network outlining activities to be considered when setting up and managing enterprise networks. The second part is corresponding addressing the setup and management of VEs.
Key network activities can be divided into 4 groups: Identify Network (IN), Network Concept and Requirements (NCR), Network Set Up (NSU) and Network Operation (NO) – in accordance with Figure 1. Last three groups include sub-groups depicted in the Figure 2:

- NCR group includes ‘Select Partners’ and ‘Determine Goal Hierarchy’ activities
- NSU group includes ‘Set Up Network’ and ‘Obtain Preparedness’ activities
- NO group includes ‘Marketing & Sales’, ‘Set of VEs’ and ‘Network Management’ activities.

All mentioned activities are explained in detailed in Trolle – Vesterager’s research report [14].
Analyzing GLOBEMEN results we identified that:

- Very small attention from system engineering viewpoint was given to the Risk Management process
- Virtual Network security management is difficult thanks to lack of a clear risk driven protection strategy

Next two paragraphs analyze these two areas.

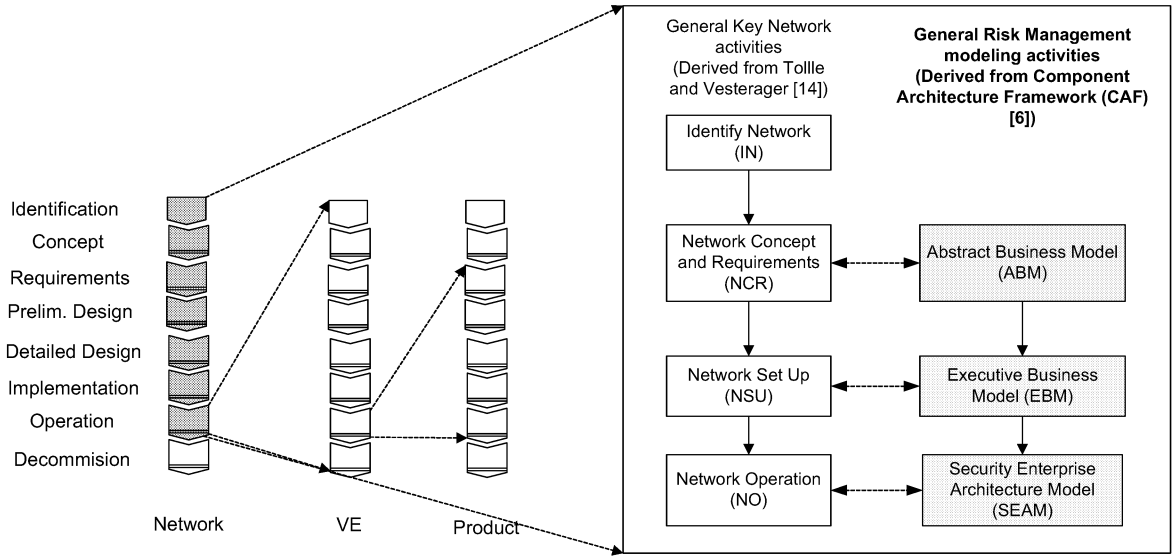


Figure 1: Virtual Enterprise General Activities Harmonized with CAF Based Risk Management Modeling

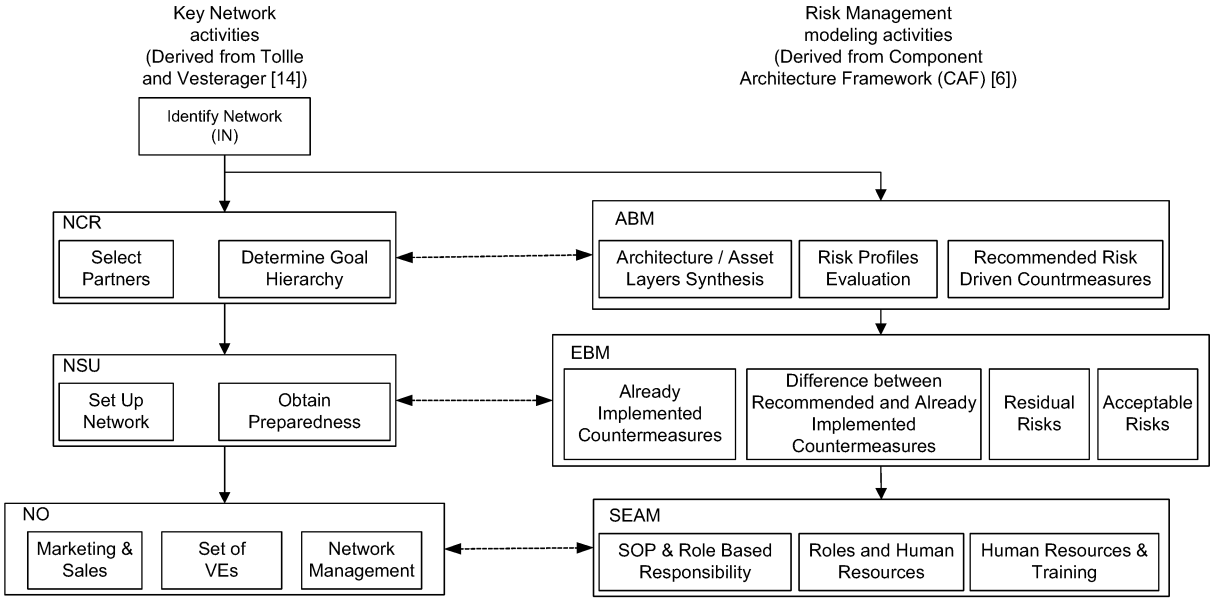


Figure 2: Detailed Description of Network Activities and their Relationship to the Risk Management Modeling Activities

2. THE REASONS FOR RISK MANAGEMENT

Figure 3 depicts VE concept. Network that has to be established for each Virtual Enterprise (A and B) includes functional links and IT delivery paths.

Functional links describe event flows among applications running in particular VE nodes. Applications are tightly coupled with business and manufacture processes running at enterprises' nodes, suppliers' nodes and customers' nodes. Business analysts who are familiar with these processes can easily evaluate an impact characterizing disruption of business process as a result of VE Network functionality losses or degradation.

Event flows among applications are situated at the seven's layer of the RM OSI or even above it. Many protocols (non tangible assets) and network distribution components (tangible and non tangible assets) must be operational strictly in accordance with pre-defined configuration – the IT delivery path. The problem is that many of these items are beyond physical security protection perimeter and therefore can be easily attacked by various threats including terrorism that becomes more and more aggressive. Second reason for Risk Management involvement relates to the lack of trust among Virtual Enterprise participants.

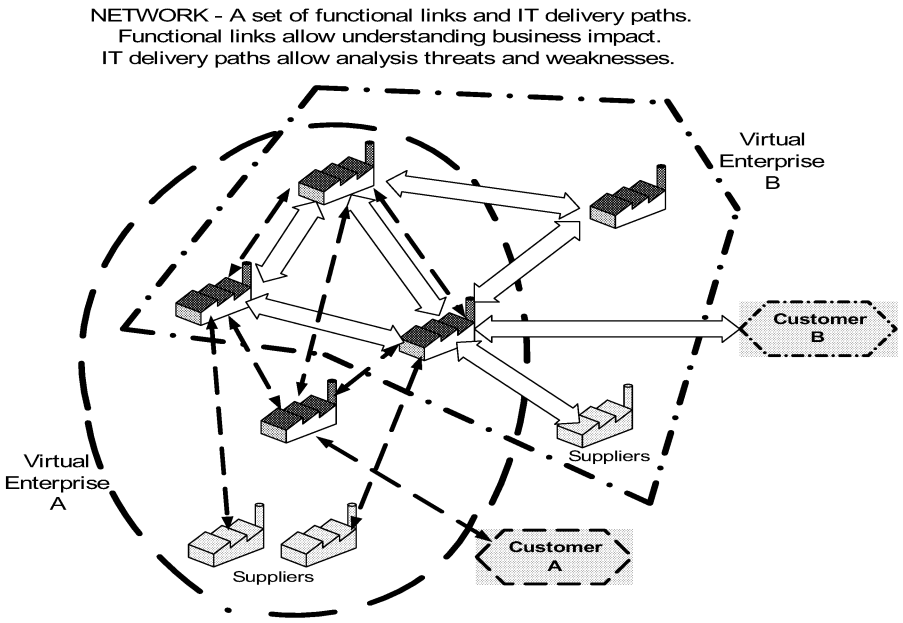


Figure 3: A Rich Picture of the VE Concept (GLOBEMEN Consortium) – Network is Identified as a Source of Risk Relevant Information

a) The threat of terrorism

Following the terrorist attacks of September 11, 2001, the American administration developed and published seven national strategies that relate to combating terrorism and homeland security. This administration also authorized General Accounting Office (GAO) to coordinate effort concerning of further development of these strategies. GAO mentioned the dramatic increases in computer interconnectivity and increasing dependence on computers to support critical operations and infrastructures in various sectors like Energy (Electric power, Oil and Gas), Water, Transportation (Aviation, Surface transportation, Trucking), Food (Mea and poultry, All other food products), and Chemical industry and hazardous materials [8,9].

GAO was also asked to identify and define the desirable characteristics of an effective national strategy and to evaluate whether the national strategies related to terrorism address those characteristics. GAO's findings on this matter were published in the report "Evaluation of Selected Characteristics in National Strategies Related to Terrorism" [10]. This document includes desirable characteristics for a National Strategy in the areas of risk assessment and risk management (Table 1).

Desirable Characteristic	Brief description	Examples of elements
Problem definition and risk assessment	Address the particular national problems and threats the strategy is directed towards	* Discussion of definition of problems, their causes, and operating environment
		* Risk assessment, including an analysis of threats and vulnerabilities * Quality of data available, e.g. constraints, deficiencies, and "unknowns"
Resources, investments, and risk management	Address what the strategy will cost, the sources and types of resources and investments needed, and where resources and investments should be targeted by balancing risk reduction and costs	* Resources and investments associated with strategy
		* Types of resources required, such as budgetary, human capital, information technology, research and development
		* Economic principles, such as balancing benefits and costs
		* Risk management principles

Table 1: GAO Desirable Characteristics of the Risk Assessment and Risk Management

b) The lack of trust

When doing business on internet every business must confront the crisis of trust, hence the companies must place a greater focus on building strong customer relationships and developing customer loyalty. It is extremely difficult for the customers to trust the websites where they don't see the person who is involved in the transaction with them [3].

To overcome trustworthiness limitations we distinguish two directions; Modular Project Development (MPD) and Risk Driven Information Security Management (RDISM).

In the rest of our paper we focus RDISM. MPD interested readers can see [15].

3. RISK DRIVEN INFORMATION SECURITY MANAGEMENT

ISO/IEC Guide 73 defines Risk Management process as set of coordinated activities to direct and control an organization with regard to risk. The aim of this Guide is to promote a coherent approach to the description of risk management activities and the use of risk management terminology. It does not provide guidance on risk management practice [12].
Various practically oriented guidances were developed in last decade but government authorities supported the only two of them. NIST (National Institute of Standard and Technology) published Risk Management for Information Technology Systems and British CCTA (Central Computer and Telecommunication Agency) developed Risk Analysis and Management Method CRAMM in response to a growing need for security in the information systems [13,4].
We have chosen CRAMM in our CAF research (see next paragraph) for three reasons.

Firstly, CRAMM is divided into two parts – the method, which provides guidance on carrying out risk assessment and risk treatment recommendations (this part is similar to the Risk Management guidelines developed by NIST [13]), and supporting software, which helps us to set up and carry out the reviews. Risk management process automation significantly reduces a time necessary for risk profiles evaluation and recommended risk driven countermeasures generation.

Secondly, it was developed in Europe and was successfully applied in various branches as well as internationally [5].

Thirdly, CRAMM includes a database of over 3 000 security controls cross-referenced to the risks they protect against and the most relevant ISO/IEC 17799 control objectives [17].
ISO/IEC 17799:2000 addresses topics in terms of policies and general good practices and it significantly simplifies overcoming “the lack of trust” problem.
Direct application of CRAMM for VE Risk Driven Information Security Management is cumbersome thanks an absence of appropriate architecture modeling interface. It was the reason why we developed new method and supporting tools called CAF – Component Architecture Framework [6, 7].

4. CAF AND VE RISK MANAGEMENT

CAF combines methodology, models and software and is compatible with GERAM Framework [1]. It can be applied in VE Risk Management modeling as a set of General Risk Management modeling activities depicted in the right part of the Figure 1:

- Abstract Business Model (ABM) is used for modeling of enterprise changing needs especially in situations when a vision has to be clearly understood from benefit, opportunity, cost, and risk perspectives.
- Executive Business Model (EBM) allows modeling of operating needs including operational risks characterizing current business processes. This model also allows monitoring differences between recommended countermeasures and already implemented countermeasures.
- Security Enterprise Architecture Model (SEAM) combines role based responsibility and security operating procedures with available human resources, suppliers, and customers.

Figure 2 depicts each General Risk Management modeling activity as a set of functional blocks:
ABM includes 3 blocks. “Architecture / Asset Layer Synthesis” is used for a development of architecture models in accordance with Zachman framework [16] and provides architecture models’ mapping to asset models, typically used for risk assessment. “Risk Profiles Evaluation” accumulates information about threats and vulnerabilities affecting asset models and calculates risk profiles associated with asset classes. “Recommended Risk Driven Countermeasures” generates recommendations using various rules that linked countermeasures with risks.
EBM includes 4 blocks. “Already Implemented Countermeasures” allows tracking of real protection status in a Virtual Enterprise. Real protection status usually differs from recommended status thanks to various constrains. “Difference between Recommended and Already Implemented Countermeasures” block supports monitoring of these differences. “Residual Risks” allows tracking of risk profiles considering already implemented countermeasures. “Acceptable Risks” have to be approved by all Virtual Enterprise participants.
SEAM includes 3 blocks. “Security Operating Procedures (SOP) & Role Based Responsibility” block associates SOPs with a particular role that has to be responsible for them. “Roles and Human Resources” maintains linkages among roles, enterprise staff, contractors and customers. Employment agreement is applied for linkages among role based responsibilities and enterprise staff. Service Levels Agreements (SLA) establish legal linkages among business entities (enterprises, suppliers, customers) participating in a Virtual Enterprise. “Human Resources and Training” block allows monitoring differences between requested qualification and already achieved. These differences are used for training plans.

Figure 1 and 2 depicts synchronization relationships between sets of Key Network activities and sets of Risk Management modeling activities. These relationships

allow smooth integration of CAF approach with VE Network establishment.

5. CONCLUSION

Virtual Enterprises seems to be very perspective direction for a development of new ways of sales and services production and delivery. Information and Communication Technology (ICT) is a very important part of this direction because it offers many powerful functions not available until last few years. Successful ICT usage will be influenced by risks that characterize Virtual Enterprise (VE) environment. This paper briefly describes reasons why Risk Management is important and also illustrates how Risk Management should be established in VE environment using Component Architecture Framework (CAF) approach.

REFERENCES

[1] Bernus,P., Nemeas,L.: "A Framework to Define a Generic Enterprise Reference Architecture and Methodology", *Proceedings of the International Conference on Automation, Robotics and Computer Vision (ICARCV'94)*, Singapore, November, 1994

[2] Bernus,P., Nemes,L., Schnidt.G.: „Handbook on Enterprise Architecture", Springer, ISBN 3-540-00343-6

[3] Bhattacharya,K.K., Saha,S.: "Trust Dimensions in E-Retailing: A Strategic Exploration, IEEE 2004 Conference, Volume 2, pp. 825 – 828, Singapore, October 2004

[4] CCTA: "The CRAMM Risk Analysis and Management Method", Crown Copyright, CCTA IT Security and Privacy Group, London, 1991

[5] CRAMM User Group, http://www.crammusergroup.org.uk

[6] Feglar,T. : "CAF Methodology Usage for Management of Information Systems Protection ", *The 8th World Multi-Conference on Systemics, Cybernetics and Informatics,* July 2004, Orlando, Florida (USA)

[7] Feglar,T., Levy,J.K. : "Protecting Cyber Critical Infrastructure (CCI) : Integrating Information Security Risk Analysis and Environmental Vulnerability Analysis ", 2004 IEEE International Engineering Management Conference, October 2004, Singapore

[8] GAO-03-121: "Protecting Information Systems Supporting the Federal Government and the Nation's Critical Infrastructures", GAO, January 2003.

[9] GAO-03-1165T: "Homeland Security: Information Sharing Responsibilities, Challenges, and Key Management Issues", GAO, September 2003.

[10] GAO-04-408T: "Combating Terrorism. Evaluation of Selected Characteristics in National Strategies Related to Terrorism", February 2004.

[11] http://cic.vtt.fi/projects/globemen/public.html

[12] ISO Guide 73: "Risk Management – Vocabulary – Guidelines for use in standards", First Edition 2002.

[13] NIST SP 800-30: "Risk Management Guide for Information Technology Systems", U.S. Department of Commerce

[14] Tolle,M., Vesterager,J.: "Developing the Business Model – A Methodology for Virtual Enterprises", pp. 291 – 308.

[15] Wang,Q., Chai,K.H., Brombacher,A.C., Halman,J.I.M.: "Managing Risks in Modular Product Development", IEEE 2004 Conference, Volume 2, pp. 815 – 819, Singapore, October 2004

[16] Zachman,J.A.: "A Framework for Information Systems Architecture", IBM Systems Journal, Vol. 26, No3, 1987, 1999

[17] ISO/IEC 17799 :2000: "Code of Practice for Information Security Management", (URL) http://www.c-cure.org/bs7799

D³FACT INSIGHT - SIMULATION OF HUGE SCALE MODELS IN COOPERATIVE WORK

Kiran Mahajan
Daniel Huber
Bengt Mueck
Wilhelm Dangelmaier

Heinz Nixdorf Institute, University of Paderborn
Fürstenallee 11
D-33102 Paderborn, Germany
E-mail: { kiran|huber|mueck|whd }@hni.upb.de

ABSTRACT

Today the lifecycle of products is shortening. The development time for new products is also decreasing rapidly. Hence, the planning of new production facilities has to be done faster and simultaneously to the development. The development of new products and their future production has to be done faster and with better quality. Ramp-up times must be short and without major throwbacks. The vision of the Digital Factory is based on these needs and offers a solution. The consequential use of digital planning, development and management assistants is the main working point. Discrete event simulation is a commonly used tool for production planning. Hence, the enlarged use of simulation reveals several deficiencies of current simulation tools.

In this paper we present our simulation tool d³FACT insight to support the efficient analysis of simulation models. The main focus of our simulation tool lies on the possibility to create and simulate huge and complex models in collaborate work. These models can be experienced in a 3D immersive environment. The simulation analysis is guided by the tool thus significant activities cannot be overseen anymore. Through dynamic detailing and a client-server infrastructure the runtime speed of complex models with highly detailed visualization is constantly on a high level. d³FACT insight offers the possibility to ease the work of simulation teams.

INTRODUCTION

Modern production systems are very complex. Simulation offers a possibility to plan these complex systems. With simulation it is possible to plan the production program, the work flow, the material flow, develop control software and much more before a new facility is built. Several such examples can be found especially in the automotive industry (e.g. Bayer et al. 2003).

The field of appliance and the demands concerning user friendliness, performance, flexibility and assistance are growing.

The state of the art and related research is presented in the next section, in sections 3, 4 and 5 the new features of d³FACT insight are described. The system architecture supporting these features is presented in section 6 followed by the conclusion and a short view on future research.

MOTIVATION

Discrete event simulation is used in increasing range of applications. Not only are production lines of several machines simulated but entire factories including suppliers and ex-works material flow. In the context of the digital factory the whole range of simulation applications (factory layout, material flow, production cell layout, etc.) should be integrated in one system. Several tools are needed to simulate all categories in the digital factory.

Thereby the size and complexity of models is growing. The effort to create and manage these huge models is too large to be done by one person. Thus simulation teams are required. Huge models are crafted modularly and one person or group is responsible for one ore more modules. The modules are joined in one model, which is run in simulation experiments. A severe problem arising is, that the common used simulation tools (like AutoMod, Flexsim, em-Plant, etc.) have no multi user support (Mueck and Dittmann 2003).

One main lack of these tools is that they don't deliver any functionality for group communication and coordination. These functions have to be realized externally and therefore additional work is needed. The work of the individual groups is not transparent to the others, so that changes in common used interfaces or newly created modules may not be known to the other groups. There is the risk that some work is done twice or that once verified activities in a simulation don't work after a change in one module.

Necessary methods for cooperative modeling and simulation are (Dangelmaier et al. 2004):

1. Locking
2. Cloning
3. Versioning

These methods have to come along with communication and coordination methods as well as centralized data storage. Locking and versioning allow parallel development on the same model, the combination of versioning and cloning leads to multiple simulation experiments based on an actual process model. Cloned models does not have be stored, only the input and output data, as well as the version identifiers of the used (sub-) models or modules.

When centralizing the data, i.e. the models, submodels, input and output data, model reuse is eased. There is a common library where all submodels are stored and everyone can make use of these. Due to versioning further changes of these submodels have no consequences on other models.

These huge and complex models, with whatever tool they are created, run slow in common simulation tools, especially when a 3D visualization is necessary.

One solution for reducing the model complexity used in practice is to hold modules of different detail. For example the module "high bay warehouse" exists in two levels of detail. The low detailed version is "merely" a source which delivers parts on demand after a stochastically distributed service time. The high detailed version includes a conveyor system, rack feeder, warehouse management system, etc. Parts which are running low are ordered, almost every element can fail, complete inventory monitoring, analysis functions and many more functions are implemented. When simulating with such structured models, only parts of interest or of high significance for the predefined experiment target, are set to high detail. All other are set to low detail. The level of detail cannot be changed during runtime also are the steps of detail big and the modules extensive, to prevent a large number of sub models. The adjusting of the detail level has to be done manually.

Both simulation and visualization run on the computer and have to share resources. One first approach is to separate these two applications (Mueck et al. 2004). A further approach is multi-resolution-modeling and dynamic detailing (Dangelmaier and Mueck 2004). The level of detail of both model and visualization is adjusted by several criteria during runtime. When modeling multi-resolution models a higher effort is necessary to allow the system to switch the detail levels, because additional information has to be implemented. The effort to create several modules of different level of detail for common simulation tools is most likely higher.

Visualization in simulation has many positive effects on simulation efficiency. It can ease the intuitive understanding of simulated processes and thus the communication of simulation results to the management. The verification and validation of models is easier and the analysis of workflow is supported. Computer graphics are evolving fast and besides 2D, 3D visualization are offered. Most simulation tools have an integrated 3D visualization of more or less good quality (Mueck and Dittmann 2003). The next logical step is an immersive 3D environment instead of 3D visualization.

Except some "set-top" approaches (e.g. Rehn and Vance 2004) there is no tool with an immersive 3D environment available. The rise of the vision of the Digital Factory calls for high quality Virtual Reality, and the introduction of an immersive simulation environment is the next evolutionary step. Immersion can facilitate the exploration of design changes and their effect on the simulation. Also it incorporates all the other positive effects of visualization.

A majority of simulation tools have unique languages to customize predefined submodels or to create own submodels. The higher the customization possibilities of a simulation tool are, the more programming is required. Users have to learn a new language every time they use a new simulation tool.

Recapitulating, these problems of current simulation tools could be identified:

1. Simulations of huge models run slow
2. There is no integrated function to adjust the model's level of detail
3. There is no tool with an integrated immersive visualization
4. Collaborate work is not supported
5. Model reuse and maintains is not efficiently possible

HUGE SCALE MODELS

As mentioned above models become more complex and more areas of production are modeled. Huge and complex models are the result. The demand for high quality 3D visualization is growing. d³FACT insight incorporates several methods to overcome the serve problem of missing computation speed. The simulation as well as the visualization changes their level of detail during runtime.

Due to the immersive environment, the user moves through the simulation. The user has a moving but defined area of view. The goal is to present a vision of a high detailed simulation and visualization. Thus objects out of focused cognition can be displayed and simulated on a lower level of detail (of course: Objects out of sight aren't displayed at all). When the level of detail in visualization is reduced the 3D objects are rendered in lower quality (less polygons, etc.) (Mueck et al. 2004). While this is relative easy to comprehend for the visualization, the system to change the level of detail in the simulation is trickier.

The criterion for changing the level of detail in visualization and simulation, i.e. a change in detail of the simulated model, is the significance for the current examination.

Several indicators for significance are suggested here (figure 1). Besides the indicators distance, occlusion and angle of view, which are commonly used for visualization, the indicator "logistical link" is used. When the user examines a certain part of a model the areas connected to this part at the input and output side may also be significant

for the examined part. E.g.: A machine in sequential production is examined. The utilization of this machine is of interest. The behavior and the status of the linked queues and machines in the near vicinity as well as the fork-lift truck system are very important for the analysis. Thus these elements are also simulated in high detail. The logistical links of an object are widely branched and cause significance in remote areas of the simulation. A further feature of d³FACT insight is that the significance of all objects is monitored and when certain measures reach untypical values, e.g. number of parts in stock, utilization of a worker, etc., the level of detail is raised for these objects.

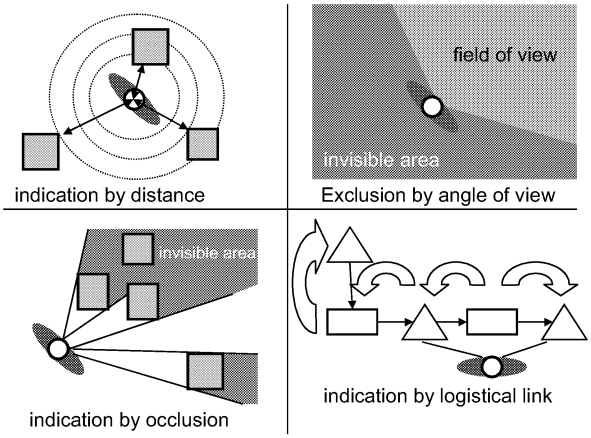


Figure 1 Methods for indicating significance

In figure 2 the structure of a multi resolution model is shown, i.e. a model that can switch the level of detail during runtime[1]. The level of detail increases with the level number. Example: An element is on detail level n. The significance of this element raises and the level of detail should be increased to level n+2. Before the new elements can be integrated into the simulation their state has to be calculated. This has to be done as precisely as possible, to minimize the inaccuracies caused by the change. The state (occupancy and event queue) of the new elements has to be calculated out of the state of the original low level element. d³FACT uses the method described in (Dangelmaier and Mueck 2004). This method is a good compromise between accuracy and computation time, but demands a higher modeling complexity.

d³FACT insight has a client server architecture, i.e. the simulation runs on one computer, the visualization server on one and each client on one, i.e. the clients are the viewers/users of a simulation. The visualization and simulation are separated and do not have to change resources. The dynamic detailing and the architecture facilitate that very large and complex models can effectively be analyzed.

[1] Dashed arrows represent the token flow. The level of detail is indicated by dots per level, i.e. transitions per level (see: Petri-nets).

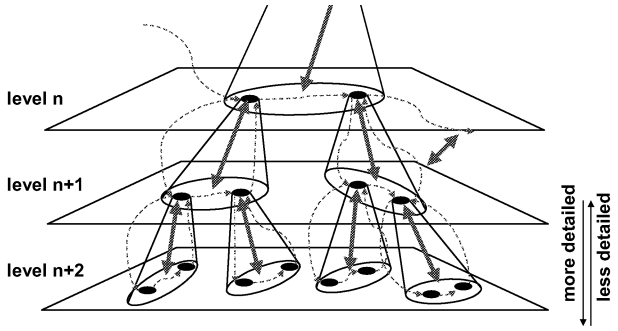


Figure 2 Structure of a multi-resolution model

COOPERATIVE WORK

Huge and complex models can only developed and maintained in teams. d³FACT insight enables groups of users to work cooperatively in an efficient way.

Several features exist to enable collaboration in modeling and simulation. The architecture of d³FACT, with its client-server structure and centralized data storing, is described in detail in the "architecture" section.

In the context of creation of huge models collaborative modeling is the main focus. In addition d³FACT insight offers a multi-user simulation environment. Several clients can connect to one visualization server and observe one simulation. To allow communication between the users a chat function is integrated in the visualization. Figure 3 shows the multi-user environment of our tool.

The bases for multi-user modeling are versioning, cloning, locking and a centralized database. All models, submodels, in- and output data and 3D representatives are stored in this database. When a model is modified it gets a new version number. All versions of a model are saved. Example: model A contains instances of sub model X and Y. The model B contains instances of X and Z. If X is changed the instances in model A and B are not changed, but the modelers of A and B get a notification about the version change. If appropriate, the modelers can update their models.

To protect the models in the model-library against misuse a user rights management system is implemented. For every object it is defined, which user has created this object and which groups of users can modify, delete, use or administer the object. The creator (owner[2]) of an object has the highest position in the rights hierarchy. He can forbid others to change the rights he set for his object. In most cases an administrator should be allowed to do this. This feature comes in handy, when models are sold. When only the *use* right is set for others than the owner, these others can't see the internal structure of the model, and can only parameterize and use it. The function of the *delete* and *administer* right is obvious. When the *modify* right is set for a user he gains full access to the internal structure of the model. The rights are cumulative, i.e. if one has the right

[2] Owner: Group of one or more users

use and *modify* for an object, he is not excluded from functions the *modify* right grants him.

The event code, which is executed when token move through the system, is written in a combination of Java and XML. To avert misuse, two classes of users are defined: users that are familiar with programming in this language and users which are not. Users which are not skilled in programming are not able to see or modify the event code.

The coordination effort is reduced by the centralized database. Every user sees which models are already composed and can reuse these in his own models, if he has the right to do so.

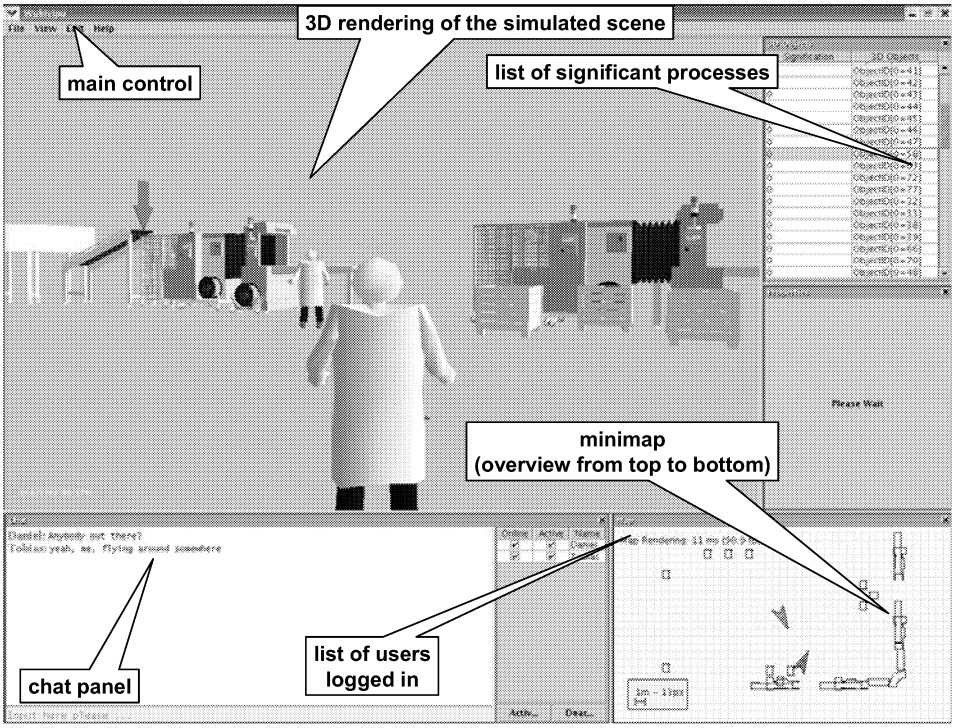


Figure 3 Multi-user environment

3D ENVIRONMENT AND IMMERSION

d³FACT insight has a state of the art 3D visualization with full immersive functionality. Every object in the simulation, e.g. machines, tokens, fork-lift trucks, etc., have a 3D representative. Object called robots, like vehicles, move free on the ground, i.e. on paths marked on the ground. Walls, machines and even the user are not passed through. Adverse to the robots, which move on a 2D plane, tokens move in three dimensions. Token paths are defined on 3D-representatives (like conveyor bands) and tokens move along these lines.

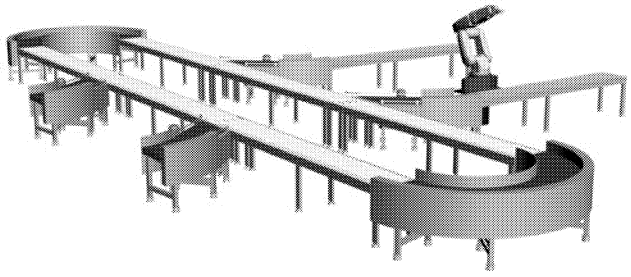


Figure 4 A 3D representation of a material flow system in d3 FACT insight

Input devices are mouse and keyboard, a standard monitor is used as output device. The movement in the 3D environment in ego-perspective is ground bound. A maximum reality is granted by not allowing the user to fly or teleport.

Many features are linked with the features of the dynamic detailing, as presented above. If an element, which is out of sight in the current position of the user, gets significant for the current analysis, its level of detail is increased and the user is offered guidance to this element. This guidance is realized by highlighting a path to the object of significance. This path will follow normal ways of traffic, so that no walls, machines, etc. have to be penetrated. This method has the positive effect, that the simulation is not analyzed by intuition or experience, but systematically. In normal conditions, the identification of critical, i.e. significant, processes (see Figure 3 for significant processes) is highly dependant on experience and intuition. It takes place in disorder and by chance. By guiding the user, information can be gathered systematically and efficiently.

To view and modify the state of an object windows are provided. When selecting an object, e.g. a machine, a window is opened and all public and observable variables are shown. The window could for instance contain data about the process time, the availability, the MTTR and the

utilization. Except for the utilization all values could be changed.

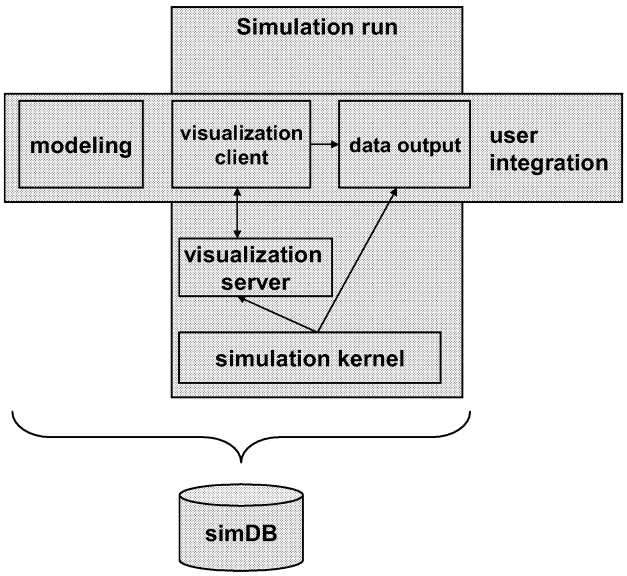


Figure 5 Architecture of d³FACT insight

ARCHITECTURE

The architecture of d³FACT insight is optimized for cooperative work and modular development. The modules are as follows (figure 5):

- modeling
- simulation kernel
- visualization 3D
- visualization 2D
- database (simDB)

The modeling is done in a 2D environment (figure 6), the models are stored in a library and can be inserted as instances in another model by drag & drop. The input and output channels of these submodels are connected via links. As said before, d³FACT insight is a token based simulation tool. The tokens are created in a submodel, e.g. a source, and are sent from one submodel to another linked submodel. One model can have none, one or more input and output channels and every output channel is linked to an input channel. While traveling through the system token activate events. When a token enters an input channel an input event is activated, when a token leaves an output channel, i.e. the output channel is free to take in a new token, a reopen event is activated. Every channel can only contain one token at a time. The event code is written in a combination of Java and XML. Next to events, variables can be defined for models. These variables can be public or private. When inserting an instance of a submodel into a model the private variables of this submodel can be seen and changed, while the private can't. Tokens can also contain variables and other tokens. When modeling is finished, the models are saved in the database. This is done in a XML-file.

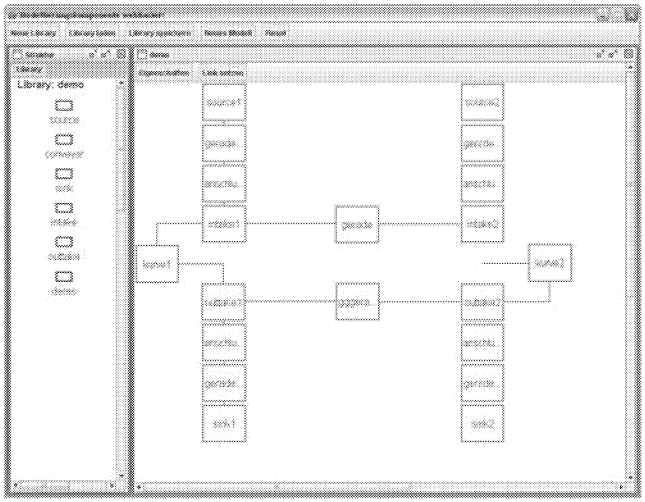


Figure 6 A typical 2D modeling environment

Before simulation the XML-file of a model is parsed by a XSL to Java code. A jar-file is created. This file is executed by the simulation-kernel and the simulation is started. The coordination and communication between the modules *simulation-kernel*, *visualization-server*, *visualization-client* and *data output* is realized by XML messages. The 3D representatives for the submodels are also saved in the database and are linked with the submodels by identifiers. The visualization server loads the 3D-objects from the database and creates the scene. Several clients, the actual users, can join the simulation and full immersion is granted (see Figure 3).

When the target of the simulation is to gather statistical data, a 3D visualization is not necessary. Therefore a less complex 2D visualization or no visualization at all can be activated. A lot of experiments can be run in shortest time. The time factor can be set to a very high level. For allowing an immersive, reality like, visualization the time in d³FACT insight evolves in discrete steps (*fixed increment time advance* (Law and Kelton 2000)). The inhomogeneous simulation time jumps of the *next-event time advance* would be very confusing for the user if the jumps are big. To reduce the inaccuracies created by the fixed time advance the incremental step is very little. The collected output data is written into the database after each simulation run and can be analyzed afterwards.

d³FACT insight is completely written in Java and therefore executable on every computer system. The visualization needs OpenGL and the database is realized in PostgreSQL.

CONCLUSION AND PERSPECTIVES

d³FACT insight offers solutions to current problems in simulation tools.

Huge models can be simulated and visualized in appropriate runtime speed. Dynamic detailing, which is made possible by multi resolution modeling (MRM) changes the level of detail of submodel during runtime. The criterion for changes in detail is the significance for the current investigation of the user. Combined with a dynamic

detailing visualization, which is done by similar criteria, the user gets the vision of a throughout high detailed simulation. Furthermore the user is guided to significant objects, so that an efficient systematical analysis is possible.

The modeling and analysis of these huge and complex models can be done in efficient teamwork. The coordination of collaborate modeling is eased by the methods *locking*, *cloning* and *versioning*. The centralized database is an important instrument to keep all users on an equal information level in reference to the modeling progress. Model reuse and maintains is greatly efficient in d³FACT insight.

The immersive photo realistic visualization improves verification and validation of models as well as the result presentation. Non-Simulation experts and non-model-creators can analyze the simulation by intuition more easily. The throughout Java/XML approach decreases the training costs for simulation newbies, because the languages Java and XML are commonly teached in schools and universities and are used in a wide range of computer appliances.

First appliance tests currently in work show the effectiveness of d³FACT insight as well as further fields of development. A quality management system is installed for improved development and bug fixing, especially in relation to the appliance tests. The modelers can report bugs and improvement suggestions to the development team. The immersive 3D environment enables further functionalities. Ergonomic examinations are imaginable, consequences from lighting, sound and heat could be experienced before the realization of the facilities. Ergonomic examinations in reference to tool and material positioning in manufacturing processes could be analyzed. The teaching of production engineering in education or the training of workers is also thinkable. Because of the ability to be analyzed by non-simulation experts a usage in production near short term planning (production program of the next shift, emergency failure bypass, etc.) is possible.

REFERENCES

Bayer, J., Collisi, T. und Wenzel, S. (Hrsg.): Simulation in der Automobilproduktion, S. 189-196, Springer-Verlag, Berlin u.a., 2003

Carson, John 2004: Introduction to modeling and simulation. In: Ingalls, Ricki G.; Rossetti, Manuel D.; Peters, Brett A.; Smith, Jeffrey S. (Hrsg.): *Proceedings of the 2004 Winter Simulation Conference* (WSC'04), 2004, S. 9-14

Dangelmaier, W., Laroque, C., Mueck, B., Renner, P., Improved Process Planning by a Material Flow Simulation with Multi-User-Support. In: Baake, Uwe F.; Herbst, Joachim; Landeghem, Rik van (Hrsg.): *11th Annual European Concurrent Engineering Conference 2004* (ECEC) EUROSIS, 2004, S. 59-63

Dangelmaier, W., Mueck, B., 2004: Using dynamic multi-resolution modeling to analyze large material flow systems. In: Ingalls, Ricki G.; Rossetti, Manuel D.; Peters, Brett A.; Smith, Jeffrey S. (Hrsg.): *Proceedings of the 2004 Winter Simulation Conference* (WSC'04), 2004, S. 1720-1727

Law, A., Kelton, W., 2000: Simulation Modeling and Analysis. McGraw-Hill, 2000, ISBN 0-07-116537-1

Mueck, B., Dangelmaier, W., Laroque, C., Fischer, M., Kortejan, M. 2004: Guidance of users in interactive 3D-visualisations of material flow simulations. In *Simulation und Visualisierung 2004*, ed. Schulze, T., Schlechtweg, S., Hinz, V., SCS-European Publishing House, Erlangen, 2004

Mueck, B., Dittmann, N. 2003: *Marktanalyse Material-flusssimulatoren*, No 11, ALB-HNI Verlagsschriftenreihe, Paderborn, 2003

Rehn, G., Vance, J., 2004: Integrating operations simulation results with an immersive virtual reality environment. In: Ingalls, Ricki G.; Rossetti, Manuel D.; Peters, Brett A.; Smith, Jeffrey S. (Hrsg.): *Proceedings of the 2004 Winter Simulation Conference (WSC'04)*, 2004, S. 1713-1718

Robinson, S., 2004: Discrete-event simulation: from the pioneers to the present, what next? In *Journal of the Operational Research Society (2004),* pp 1-11

Taylor, S., 2004: Panel on future challenges in modeling methodology. In: Ingalls, Ricki G.; Rossetti, Manuel D.; Peters, Brett A.; Smith, Jeffrey S. (Hrsg.): *Proceedings of the 2004 Winter Simulation Conference (WSC'04)*, 2004, S. 327-332

AUTHOR BIOGRAPHIES

KIRAN MAHAJAN studied Mechanical Engineering at the Delft University of Technology in the Netherlands with a specialization in Production Engineering. Since 2004 he is a research assistant at the group of Prof. Dangelmaier and a PhD student at the International Graduate School of Dynamic Intelligent Systems at the University of Paderborn. His research interests are interactive analysis of material flow simulations in virtual environments. He can be reached at kiran@hni.upb.de.

DANIEL HUBER studied industrial engineering at the University of Paderborn, Germany. Since 2005 he is a research assistant at the group of Prof. Dangelmaier, at the Heinz Nixdorf Institute of the Univerity of Paderborn. His main research interest is the simulation of industrial processes. His e-mail address is huber@hni.upb.de.

BENGT MUECK studied computer science at the University of Paderborn, Germany. Since 1999 he is a research assistant at the group of Prof. Dangelmaier, at the Heinz Nixdorf Institute of the University of Paderborn. He received a PhD in 2005 for his work on dynamic multiresolution modeling of complex simulation systems. His main research interests are logistics systems and tools to simulate those systems. His e-mail address is mueck@hni.uni-paderborn.de.

WILHELM DANGELMAIER studied Mechanical Engineering at the University of Stuttgart, Germany. Since 1981 he was director and head of the Department for Cooperate Planning and Control at the Fraunhofer Institute

for Manufacturing (Fraunhofer IPA). In 1990 he became Professor for Facility Planning and Production Scheduling at the University of Stuttgart. In 1991, Dr. Dangelmaier has become Professor for Business informatics at the Heinz Nixdorf Institute; University of Paderborn, Germany. In 1996, Prof. Dangelmaier founded the Fraunhofer Center for Applied Logistics (Fraunhofer ALB). His e-mail address is whd@hni.upb.de. The web address of his working group at the Heinz Nixdorf Institute is www.hni.upb.de\cim.

MANUFACTURING OPTIMIZATION AND SET-UP

BETASIM EXTENDED FRAMEWORK FOR GENERAL DISCRETE EVENT SIMULATION

František Huňka
University of Ostrava
Institute for Research and Applications of Fuzzy Modeling
30. dubna 22, Ostrava 1, Czech Republic
E-mail: frantisek.hunka@osu.cz

KEYWORDS
Computer software, Discrete simulation, Object-oriented, Model design

ABSTRACT
BetaSIM is a discrete event simulation framework (Kreutzer and Osterbye 1998), (Osterbye and Kreutzer 1999) based on the BETA object oriented language. The framework is designed to support queuing scenarios based on light-weight processes, in which a special attention is paid to providing abstractions that allow to specify synchronization between processes in a simulation model. The aim of this paper is to describe extension of the original framework to cope with general discrete event simulation requirements. The purpose of this extension is to exploit the framework for more complex simulation tasks such as cellular systems and systems with nested simulation, described in (Kindler et al. 2003, 2004), using in this way many BETA's unique features.

BETA OBJECT ORIENTED LANGUAGE CHARACTERISTICS

BETA is a programming language from Scandinavian school of object-orientation, which may be characterized as an object oriented, process or agent oriented and block oriented language (Madsen et al. 1993). BETA is among others characterized by powerful abstraction mechanisms. This includes the pattern concept, which e.g. unifies class and method. The notion of virtual pattern generalizes virtual procedure (method) and provides a virtual class concept that supports generic classes. Native compilers for BETA exist for a number of platforms. Recently work has been going on to port BETA to the bytecode based platforms of SUN Java and Microsoft .NET.

One of the main benefits of BETA is that its linguistic features and light-weight processes can be used to create high level synchronization abstractions. BETA is based on a small number of very general concepts, which makes it highly general and orthogonal. These concepts support a wealth of additional machine and application specific features through suitable libraries. These libraries in BETA support seamless integration of the built in abstractions into the whole application.
BETA allows procedures as well as classes to be specialized. Syntactically this is achieved by prefixing definition of a procedure body with the name of the "super procedure"; i.e. in exactly the same way in which this is done for classes.

BETA supports the notion of virtual procedures as well, and as in other object oriented languages virtual procedures are bound dynamically. However, in BETA a virtual procedure cannot be redefined, but only specialized. This means that execution will always start at the most general level and be delegated recursively to more specific ones through use of the INNER statement. This feature can be used with benefit in object oriented modeling.
Block structure allows to declare a procedure as local to any other. BETA goes further in that procedures can be specialized in the same way as classes in other languages. Therefore, procedures can contain local virtual procedures, which can be redefined in procedure specialization.
From the simulation perspective BETA supports co-routines (or multiple threads), a prerequisite for building process-oriented models. These co-routines (threads) differ in BETA from standard objects in declaration and there are also some small differences in manipulation (e.g. in storing in the lists). On the other hand BETA offers the possibility of including parameters when calling co-routine, which may be used in the construction of generators.

BETASIM FRAMEWORK CHARACTERISTICS

Modularization and hierarchies play an important role in the process leading to a layered design as a basis for model and program development. Object-oriented description offers an elegant and powerful metaphor for organizing such layers of knowledge through locality of description. BetaSIM framework (Kreutzer and Osterbye 1998) is based on a layered design where the concepts provided at one level may be employed to compose more specialized concepts at next layer. The framework can be extended by adding new layer or adopting current layer.

BetaSIM framework introduces high level synchronization abstractions such as re-entrant monitors, server abstractions, interruptible servers and nested servers (Osterbye and Kreutzer 1999). The key abstraction mechanism used in creating these abstractions is procedure specialization. All these synchronization issues can be further exploited in making general extension of the framework. This makes the framework more powerful,

suitable for more complex modeling and simulation applications.

The drawbacks of the framework come from being designed for queuing simulation scenarios. The framework does not have means for taking over control directly between processes as well as does not use the *main program* process in simulation and so on.

BETASIM EXTENDED FRAMEWORK CHARACTERISTICS

The extension is designed to enlarge exploiting of the simulation framework to general use. In this way the design of the extended framework was inspired by the Simula interface for simulation as it proves its suitability for general simulation purposes. Our aim was to change bottom layer by modification current classes of the layer and by adding some new ones to fulfill requirements. On the other hand we try to leave all functionality of synchronization abstraction layers and other advanced feature of the original framework.

First and the most important task was to regulate co-routines to behave as "symmetric co-routine". Co-routines in BETA behave like "semi-co-routines". They are so called because there is an asymmetry between the calling co-routine and the co-routine being called. The caller explicitly names the co-routine to be called, whereas the called co-routine returns to the caller by executing *suspend* command, which does not name the caller explicitly. The *suspend* command always suspends the active co-routine and not the lexically surrounding one.

There is another kind of co-routines, called *symmetric co-routines*, which explicitly calls the co-routine to take over control. It does not return to the caller by means of *suspend* command, giving a symmetric relation between the co-routines. Symmetric co-routines can be modeled in BETA through abstract class (Madsen et al. 1993). This abstract class must be used as a superclass for all co-routines that are taking part in the *symmetric* scheduling so it is embedded in the bottom layer.

The abstract class that makes it possible for "semi-co-routine" to behave as a "symmetric-co-routine" is implemented as a procedure *Start* at the simulation framework level and a method *resume* in the *Process* class. The code of this procedure follows:

```
Start:
 (proc enter next: instance Process;
          active: instance Process;
  do while next <> none
    begin active : = next; next:= none;
      act : = true;
      active;  (* start running the process *)
    end;
 proc)
```

Figure 1: Procedure *Start* code

The body of the *Start* procedure is created by a loop, which can be interrupted only at the end of application or when the *main program* takes over control. It is of course possible for the main program to transfer control to the next process to continue in executing or finish the application.

When the method *resume* is called, control is transferred to the receiver of the method. In a more detailed way *this(Process)* is stored to the *next* variable and caller process is suspended. That means that the application continues in the body of the *Start* procedure (actually inside the loop). The code of the resume method follows:

```
Resume:
  (proc
     do next : = this(Process); SUSPEND;
   proc)
```

Figure 2: Method *Resume* code

At the beginning of the application it is necessary for the introductory process to be sent to the *Start* procedure.

Extended simulation framework also requires to declare *main program* process for transferring control to the *main program*. It is also possible to transfer control back to the next *Process*. For this reason the *CALL* procedure was declared.

For standard behaviour the following methods *HOLD*, *PASSIVATE*, *WAIT* and *CANCEL* at the bottom layer were declared. All these methods behave in a Simula way (interface).
In the original BETASIM framework there are *HOLD*, *PASSIVATE* and *WAIT* methods but they are not directly accessible to processes and have slightly different code just suitable for queuing simulation scenarios.

In the bottom layer of the framework a set of methods for running processes is declared. The set covers direct execution (*run*), execution with some delay (*runAt*) or execution before a given process (*runBefore*) and execution after a given process (*runAfter*) and so on. In the original BetaSIM framework processes take over control only by getting at the top of the process queue. Explicit take over control between two processes is not possible.

The original framework declares the *SourceProcess* class (subclass of *Process*) as a Process generator. The *SourceProcess* class is parameterized by a type of process to be created. In addition, this class ensures unique identification of a new process for tracing purposes and creates new processes at time with a given probability. Initial parameters cannot be transferred directly to a new process as it is impossible in BETA. They are transferred using *Init* method that always runs at the creation of a new process.
The extended framework enables to generate new processes directly from subclasses of the *process* class. Instances created in this way hold their class name and a unique number identification and random distribution in the same way as in the previous case.

BETASIM EXTENDED FRAMEWORK STRUCTURE

The simulation framework can be used directly as a prefix for a specific application or as a prefix for further specialized
layer. The framework is composed of:

- simple attributes that hold simple values e.g. status of the processes,
- essential processes declaration,
- methods of the framework,
- classes or singular objects.

In the framework there are four essential processes that are used during the simulation. *Main* process that refers to the main program, *current* process that refers to the active process, *next* and *nextDt* processes that are used in the *resume* operation for keeping reference to the next process that takes over.

There are two kinds of the methods in the framework. The first kind of methods represents these methods that implicitly operate on the *current* (active) process. Example of these methods are methods *HOLD* that holds *current* process by a given delay, *PASSIVATE* that removes *current* process from the process queue and cancel process reference in the process queue, *WAIT* that puts the active process into the waiting queue and applies *passivate* method on it.
The second kind of the methods provides specific operations on a process, whose reference must be explicitly given. Example of these methods is a method *CANCEL* that cancels any process, *Start* that starts cycle of simulation, *call* that restart simulation cycle after return from the *main* program, *run* that puts the process at the first place in the process queue, *runAt* that puts the process at given place in the process queue and so on.

The most important class in the framework is the *process* class. New attributes and methods were added to this class to fulfill extended requirements. Earlier mentioned *resume* method is used for suspending *current* process and taking over to explicitly named process. *NextEv* method returns reference to the very next process in the process queue. This method enables *current* (running) process to have reference of that process without necessity to do it through another facility. The *firstEv* method returns a reference of the first process in the process queue. *pCounter* is an example a singular object that provides each newly created process with the name of its class and with a sequence number that is unique identifier of the process. This notion is helpful in debugging and tracing operations. The *pCounter* singular object is used when there is no explicit process generator but the process creates further processes.
The *EventNotice* class declares elements of the process queue. Its attributes are references to the process and simulation time for taking over control in the simulation.

All these additional attributes, process declaration, methods and classes are essential part of the simulation framework but the most important part of the extended framework is the mechanism that enables the active process to explicitly take over another process, which can't be inevitable in the first place in the process queue. The extended simulation framework was tested on experimental simulation models of the cellular systems and queuing systems.

CONCLUSION

BETASIM framework, drawing from the unique BETA features, introduces high level synchronization abstractions arranged in layered design. This design is clearly arranged and enables thus not only adding new layers but also making adaptations of the given layer. Extended framework enlarges capabilities of the original framework more generally. New extensions were done with great understanding to the original framework that means all the sooner designed synchronization abstractions such as re-entrant monitors, server abstractions, interruptible servers and nested servers can be used together with extended features of the new framework. Extended framework can be further extended and structured to independent domain oriented modules using prefixes.
Layered design and extended features of the framework makes the framework more powerful and suitable for more complex modeling and simulation purposes. Further research will continue mainly in exploiting the extended framework for real world tasks with focus on nested simulation.

ACKNOWLEDGEMENTS

The paper was supported by the research scheme No: MSM6198898701

REFERENCES

Kindler, E., Křivý, I., Tanguy, A.: Formalized World Viewings - a Way to Nested Simulation. In Second International Industrial Simulation Conference. Malaga. Gent : Eurosis, 2004. pp. 20-24. ISBN 90-77381-12-0.

Kindler, E., Křivý, I., Tanguy, A.: Automatisation de la construction de modeles pour la simulation reflective. In Conférence Francophone de Modélisation et Simulation. 2003 Toulouse. Toulouse : SCSI , 2003. pp. 379-384. ISBN 3-936150-24-9

Kreutzer, W., Osterbye, K.: BETASIM A framework for discrete event modeling and simulation. In Simulation Practice and Theory 6 1998. pp. 573-599

Madsen, O. L., Moller-Pedersen, B., Nygaard, K.: Object-Oriented Programming in the BETA Programming Language. Addison-Wesley, 1993, ISBN: 0-201-62430-3

Osterbye, K., Kreutzer, W.: Synchronization abstraction in the BETA programming
language. In Computer Languages 25 1999 pp. 165-198

Achieving efficiency through simulation in a declining manufacturing market

Rob Walker
Habtom Mebrahtu
Caroline Strange, Professor
Faculty of Science & Technology
Anglia Polytechnic University
Bishop Hall Lane, Chelmsford
Essex CM1 1SQ, U.K.
R.W.Walker@apu.ac.uk
H.Mebrahtu@apu.ac.uk
C.M.Strange@apu.ac.uk

Abstract

Using simulation to model manufacturing processes, to increase understanding and enable economic management of the system, is now well established. Frequently the improved understanding gained from such modelling is used to increase production output; however, simulation can also assist in planning a reducing output in cases where there is a declining market.

In this case study, an electronics assembly line is studied, where a steadily reducing output is required. One approach in achieving an economic solution is to reduce the number of workers on the assembly line. The simulation demonstrates that analysis of worker's shift patterns can achieve cost-effective and efficient solutions in achieving the required reduction in output.

Introduction

There is now an increasing need to manage in the most economic way the declining output requirement of a manufacturing system. Component obsolescence, for example, can have considerable consequences for the designer. Flaherty (2005) states that 'the majority of (component) obsolescence problems are solved by finding alternative parts. Flaherty argues that a strategy for coping with component obsolescence is vital. Equally a downturn in demand for a product can have a profound impact on the manufacturer. This declining demand can result from, for example, redundant technology, changing consumer demand, new improved design and the transfer of production to another location. If the manufacturer faces the prospect of reducing demand for his products the options may include the following:

- A reduction in the hours worked;
- A reduction in the number of people employed on each shift;
- A reduction in the number of machines available;
- Stockpiling the finished product until it is required, often called an all time buy;
- Considering alternative markets.

In practice all the above must be employed to match as closely as possible the system output and the demand. Each of the above impacts the economics of the operation and all the above options need to be considered as a whole.

Reducing output may result in:

- More time wasted starting the line from empty, for example the line has to be cleaned;
- Changes in machine reliability due to an increased number of stops and starts;
- An increase in fixed overhead burden per hour of operation.

The consequences of reducing the number of people include:

- Inefficiency, particularly when the relationship between output and the number of people employed is not linear for highly automated lines;
- The inability to reduce manning beyond a certain limit;
- The likelihood that output may become very unreliable if a number of breakdowns occur simultaneously;
- An increase in lost manufacturing time, particularly on highly automated lines where the time wasted moving between stoppages will increase;
- A loss of worker motivation as redundancies are made;
- The sapping of moral of the people involved in the change as Maslow's 'safety needs' are undermined (Handy, 1993).

When considering a possible reduction in the number of workers, the variation in worker performance should ideally be reflected. Baines, Hadfield, Mason and Kay (2003) argue that 'the assumption that worker performance can be easily predicted is incorrect'. Further information would be required, for example, worker attitudes and the effect of downsizing on worker morale. Additionally, when considering a possible reduction in the number of shifts, the variation of human performance due to time of day effects may be another contributing factor. In his pilot study, however, Lawrence (2004) found that the introduction of circadian rhythm and individual efficiency variations to models did not have a significant effect on system output. Lawrence acknowledges that a greater understanding is needed of the effect of these human factors in relation to system size and complexity.

Reducing the number of machines, where there are parallel facilities may be a viable option, but on highly automated lines the cost of changing the control software may be prohibitive. It is often unlikely that space generated, through the removal of redundant machinery, can be productively employed again; this then leads to an increase in the fixed overhead burden. Reducing manufacturing output can also result in systems becoming unbalanced.

Alternatively, the manufacturer may consider stockpiling the product. This has obvious cost implications and the following factors must also be taken into account:

- The products may not be suitable for long-term storage;
- The number of parts needing to be stored may not be known.

Computer simulation is an essential management tool in determining the most economic course of action throughout a period of declining output and possible eventual shutdown of the system. The ability to model a system with a large number of variables, including non-linear relationships between the resources deployed and the output achieved, provides a valuable insight into the consequences of system change. Other methods, for example, spreadsheets and flowcharts could not take into account the complexity of a real system (Baudin et al, 1992). Simulation is one of the most powerful analysis tools for understanding the operation of complex processes and systems (Law & Kelton, 2000). Mebrahtu et al (2004) have shown, however, that where the output requirements of the system change in a more complex manner, the use of an expert system to determine the most economic system configuration is becoming a viable proposition.

Model

The model used to demonstrate a downturn in manufacturing output is a line assembling electronic printed circuit boards. Due to changing consumer preferences the demand for the product is declining and after a period of time the company intends to close the system and transfer the production to another of its factories.

The objective of the model is to manage this process as economically as possible.

The line starts from empty following any prolonged (i.e. more than 2 hours) shutdown to maintain cleanliness. The line is fully automatic and two people are employed to repair the machines and carry out the rework operation. The two people have staggered lunch breaks so the line does not stop. The machines, see Table 1, in the line have the following cycle times and breakdown histories included in the Witness simulation model. All the people are trained to be able to carryout all the repairs required in the Witness model and the rework operation.

The PCB's are prepared by washing, screen printing with solder paste and assembled with bought in parts by the HSP (high speed placement) and the robots. After assembly the solder is reflowed and then the parts are tested and if accepted dispatched to the customer. Currently 10% of parts fail the test on the first attempt. The reject assemblies are reworked and retested. Currently 20% of the reworked parts are rejected on test and they are scrapped. The magazines on the HSP machine are reloaded every 500 parts; this requires the machine to be stopped for 300s.

Projected demand

The demand is currently 5200 assemblies per day this is expected to fall to 3500 per day in the near future and then 2000 per day for a period and then cease.

Results from Witness - Daily output

Number of operators per shift				
Day shift	Late shift	Night shift	Output	Scrap
2	2	2	5238	162
2	2	0	3711	121
2	0	0	1822	68
1	1	1	4873	151
1	1	0	3171	101
1	0	0	1388	54
2	2	1	5177	157
2	1	1	5122	156
2	1	0	3547	117

A consistency test was carried out with the aid of a cusum chart to ensure that the model is stable and the results are valid (Walker et al 2004). The consistency check was carried out by changing all the random number streams in the Witness model by a macro in Excel and rerunning the model and recording the results in Excel. Excel was then used to construct a cusum control chart. This showed the model is stable and gives consistent results as shown in the cusum chart in figure 2. The 95% confidence limits for fully manned 3 shift option was 5005 and 5471 based on a standard deviation of 113.

The cusum chart in figure 2 requires two pieces of information to initiate the system they are the target value and the sample standard deviation. We suggest that the mean of the initial results is the target value and the sample standard deviation is the s value used to calculate the mask limits. A minimum of 25 readings should be used to calculate these variables.

The calculation of the cusum is a two stage process; the first stage is to calculate the difference between the target value and the data point (x-T) and then to add this result to the previous cusum. The first cusum is simply the first x-T value.

A cusum chart in essence is about slope: if the cusum points are approximately, horizontal then the readings are near the target value and the process is probably under control. If the slope is upward then the readings are above target and the process may be out of control. If the readings are less than target, the slope will be downward.

A mask is used to determine if the slope is too great for the process to be considered under control. The focus of the mask is placed on the last reading; all the previous readings should fall within the arms of the mask; if they do, then the process is under control.

The semi-parabolic mask that we recommend is equivalent to a 2.65σ action limit and 1.65σ warning limit (BS 5703, 1981). Other mask designs are available; see BS 5703 (1981) for further details.

The value of s used to draw the mask that is the sample standard deviation (Coleman, 1996) of the x-T values for the results from the initial results.

Validation of the model can also be carried out by comparison with actual output figures. It may also be possible to validate the performance of the system by finding the output from the line when an operator was absent and hence confirm the predicted output.

Analysis

From the above results a strategy can be formulated that 3500 assemblies per day can be achieved by operating a two-person day shift and a one-person late shift. The simulation clearly shows that this is achievable and takes into account the need to restart the line from empty each morning. Working a four-day, three shifts per day, week is not an option, as this would result in 3000 too many assemblies per week. To achieve 2000 per day a two person day shift could be used and 1.25 hours overtime per day worked to achieve the output required. This strategy would not only save the cost of employing the people but also the night shift allowance. The most important saving is not manufacturing parts that are not required for immediate sale.

The machine reliability would need to be constantly monitored to ensure that the more frequent line stops and restarts from cold do not adversely affect the machine performance.

Conclusions

In a case of reducing demand, the ability to accurately predict the effect of changes is crucial in order to maintain an economically viable production line. This model demonstrates the detail and clarity of the simulation process with the consequent confidence in the recommendation made. More complex systems than the model used in this paper would make setting a clear strategy almost impossible without extensive trial runs of the shift patterns using the line in question.

The reduction in manning level is particularly important as management must now when and how to adjust the manning level as this will involve breaking up the production team in some way. This would be either by redeployment elsewhere in the business, redundancy or some package of inducements to leave the company voluntarily. All these changes must be planned in detail if the moral of the remaining work force is not to be needlessly undermined and unnecessary costs incurred.

The use of cusum charts is a powerful way of determining the stability of the data from a simulation and should be widely adopted particularly in the manufacturing sector where control chart should be understood by the staff involved in the interpretation of the results. It should prevent action being taken on one set of results and inappropriate decisions being made on unrepresentative data sets.

References

Baines, T, Hadfield, L, Mason, M and Kay, MK (2003) *Human Performance Modelling within Assembly Line Simulations*, Proceedings of the Industrial Simulation Conference, 9 – 11 June, Valencia, Spain, p326 – 330.

Baudin et al (1992) *A Comprehensive analysis of Methods for IC Manufacturing Performance*, Int'l Semiconductor Manufacturing Science Symposium, IEEE

BS 5703 Part 3 (1981*) Data analysis and quality control using cusum techniques*

Coleman, Greenfield, Jones & Puzey, *The Pocket Statistician* , Arnold 1996

Flaherty, N (2005) *New Parts for Old*, IEE Review, vol 51, No. 1 p34-37

Handy C (1993) *Understanding Organizations,* Penguin

Law & Kelton (2000) *Simulation Modelling and Analysis*, McGraw Hill Inc.

Lawrence, P (2004) *On the Need for Human Performance Modelling in the Simulation of a Balanced Production Line*, Proceedings of the Industrial Simulation Conference, 7 – 9 June, Malaga, Spain, p415 – 419.

Mebrahtu H, Walker R and Mileham T, (2004) *Performance enhancement using an expert machanism in a manufacturing simulator*, Journal of Engineering Manufacture, Proc. I Mech E Vol. 218 Part B p245-249

Walker R, Mebratu H and Mileham T, (2004) *Automating part of the model verification process* ISC 2004 p384-388

Table 1 Cycle time and breakdown records

Machine	Cycle time (s)	Frequency (mean time to failure [s])	Duration
Wash	300 for 60 parts in a batch	Negative exponential 10000	Normal mean = 180s, SD = 30s
Screen print	8 per part	Negative exponential 3500	Normal mean = 500s, SD = 50s
High Speed Placement (HSP)	4 per part	Negative exponential 7500	Normal mean = 180s, SD = 30s
Robots 1 & 2	17 per part	Negative exponential 10000	Normal mean = 180s, SD = 30s
Reflow	5 per part	Negative exponential 10000	Normal mean = 180s, SD = 30s
Test	18 per part	Negative exponential 5000	Normal mean = 300s, SD = 20s
Rework	Uniform distribution between 30 and 90	None	
All conveyors	3 per part index time	Negative exponential 10000	Normal mean = 180s, SD = 30s

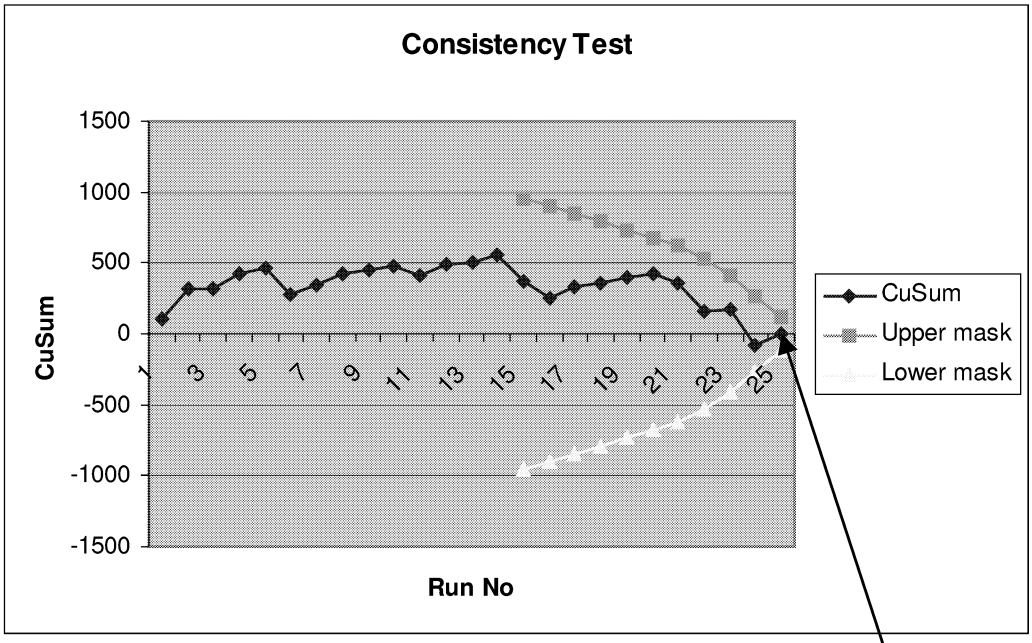


Figure 2 Consistency check on 3 shifts fully manned model

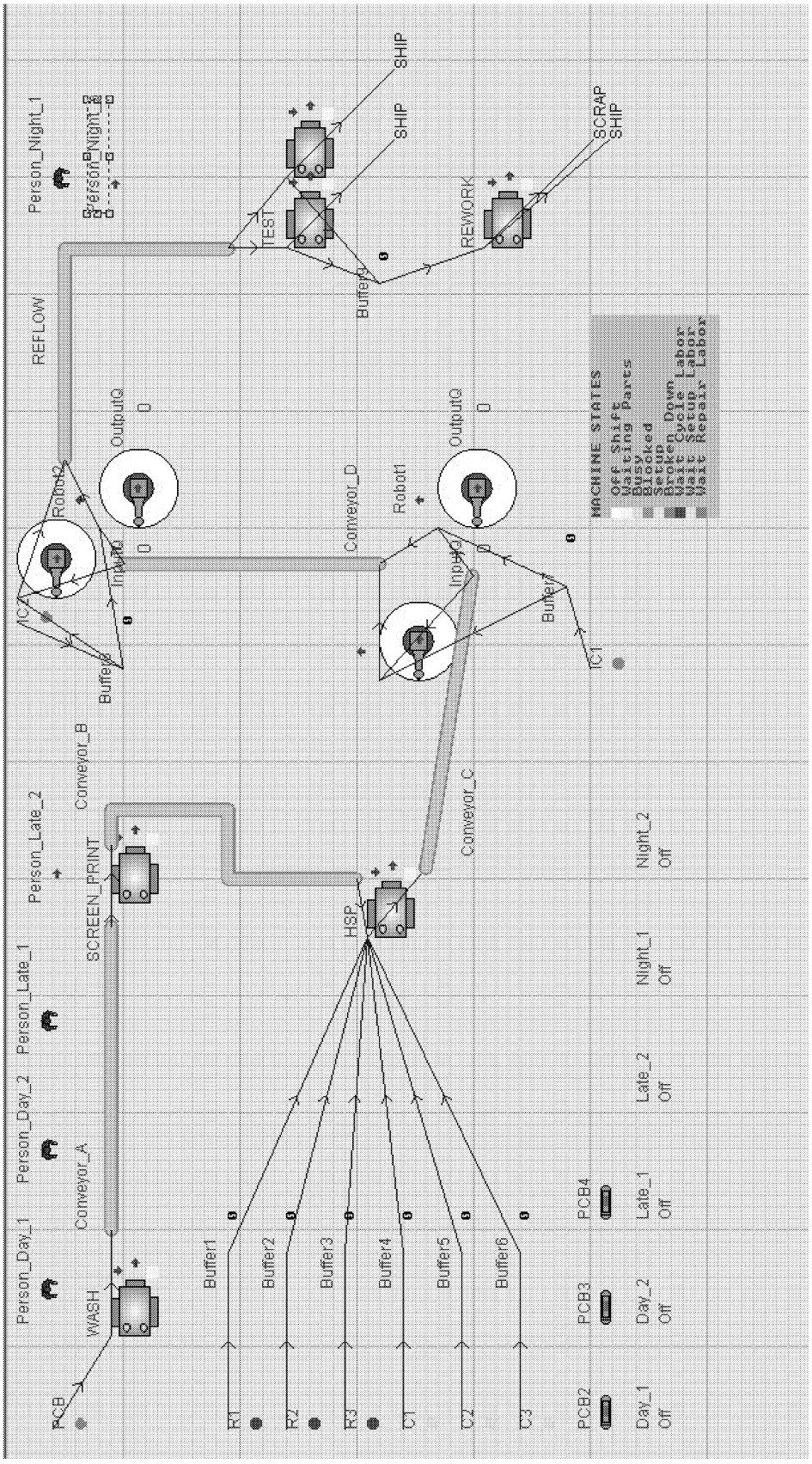


Figure 1 The Witness layout

Visualizing Lean Effects with Simulation-Based Value Stream Mapping

Yang-Hua Lian, Hendrik Van Landeghem
Department of Industrial Management
Ghent University
Technologiepark, 903, B-9052, Ghent, Belgium
email: yanghua_lievens @yahoo.com.tw, hendrik.vanlandeghem@ugent.be

KEYWORDS
Lean, Value stream mapping, Product families.

ABSTRACT

Value Stream Mapping (VSM) has become a popular lean implementation method in recent years. However, its drawbacks such as being time-consuming, unable to detail the dynamic behaviour of production processes and handle the complexity, have stimulated some researchers to look for tools to enhance VSM. Simulation, which creates consensus by its model-building interface, visualization of dynamic views and output analyzing abilities has been proposed as such a tool. We have tried to compensate the shortcomings of static VSM by using simulation tools in the previous works and proposed a 4-phases simulation-based value stream mapping method for lean implementation as well as developed a simulation modelling formal method based on the VSM paradigm. In the formal method, atoms, database structure and model generator are designed and created. The idea is that all the VSM icon functions can be corresponding by composing a set of predefined simulation modules (called atoms in our case), which can be reused. Then, simulation models of different VSM scenarios can be generated quickly and automatically by a model generator with the set of atoms and the designed database. The model generator is designed to create discrete-event simulation models for a predetermined simulation language from the neutral description of the system and the actual data in the database so that the system is data-driven. Algorithms of converting the raw ERP data and information of a VSM map into generator tables of the structured database were also developed within our formal method. In this paper, we simply apply our formal modelling method to a real case. A current state model of the real manufacturing system case and 3 scenarios of VSM future states are generated to see the effects of lean by transforming part of the system to pull kanban, pull kanban with priority and pull kanban with priority plus order release by takt time. We include also an additional aspect, the aggregation of the data for real world to a representative simulation model in scenario 4 and 5, which to our knowledge was not considered in research up to now.

I. Introduction

Lean transformation, such as physical factory layout redesigns or changes of the supply chain infrastructure, involve high costs. In recent years, VSM has emerged as the preferred way to implement lean. VSM is a mapping tool used to describe supply chain networks. It maps not only material flows but also information flows that signal and control the material flows (Rother and Shook 1999). However, since VSM is a paper and pencil tool of lean manufacturing, it is time-consuming. Furthermore, VSM is a static tool which has limitations for detailing dynamic behaviour as well as to handle the complexity and uncertainty, causing barriers to effectively apply the improved situations. We have tried to tackle these drawbacks and proposed a 4-phases simulation-based value stream mapping method for lean implementation in our previous research (Lian and Van Landeghem 2002). In Lian and Van Landeghem 2003 to 2004, we further developed a simulation-based VSM formal simulation modelling method based on the VSM paradigm (called simulation-based VSM formal method). VSM atoms formalism and manipulation, database structure and model generator were elaborated and the formal method was applied to a hypothetical example. Lian and Van Landeghem, in the beginning of 2005, detailed the formal modelling procedure. Algorithms of converting the raw ERP data and information of a VSM map into generator tables of the structured database were developed. A real case was applied to the formal method and an overview of its current state was given. Now in this paper, we apply the method to the case again and measure the lean effects by simply transforming part of the current system according to some guidelines of a lean value stream like pull kanban and takt time. Models of 3 future state scenarios are generated by the same generator with 3 different data set. In addition, scenarios 4 and 5 are produced for the data aggregation aspect. We describe the manufacturing systems of 3 future states in section II of the paper and analyze the outputs in section III. In section IV and V, we look at the sight of data aggregation / product families. Given the large number of manufactured components, it is difficult to map each of their unique flow paths on the paper. To address this problem and in order to map multiple flows in a value stream, Rother and Shook suggest that “draw such flows over one another, but do not try to draw every branch if there are too many.” “Choose the key components first, and get the others later if you need to” (Rother and Shook 1999). One would normally start by mapping a large-quantity and high-revenue *product family*. A product family is defined as “a group of products that pass through similar processing steps and over common equipment in your downstream processes” (Rother and Shook 1999). We give an insight of this issue by running the current state mode with aggregated order data (scenario 4) and disaggregated order data (scenario 5).

II. Case Description

The company of the case study, located in Belgium, is a major manufacturer of poultry and pig raising equipment, including feeding, drinking, storage and feed transportation systems. The current manufacturing system of their division "Components Production" encounters a capacity problem on a machine (HYDRAP). The current VSM of the company is shown in Figure 1. This manufacturing system consists of 31 main atoms (entities). Over the reference period to be simulated, the production control center (atom ID 18) releases 975 production orders of 382 different components to work stations. There are a total of 8 entities of workstations in the simulated system: (1) HYDRAP, (2) metal shear, (3) laser-punching machine, (4) MULLER, (5) bending<2m, (6) bending<4m, (7) Outsourcing, and (8) Others. Work station 1, 2, 4 are single-server workstations. Workstation 3, 5 and 6 are stations with parallel machines (capacity 3, 2, 2). Workstation 7 is an outsourcing process whose cycle time is not considered in the simulation and workstation 8 is a general workstation representing all other machines in the system. Before each workstation, there is an entity of staging inventory (ID 10-17) presenting the work-in-process. Some other entities are: order staging (20), warehouse (9), measurements (21), Order / flow type / cycle time / setup time data tables (19, 24, 25-31). The 975 production orders are also represented through atoms in the system, but they are only produced during model execution, not in the stage of model generation. Therefore, we don't need to consider them here. The released orders are only influenced by the data in the order data table (ID 19).

In the production process, each component has a routing (flow type) according to a routings table, which indicates the sequential steps of each specific component. For example, when production control releases the production order of a component with flow type **1**, this component will be manufactured through the system with a sequence "**16-8-15-1-17-7-9"**. It means this component will go through the manufacturing system with following steps: 1) going to WIP inventory **16** and wait for the process of workstation 8, 2) processed by workstation **8**, 3) going to WIP inventory **15** and wait for the process of workstation 1 of HYDRAP, 4) processed by workstation **1**, HYDRAP, 5) going to WIP inventory **17** and wait for outsourcing process, 6) outsourcing process **7**, 7) going to the warehouse (magazijn, atom ID **9**). The routing table is then converted into a routing matrix and atom connections table via a conversion algorithm that we had developed.

There are 7 guidelines of a lean value stream according to Rother and Shook 1999. In the 3 future states, we simply adopt some guidelines: install supermarkets before HYDRAP, pull the system and synchronize pace of the production to match the pace of demand. As the use of the supermarket atom is specific to each product type and must combine with other atoms such as withdraw (VSMOPRO01) and storage (VSMSTOR01) atoms, and the computation time of generating routing matrix and atom connections table increases tremendously when the number of entities (atoms) increase in the model, in the future states, we simply choose 10 products and compare the lead times of orders from these 10 products, queuing time and machine utilization when the system of them switching from push to pull. Thus, in the current state, there are only push orders, but push and pull orders are combined and flowed through the system in the future states. Totally, there are 31 atoms in the current state model and 84 atoms in the one of future states. The future state value stream map of the 3 scenarios is shown in Figure 2. We explain the scenarios as follows.

- Current State: 975 push orders.
- Scenario 1: Simply install supermarkets for 10 product types before HYDRAP machine. 946 push orders and 29 pull orders.
- Scenario 2:Install supermarkets for 10 product types before HYDRAP machine and the kanbans released from the supermarkets are with priority for processing over other push orders. 946 push orders and 29 pull orders.
- Scenario 3: Install supermarkets for 10 product types before HYDRAP machine and the kanbans released from the supermarkets are with priority for processing over other push orders. Additionally, push orders are released in the pace of takt time which is defined as the available production time divided by the rate of customer demand (Womack and Jones 1996; Rother and Shook 1999). For example, if the customer demand is 16 and the production capacity is 2 shifts (16 hours), the production control will only release one production order per hour. 946 push orders and 29 pull orders

III. Summary of Simulation Results: Current State v.s. Future States

The simulation results are shown in Table 1-4. The machine utilizations have no huge fluctuations except the laser-punch machine in scenario 3 (Table 1). The queuing time before machines are reduced significantly, specially for the high utilization machines like laser-punch and HYDRAP (Table 2). Tables 3 and 4 are the results of 29 orders from the 10 chosen products in the current state and 3 future states. Average lead times are decreased nearly half from current to scenario 3 of the future state and the average value-added ratios are raised about 15% (Table 3-4). The effects of lean are plainly shown by applying our simulation-based VSM to this case study.

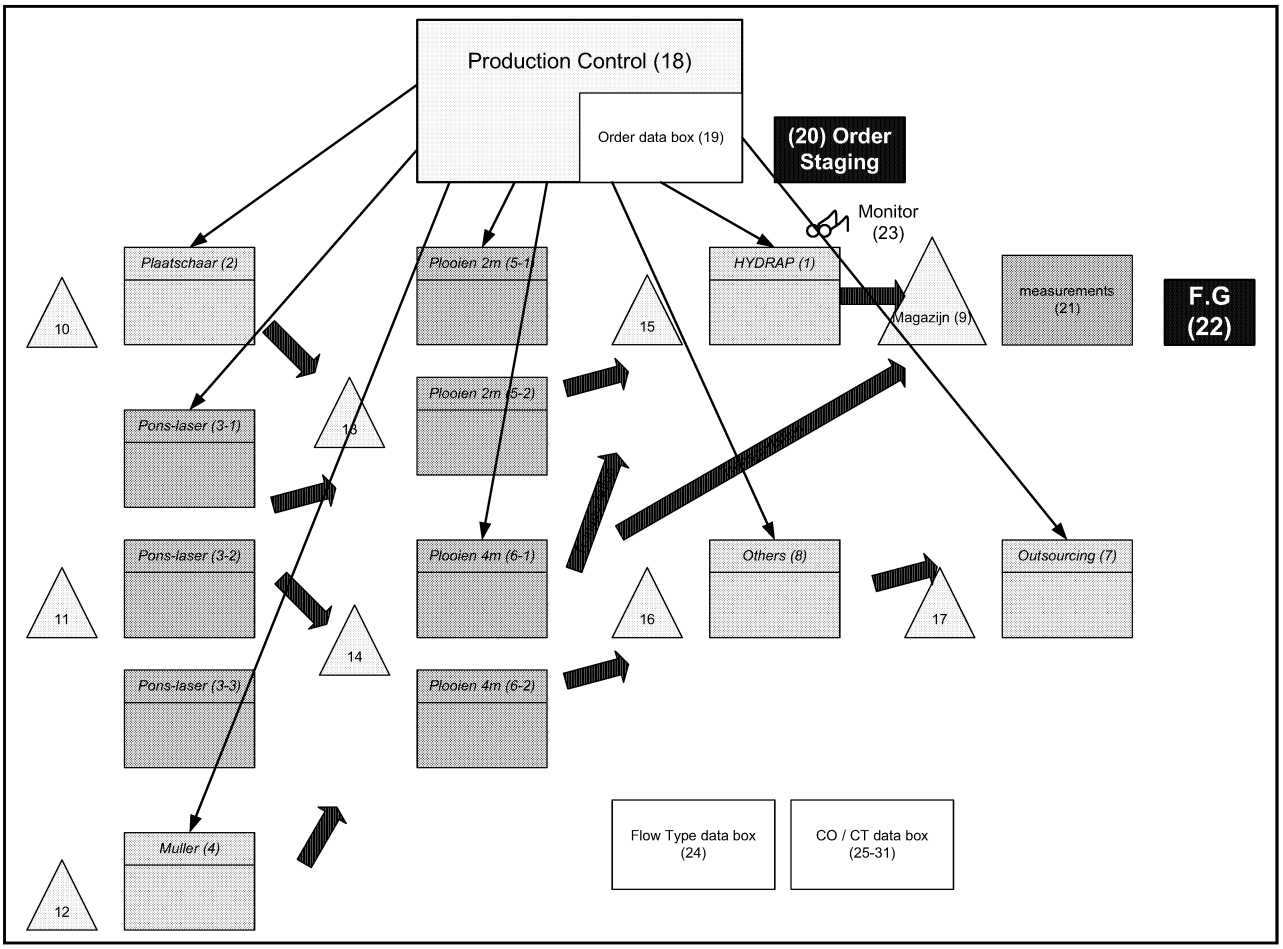


Figure 1. Current State VSM of the Case Study

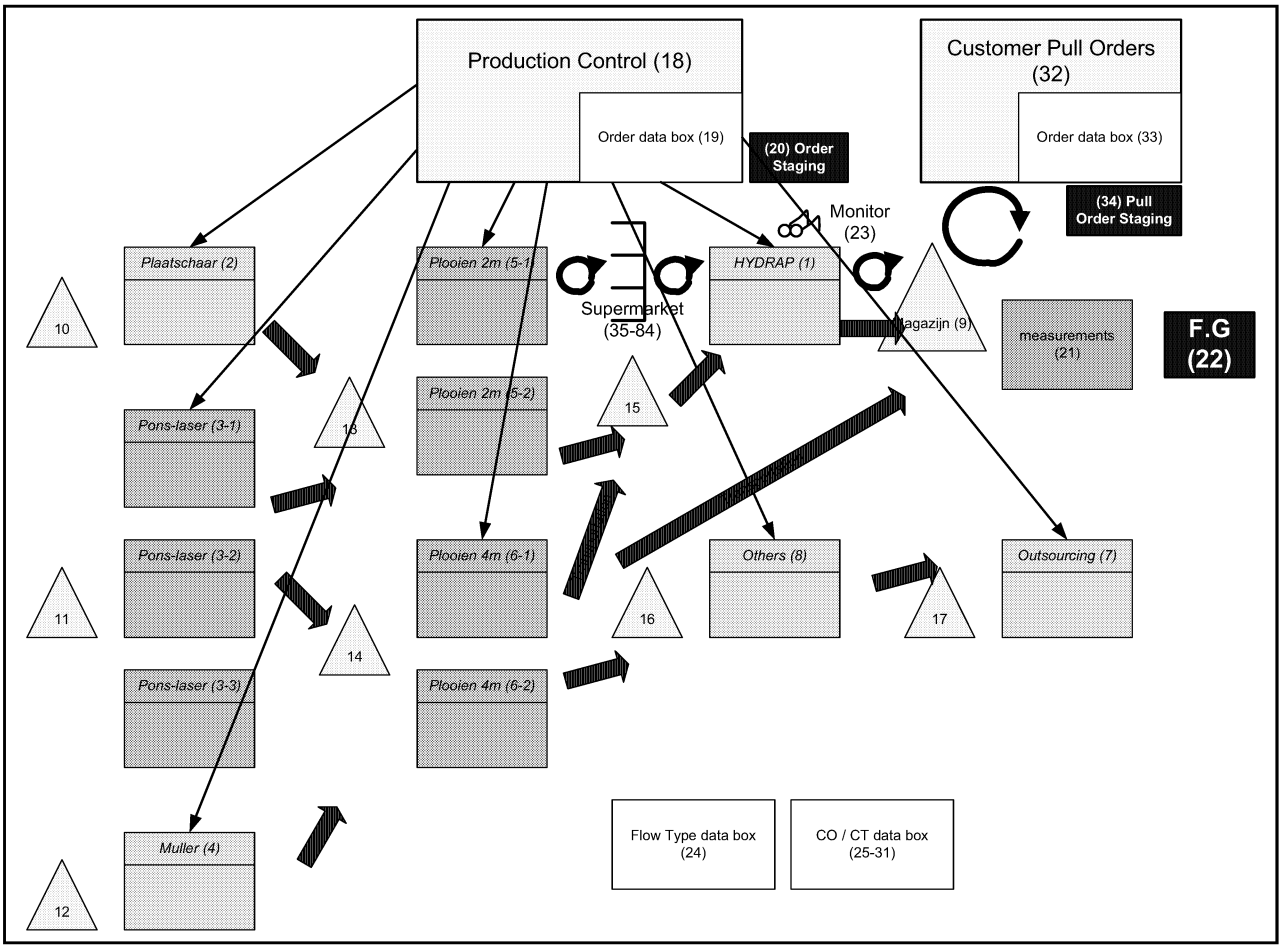


Figure 2. Future State VSM of the Case Study

Table 1. Machine utilization (%): Simulation results (1) *

Work Station	Shear	Laser-Punch	MULLER	Bending<2m	Bending<4m	HYDRAP
Current State	2.1	53.2	19.1	55.3	10.6	55.3
Scenario 1	2.1	51.1	19.1	55.3	10.6	55.3
Scenario 2	2.1	51.1	19.1	55.3	10.6	55.3
Scenario 3	2.1	72.3	19.1	57.4	10.6	55.3

* Based on 2 shifts (16 hours)

Table 2. Waiting time (minutes): Simulation results (2) *

Work Station	Shear	Laser-Punch	MULLER	Bending<2m	Bending<4m	HYDRAP
Current State	8.41	265.25	159.04	122.85	8.77	347.22
Scenario 1	8.41	255.65	159.34	139.14	10.22	339.96
Scenario 2	8.41	251.93	159.34	122.80	9.81	331.01
Scenario 3	0.00	63.57	82.70	112.95	12.34	291.87

* Based on 2 shifts (16 hours)

Table 3. Lean Indicators (minutes): Simulation Results (3)

Indicators	Average Lead-Time	Average Value-Added Ratio (%)
Current State	1810	43.80
Scenario 1	1479	48.40
Scenario 2	1116	51.95
Scenario 3	973	58.49

Table 4. Lead Time Decreasing Rate (%): Simulation Results (4)

From / To	Current State	Scenario 1	Scenario 2	Scenario 3
Current State		18.30	38.36	48.23
Scenario 1			24.56	36.64
Scenario 2				16.02
Scenario 3				

IV. Product Families

In the book "Learning to See" (Rother and Shook 1999), it is mentioned, "One point to understand clearly before starting is the need to focus on one product family. Your customers care about their specific products, not all your products…". One would normally start by mapping a large-quantity and high-revenue product family. A product family is defined as "a group of products that pass through similar processing steps and over common equipment in your downstream processes". Here, we create 2 scenarios to have an insight of simulation results between aggregated and disaggregated data. In the study case, there are hundreds of products with 46 different routings. However, more than 80% of product sales can be classified into only 10 routings (Table 5). Thus, we eliminate the products from other routings (975 push orders reduced to 798) and categorize the product type by its routing. Hundreds of product types reduce to 10 at once. 798 push orders are run in 2 ways, aggregated and disaggregated, respectively (scenario 4 and scenario 5). We explain them as follows.

Table 5. Product Families

Family Nr = Product type	Route Nr	Routing Steps	% of 975 orders	% of 798 orders
1	34	3-5-9	30.87	37.72
2	39	3-9	12.62	15.41
3	30	3-5-8-9	9.95	12.16
4	28	3-1-9	8.21	10.03
5	37	3-6-9	5.54	6.77
6	19	3-8-9	3.69	4.51
7	32	3-5-1-9	3.59	4.39
8	46	4-9	2.97	3.63
9	5	1-8-9	2.36	2.88
10	38	3-7-9	2.05	2.51
Total			81.85	100.00

- Scenario 4: Aggregated data. All the product types belonged to the same routing become one product. Totally, 798 push orders consist of 10 products. Weight average processing times are used on each machine. For example, suppose there are 3 articles belonged to routing 1 and we cluster them as Product 1. If the total order quantities of these 3 articles during the experiment period are (20,10,40) and their specific processing times on the machine HYDRAP are (10,30,5), now the processing time of Product 1 on the machine HYDRAP will be equal to 10 ((10×20+30×10+5×40)/(20+10+40)).

- Scenario 5: Disaggregated data. All the product types remain. 798 push orders consist of 308 products. Each product type has its processing time on each machine.

Because the total released orders (total workload of the machines) reduced from 975 to 798 (about 18%), we also decrease 18% of the machine capacity from 2 shifts / 16 hours to 13 hours of the 2 scenarios when generating the simulation models.

V. Summary of Simulation Results: Aggregated Data v.s. Disaggregated Data

The simulation results of scenario 4 and 5 are shown in Table 6-8. As we can see, there is no difference in the result of machine utilization between aggregated and disaggregated data. The utilization of HYDRAP falling down from 55.3% of the current state to 28.9% of scenario 4 and 5 is due to the reason that the large-quantity and high-revenue product families (the 798 orders chosen) do not need to be processed by HYDRAP. In another word, HYDRAP deals with quite amount of eliminated orders (small-quantity families) in the current state.

In the simulation results of queuing time and lean indicators, it turns out that there is no big distinction between 2 scenarios. If there is no significant difference of running simulation model by aggregated data or disaggregated data, then we can diminish the complexity of the simulation models by just running the aggregation data. For instance, in the scenarios of future states, we only can see the simulation results of 29 orders switching from push system to pull system because of high complexity of product types. The use of supermarket atom is specific to each product type and high complexity of product types means tremendous atoms in the model and long computation times. If the aggregated data are sufficient to represent disaggregated data, this will solve the difficulties that we are confronted with in the model generation of the future states.

VI. Conclusions

In this paper, the current state model of a real case and 3 scenarios of VSM future states are generated to see effects of lean by transforming part of the system. Lean effects are simply shown and proved by our simulation-based VSM formal method. Another important aspect, the aggregation of the data from real world to a representative simulation model is considered in our paper and which to our knowledge has not yet been dealt with by other researchers till now.

REFERENCES

Lian, Y.H. and Van Landeghem, H., 2002. "An Application of Simulation and Value Stream Mapping in Lean Manufacturing". *Proceedings of 2002 European Simulation Symposium*, October 23-26, Dresden, Germany. pp 300-307

Lian, Y.H. and H. Van Landeghem, 2003. "Modeling Logics of Simulation Generator for Lean Manufacturing", *Proceedings of 2003 Industrial Simulation Conference*, June 9-11, Valencia, Spain. pp 258-262.

Lian, Y.H. and H. Van Landeghem, 2004. " A Simulation-Based Value Stream Mapping Formal Method for Lean Implementation: Atoms, Database and Model Generator", *Proceedings of 2004 European Simulation and Modelling Conference*, October 23-25, Paris, France. pp 58-65.

Lian, Y.H. and H. Van Landeghem, 2005. "Simulation-Based Value Stream Mapping: The Formal Modelling Procedure", *European Concurrent Engineering Conference 2005*, April 11-13, Toulouse, France.

Rother, M. and J. Shook. 1999. Learning to See: Value Stream Mapping to Add Value and Eliminate Muda, Lean Enterprise Institute, Brookline, MA.

Table 6. Machine utilization (%): Simulation results (5) *

Work Station	Shear	Laser-Punch	MULLER	Bending<2m	Bending<4m	HYDRAP
Scenario 4	0.0	60.5	18.4	50.0	7.9	28.9
Scenario 5	0.0	60.5	18.4	50.0	7.9	28.9

* Based on 13 hours

Table 7. Waiting time (minutes): Simulation results (6) *

Work Station	Shear	Laser-Punch	MULLER	Bending<2m	Bending<4m	HYDRAP
Scenario 4	0.00	271.65	39.92	71.37	2.44	53.41
Scenario 5	0.00	274.96	52.66	65.84	0.00	54.60

* Based on 13 hours

Table 8. Lean Indicators (minutes): Simulation Results (7)

Indicators	Average Lead-Time	Average Value-Added Ratio (%)	Orders not on time
Scenario 4	724	35.57	128
Scenario 5	686	38.67	135

OEE SIMULATION IN PRODUCTION RAMP-UP

Jürgen Fleischer
Gisela Lanza
Thomas Ender
Institute of Production Science (wbk)
Universität Karlsruhe
Kaiserstr. 12
D-76128 Karlsruhe, Germany
E-mail: Lanza@wbk.uka.de

KEYWORDS
Industrial control, Mechanical engineering, Decision-making, Forecasting, Process-oriented

ABSTRACT

Ramp-up of a new production system has become a decisive factor influencing profit. The reason is an incessantly growing product range with constantly decreasing life cycles therefore ramp-ups have to become more efficient and faster. Characteristic for ramp-up processes are instable business processes, leading to a large number of quality problems. The approach of Overall Equipment Effectiveness (OEE) simulation presented in this paper allows simulating the instable state of a production system during ramp-up. Thus, the production planner is supported during ramp-up planning and the actual ramp-up period. Using OEE Simulation the planner can identify deviations from the target ramp-up curve at an early stage.

INTRODUCTION

Due to decreasing product life cycles and the constantly increasing diversification of the product range, the ramp-up phase poses an important challenge. As a result, companies especially in mass production have to successfully accomplish an increasing number of ramp-ups within a short period of time (Terwisch 1999). The relevance of fast ramp-up is outlined in the following paragraphs.
Ramp-up is the phase between development and production. Commonly, product development and production development run almost parallel, which is called Simultaneous Engineering. While the two areas of planning are still kept separate at this point, they meet directly during ramp-up. The product is still modified after the first machines for production have been set up, so that the production system has to be adapted accordingly.
In the last phase of ramp-up, the so-called time-to-volume, the complete product is produced under commercial production conditions for the first time. Starting with Job No.1, which is the first unit for sale, the daily output is increased gradually until reaching the full scale of the production. With regard to automobile production, the time-to-volume period is currently up to six months (Schmahls 2000).
A simulation-based approach to support a controlled fast ramp-up will be presented in this paper. With simulation, the complex interrelations of highly networked production processes and a multitude of time-dependent stochastic variables can be taken into account (Kuhn 1997). Also, the related dynamic effects, which are only insufficiently described with conventional analysis methods, can first be modeled and then optimized (Feldmann 1997).
The approach shown in this paper is structured as follows: The next section outlines the importance of "Time to market" and quality during production ramp-up. Then the OEE is introduced and transformed to an time-variant key figure to measure the ramp-up success. In the next paragraph, the simulation based OEE and the simulation technics are introduced. The paper concludes with the implementation example of an axle production of a large scale manufacturer and an outlook on future research.

IMPORTANCE OF RAMP-UP

The aim of production ramp-up is a fast launch of new products, that means reducing the Time-to-market. The reduction of Time-to-market can be achieved by increasing the effectiveness and efficiency of production ramp-up (Figure 1). Effectiveness means increasing output per time and therefore to achieve the planned output at an earlier stage. Efficiency means an optimal input-output ratio.

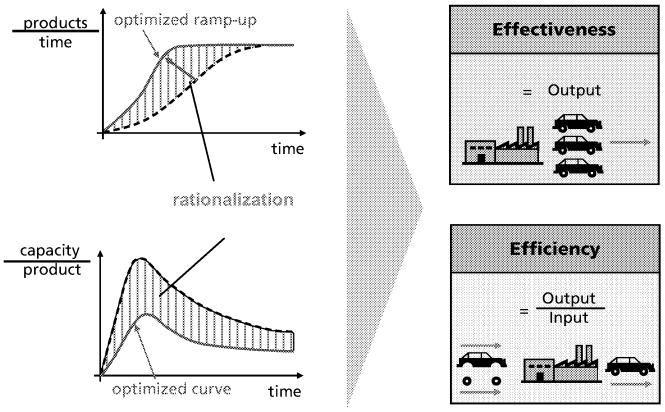


Figure 1: Time-to-market during production ramp-up

Product quality is the most influencing factor for these two targets. Base of reaching these targets is high system quality consisting of technical and organizational components. Just a high product quality and a controlled production process guarantee the necessary effectiveness and efficiency.

Based on this insight in production ramp-up time-variant figures are defined on the OEE, (Formula 1).

OEE = Availability(t) * Performance(t) * Quality(t)

Formula 1: OEE-calculation

The OEE belongs to the measurement methods of the Total Productive Maintenance (TPM)-theory. The basic idea of TPM is that each machine and manufacturing system has an optimal point of work. This point hardly can be reached because there are different sources of losses. These TPM-sources of loss (Figure 2) can be divided into three different categories: availability losses, quality losses and performance losses (Nakajima 1988).

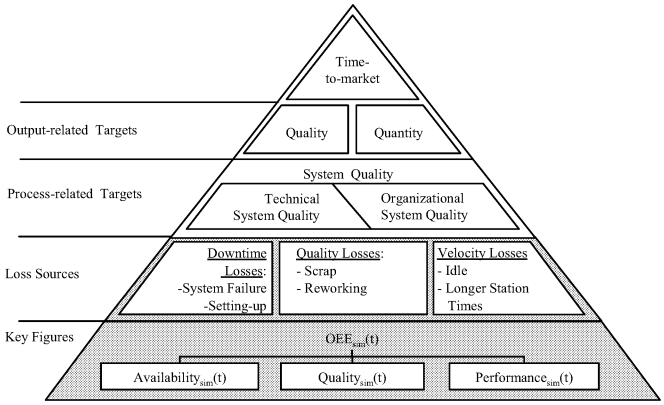


Figure 2: TPM-sources of loss

To realize a fast Time-to-market the output-related targets quality and quantity have to be reached. These targets depend on the system quality, which is a process-related target and can which be found in the technical system quality and the organizational system quality. The system quality is influenced by three kinds of losses. Availability losses result from system failures of the production process, the setting-up and rigging processes, e.g. the exchange of tools. Quality losses appear through scrap and rework, as well as the time losses after a machine stops. Performance losses are idle runnings, longer station times and short stops of the production system as well as a reduced cycle time. The OEE-Key-Figure is calculated by multiplying the availability-, quality- and performace-figures and is used to show the best possible production potential. This point is the 100% benchmark. In the following approach, the OEE-Key-Figure is used during production ramp-ups to measure the ramp-up success. This key figure allows the measurement of the two targets quality and output at the same time. (Nakajima 1988).

TIME-VARIANT OEE-KEY-FIGURE FOR BENCHMARKING THE RAMP-UP SUCCESS

The common OEE-aproach is used during stable production processes. The following approach extends the OEE in two directions. First, to a time-variant key figure, that describes the development of the production system over time and its instable conditions during production ramp-up. It is possible to forecast the OEE and so the development of the ramp-up. This information is used to analyze the current situation and to control the production ramp-up.
Moreover, the OEE normally just considers one single machine.
The expanded key figures used in this approach assess the whole production system consisting of interlinked machines and manufacturing systems. The procurable total value of the OEE depends on the OEE values of the single machines and manufacturing systems and moreover on the interlinked material flow. The basic options to interlink two consecutive machines and manufacturing systems are the loose and the rigid one. In practice often a mixture of interlinkings is used, e.g. sections of rigid interlinked work stations are interlinked by an interposed buffer memory in a loosely way (Etteldorf 1999). A loosely interlinked material flow is taken into account with the loss of time, idle runnings as a result of empty input buffers, which should supply the work station or full output buffers which block the work station. The OEE of a production system with a loosely interlinking principle is not analytically definable. This depends on the used buffers, which limit the system failures (Etteldorf 1999). Therefore, the enhanced key figure $OEE_{sim}(t)$ is established to determine the OEE of the production system with time-variant system states. Simulating the $OEE_{sim}(t)$ brings the benefit to include the instable system states into decision-making and to forecast the $OEE_{sim}(t)$ development. This is not possible in an analytical way.

SIMULATION BASED OEE-PROGNOSIS DURING PRODUCTION RAMP-UP

According to the approach, this enhanced key figure consists of the three following single values: $availability_{sim}(t)$, $performance_{sim}(t)$ and $quality_{sim}(t)$ (Figure 2). Main task of the approach is to forecast the time-variant $OEE_{sim}(t)$ by implementing the created simulation model. Therefore, all influences on these three relevant parameters have to be considered:

- $Availability_{sim}(t)$: The time-variant availability of resources are modeled by a time-variant distribution of the interruption frequency and interruption duration. The early failures of the mechanical components for example are represented by a Weibull distribution (Formula 2).

$$\text{Availability}_{sim}(t) = \frac{\text{planned operating time} - \text{down time}}{\text{planned operating time}}$$

Formula 2: Calculation $Availability_{sim}(t)$

- $Performance_{sim}(t)$: The variant cycle times during ramp up have to be modeled as well. The effects of the time-variant parameters on idle runnings and longer station times are calculated by the simulation (Formula 3).

$$\text{Performance}_{sim}(t) = \frac{\text{planned cycle time}}{\text{real cycle time}}$$

Formula 3: Calculation $Performance_{sim}(t)$

- $Quality_{sim}(t)$: The number of scrap and rework is determined by quality development of processes during production ramp-up and is modeled by quality capability curves of enterprise activities (Fleischer 2003). Enterprise activities are defined as elementary production processes like drilling, grinding or milling that can not be divided in more detailed units. Central idea of this approach is the term "quality capability curve". The term "curve" expresses the fact, that the quality capability of a production process is not stable during production ramp-up. As a consequence, quality capability of an enterprise activity is not a constant value. Due to the impact of influencing factors on enterprise activities, the quality capability of an enterprise activity is developing (Formula 4).

$$Quality_{sim}(t) = \frac{\text{high quality parts}}{\text{produced parts}}$$

Formula 4: Calculation $Quality_{sim}(t)$

These key figures allow to evaluate the instable production system during production ramp-up. Therefore a simulation is used, which handles the quality aspects and the time-variant parameters. The process-oriented OEE simulation tool QSim creates analyses in form of simulated $OEE_{sim}(t)$ key figures, which represent different system states in production ramp-up. The base of the simulation and OEE-prognosis is a process view on the company during production ramp-up. Processes, which have a clear impact on availability, performance and quality, will be identified. The structure of simulation processes are described by three characteristics: the process member, the order of interaction between the process members and the process results, which are key-figures, that influence the components of the $OEE_{sim}(t)$. The process members are characterized as resources. Figure 3 shows processes, process members and their interactions. Maintenance processes like emergency or standard maintenance processes have a strong influence on availability. A lack of availability has two reasons. First, a technical interruption leads to a machine stop and reduces the availability. Second, a lack of resources like a badly installed maintenance process or a lack of qualified workers also influence the availability in a negative way. The aspect of quality results from technical processes. Performance is paid attention by changing cycle times during ramp-up. Some processes are identified with an direct impact on production quality what is measured with quality rates and modeled with quality capability curves. Quality capability curves collect the quality ramp-up experience from different ramp-ups for one enterprise activity. Enterprise activities are used to concentrate quality knowledge of similar reference processes to standardize them and to support an easy modeling technic for the enterprise specific production (Lanza 2004). So it is possible to forecast the quality development with statistical technics. Figure 7 shows an exemplary development of a quality capability curve for an drilling operation during the ramp-up process. At beginning of the ramp-up, the statistical spread is quite high and, after a certain number of produced drill holes, opimization of operation and gain of experience of process members, the process becomes more stable and the statistical spread decreases. This development is typical for ramp-up processes and can be stored in quality capability curves and can be used for a simulation-based OEE-prognosis.

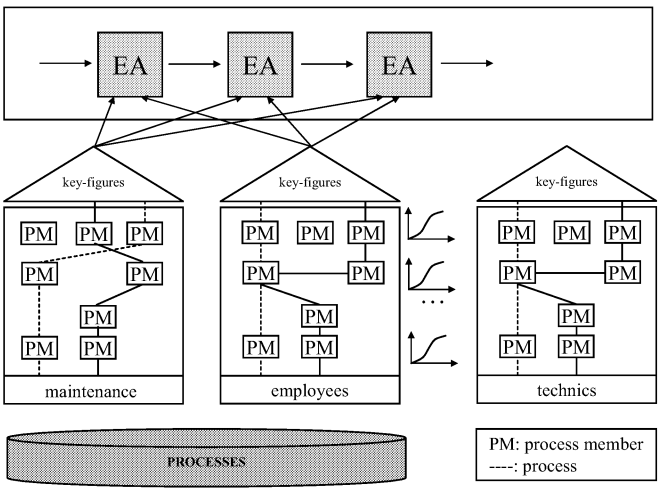


Figure 3: Collection of process information by interaction of processes and process members

Figure 3 also shows the time-variant aspect of the simulation approach. The process members gain knowledge during ramp-up and influence the process key figures through their experience. The modeling and implementation of this growing experience and knowledge is the basic idea of the simulation.

IMPLEMENTATION OF THE OEE SIMULATION

Implementing the simulation in a preliminary way, four different basic technics are used. The modeling of QSim is done with MS Visio. The storage of the model, of temporary data and of the simulation results is realized with an MS SQL-Server. The simulation algorithm, which represents the simulation logic, is implemented with C++. Analyses are done and are visualized with MS Excel.
For modeling the processes, an application in MS Visio, so called SimVisio was developed, which allows to design the processes and to parameterize the process. The process members can be chosen by drag and drop and can be connected to processes in the same way. The implementation is realized at this stage by macros and MS Visio standard functions. At this level, it is also possible to describe the processes with the necessary parameters, like key figures, process names, names of process members and their estimated influence on the process. This information is transferred to the MS SQL-Server, which is used to store data from SimVisio. The database is next to the simulation algorithm the central element of the simulation process (Figure 4).

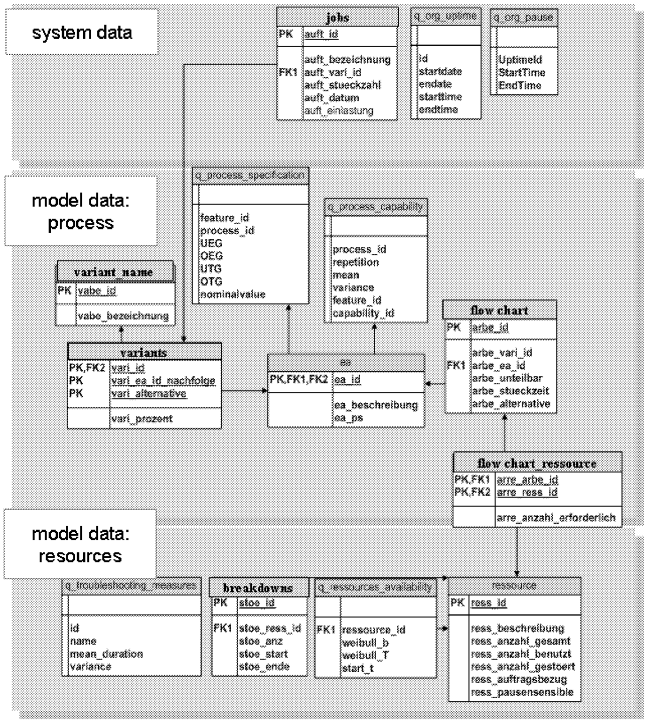


Figure 4: Entity-relationship model of the database

According to the SimVisio configuration of the simulation model, the information in the database is divided into system data, model data process and model data resources. System data are orders, shift time and breaks. Model data process includes e.g. information about the production process, process members, process key figures and their attributes as well as the order of the production processes, if different versions of one product are produced. Model data resources consist of tables, which include information about the risk of technical breakdowns or the unavailability of resources. This is realized by different parameters of the Weibull distribution for different ramp-up phases.
The simulation logic is stored in the database and in a C++-application, which runs the simulation algorithm. During the simulation, the C++-application uses the information in the database and writes temporary data into the database. Figure 5 shows a simulation run. Jobs are executed with enterprise activities.

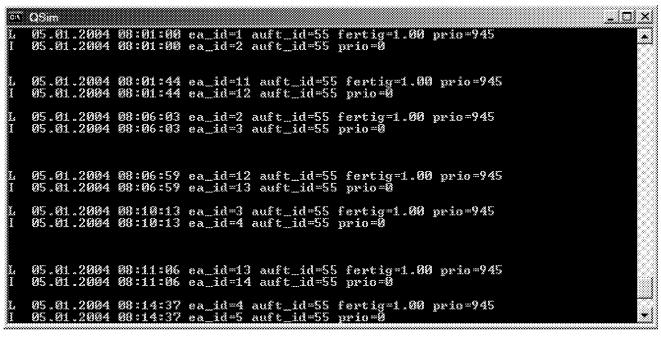


Figure 5: OEE simulation run

Result of the simulation is an OEE-prognosis calculated by the key figures $availability_{sim}(t)$, $performance_{sim}(t)$ and $quality_{sim}(t)$. It summarizes the results in the key figure $OEE_{sim}(t)$. A job is defined as a product, that has to be produced. The attribute ea_id identifies the current enterprise activity, that is done. The attribute auft_id marks the current product, that is produced and the attribute prio is used to prioritize the jobs. For these jobs and enterprise activities, the components of the $OEE_{sim}(t)$ are calculated and stored in the database for later use. Whereas, the database is static the C++-simulation code is the dynamic part of the simulation. The availability, performance and quality information for the enterprise activities result from the process analyses done during the simulation run. The order of the jobs and the use of the enterprise activities are determined during the configuration phase and stored in the database. If the simulation algorithm is started, five steps are executed until all jobs are finished and the OEE components are calculated (Figure 6).

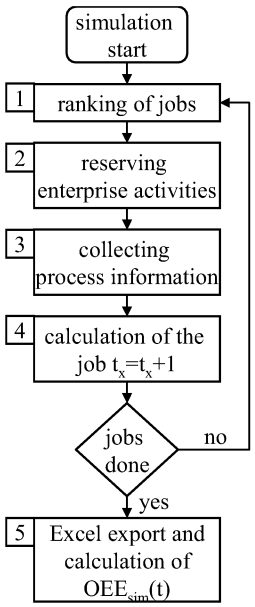


Figure 6: simulation logic

First, the simulation algorithm ranks the jobs by using the priorities, which are given by the production planner. That means, that there are variants of the product which have to be produced more urgent then other variants. In the second step, necessary enterprise activities are reserved. It is possible that different variants of a product use different enterprise activities, therefore the necessary ones have to be identified. The process results are collected process information by simulating the capabilities of the process members and their influence on the OEE components. The information about availability, performance and quality are saved for a certain job at a certain time. Fourth, the process results are simulated for the needed enterprise activities. Now the priorities are calculated again and the enterprise activities are reserved for the next job. After all jobs have been done by this logic, the stored information about all jobs is summarized. At the end, $availability_{sim}(t)$, $performance_{sim}(t)$ and $quality_{sim}(t)$ values and therefore the $OEE_{sim}(t)$ for the executed jobs is determined. After the simulation was successfully finished, the simulated results are exported in a MS Excel file for visualization. MS Excel standard analysis functions are used to summarize the simulation results and to create charts and plots. Figure 8

shows an example visualisation of the following exemplary application in the field of axle production.

EXAMPLARY APPLICATION OF THE OEE SIMULATION

The validation of the theory of time-variant OEE simulation during ramp-up processes and instable production systems as well as the use of the implementation QSim is shown in the following example. There the diameter of a steering knuckle and the drilling operation of an axle supplier in the automotive industrie. The product as well as the enterprise is typical for a large-scale manufacturer with a small number of variants. Clearly installed processes and necessary measuring data are the pre-requisites for implementing and validating the $OEE_{sim}(t)$-prognosis with QSim. Especially the circumstances , that a new product was produced on existing machines in a stable environment support the creation of the simulation model. The modeling of the quality capability curves is demonstrated exemplarily at the diameter of the steering knuckle. The technical factors and their resulting C_p- and C_{pk}-values are well-determined. The C_p-value, process capability, is a simple and straightforward indicator of process capability, whereas the C_{pk}-value, process capability index is an adjustment of C_p for the effect of non-centered distribution. Main input factors are diameter, rotary speed, feed rate, geometry of cutting edges as well as tools and material. Based on these input factors, influencing parameters and process members on a higher level like worker qualification, maintenance and the machines are regarded. Further, the development of the quality, the availability and the performance has to be taken into account. This is realized by defining the quality capability curves based on empirical data. The quality values of the reference process during ramp-up is stored in detail and correlated with the estimated process. The impact on variable consumption of resources will be simulated. Figure 7 shows the nearly constant development of the quality characteristics of the enterprise activity “drilling”.

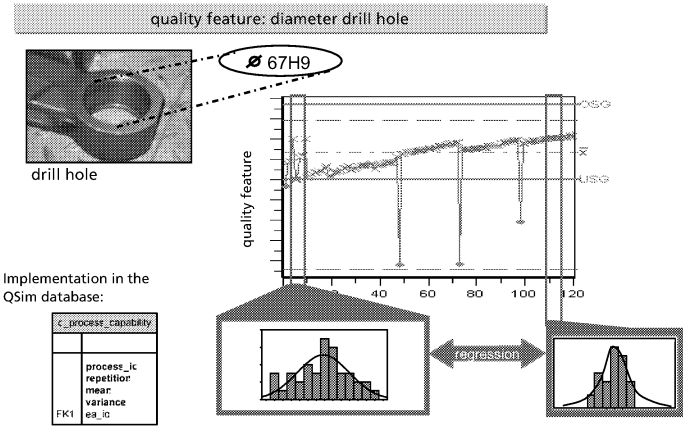


Figure 7: Example quality capability curve for a drilling process

Figure 8 describes the components of the $OEE_{sim}(t)$ result of the simulation. The $OEE_{sim}(t)$ visualization shows a typical quality development of the $OEE_{sim}(t)$ components $availability_{sim}(t)$, $performance_{sim}(t)$ and $quality_{sim}(t)$ during the ramp-up process.

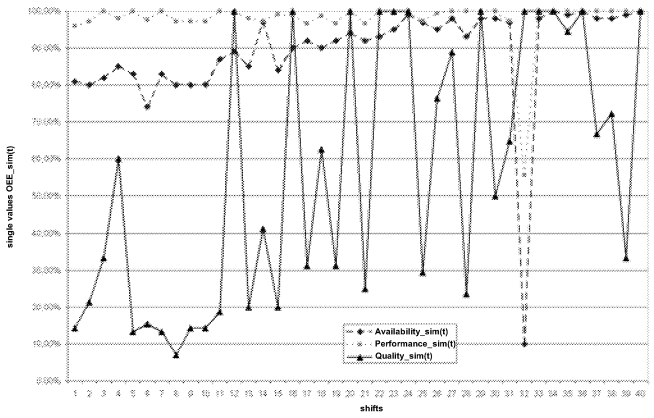


Figure 8: MS Excel visualization of the $OEE_{sim}(t)$

The illustrated implementation example enables an early evaluation of planning and makes sure that planned processes will be capable on time.

SUMMARY AND OUTLOOK

The challenge of fast ramp-up is considerably gaining importance as the frequency of ramp-up processes is constantly increasing. This paper outlines a concept, which applies OEE simulation to enable a controlled transition of the production system into a new state of a serial production. Further research should be done in continuation of quality-forecast, based on the quality capability curves as well as in the creation of a ramp-up-control with a holistic support for ramp-up. In addition the prototype-simulation-tool QSim has to be widely completed and optimized, referring to the usability of the software and its organizational integration.

REFERENCES

Terwiesch, C; Chea, K.S.; Bohn, R.E.: *An Explanatory Study of International Product Transfer and Production Ramp-Up in the Data Storage Industry.* University of California: San Diego. p. 29., The Information Storage Industry Centre, 1999

Schmahls, T.: *Beitrag zur Effizienzsteigerung während Produktionsanläufen in der Automobilindustrie.* Chemnitz: Institut für Betriebswissenschaften und Fabriksysteme, Dissertation der Technische Universität Chemnitz, 2001

Kuhn, A.; Reinhart, G.; Wiendahl, H.-P.; (Eds.): *Handbuch Simulationsanwendungen in Produktion und Logistik*. Carl Hanser Verlag, Munich, 1997

Feldmann, K.; Reinhart, G.: *Simulation - Schlüsseltechnologie der Zukunft?* Herbert Utz Verlag, Munich, 1997

Lanza, G.: *Simulationsbasierte Anlaufunterstützung auf Basis der Qualitätsfähigkeiten von Produktionsprozessen*. Karlsruhe: Forschungsberichte aus dem Institut für Produktionstechnik (wbk), Dissertation der Universität Karlsruhe (TH), 2004

Fleischer, J.; Spath, D.; Lanza, G.: *Quality Simulation for Fast Ramp Up (CIRP Best Paper Award)*. Saarbrücken: Proceedings of the 36th CIRP International Seminar on Manufacturing Systems, 2003

Nakajima, S.: *Introduction to TPM*. Cambridge: Productivity Press, 1988

Etteldorf, J.: *Analyse und Verbesserung der Gesamtanlageneffektivität an automatisierten Produktionsanlagen*, Düsseldorf: VDI-Verlag Fortschritt-Berichte VDI Reihe 2 Nr. 541., Dissertation der Universität Siegen, 2000

BIOGRAPHY

PROF. DR.-ING. JÜRGEN FLEISCHER is head of the Institute of Production Science (wbk) at the Universität Karlsruhe (TH). Beforehand he was working in different leading functions in research and development, planning and production with DaimlerChrysler and Bombardier.

DR.-ING. GISELA LANZA is head of Production System at the Institute of Production Science (wbk). She researches in the field of plant planning, methods for Life Cycle Performance und quality management.

DIPL.-WI.-ING. THOMAS ENDER works as scientist in the area of Production Systems at the Institute of Production Science (wbk). He researches in the field of production ramp-up and simulation.

PRODUCTION PLANNING AND JOB SHOP SIMULATION

THE IMPACT OF HUMAN PERFORMANCE VARIATION ON THE ACCURACY OF A JOB SHOP SIMULATION

Peter Lawrence
School of Business Systems
Monash University, Clayton Campus
Victoria 3800 AUSTRALIA
Email: peter.lawrence@infotech.monash.edu.au

ABSTRACT

It has been observed that stochastic discrete event simulations of systems involving human workers demonstrate considerably less accuracy than simulations involving deterministic automated work cells and that this highlights the inadequacy of current methods of modelling human performance (Baines and Kay 2002). This study investigates the impact of two proposed sources of human performance variability on production time in a small job shop, time of day effects and individual worker efficiency, and assesses the accuracy of a traditional, time- and worker-invariant simulation modelling approach on such a system. In the system studied here it was found that averaging human behaviour using traditional discrete event simulation techniques did not produce significant errors, even for large performance variations. In fact, errors appear to be smaller than those found in a similar study involving a balanced manufacturing flow line (Lawrence 2004).

INTRODUCTION

For close to half a century discrete event simulation has been used to model factory production lines and improve manufacturing operations management (Kelton, Sadowski and Sadowski 1998:13, Robinson 1994:3). While the accuracy and reproducibility of fully automated work cells allows simulation models to be deterministic the involvement of human workers creates complications. However, the flexibility afforded by human workers assures their presence in the factory for some time into the future and so addressing their impact on factory operations remains important.

The primary consequence of human involvement in discrete event simulation is the need to model the irregular timing of human activity such as arrival times and work cell cycle times. This is generally achieved by replacing the fixed timing characteristic of automated systems with a stream of pseudo-random numbers generated by an appropriately chosen statistical distribution and thus developing a stochastic rather than deterministic discrete event simulation model.

However, while this is the standard approach it has been claimed by Baines and Kay (2002) that while discrete event simulation results for automated systems are generally good they are very often inaccurate for systems involving human workers.

Baines and Kay explore a possible explanation for this discrepancy suggesting that human behaviour, in particular work performance, is neither stationary nor consistent between individuals. In terms of discrete event simulation methodology this implies that, rather than the traditional stochastic modelling approach of a single time-invariant distribution for each work cell, the model should incorporate a different, time-dependent distribution for each individual worker for each work cell.

A considerable amount of effort has gone into developing an understanding of, and even simple models for, human performance variation. This is particularly prevalent in the Psychology literature. However, though Baines et al. (2003a, 2003b) cite some useful work in the area, a reliable model of human performance suitable for discrete event simulation work cell applications appears still to be some time away.

Given the difficulty involved in modelling these variations in human performance a study of the likely level of errors introduced by this human variation was conducted and reported last year (Lawrence 2004). That study looked at the impact on the modelling of a balanced manufacturing flow line of averaging human variation into a single, time-invariant probability distribution, as is normal practice in discrete event simulation methodology. The results reported in that study showed that, while some errors were introduced, the averaging process did not appear to have a significant effect upon simulation accuracy.

A balanced flow line, is however, a very simple and orderly system where all work pieces pass along a single path each undergoing identical processing in turn. In an effort to determine whether human variation may be more disruptive in a less orderly system this paper extends that work by applying the methodology to a different type of manufacturing system, the job shop. In the job shop model there are numerous different job types each taking different paths through the system's work cells and hence creating a seemingly more chaotic work flow. As with the previous study, this study aims to look at how important some of these human variations are in affecting the accuracy, and hence utility, of a discrete event simulation model in this less ordered system.

HUMAN PERFORMANCE VARIATION

There are many ways in which the behaviour, and hence performance, of human workers may vary from that of machines but this study concentrates on two particular characteristics that are both well accepted and have a direct impact on the time taken to perform a task. They are time of day effects and individual worker efficiency (Smith 1989, 1992).

Time of Day Effects

It is known that many human biological functions and characteristics, such as body temperature, follow a natural cycle and it is generally accepted that this impacts behaviour and performance in some manner. Work in this area appears to have begun early in the twentieth century and was aimed at establishing the optimum time of day for academic teaching (Carrier & Monk 2000).

Aschoff & Wever (1962: cited in Scott 1994) established that these biological rhythms have a period of about 25 hours but they are entrained to the 24-hour day by external time cues. This natural chronobiological cycle has become commonly known as the circadian rhythm (Scott 1994).

Scott cites much evidence to suggest a connection between this circadian rhythm for body temperature and human performance. Tasks with a small cognitive load seem to have a performance variation that tracks the circadian body temperature cycle (Kleitman 1963) while high memory load tasks are phase-shifted from this cycle. Folkard *et al.* (1976) suggested high memory tasks to be in anti-phase with the temperature rhythm and Monk's (1982) analysis used Folkard's data to postulate a medium to high memory load performance cycle that peaks 3-hours ahead of the low load performance cycle.

Monk also noted that superposed on this circadian effect is the additional factor of number of hours worked. Smith (1989, 1992) also discusses the impact of meals on performance including the so called "post-lunch dip".

All of this suggests that, if human performance measures such as cycle times can be adequately modelled by pseudo-random numbers from some statistical distribution the parameters of that distribution are unlikely to be time-invariant.

Individual Worker Efficiency

Clearly not all workers are identical and so it is not expected that they will perform with equal efficiency. Baines et al. (2003a, 2003b) postulate that this will have an impact on model performance.

It is often common practice in manufacturing facilities for many different workers to operate a single work cell over time. Thus, empirical data gathered over several days at a particular work cell, to be used in fitting appropriate statistical distributions for the discrete event simulation model, will aggregate across numerous workers. While this is not ideal it is generally not possible to model individual workers, not just because of the extra effort required but often also for industrial relations reasons. The Baines *et al.* (2003) study installed an automated system that gathered work cell cycles times over a prolonged period on a major automotive production line. To gain worker acceptance, it was necessary to ensure that data pertaining to individual workers could not be identified. Thus the distributions used in a stochastic discrete event simulation model will usually be integrated over the entire workforce employed at the time data gathering occurred.

Other Factors

While a number of other factors may differentiate the performance of human workers from that of machines the aim of this study was to concentrate on just two factors in human performance, circadian time-variance and individual worker efficiency.

METHODOLOGY

Two models were created of an identical job shop. The only difference between the two models was that the first, referred to as the Ψ model had work cell cycle time calculations that included a circadian component (of amplitude γ) and a variation in efficiency (η) of the individual worker carrying out the task. Variations on this model are thus referenced to as $\Psi_{\gamma,\eta}$. The Ψ model, in a sense, represents the "real" system.

The second model, referred to as the M model, had work cell cycle times that were best-fit, time-invariant statistical distributions based on empirical data taken from 5 complete replications of the corresponding Ψ model. Thus $M_{\gamma,\eta}$ is the model with time- and worker-invariant cycle time distributions based empirically on the 5-replication performance of $\Psi_{\gamma,\eta}$.

To assess the M model's success at predicting the performance of the Ψ model each was run for 200 replications and confidence interval estimates made of the elapsed time required to complete 500 units of production based on the null hypothesis that there is no difference between the required production time of the M and Ψ models.

All models were built using the Arena simulation package and empirical distribution fitting was performed using Arena's inbuilt Input Analyser.

The Job Shop

The facility modelled in this study was a hypothetical job shop consisting of six work stations each set up to perform a separate operation and each with the capacity to support only one job at a time. The shop contains four workers who are responsible for producing finished goods to order by visiting a series of these work stations. Each job requires a randomly determined series of operations (and hence work station visits) within the following constraints:

- Each job requires no less than one and no more than six operations,
- A job may require visiting the same work station more than once but never twice in succession,
- The probabilities of requiring 1, 2, 3, 4, 5 or 6 operations are, respectively: 0.2, 0.4, 0.2, 0.1, 0.07 and 0.03,
- The work time required to complete each operation is stochastically determined,
- The work time can be dependent upon which worker performed the operation but not upon the operation (work station), that is, even though the six work stations perform different operations their timing characteristics are identical.

At the commencement of work, each worker moves to the job dispatch computer and receives his first job. He then performs the operations required for that job by visiting each of the specified work stations in turn and occupying that work station for the duration of the required work time. If the required work station is already occupied then he, and any subsequently arriving workers, queue for access to the work station on a first-come-first-served basis.

For simplicity, it is assumed that the time taken to travel between any two work stations, or between a work station and the job dispatch computer, is exactly 9 seconds in all cases, for all workers, independent of any individual human efficiency considerations.

When the job is complete the worker returns the finished product to the job dispatch area and picks up his next job. Once at the job dispatch area the time taken to drop off the finished job and pick up the next job order is assumed to be negligible. It is also assumed that all raw materials are always available wherever and whenever they are required.

Modelling Human Performance

The objective of this study is to determine the impact of some typical human performance variability on the accuracy of simulation modelling and so the aim here is not to produce an accurate model of human performance in any given situation but one that incorporates plausible, representative variations in work cell cycle time as a function of time of day and individual worker efficiency.

As already mentioned, the processing times at each work station were pseudo-random variates, Specifically, they were taken from a lognormal distribution with a mean (μ) of 9.0 minutes and a standard deviation of 0.63 minutes (7% of the mean). This gives a 3σ range of ±21% which accords with the 20% variability target from the company work-study convention reported by Baines et al. (2003) and that was used in the earlier balanced flow line study.

The time of day component was modelled by varying the mean of the lognormal distribution over the course of the shift as a sinusoid, as is commonly the case with modelling circadian rhythms (Scott 1994), with an amplitude γ about a mean of 9.0. More specifically,

$$\mu = 9.0\left[1+\gamma\cos\left(\tfrac{2\pi t}{1440}\right)\right]$$

where t is time measured in minutes.

To model worker efficiency the mean of the lognormal distribution was further modified by multiplication by the efficiency coefficient (η_i) of the individual worker involved. Thus,

$$\mu = 9.0\left[1+\gamma\cos\left(\tfrac{2\pi t}{1440}\right)\right].\eta_i$$

The choice of a cosine function with zero phase shift ,rather than a sine function are any other fixed phase shift was arbitrary. The effect of a phase shift in terms of circadian cycles is to change the time of day at which the simulation begins or, as discussed earlier, represents a variation in the cognitive load of the tasks involved. This may have some effect on the overall results due to simulation runs not taking an integer number of circadian cycles. However, the objective of this study is to determine whether or not the averaging process of conventional discrete event simulation data gathering significantly alters the results of the simulation model and so the phase angle of the original cycle is unlikely to be of importance.

Experiments

The time taken to complete 500 jobs, T_{500}, was taken as the primary measure of system performance. For each replication of the model the shop was allowed to operate until 500 units had been produced, at which point the replication was stopped and the total simulation time recorded. As an additional simplification, the simulation period was neither broken nor divided in any way to represent any form of shutdown or change of workers at shift completion. The time required to complete 500 jobs was of the order of two to three days and so individual worker efficiencies were held constant over that period. In addition, circadian rhythms were allowed to run full cycle, unlike the balanced flow line study where each 8-hour shift restarted the cycle.

As a control, to test that the basic methodology was sound, the first Ψ model pair built was $M_{0,1}/\Psi_{0,1}$, that is, where there is no circadian amplitude and all workers have unity efficiency.

Three further model pairs were then built and tested as follows:

- $M_{0.1,1}/\Psi_{0.1,1}$ to determine the effect of circadian variation In this case worker efficiency remained invariant at unity but the circadian cycle imposed a variation with 10% amplitude on the mean cycle time of each work station. This meant that, over the course of the production run, the mean cycle time varied from 9.9 down to 8.1 mins.
- $M_{0,1\pm0.1}/\Psi_{0,1\pm0.1}$ to determine the effect of variable human work efficiency. In his case there was no circadian variation but the 4 workers were

assigned different values of η_i symmetrically distributed about unity with a maximum variation of 0.1. Specifically, the chosen values were {0.9, 0.95, 1.05, 1.1}.

- $M_{0.1,1\pm0.1}/\Psi_{0.1,1\pm0.1}$ to determine the combined effect of both the circadian and worker efficiency variations described above.

Each replication involved a different job-set (500 jobs) but the same 200 job-sets were used for the each of the model variations to reduce variance in pairwise confidence interval calculations. Additional variance reduction was achieved by using separate random number streams for each work station and then common random numbers between models.

RESULTS

Table 1 shows the resultant production figures for each of the models expressed as a 99% t-statistic confidence interval. The use of 200 replications to determine the mean production volume justifies, via the Central Limit Theorem, the use of the t- (or z-) statistic and also explains the small size of the intervals.

Table 1: Mean Production Times for 500 Random Jobs in Various Ψ & M Models Including 99% Confidence Interval Half-Width.

Model	T_{500}	*½ width*
$\Psi_{0,1}$	3813	16.0
$\Psi_{0.1,1}$	3780	15.0
$\Psi_{0,1\pm0.1}$	3800	15.9
$\Psi_{0.1,1\pm0.1}$	3767	15.0
$M_{0,1}$	3815	16.2
$M_{0.1,1}$	3793	16.0
$M_{0,1\pm0.1}$	3811	16.2
$M_{0.1,1\pm0.1}$	3785	16.0

Effect of Human Variation on System Performance

Considering just the Ψ models, it looks at first view that the introduction of circadian variation produces a small but statistically significant drop in production time within the system. However, this must be treated with some caution because these run lengths did not involve an integer multiple of the circadian period and hence will not have the same mean work station time as the time invariant model $\Psi_{0,1}$. Table 2 shows the mean (and standard deviation) for cycle times from each of the Ψ models aggregated across all replications and work stations (approximately 250 000 data points in each case).

Table 2 Mean and Standard Deviation of Cycle Times for the Ψ Models

Model	μ	σ	*Scaled* T_{500}
$\Psi_{0,1}$	9.00	0.631	3813
$\Psi_{0.1,1}$	8.92	0.900	3814
$\Psi_{0,1\pm0.1}$	8.95	0.952	3821
$\Psi_{0,1\pm0.1}$	8.87	1.14	3822

As can be seen from this table, the circadian variation produces a small decrease in the mean cycle time of work station operations and this must be, at least in part, responsible for the reduction in production time, T_{500}. As a simple transformation to try to remove this effect the production volumes reported in the last column of table 2 have been linearly scaled to allow for this variation in mean cycle time.

An alternative approach would be to run each model for a fixed number of circadian cycles and measure system output in number of completed jobs. This would also be a reasonable approach but has the disadvantage that each run would no longer involve the exact same sequence and mix of jobs and would introduce a small amount of extra variance into the results. This variance may or may not be greater than the variance introduced via slightly different run lengths involving slightly different mean processing times due to the incomplete portion of the last circadian cycle.

Comparison of the scaled data suggests that the circadian rhythm variation may have little or no effect on overall production time. This is perhaps not surprising since the variation is smooth and is driven by time of day. It is hence uniform across the shop and simply slows down and speeds up the operation cyclically without significantly impacting internal interactions.

Worker efficiency variation also produces a small drop in average cycle time, despite the symmetrical variation of η that was used. This is most likely due to efficient workers completing a slightly greater proportion of the work than their slower colleagues.

The net effect of human variation on T_{500}, if any, is extremely small and certainly beyond the statistical resolution of these experiments. This contrasts with the balanced flow line study where human variation was responsible for as much 8% variation in production volumes over a fixed shift time.

The Accuracy of the M Models

Table 3 shows the pairwise t-statistic 99% confidence interval estimates of the difference in T_{500} between each M model and its corresponding Ψ model, thus testing the Null Hypothesis that there is no difference between their production rates.

Table 3: Production Time Variation between M and Ψ Models Including 99% Confidence Interval Half-Width

Model	ΔT_{500}	*½ width*	$\Delta T_{500}/T_{500\Psi}$
$M_{0,1}-\Psi_{0,1}$	1.66	2.16	0.04%
$M_{0.1,1}-\Psi_{0.1,1}$	13.2	2.91	0.35%
$M_{0,1\pm0.1}-\Psi_{0,1\pm0.1}$	10.6	5.48	0.28%
$M_{0.1,1\pm0.1}-\Psi_{0.1,1\pm0.1}$	18.6	5.62	0.49%

The Confidence interval on T_{500} for $M_{0,1}-\Psi_{0,1}$ includes zero and hence the Null Hypothesis cannot be rejected. This, combined with the fact that we can be 99% confident that these models differ by no more than 3.7 minutes (which is less than 0.1%) strongly supports the validity of the methodology.

In each of the remaining cases the corresponding Null Hypothesis can be rejected meaning that there is a statistically significant difference between the production time of the Ψ model and its M model simulation. However, the size of the error is extremely small, still below 0.5% for the worst case of combined circadian and worker efficiency variation. This is well within acceptable experimental error tolerances for the vast majority of discrete event simulation applications and hence suggests that this form of human variation is not an important factor in ensuring the accuracy of a job shop simulation such as this.

In fact, the errors relating to worker efficiency variation found in this study appear to be an order of magnitude smaller than those found in the flow line study. This may reflect the fact that a slow worker in a flow line is more strongly coupled to the performance of other workers than in the job shop. A bottle neck caused by a slow worker in a flow line has higher potential to disturb all workers in the line than the more chaotic flow of a job shop.

Larger Variations

It appears from these results that human performance variations produce modelling errors of substantially less magnitude than the original variations. In this system, circadian and efficiency variations of ±10% produce modelling errors of less than half a percent. To investigate the sensitivity of these errors to magnitude of human variation a further set of models was produced where the amplitude of the variation was increased to 20%. These results are presented in table 4.

Table 4: Production Time Errors for Extreme Human Variation (Including 99% Confidence Interval Half-Width)

Model	ΔT_{500}	*½ width*	$\Delta T_{500}/T_{500\Psi}$
$M_{0.2,1}-\Psi_{0.2,1}$	9.57	4.96	0.26%
$M_{0,1\pm0.2}-\Psi_{0,1\pm0.2}$	-60.6	7.11	-1.61%

This represents extreme variation where the most efficient worker at the top of his cycle performs at 2.25 times the speed of the least efficient operator in a circadian slump. However, the modelling errors produced remain small. For the case of circadian variation the $M_{0.1,1}-\Psi_{0.1,1}$ and $M_{0.2,1}-\Psi_{0.2,1}$ confidence intervals overlap and so this doubling of amplitude produces no statistically significant impact on the mean error. Doubling the size of the worker efficiency variation does, however, more than double, perhaps even triple, the error, although it still remains small at well below 2%.

It could be that this error is, at least in part, due to inadequacies in the initial distribution selection for the M models. The distribution fitting was based on selecting the best fit from a variety of standard simulation distributions such as normal, lognormal, beta, gamma and triangular. Looking at the cycle time data produced by the 5 replications of the $\Psi_{0.2,1}$ model, an example of which is given in figure 1, shows that this may be inadequate. The sharp split between high (1.2 & 1.1) and low (0.9 & 0.8) η values produces cycle times that would be better modelled with some form of bimodal distribution.

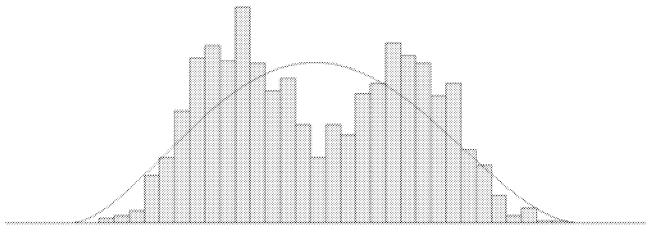


Figure 1: Cycle Time Distribution for Work Station 1 After 5 Replications of $\Psi_{0.2,1}$

Examination of the utilisation of each of the work stations also gives some insight into the system's insensitivity to human variation. Table 5 shows the mean work station utilisations for each of the 6 stations for both the $\Psi_{0,1}$ and $\Psi_{0.1,1\pm0.1}$ systems across 200 replications.

Table 5: Mean Utilisation of Work Stations Across 200 Replications for □0,1 and □0.1, 1±0.1 Models

	Mean Utilisation	
Work Station	$\Psi_{0,1}$	$\Psi_{0.1,1\pm0.1}$
1	56.2%	56.0%
2	49.1%	49.0%
3	50.0%	49.9%
4	49.9%	49.9%
5	49.9%	49.8%
6	44.2%	44.1%

It can be seen that all work stations are running at about 50% utilisation and so there is plenty of capacity and little chance of bottlenecks developing. Although this system is

ostensibly more chaotic than the flow line it is in fact running well below capacity and in that sense is less stressed than the flow line system. Any variation in worker performance is not likely to cause significant disruptions to work flow within the system. In effect, this means that the degree of coupling between workers is likely to be less than was the case with the flow line and efficient workers are able to take the slack left by inefficient workers without other negative effects. The results in Table 5 show that the 10% human variation has little effect on work station utilisations for the low utilisation levels found in this system.

It may be that a system operating close to capacity, with at least one work station running close to 100% utilisation, could be adversely affected by human variation due to slow workers holding up their more efficient counterparts in queues. However, further work with systems running close to capacity would be required to establish whether or not the effect of plausible human variation would become significant.

CONCLUSIONS

This study has suggested that, the effect of system wide temporal variation in cycle time has little effect on the production rate of a simple job shop, even less than was the impact on a balanced flow line. Hence traditional discrete event simulation methods should successfully model such a system accurately.

While this work cannot conclude that human variation is irrelevant to discrete event simulation it does suggest that there are many circumstances in which quite substantial variation need not be specifically modelled to produce accurate results. Nonetheless, it remains possible that many other systems may exist where failure to account for human variation will produce larger errors. It remains to be investigated whether this variation is more important in systems where one or more components are operating close to capacity.

REFERENCES

Aschoff, J. and R. Wever. 1962. "Spontanperiodik des Menschen bei Ausschluß aller Zeitgeber." *Naturwissenschaften* 49, 337-42.

Baines, T.S. and J. M. Kay. 2002. "Human Performance Modelling as an Aid in the Process of Manufacturing System Design: A Pilot Study." *International Journal of Production Research* 40, No. 10(July), 2321-34.

Baines, T.S., L. Hadfiled, S. Mason, J.M. Kay and J. Ladbrook. 2003a. "Human Performance Modelling within Assembly Line Simulations." In *Proceedings of the Industrial Simulation Conference 2003* (Valencia, Spain June 9-11) Eurosis, Ghent, Belgium, 326-30.

Baines, T.S., L. Hadfiled, S. Mason and J. Ladbrook. 2003b. "Using Empirical Evidence of Variations in Worker Performance to Extend the Capabilites of Discrete Event Simulations in Manufacturing." In *Proceedings of the 2003 Winter Simulation Conference* (New Orleans, USA Dec. 7-10) IEEE, Picataway, N.J, 1210-6.

Carrier, J. and T.H. Monk. 2000. "Circadian Rhythms of Performance: New Trends." *Chronobiology International* 17. No. 6, 719-32.

Folkard, S., P. Knauth, T.H. Monk and J. Rutenfranz. 1976. "The Effect of Memory Load on the Circadian Variation in Performance Efficiency Under a Rapidly Rotating Shift System.", *Ergonomics*, 19, 479-488.

Kelton, W.D., R.P. Sadowski and D.A. Sadowski. 1998. *Simulation with Arena.* McGraw-Hill, Boston.

Kleitman, N. 1963. "Sleep and Wakefulness (Rev. Ed.)." Uni Chicago Press, Chicago.

Lawrence, P.J. 2004. "On the Need for Human Performance Modelling in the Simulation of a Balanced Production Line.", *Proceedings of the Industrial Simulation Conference 2004*, (Malaga, Spain, June 7-9), Eurosis, Ghent, Belgium, 415-9.

Robinson, S. 1994. *Successful Simulation.* McGraw-Hill, London.

Smith A.P. 1989. "Diurnal Variations in Performance." In *Acquisition and Performance of Cognitive Skills*, A.M. Colley and J.R. Beech (Eds.). John Wiley & Sons, Oxford, 301-25.

Smith A.P. 1992. "Time of Day and Performance." In *Handbook of Human Performance,* A.P. Smith and D.M. Jones (Eds.). Academic Press, San Diego, 217-35.

Scott, A.J. 1994. "Chronobiological Considerations in Shiftworker Sleep and Performance and Shiftwork Scheduling."

A SIMULATION APPLICATION FRAMEWORK FOR PRODUCTION PLANNING AND SCHEDULING

Hans P.M. Veeke, Gabriel Lodewijks
Delft University of Technology, The Netherlands
e-mail: H.P.M.Veeke@wbmt.tudelft.nl; G.Lodewijks@wbmt.tudelft.nl

KEYWORDS

Model design, Discrete simulation, Decision support, Production control

ABSTRACT

This paper presents a general framework for the application of simulation during the different phases of decision making in production control. Each phase of decision-making requires a different simulation model with different goals and characteristics. The modeler should be aware of these characteristics in order to take full advantage of the use of simulation. The paper shows the decision-making structure and explains the way how simulation could support and improve this. Finally a general framework results with different interacting models, reflecting the decision-making structure of a company.

INTRODUCTION

Currently, simulation is commonly being used for decision support in the design of transport, production and service systems. Models for this type of use were disposable until the last decade. Nowadays, the use of simulation for operational decision support in project planning, production planning and scheduling is growing, although it is not common yet. The general trend is to directly use control algorithms, developed during design, for the real operations (e.g. [ConCannon et al, 2003]). Also some isolated models exist for project planning and production planning /scheduling (e.g. [Chong et al., 2003]), which cover a part of operational decision making. In both cases the major difference with the traditional use of simulation modelling is that the model software is not meant to be disposable, but should operate frequently, user-friendly and integrated in the operational decision making structure.
In this paper a framework is presented for the integrated use of simulation in production design and control. The framework is restricted to customer order controlled production environments.
First a conceptual view of decision-making in production environments will be defined. It will be shown that there are three different levels, each with its own goals, but all are connected in a hierarchical way. In this three level concept, five model types can be distinguished. The relation between these model types will be explained in terms of interface data. The functionality of these models, required to support (and improve) production planning and scheduling will be discussed. Based on this concept, a structured framework for simulation modelling will be derived. The resulting framework is now being used as the basis for further research into simulation integrated production planning and scheduling.

PRODUCTION CONTROL

Decision-making (human or automated) is an essential part of production control. The requirements, possibilities and restrictions of operational production control are defined during the (re)design of a production system.
According the Delft Systems Approach [in 't Veld, 2002], effective production control can only be achieved if and only if all of the next conditions are met:

1. *there is a clear goal*. For a production process the general goal is to deliver the right products at the right time and location in the best economical way.
2. *The goal is feasible.* If a goal is unfeasible then there is no sense in pursuing it.
3. *There are possibilities to intervene*. If one can't intervene then there are no possibilities to react to disturbances.
4. *There is a perception of the consequences of interventions.* Without perception of the consequences, interventions may easily move the production farther away from the goal; in the best case one simply doesn't know what will happen.

The second condition states that a goal has to be feasible. From the viewpoint of the simulation domain, only quantitative feasibility is considered, for example production rates, lead times and stock volumes. During design, a system structure and capacities have to be derived that are capable of achieving the goal in general. The conditions are defined under which the system should operate and the structure and capacity of the system are established to suit these circumstances. The structure is expressed in terms of operational "functions" to be fulfilled and they are combined into a production "organisation". If the original conditions change structurally (when the system has become operational), the system may have to be redesigned. This process will be called 'function design' from now on.
Operational decisions should guarantee the goal's feasibility for specific situations. For example, during operation the order stock is known (real or by forecast) and the state of capacities is known. Feasibility means in this

case that a match can be made between orders, capacities and materials. In other words: will the system function (fulfil its requirements) under these circumstances and how can this be achieved? This is "function control", which delivers a prescription or "standard" for the near future.

The third condition refers to the fact that interventions are required if disturbances cannot be excluded. This is the case in every customer order controlled environment. Function control cannot take specific disturbances into account by definition; it can only provide the capability to deal with future disturbances (e.g. by providing "slack"). In order to deal with it, the standards from function control are used to define the correct intervention. Concrete disturbances in the process, input or output should be solved. This is called "process control".

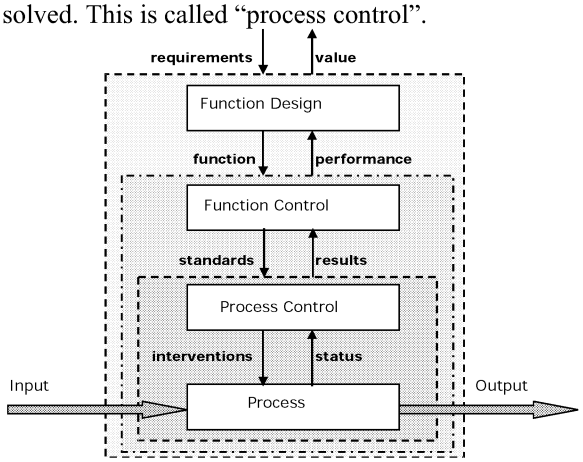


Figure 1. A functional view of production control

The fourth condition, perception of consequences of interventions, is usually based on experience of the decision maker, but it will be later explained that simulation is extremely useful for this purpose, specifically for production planning and scheduling.

The requirements entering the system on top in figure 1 are the result of policy making. Fulfilling these requirements should result in the required company- and/or customer-value, leaving the system at the top. Function design has a lot of possibilities to fulfil them, but the selected function should guarantee feasibility at least. Function design decides on the functions to be fulfilled and resources to be used. The result of the developed system is called the "performance". Function control includes specific circumstances into the general function structure and formulates standards for the coming operation period. These standards are used to prescribe the expected progress of the processes and to decide on interventions if necessary. The results of repetitive executions of the process are being reported to function control, which (re)defines new standards for the near future. Finally, the whole system of function control, process control and the process itself determines the performance of the production system. This conceptual view will now be used to define the position of production planning and scheduling.

PRODUCTION PLANNING AND SCHEDULING

The function of production planning is defined as:

"to provide the best economical way to produce the right products at the right time at the right location for an expected situation".

Production planning usually fulfils its function by determining the orders to be executed and by determining the required capacities and materials for these orders in quantity and time.
The function of production scheduling on the other hand is:

"to provide the release and execution of orders according the conditions of production planning in a certain situation".

For this function, materials and work instructions are collected and the execution sequence is determined. The phrases "for an expected situation" and "in a certain situation" in the definition of planning and scheduling respectively, are essential. The production planning function always uses an idealized image (a model) of the future, where production scheduling has to deal with real (often disturbed) situations. From this it becomes clear that production planning belongs to function control while production scheduling is a function within process control.

Function control monitors the results of execution and checks their accordance to the requirements; it derives standards (makes a production plan) for the process itself that will hold for a certain period of time (the planning horizon). Process control continuously monitors the real state of the process (progress according the production plan, errors, flow rates etc.) and eventually intervenes to keep the process within the defined production plan.

Zooming further into 'Function Control', makes it clear that the production plan is the final result of echelons of similar planning functions with ever increasing levels of detail. These echelons can be found in the usual structure of MRP-II systems (see figure 2).
In this structure the master planning is the leading plan for the material requirements planning and the capacity requirements planning. It shows however the (intermediate) results of the functions, but in no way the processes to achieve them. Above that, there is no connection with the results of function design, which reflect the feasible requirements of the production facility.
For planning purposes the basic element is taken to be an indivisible work unit: "a task". Standards for tasks are required in order to deliver a production plan. These task standards express processing time, setup time, lead-time,

required capacity etc. Combinations of tasks represent complete orders.

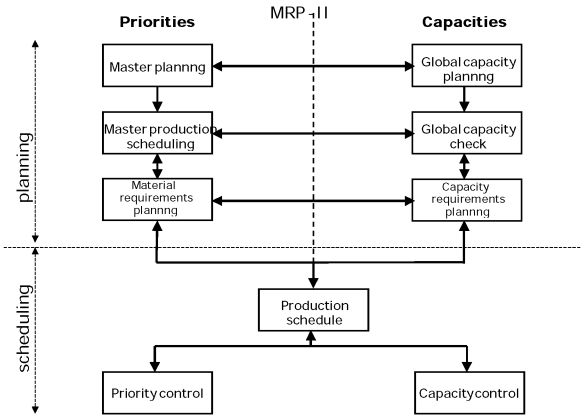


Figure 2. MRP-II planning structure

Orders can be very complex and may require a standard for sequencing, capacity assignment and material use. These complex orders are considered "projects" and all usual methods like CPM, Pert etc.[Wild,1999], may be applied. Therefore the standardization of projects (complex task structures) is considered a separate function, the results of which are used during the formulation of the total production plan. Figure 3 shows the general functions required to make a production plan.
Function design defines both initial data and technological / organizational changes. The results are data included in the "function" description.
Finally it is noted that the interaction between function control and process control should be balanced very carefully. Too much detail in the production plan may result in an ineffective scheduling function, because there is no room left for interventions.

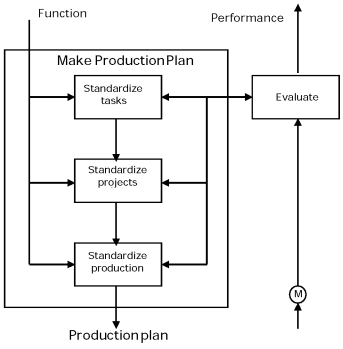


Figure 3. Production planning functions

The result is that the production plan has to be recreated with every disturbance adopting the role of production schedule. On the other hand, production scheduling should not be used to adopt the role of production planning. In this case the principles of modelled future and hard reality are mixed up and the relation with the original goals of the system are lost. A modelled future usually uses averages and expectations; the hard reality fully consists of individual (i.e. non-statistical) data. The hierarchical functional approach as presented here is a precondition for the preservation of a correct view for effective control.

MODELLING AND SIMULATION IN PRODUCTION PLANNING AND SCHEDULING

The previous paragraph explained the functions of plans and schedules to be used for production control. When simulation is being used, these functions also determine the goals of the simulation models.
As explained, there are three function levels where decision-making for production processes takes place:

1. function design
2. function control (planning)
3. process control (scheduling)

At the function design level, simulation is applied in the traditional way. This means that simulation is used to construct a model for a new or improved situation and to investigate the results to be expected. Simulation models at this level are primarily used once only, but a tendency is becoming evident to use the same control software in a real-time environment that has originally been developed for the simulation model. The models at this level will be called "system models" from now on and mostly they are concerned with the layout, organization and productive use of resources assuming order arrival patterns and order compositions. The primary goal is the correct functioning of the system. The results of these models are structures, behaviour descriptions, capacity definitions, stocking requirements and –last but not least- control concepts and algorithms (such as a planning principle), all in order to realize the required performance (see figure 4).

In function control i.e. planning, simulation should be used as an operational tool. The goal of planning (and thus of the models used) is to provide conditions in which the required performance can be reached with a given order stock and/or with forecasted orders. Restrictions that conform to priorities, availability and technical feasibility, are in fact added here to the system definition of function design. A simulation model here represents the real system and focuses on required due dates and minimum costs. Results are (required) available capacity, order and task release dates and material requirements. They all should be checked for feasibility at least. This is the primary goal of simulation models at this level. They are called "planning models" from now on. It is already shown [Veeke,2003] that simulation is able to deal with combinatorial problems that cannot be solved (in time) with analytical approaches.
Orders are task structures. When orders are very complex or when they become repetitive or when they have very strict due date requirements, they can be considered "projects". For these projects the usual project planning methods are applied in order to define optimal sequencing, capacity assignment and lead times. This type of projects is also often encountered in maintenance environments. As shown in figure 3, these project plans precede the

production planning. Actually, a project plan is a standard to be used in the production planning. It has already been shown that project planning with simulation leads to new interpretations of a traditional project characteristic like the critical path [Ottjes, 2000]. Above that, the interaction between project models and production models may lead to new ways of using priorities, capacities and materials. The interaction fits perfectly to the matrix organization principle, and supports decision-making between project managers and department managers. Although there are many multi-project planning models already, they don't distinguish project and production planning goals.

At the scheduling level (process control), the use of simulation is also widespread. It is especially at this level that simulation models try to find optimum solutions for detailed scheduling problems. Nevertheless it is generally acknowledged that "production scheduling in industrial practice is still primarily a manual task, despite the immense research efforts and the fact that manual scheduling is a very complex task" [Wiers et al, 1996]. Therefore the main goal of these "scheduling models" should not be to present optimal solutions from a complex algorithm, but rather solutions that would be found by the human decision maker. The model should primarily represent the behaviour of the decision-maker and show the consequences of the decisions. Here the use of simulation offers the added value of giving perception and understanding of consequences, the fourth condition for effective production control. It is also not useful to implement algorithms that use data that are not available in reality. For example, a rather simple scheduling algorithm like "minimum slack" is based on the slack of tasks at the moment of decision making; It assumes almost automatically that the slack is continuously known. However, most small to medium sized companies register progress once a week and therefore the decision-maker does not have this attribute and will not use it. This example makes it also clear that simulation can be used to show the consequences of better decision-making when new technologies like barcodes and RFID (Radio Frequency Identification) enable the continuous registration of progress.
Figure 4 shows the resulting structure of models and relations, derived from the decision-making structure of figure 1. At the bottom the term "process model" is used for simulation models that describe the real processing.

THE USE OF THE FRAMEWORK

The framework represents the 5 different types of simulation models and their relations that can be used during design and operation of a production system.
The contents of each of these model types will be discussed shortly.
The *System models* should offer the facilities to support the definition of production systems according the production requirements and an expected or known disturbance behaviour. For this purpose they should be able to calculate the occupation of a system stochastically, where a workload is assumed. Production patterns, a maintenance program and the disturbance behaviour of the equipment usually specify this workload. Above that they should be able to determine stock levels and tool demands, in order to perform the production process.

A *project model* is part of function control. It is capable to define repetitive complex orders. It takes available capacity into account, materials, tools, job sequences and processing times.
In general these project models offer the following facilities:

- Stochastic calculation of a project structure, with pessimistic, average and optimistic task times
- Determination of a realistic lead-time with finite and infinite capacity
- Determination of critical activities with different acceptable project lead times.
- Determination of project costs
- Determination of the tools and materials requirements.

Project models are also very useful in maintenance environments for the determination of a preventive maintenance program.

Planning models should offer the same facilities as project models, but for non-related orders and projects in restricted capacity environments.
Scheduling models show the consequences of different scheduling rules based on the behaviour of the decision-maker(s). Such a model is able to:

- Deal with capacity calendars and all kinds of shift schedules
- Deal with capacity restrictions for each task
- Deal with job structures and sequencing for projects and activities
- Establish orders and projects in predefined periods and to show the consequences of deviations therein
- Fix tasks in predefined periods (both automatic and manual) in the planning schedule
- Add capacity automatically under predefined conditions
- Take material and tool requirements into account.

Process models can be made as detailed and specific as required. The models can be logistic oriented, and in that case model the operational specifics very globally, because the main interest lies in the sequence of operations. There are also numerous technological models that describe a process in detail for example by means of differential equations.
Figure 4 finally shows the most important interface data between the models that can be used.

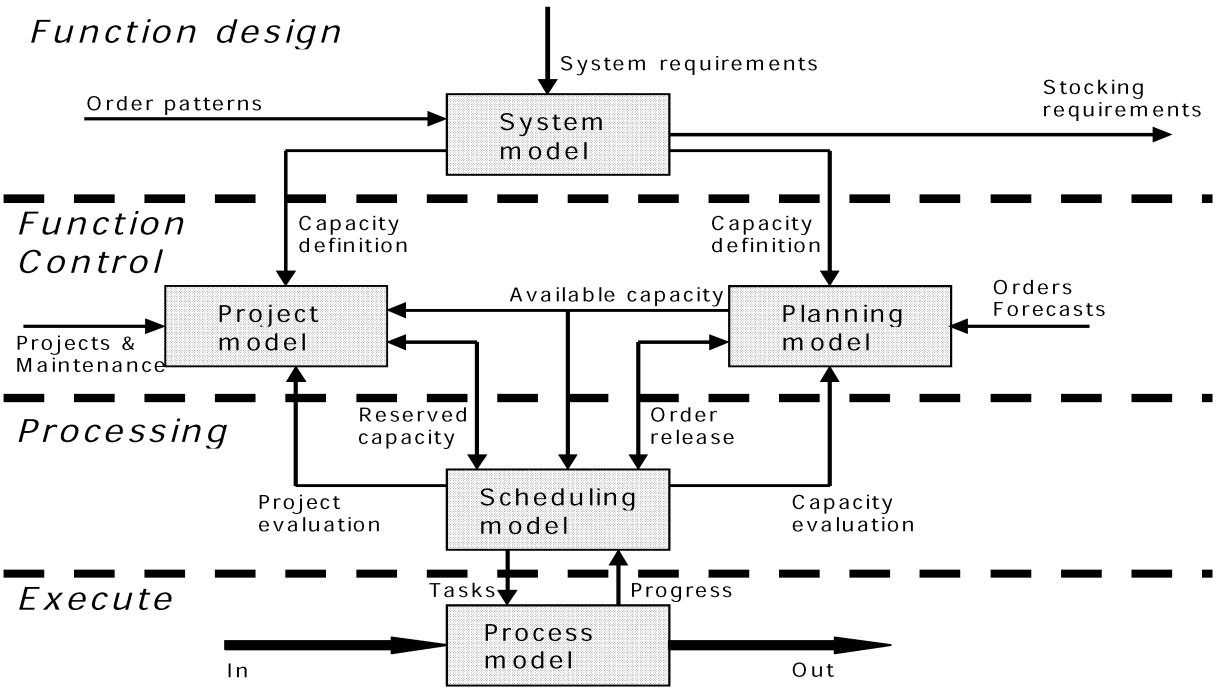


Figure 4. The simulation application framework

CONCLUSIONS

The framework shows that;

- the goals of the model types differ, each representing only a part of the functions that are required for production design and control.
- A model originally meant for one of these functions cannot be automatically extended to serve another one. Including a lower function level in an already developed model requires extra data and more detail; including a higher level would introduce more degrees of freedom.
- The models serve functions, which are part of a well-defined decision-making organizational structure. If the organization has separated these functions, one should be careful to combine them into one single model.

The model structure shows a number of interesting aspects for the application of simulation in production design and control.

- Each model can use a model of the lower level. A planning model may use a scheduling model in order to make a realistic planning; a system model may use a planning model to simulate control during layout and organization research etc. The structure represents the use of simulation within simulation.
- The information flows between the models are guidelines for a distributed simulation network, which conforms to the decision-making structure within the company. For example a project model can be connected to a planning model in order to interactively harmonize decision-making within a single project and a multiple project environment. They influence each other with respect to capacity reservation and availability.
- The structure enables the correct positioning and thus the correct goal-definition of simulation models to be constructed. Although advanced algorithms can be defined within simulation models, if they don't fit the decision-making structure of the company then the algorithms are useless.
- The structure shows the relation between logistic research and technological research explicitly. The simulation world can be roughly divided into:
  - o The technological world that models technological (and often continuous) processes in great detail
  - o The logistic world that models operational decision making in great detail, thereby modelling technical processes just as time-consuming activities.

  In practice there is a tendency in both worlds to extend to the other world. It is however quite difficult to implement the achievements of each world completely. The structure expresses how to connect both simulation worlds and so how to take advantage of the achievements of both worlds.

The model structure is now being used to define several simulation research projects for the development of a virtual industrial environment. This environment will be used to investigate new operational and control structures in highly automated production systems. Above that, prototyping of new technology will be a major topic which is supported by this structure (hardware-in-the-loop simulation).
The functional view as presented here ensures that a variety of industrial organizations can be modelled within the model structure. This is also the main requirement to investigate the use of simulation itself within production control.

REFERENCES

K.H. Concannon, K.I. Hunter, J.M. Tremble, 2003, "*SIMUL-8 Planner Simulation-based Planning and Scheduling*", Proc. of the 2003 Winter Simulation Conference, pp. 1488-1493

C.S.Chong, A.I.Sivakumar, R. Gay, 2003, "Simulation-based scheduling for dynamic discrete manufacturing", Proc. of the 2003 Winter Simulation Conference, pp. 1465-1473

J. in 't Veld, 2002, "*Analyse van organisatieproblemen*"(in Dutch), Stenfert-Kroese, ISBN 90-207-3065-7

R. Wild, 1999, "Production and Operations Management", Cassell Educational, ISBN 0304-70403-2

H.P.M. Veeke, 2003, "Simulation integrated design for Logistics", PhD thesis TU Delft, Delft University Press, ISBN 90-407-2417-2

J.A.Ottjes, H.P.M. Veeke, 2000, "Project Management with Simulation", Proc. of the ICSC Symposium on Intelligent Systems & Applications (ISA 2000), Wollongong, Australia

V.C.S.Wiers, T.W. van der Schaaf, 1997, "*A framework for decision support in production scheduling tasks*", Production Planning & Control, vol. 8, nr. 6, pp. 533-544

AN EVENT DRIVEN APPROACH TO WORKFLOW SIMULATION

Agustín Yagüe[1]
Pedro Pablo Alarcón[2]
Juan Garbajosa[3]
Jessica Díaz[4]
System and Software Technology Group
Universidad Politécnica de Madrid
Postal address: E.U. Informatica. Ctra. Valencia Km. 7
E-28031 Madrid
Spain
E-mail[1]: jgs@eui.upm.es
E-mail[2]: pcavero@eui.upm.es
E-mail[3]: ayague@eui.upm.es
E-mail[4]: jessi@syst.eui.upm.es

José Luis Esteban
Investigación y Programas S.A. (IPSA)
Postal address: Paseo de la Habana 12
E-28036 Madrid
Spain
E-mail: jotaele@ipsa.es

KEYWORDS
Simulation, event driven, workflow, system modelling, validation, model-based, model-driven, XML, technical networks, XNETMOD

ABSTRACT

This paper presents an approach to workflow simulation based on event driven message queuing. Workflow systems can be considered from two perspectives: topological and logical**.** Each one needs the other to play the adequate role in the global system model. Our approach is based on some events that appear in a real workflow model. These events represent the workflow components behaviour and are domain independent. Therefore, this approach is general and could be used to develop a domain-independent simulator kernel that simulates resource moving across a workflow system. The approach has been validated in a case of data capturing industry, and using a technical platform that consists in a modelling language, the XNetMod language, and a model interpreter.

1. INTRODUCTION

The value of simulation for design and analysis of complex systems has been well described in literature (Pedgen et al. 90). Event driven programming is an important conceptual and practical programming paradigm. Event driven programming has as more important basis the following: loose coupling, state based control, concurrent and distributed computing (Hansen 2004).

A workflow model can be used to design automated or semi-automated solutions for certain business processes within an enterprise (Miller, 1995). Workflow technology can benefit simulation in a very important way. A workflow model can be understood as composed of multiple tasks, and This approach could be applied to the processing of bank documents such as receipts, and bills of exchange. The technology described within this paper has been validated within this application domain.

Some works have been published based on formalization and verification of event driven processes (Van der Aalst, 1999). Some works have been performed related to workflow composition, and specifically using artificial intelligence (Kim, 2004). Workflow system models facilitate the execution and control work in real time (Perakath, 1995). This may help understand why it may be interesting to develop an independent application domain event-based simulator.

The approach described within this paper is based on modelling a number of events that appear in a real workflow model. These events represent the workflow components behaviour and are domain independent. Therefore, this approach is general and could be used to develop a domain-independent simulator kernel that simulates resource moving across a workflow system. The approach has been validated in a test case of data capturing industry, and using a technical platform that consists in a modelling language, the XNetMod language, and a model interpreter (Alarcon et al, 2003).

The paper is structured as follows: Section 1, Introduction, this one. Section 2, Presents the workflow model language. Section 3, Presents the concept of Workflow Views as understood by the authors. Section 4 describes the Workflow Event Approach; Section 5, Workflow Definition; Section 6, Simulation Process. Section 7, Presents an example of a Complete Simulation Session. Section 8, Conclusions.

2. WORKFLOW MODEL LANGUAGE

The language was called the XNetMod language. Because of the design criteria, XNetMod family language can support the modeling of a system using concepts of its application domain.

The structure of the language can be described according to the following data model, Fig. 1. The main class of the model is the class "model". This model has defined hierarchical node relations. From model point of view, in

XNetMod language hierarchical node relations are not limited and are supported using "nodeLevel1" and "nodeLevel2" classes.. The number of hierarchical levels is not limited, and a hierarchical node organization is supported. In the end a tree structure is obtained.

Different types of net topologies in the same model are represented by classes "net". The "pin" elements refer to the nodes which are directly connected by means of the related network. Additionally, "rule"-elements, which may define legal and illegal network patterns for the relevant network, are also included in the model. In general, XML-based grammars are appointed to represent data models with a tree structure. Therefore, the network topology is described in conjunction the node characteristics. Node descriptions can be extended for specific applications. Obviously, the depicted class model may be also depicted by means of an XML-based language. Using the natural structure of the XML grammar, the model structure can be easily expressed in a form with four main semantic parts, as shown in Fig. 1:

1. A set of functional elements of a network ("node"-elements) - nodes in a graphical representation of the model
2. A set of networks ("net"-elements) with their connections ("pin"-elements) - edges in graphical representation of the model
3. A set of verification rules ("rule"-elements) which may define legal and illegal net-work patterns for the relevant network
4. A set of attributes ("attr"-elements) related to node elements, network definitions, and connection descriptions.

The modeling language provides a rather conservative set of conventions for the definition of the set of "node"-elements and the topology of the related networks. The definition of the internal structure of the "node"-elements and the "attr"-elements can be provided in an external file. This possibility can be exploited in cases when the origin of "node"-elements is foreign - the information about the elements is provided by an element supplier, the model engineering task may have access to external data bases. Herewith, local variations of the nodes' functionality specification result only in a local adaptation of the network model. Moreover, the separation of the model elements description and network definition allows a rather independent access to the model functional elements and the model topology on the model processing level.

One of the mechanisms for verification is rules. These rules can be used for the verification of the network structure involving pattern-based mechanisms ("rule"-elements) will be also defined. Even at this stage, without definition of the "node"-elements internal structure, the model topology can be interpreted in a formal way involving the patterns defined by the "rule"-elements. The network modeling language defines, on the one hand, the structure – topology – of the technical facility or process and, on the other, provides the highest level of abstraction for the description of the process functional relations.

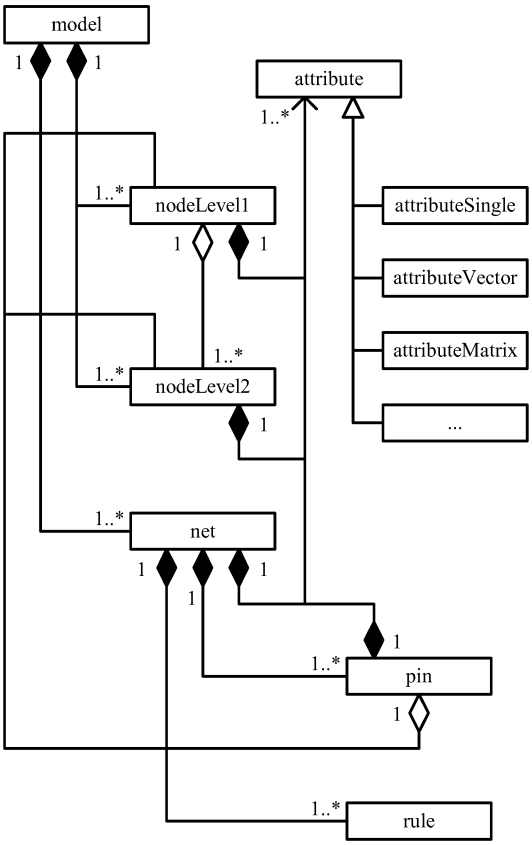


Figure 1. Language basic concept.

Therefore, the modeling language possesses the highest level in the language hierarchy. Models are very much associated with the domain they represent (Savolainen et al. 1995). The XNetMod language uses its modeling features, e.g. node descriptions, so that the model schema may be defined to each domain features. That is why it is possible to speak of a language family.

3. WORKFLOW VIEWS

Workflow structures involve:

- a set of nodes representing the network topology, each node has its own fan-in and fan-out values.
- a set of activities related to the topology view representing all the elements that forms the workflow system.
- in addition, resources, another kind of workflow elements, and that should be included in a workflow simulation model.

Workflow systems can be considered from two perspectives or views: topological and logical. Each one needs the other to play the adequate role in the global

system model. The following sections describe more in detail the model basis further developed in the rest of the paper.

Topological view

In the topological view two kind of nodes represents the network topology: Active and passive nodes. A passive node could be source to different active nodes and could be a targeted node from more than one node. That means multiple cardinality between nodes, as it is shown in the figure 2.

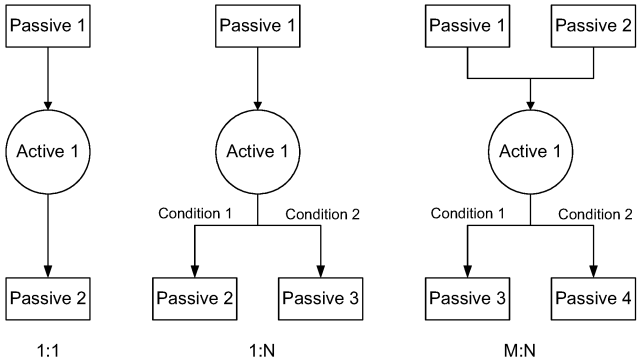


Figure 2. Topological workflow Passive-Active cardinality.

These links represent permanent connections between workflow nodes, but there are no events raised from topological nodes

Logical view

This view is really the workflow user view, because it is the part of the workflow definition where activity types and resources are defined.

From this point of view, passive nodes are priority queues where items are stored waiting to be required by activities, and active nodes represent workflow activities. An activity type could receive items from different queues and could send items to different queues depending on the activity execution result. A queue could produce items for any active node linked with it. This means that there is a 1:1 relation between active nodes and activity type and 1:1 relation between passive nodes and queues.

But there are more elements in this logical view. In simulation time, any activity type may be executed by more than one activity instance, so activity type nodes are really representing one or more activity instances. Another important aspect to be considered is the relation between the roles that an activity type may represent and the resources skills. In simulation time the most important issue is to deliver all the activities in the minimum slice of time with the minimum resources requirements. In consequence, the workflow simulator should use this resource skill information and the requirements about activity skills.

Figure 3 shows two different workflow logical views.

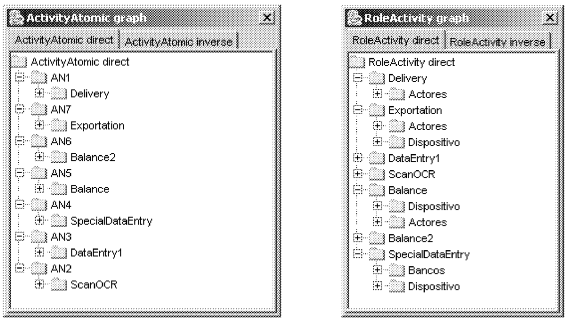


Figure 3. Active type Node - Activity and role activity association

4. WORKFLOW EVENT APPROACH

Once the workflow topology has been defined, it is time to describe the events that may be raised from the elements. Elements are responsibility of the events in exclusivity.

There are two different kind of events, events raised from passive nodes and events raised from active nodes.

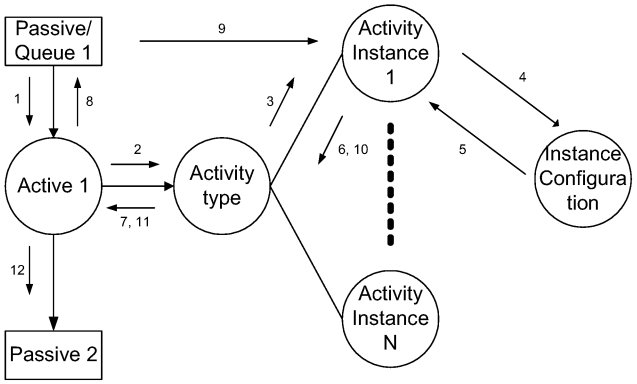


Figure 4. Workflow simulation sequence event.

Passive node event

This event is launched when an item is received from somewhere, and it represents that there is, at least, one item waiting for a request from some activity.

As a response to this event, the active node linked to this passive node, should create the activity instance to process the items when required.

Active node events

These events are related to two different situations. The first type of event is launched when an active node receives an event from a queue. The main goal of the event is to create so many activity instances as the workflow configuration has defined.

Once an activity instance has been properly configured, it means that the activity has allocated all the resources needed to be performed, and a *getItem* event is launched. This second event is targeted to a queue in order to request to process an item.

5. WORKFLOW DEFINITION

Once views and events have been defined, it is possible to describe the workflow model. There are seven different connections to configure the workflow:

1. Topological connections to describe the network topology.
2. Role hierarchy to describe hierarchical relations between different roles needed to represent activities.
3. Resource skills to assign some roles to workflow resources.
4. Aggregation hierarchy to describe some aggregation levels between items.
5. Role activity relationship to describe which kind of roles are needed to create an activity instance.
6. Document activity relationship to describe the kind of item to be processed by an activity.
7. Activity atomic to describe relationship between activities and atomic nodes.

Node connections mentioned above can be expressed using the XNetMod language specifically developed to model technical networks such as pipe networks, road networks, and also workflow systems (Alarcon et al.,2003)

6. SIMULATION PROCESS

Once the workflow is defined, it is possible to produce a system simulation. This section describes how the simulation process works. This simulation starts when a start simulation event is launched. The first activity to be done is to create item instances for the simulation. Figure 4 shows the simulation process. These items will be allocated in the first queue of the system. This process is not represented in the Figure 4. Once the items have been allocated in the main queue, a wake up event is raised to the activity type, see annotation number 1 in Figure 4.

This event is captured by the active 1 node and automatically passed to the activity type node (2). When the activity type node receives the event, it first checks if there is any activity instance node waiting for any item. If there is no activity instance already created, the activity type node, will create so many activities as defined in the workflow configuration information. When the activity instance is running (3) it is time to request information on resources available to be associated to this instance (4), information provided by the workflow control system. If the workflow control system does not have enough available resources, the activity instance will wait for resources. Once the system has enough available resources, it will wake up the activity instance (5) and now the simulator is prepared to process items.

When the activity instance is ready to work, an event to request for items (6) is sent. This event passes across the activity type node (7) and across the active node **until** it reaches the passive node target of the event (8). When the queue receives the event raised by the activity, it knows that the activity *id* is waiting for items, so this is the right time to send the item (9).

After the activity receives the item, it is processed to generate the appropriate values, and including some time delays. An activity receives documents from input queues. Also, an activity can generate new documents and with no input queues. Therefore there will be zero or more input queues for a given activity. Each input queue of an activity has a priority assigned, and the activity will receive documents from high priority queues first. Note: a certain queue can be a high priority queue for an activity but a low priority queue for another activity.

After processing a certain document, an activity will evaluate a number of rules to decide a final state or termination condition for the document. This condition will be used by the workflow system to choose one out of the available output queues for the activity. There will be at least one output queue for a given activity.

Modified items are sent to the next queue passing through the activity type node (10) and the active node (11), to reach the targeted queue (12)

7. COMPLETE SIMULATION SESSION

The workflow simulation is divided in the following steps:

1. Workflow item definition. This step is used to describe the workflow model using the XNetMod language. It is possible to use any text editor to define this XML file with the workflow configuration.
2. Workflow load process. This is considered as the first simulation process, and its main activity is to validate the workflow model and to build the memory workflow model representation. Figure 5.

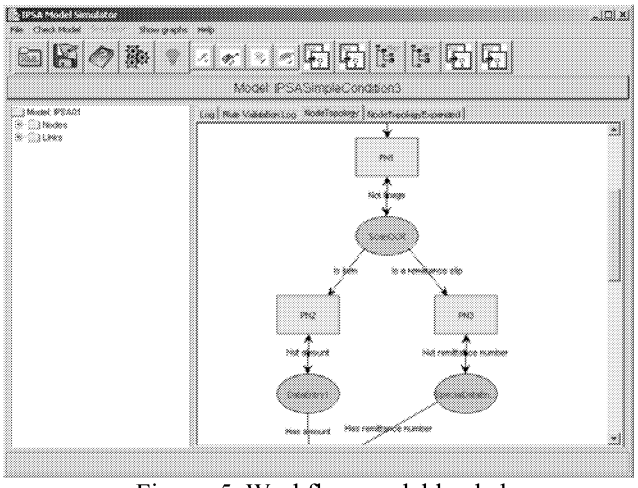


Figure 5. Workflow model loaded.

3. Workflow Activity definition. This step is included to configure the number of instances of each activity type and it must be instanced in simulation time. Figure 6. Depending on values, the workflow system detects if there are resources enough to do the simulation.



Figure 6. Activity quantity definition.

4. Model Item quantity definition, where is defined the number of items that should be created to run the simulation. Figure 7.
5. Finally, simulation process. Using the information given before, the simulation process starts to move items along the workflow Figure 8. Important item simulation information is registered in a database to be processed as simulation results Figure 9.

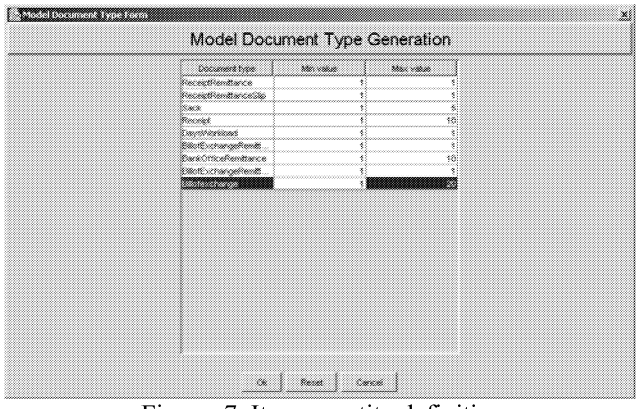


Figure 7. Item quantity definition.

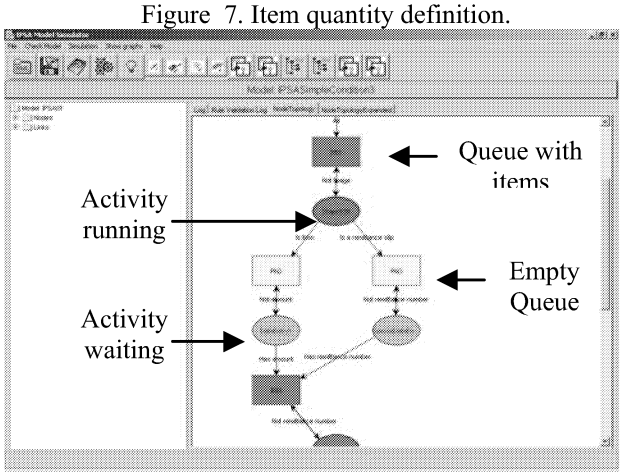
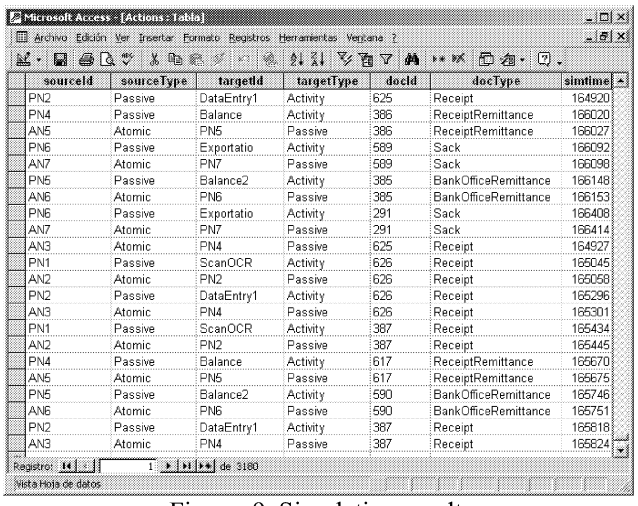


Figure 8. Simulation running.



sourceId	sourceType	targetId	targetType	docId	docType	simtime
PN2	Passive	DataEntry1	Activity	625	Receipt	164920
PN4	Passive	Balance	Activity	386	ReceiptRemittance	166020
AN5	Atomic	PN5	Passive	386	ReceiptRemittance	166027
PN6	Passive	Exportatio	Activity	589	Sack	166092
AN7	Atomic	PN7	Passive	589	Sack	166098
PN5	Passive	Balance2	Activity	385	BankOfficeRemittance	166148
AN6	Atomic	PN6	Passive	385	BankOfficeRemittance	166153
PN6	Passive	Exportatio	Activity	291	Sack	166408
AN7	Atomic	PN7	Passive	291	Sack	166414
AN3	Atomic	PN4	Passive	625	Receipt	164927
PN1	Passive	ScanOCR	Activity	626	Receipt	165045
AN2	Atomic	PN2	Passive	626	Receipt	165058
PN2	Passive	DataEntry1	Activity	626	Receipt	165296
AN3	Atomic	PN4	Passive	626	Receipt	165301
PN1	Passive	ScanOCR	Activity	387	Receipt	165434
AN2	Atomic	PN2	Passive	387	Receipt	165445
PN4	Passive	Balance	Activity	617	ReceiptRemittance	165670
AN5	Atomic	PN5	Passive	617	ReceiptRemittance	165675
PN5	Passive	Balance2	Activity	590	BankOfficeRemittance	165746
AN6	Atomic	PN6	Passive	590	BankOfficeRemittance	165751
PN2	Passive	DataEntry1	Activity	387	Receipt	165818
AN3	Atomic	PN4	Passive	387	Receipt	165824

Figure 9. Simulation results.

8. CONCLUSIONS

This event driven approach is adequate to simulate workflow systems because the way to represent and to operate the system is quite similar to the real workflow system operation.

Based on this approach, the simulated system is continuously working with the resources available in each moment removing active waiting time.

Of course, this approach has some difficulties to be applied because is not easy to detect deadlocks.

REFERENCES

Hansen S. and Fossum T. “Events Not equal to GUIs”. Proceedings of the SIGCSE’04, Norfolk, Virginia. USA. 378-381.

John A. Miller et al. 1995. ”Simulation modeling within workflow technology”. Proceedings of the 1995 Winter Simulation Conference. 612-619

Perakath, C. Benjamint, Marshall C. and Mayer R. J. 1995. ”A workflow analysis and design environment (WADE)”. Proceedings of the 1995 Winter Simulation Conference. 597-603.

Birtwistle, G. M. 1979. Discrete Event Modeling in SIMULA. MacMillian, London.

Schruben L. 1983. Simulation modeling with event graphs. Communications of the ACM 26: 957-963

Kim J., Sprangen M., Gil Y. 2004. “An Intelligent Assistant for Interactive Workflow Composition.” Proceedings of the International Conference in Intelligent User Interfaces (IUI-2004). Madeira. Portugal.

Pedgen C.D., R.E. Shannon, and R.P. Sadowski. 1990 “Introduction to simulation using SIMAN”. McGraw-Hill; New York.

Van der Aalst, W.M.P. “Formalization and Verification of event-driven process chain”. Information and Software Technology, 41(10):639-650, 1999.

Alarcon, P.P. , Garbajosa J., Yague, A., Lisounkin, A., Schreck, G., Esteban, J.L. "Modeling Language Interpretation for Validation". STEP '03: Proceedings of the Eleventh Annual International Workshop on Software Technology and Engineering Practice (STEP'03). IEEE Computer Society, ISBN 0-7695-2218-1,2003.

T. Savolainen, D. Beeckmann, P. Groumpos, and H. Jagdev, “Positioning of modeling approaches, methods and tools”, Computers in Industry, 1995 Vol.25, pp.255-262.

ACKNOWLEDGEMENTS

The work described has being partially developed in the framework of a European research and development project XNetMod "XML Based Modelling Language for Simulation of Technical Networks" (CRAFT-IST Contract No. 2001-52057). The consortium includes two RTD performers – the Fraunhofer IPK (Berlin / Germany) and the Universidad Politécnica de Madrid, Escuela Universitaria de Informática (Madrid / Spain). The first users of the technology are four industrial companies – ELPRO Prozessindustrie- und Energieanlagen GmbH (Berlin / Germany), SimPower Simulator Systeme GmbH (Karlstein am Main / Germany), Investigacion y Programas S.A. (Madrid / Spain), and ACEBO, C.B. (Arcas de Villar - Cuenca / Spain).

This work has been partially sponsored by project AGMOD TIC2003-08503. Ministerio de Educación, Ciencia y Cultura de España.

AUTOMOTIVE PRODUCTION SYSTEMS SIMULATION

SIMULATION VALIDATES DESIGN AND SCHEDULING OF A PRODUCTION LINE

Vidyasagar Murty
Neelesh A. Kale
Rohit Trivedi

Production Modeling Corporation
Suite 1006 Parklane Towers West
Three Parklane Boulevard
Dearborn, Michigan 48126 U.S.A.

Onur M. Ülgen
Edward J. Williams

Industrial & Manufacturing Systems Engineering Department
University of Michigan – Dearborn
4901 Evergreen Road
Dearborn, Michigan 48128 U.S.A.

KEYWORDS

Manufacturing, process design, discrete-event process simulation, scheduling

ABSTRACT

Discrete-event process simulation has historically enjoyed its earliest, most numerous, and many of its most conspicuous successes when applied to the design and/or the scheduling of production processes. In this paper, we describe an application of simulation to the design, layout, and scheduling policies of a production line in the automotive industry. Specifically, the production line in question was and is vital to the operations and profitability of a first-tier international automotive supplier. In addition to describing the process itself, the simulation model, and its results, we discuss some complex challenges of input data collection and interpretation.

INTRODUCTION

Discrete-event process simulation has a long pedigree of success in many fields of application; indeed, one of its earliest and still very frequent areas of application is in the manufacturing sector (Law and McComas 1998). The automotive industry, a major component of the manufacturing sector on most continents, has not only become increasingly competitive in recent years, but has developed longer and more complex supply chains. A chain is no stronger than its weakest link – all components of a supply chain must function reliably and efficiently to provide high consumer and shareholder value (Chopra and Meindl 2004). In this paper, we describe the application of simulation to the design, layout, and establishment of scheduling policies pertinent to a manufacturing line of a first-tier automotive supplier (i.e., a supplier who sells automotive components directly to a manufacturer of vehicles). Due to the increasing competitiveness throughout this industry, first-tier (not to mention second-tier, third-tier, etc.) automotive suppliers must constantly increase their efficiencies to withstand competitive pressures on price, timeliness of delivery, and flexibility (Walsh 2005). Given the extensive history of simulation successes in improving manufacturing processes and operations, an extensive simulation analysis was a logical weapon of counterattack against these pressures. Representative examples of these successes appear in (Graupner, Bornhäuser, and Sihn 2004) relative to the processed-foods industry; (Steringer et al. 2003), who examined the logistics and material-handling strategies within diesel-engine assembly; and the application of simulation to scheduling interactions among raw material suppliers and an automotive stamping plant described by (Grabis and Vulfs 2003). Additionally, (Ülgen and Gunal 1998) discuss several applications of simulation in both automotive assembly plants and in plants which manufacture automotive components, taking care to note the extensive commonality of both concepts employed and benefits realized. In this paper, we provide an overview of the manufacturing process we analyzed collaboratively with the client, describe the construction, verification, and validation of the model, and present results and conclusions emerging from the study. We give particular attention to complexities arising from the collection and interpretation of input data. Whereas the newcomer to simulation methodology is likely to view the seemingly exotic step of model construction as most pivotal, experienced analysts know that "data collection is one of the initial and pivotal steps in successful input modeling" (Leemis 2004); note that the *input* is *modeled*.

PROCESS OVERVIEW

The first step of the simulation project, as in any simulation study, was defining the project objective. Once the objective was defined, the complete process was mapped and all relevant details were documented. The process description is as follows:

The production facility consists mainly of four compaction presses (P1, P2, P3 and P4), two assembly robots, and four sintering furnaces (F1, F2, F3 and F4), as shown in Figure 1 (last page). The facility produces four different types of metal powder precision components ("carriers"), denoted A, B, C, and D; and each component consists of two part types (symbolically $[A_1, A_2]$, $[B_1, B_2]$, $[C_1, C_2]$, or $[D_1, D_2]$). The presses typically run in pairs; for example, if P1 is producing B_1 part types, press P2 is producing B_2 part types. Presses P3 and P4 cannot produce part types for carrier type C. Hence each press has two die sets. At any given time, a given press is using one die set while the other die set is being set offline for the next part type. (Accordingly, at any given time, the press compacts

one kind of the part). After each die changeover on a press, there is a two-hour start up time for quality checks.

The carrier parts are then routed from the press to the buffer with negligible travel time. The assembly robots pick parts from the buffer, assemble them to form a carrier and place them on the furnace conveyor using a round-robin discipline. The part picking is done using an "oldest individual part" discipline. For example, suppose an A_1 part has waited 10 minutes, an A_2 part has waited 1 minute, a B_1 part has waited 20 minutes, there is no B_2 part in the buffer, a D_1 part has waited 9 minutes, and a D_2 part has waited 8 minutes. Then, since assembly of a B carrier is at the moment impossible, an assembly robot will pick the A_1 and A_2 parts for assembly next. The robots assemble carriers and feed the furnace conveyors, which run continuously through the furnaces, as long as there are parts in the buffer. By policy, the fourth furnace is fed only when all the other three are full; indeed, the client wished to examine the possibility of "mothballing" (entirely abandoning use of) the fourth furnace. Carriers are sintered (i.e., the powdered mixture of metal they comprise is heated to just below the fusing point of the most easily fused ingredient, causing coalescence into a strong component (El Wakil 1998)) as they travel through the furnaces, and upon exit are ready to move to the finished goods storage area. The entire process is fully automated. Scrap rates for the presses and furnaces are assumed to be 2% and 1% respectively. Additional modeling assumptions, discussed with and approved by the client, and documented, were:

1. Raw material is always available
2. Operators are modeled as resources that are always available.
3. The production is fully automated and continuous.
4. Robots have 5% of downtime with one hour as mean time to repair (MTTR)
5. Travel time between the buffer and furnace conveyor is zero.
6. There is no blocking of parts upon exiting the furnace conveyor.
7. Die setup takes eight hours and startup takes two hours. Both add to a total of ten hours for the changeover of die.
8. The robot assembles the parts on a FIFO [first-in, first-out] basis.
9. There is no delay when production is switched from presses P3 and P4 to presses P1 and P2, provided P3 and P4 have been running for more than 12 hours.

MODEL CONSTRUCTION, VERIFICATION, AND VALIDATION

The analysts and clients agreed upon the use of the SIMUL8® software for this project. This software is relatively easy to use. In addition to provision of standard constructs such as Work Entry Points, Storages (queues or buffers), Work Centers, Resources, and Work Exit points, SIMUL8® allows construction of the simulation model logic and its animation to proceed concurrently. Additionally, SIMUL8® provides features such as Schedules for Resources, plus the ability to "profile" a model to discover where most of the model execution time is spent (Hauge and Paige 2004). To improve model run-time performance, the analysts then concentrated their efforts on those portions of the model logic consuming the largest percentages of execution time.

To aid in model verification, the complete model was built in two stages. One model contained the presses; the other, the robots and furnaces. After verifying each of these models, the analysts linked them into one larger model, hence using the principle of modular design well known to software engineers and practitioners (Deitel and Deitel 2003). Additionally, these originally separate models confirmed that the presses (not the assembly robots, nor the furnaces) were the system bottleneck. Since the client already firmly believed this, its early corroboration by the study increased the credibility of the analysis.

A significant step in model construction and validation was distribution fitting for the raw downtime data. Downtime data included repair time (TTR) and time between failures (TBF) for four types of downtime (mechanical, electrical, hydraulic and miscellaneous) for each of the four presses. The client provided TTR and TBF data for a year, and remarked "each press is down about 25% of the time." Since SIMUL8® considers MTTR and MTTF as input, the given TBF data was converted to TTF by subtracting TTR from TBF for each downtime event. Distributions were fitted to each MTTR and MTTF using the Stat::Fit® distribution-fitting tool. The fitted distributions were analyzed with Kolmogorov-Smirnov and Anderson-Darling goodness-of-fit tests, with greater reliance placed upon the Kolmogorov-Smirnov test results (the test which, due to familiarity, the client found more credible). Use of the best fitting MTTR and MTTF distributions in a preliminary test run (these distributions were gamma and Weibull with parameters implying long tails) produced results implying the machines would be down more than 50% of the time, a severe mismatch with direct observation. The analysts next discussed this problem with the clients at length. The discussion revealed that the original data set of TTFs and TTRs contained very long downtimes because if, for example, a repair began just before quitting time on a Friday, and was completed the following Monday morning after a weekend hiatus, the entire weekend was wrongly included in the downtime (Williams 1994). After cleansing the data, Stat::Fit® was rerun and the new distributions obtained (exponential) yielded test data closely matching the client's newly gained understanding of TTF and TTR.

After the above data cleansing was completed, model verification and validation were successfully undertaken using generally recognized techniques such as checking hand calculations against deterministic runs, examination of traces and of the animation, structured walkthroughs of the model logic, and Turing tests undertaken cooperatively with the client (Sargent 2004).

RESULTS AND CONCLUSIONS

The client's primary performance metric was the "makespan of a production cycle." In the client's

terminology, a "production cycle" is the production of all carrier varieties in the amounts demanded by the marketplace in one week and its "makespan" is the time required for that production. Hence, the basic target makespan is 7.0 days or one calendar week. The client was particularly interested in comparing the merits of sequential scheduling (involving production of parts at only two presses, and hence producing only one type of carrier at any given time) versus batch scheduling (in which presses P1 and P2 run throughout the week [unless down] and presses P3 and P4 run as needed). Therefore, model experimentation focused upon (a) comparison of these scheduling disciplines, (b) assessing the sensitivity of system throughput to downtime, and (c) assessing the sensitivity of system throughput to buffer sizes. Accordingly, five scenarios were explored in detail, as summarized in Table 1 (last page). All five scenarios were run seven days a week, three shifts per day, for seventy weeks (ten-week warm up time and sixty-week run length). The 95% confidence intervals for the makespan performance metric are based on six replications. In this table, downtime data set 1 represents expected downtime of the presses, whereas downtime data set 2 represents severe downtime ("worst-case analysis"). The "overall buffer capacity" represents a physical constraint on the buffer immediately downstream from the presses, whereas the "buffer limit per part type" represents an operational constraint on the number of any one part type allowed to reside in the buffer at any given time. As indicated by the table, a configuration using batch scheduling, a 24,000-capacity buffer permitting 6000 parts of one type to reside therein, and simultaneous use of three furnaces meets the makespan target even under robustly – even under the stress of very pessimistic downtime assumptions (scenarios 4 and 5).

In addition to the clear superiority of this alternative (which permitted the client to achieve operational savings by using one fewer furnace than anticipated), other significant insights gleaned from this simulation study were:

1. Increased press downtime leads to increased press blockage because when a press goes down more frequently its paired press, which is compacting the corresponding part, fills its share of the buffer and becomes blocked more frequently. Concurrently, increased press downtime increases the system sensitivity to the buffer limit per part.
2. The expenses of increased buffer size (these expenses include capital investment, use of floor space, and increased work in process) are justified not only to achieve the required makespan, but also to improve press utilization.
3. Neither robot can begin assembling a carrier unless $\min(X_1$ parts available, X_2 parts available) [X e {A,B,C,D}] = y; currently $y = 1$. Increasing the value of y will improve furnace utilization, and evaluating various plausible values of y will be the object of further study.
4. Batch scheduling is significantly superior to sequential scheduling.

ACKNOWLEDGMENTS

All five authors take pleasure in commending anonymous referees for their valuable suggestions to improve the organization and clarity of this paper.

REFERENCES

Chopra, Sunil, and Peter Meindl. 2004. *Supply Chain Management: Strategy, Planning, and Operation*, 2nd ed. Upper Saddle River, New Jersey: Pearson Education, Incorporated.

Deitel, H. M., and P. J. Deitel. 2003. *C++ How to Program*, 4th ed. Upper Saddle River, New Jersey: Pearson Education, Incorporated.

El Wakil, Sherif D. 1998. *Processes and Design for Manufacturing*, 2nd ed. Boston, Massachusetts: PWS Publishing Company.

Grabis, Janis, and Girts Vulfs. 2003. Simulation and Knowledge Based Scheduling of Manufacturing Operations. In *Proceedings of the International Workshop on Harbour, Maritime and Multimodal Logistics Modelling & Simulation*, eds. Yuri Merkuryev, Agostina G. Bruzzone, Galina Merkuryeva, Leonid Novitsky, and Edward Williams, 177-182.

Graupner, Tom-David, Matthias Bornhäuser, and Wilfried Sihn. 2004. Backward Simulation in Food Industry for Facility Planning and Daily Scheduling. In *Proceedings of the 16th European Simulation Symposium*, eds. György Lipovszki and István Molnár, 47-52.

Hauge, Jaret W., and Kerrie N. Paige. 2004. *Learning SIMUL8: The Complete Guide*, 2nd edition. Bellingham, Washington: PlainVu Publishers.

Law, Averill M., and W. David Kelton. 2000. *Simulation Modeling and Analysis*, 3rd edition. Boston, Massachusetts: The McGraw-Hill Companies, Incorporated.

Law, Averill M., and Michael G. McComas. 1998. "Simulation of Manufacturing Systems." In *Proceedings of the 1998 Winter Simulation Conference*, eds. D. J. Medeiros, Edward F. Watson, John S. Carson, and Mani S. Manivannan, 49-52.

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.

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.

Steringer, Robert, Martin Schickmair, Johann Prenninger, and Maximilian Bürstmayr. 2003. Simulation of Large Standard Stillage Placement on a Diesel-Engine Assembly. In *Proceedings of the 15th European Simulation Symposium*, eds. Alexander Verbraeck and Vlatka Hlupic, 425-429.

Ülgen, Onur, and Ali Gunal. 1998. Simulation in the Automotive Industry. In *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*, ed. Jerry Banks, New York, New York: John Wiley & Sons, Incorporated, 547-570.

Walsh, Tom. 2005. For Pinched Parts Suppliers, It's Life on the Edge. *The Detroit Free Press* 174(272):1A, 9A (3 February 2005).

Williams, Edward J. 1994. Downtime Data -- its Collection, Analysis, and Importance. In *Proceedings of the 1994 Winter Simulation Conference*, eds. Jeffrey D. Tew, Mani S. Manivannan, Deborah A. Sadowski, and Andrew F. Seila, 1040-1043.

AUTHOR BIOGRAPHIES

VIDYASAGAR MURTY holds a bachelor's degree in Mechanical Engineering (Jawaharlal Nehru Technological University, India, 2000) and a master's degree in Industrial Engineering (University of Cincinnati, 2003). He joined Production Modeling Corporation in 2003 as an Applications Engineer. He mainly uses Enterprise Dynamics®, WITNESS®, SIMUL8® simulation packages and manages simulation projects. He is a member of the Institute of Industrial Engineers [IIE] and has served as Vice President of Administration on the IIE - Greater Detroit Chapter board since 2004. His e-mail address is vmurty@pmcorp.com

NEELESH A. KALE received a Bachelor of Engineering degree in Production Engineering from the University of Pune, India (2000) and an M.S. degree in Industrial Engineering from Oklahoma State University, USA (2003) with a concentration in operations research and statistics. Currently he is working as a junior simulation analyst with Production Modeling Corporation, Dearborn, Michigan. His interest areas are simulation modeling and analysis, and traditional industrial engineering techniques for performance improvement. He frequently uses Enterprise Dynamics®, Simul8®, and WITNESS® simulation packages for modeling and analysis. His e-mail address is nkale@pmcorp.com.

ROHIT TRIVEDI earned his bachelor's degree in the field of Mechanical Engineering (Maharaja Sayajirao University of Baroda, Gujarat, India, 2001) and completed his master degree program in Industrial Engineering with concentration in the field of Engineering Management Program (Wayne State University, Detroit, Michigan, USA). He is currently pursuing his master degree program in the field of Business Administration (Wayne State University, Detroit, Michigan, USA). He is working as an Engineering Consultant with primary focus in the areas of Process Management, Simulation, Lean Manufacturing and traditional Industrial Engineering. He enjoys teaching as an external faculty member for University of Michigan -- Dearborn. He was awarded the Graduate Professional Scholarship from Wayne State University Graduate School, 2004-2005. He received second prize at the national level for Technical Paper Presentation Contest. (TKIET, Warananagar, Maharashtra, India, 2000). He was a member of ISTE (Indian Society for Technical Education, 1997-2001). His e-mail address is r.trivedi@pmcorp.com and his personal web page is http://rohittrivedi.tripod.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 that 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).

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 Production Modeling Corporation, Dearborn, Michigan, as a senior simulation analyst. Also, since 1980, he has taught evening 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; Istanbul, Turkey; Genova, Italy; and Riga, Latvia. He has just served as Program Chair of the 2004 Summer Computer Simulation Conference, and is serving as Program Chair for the 2005 IIE Simulation Conference and the 2005 Summer Computer Simulation Conference. His e-mail address is: ewilliams@pmcorp.com and his university web page is <http://www-personal.umd.umich.edu/~williame>.

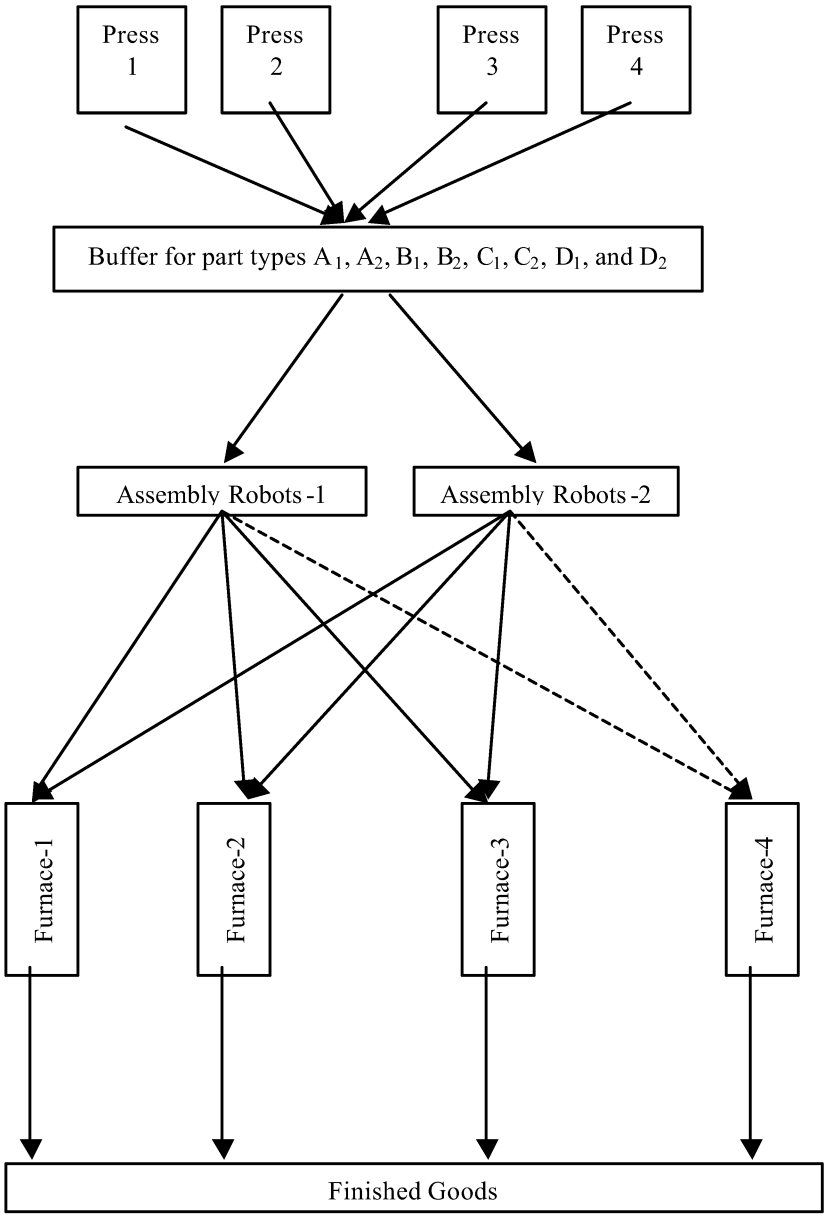


Figure 1. Schematic Layout of Compaction – Sinter Line Area

Scenario	Scheduling Method	Downtime Data Set	Overall Buffer Capacity	Buffer Limit per Part Type	Furnaces Used	Makespan (95% Confidence Interval)
Base	sequential	1	2880	1440	4	8.96 – 9.34
2	sequential	1	24,000	12,000	4	8.44 – 8.72
3	sequential	2	2880	1440	4	10.34 – 10.72
4	batch	1	24,000	6000	3	5.94 - 6.29
5	batch	1	24,000	6000	3	6.60 – 6.91

Table 1. Results of Various Scenarios

Bezier Curves for Representing Clothoids, Circles and Straight Lines

N.Montes, J.Tornero, L.Armesto
Dpto. Ingeniería de Sistemas y Automática
Universidad Politécnica de Valencia
46022 Valencia (Spain)
Email: nimonsan@upvnet.upv.es jtornero@isa.upv.es leoaran@isa.upv.es

Abstract- This paper generates Bezier curves for path planning vehicle navigation, in particular for computing free-collision continuous paths. Generic continuous paths are obtained by combining straight lines, circles and clothoids. A set of three clothoidal segments has been the selected solution and approximated by Bezier control points. The proposed method is very fast so can be implemented in real-time applications. In addition and in order to adapt sampling frequencies between the planner layer and the vehicle control layer, a dual-rate also based on Bezier functions have been presented. That is the interface is used to generate a continuous interpolating trajectory function using control points at a slow frequency.

Keywords: Clothoid, s-series, Bezier approximations, Dual-rate control, collision-free path, three clothoids

1. Introduction

In autoguided vehicles, it is required to implement fast algorithms to find free-collision continuous paths. In fact, for path planning in mobile robot, a simple and fast method is the use of *RS paths*. It concatenates straight lines and circular arcs [6],[7]. The disadvantage of this technique lies on the fact that the resulting curvature profile has discontinuities at the connecting points.

In that sense, clothoids are ideal for generating smooth transitions among straight lines and circles. We can define any path combining at most 7 pieces of clothoids, circles and straight lines. This is denominated *scc-paths (simple continuous-curvature paths* [8],[9].

We can also obtain algorithm to find free-collision continuous paths by the use of *CC-turns* based on straight lines and circles [10].

In [11] two equal clothoids are used to build an *elementary path.* If we combine two *elementary paths,* we have a *bi-elementary path*, and we can use it to avoid any obstacle on the plane [12] or even with only an *elementary path* and a straight line[13].

The main problem lies on the computational complexity to solve clothoids using Fresnel integrals. The solution is based on approximating clothoid by parametric functions. In general, a high order polynomial degree to approximate a clothoid with a low error.

However in this paper, a new, fast, low order and accurate polynomial approximation to clothoids is presented theoretically and validated in the context of mobile robotics.

We have introduced a new clothoidal combination to find free-collision continuous paths. This clothoidal combination is based in three clothoids, joining a circle with a straight line.

The proposed method is usually executed at a lower sampling rate than that corresponding to the vehicle control one. In this sense a dual-rate interface is used to adapt both sampling frequencies. This interface generates a continuous interpolation function from low sampling control points, which is afterwards discretized at high-sampling rate.

2. Introduction to clothoids

The main characteristic of clothoids or Cornú spirals, as shown in Fig.1, is that the variation of its radius along the whole curve is constant; clothoid radius is infinite at the start of the curve. The equation is:

$$r \cdot l = A^2 \tag{1}$$

where:
A : is the characteristic parameter of the clothoid.
l : is the length variable of the curve.
r : is the radius of the curve at the length *l*.

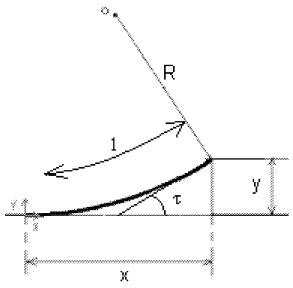


Fig 1 Clothoid

The clothoid orientation (τ) is defined as:

$$\tau = \frac{A^2}{2R^2} \tag{2}$$

To obtain any point (x, y) on the curve it is necessary to solve the following integrals:

$$x = A\sqrt{2}\int_0^{\xi}\cos(\xi^2)d\xi \qquad y = A\sqrt{2}\int_0^{\xi}\sin(\xi^2)d\xi \tag{3}$$

where $\xi = \frac{l}{A\cdot\sqrt{2}}$

These integrals are Fresnel integrals of the form

$$C(\xi) = \int_0^{\xi}\cos(\xi^2)d\xi \qquad S(\xi) = \int_0^{\xi}\sin(\xi^2)d\xi \tag{4}$$

which are analytically impossible to be solved and this constitute the main problem of working with clothoids.

In the past, researchers used Taylor series to approximate Fresnel equations, but to obtain low error it is necessary to use a high order polynomial equation. Researches propose a lot of methods based on B-splines, rational approximations, Chevishev equations, etc., but all of these methods have the same problem: they are too complicated to build and the error is being accumulated as it approach the end of the clothoid.

A new method to approximate the clothoid was proposed in [3]. It is based on Hermite expansion around two points instead of Taylor expansion around one point. The main advantage of this approximation is that it guarantees the contact with the real clothoid at the beginning and at the end of the curve, in spite of the fact that real and approximated clothoid may diverge along the path. In this approach, a low error can be obtained with a low polynomial degree. In particular, for a third order polynomial the error is insignificant. This approximation is called *s-series* and for a *k-order* is described as follow,

$$a(u) = \sum_{k=0}^{\infty} a_k(u)\cdot s^k \tag{5}$$

where:

$s = (1-u)^k \cdot u^k$

$a_k(u) = (1-u)\cdot a_k^0 + u\cdot a_k^1$

for $u \in [0,1]$

In order to calculate the clothoid with length L, it is necessary to know the characteristic parameter of a clothoid (A) and the final radius (R). In that case, for the Fresnel integrals,

$$\xi_0 = 0, \quad \xi_L = \frac{L}{A\cdot\sqrt{2}} \tag{6}$$

where L and R are related according to equation (1).

For instance, the simulation results of a polynomial approximation with parameters A=50, R=27 is shown in Fig 2:

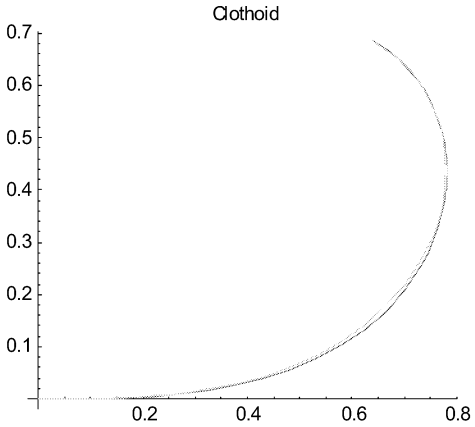


Fig 2 3rd order s-series

It can be seen that both real and approximated curved are very closed for this third order *s-series.* This approximation guarantees also the same start and final point of both clothoids.

Simulation shown in Fig. 2 is affected by a scaling factor, which can be corrected with the homothetic factor equal to $\sqrt{2}\cdot A$

3. Generating Bezier curves

Bezier curves are widely extended for representing functions in CAD applications. In that sense, many researchers are focusing their activities in translating specific curves to Bezier equations.

Bezier approximations are based on a set of control points, as follows,

$$B(u) = \sum_{k=0}^{N} p_k \cdot \frac{N!}{k!\cdot(N-k)!}\cdot u^k (1-u)^{N-k} \tag{7}$$

where:

p_k : Bezier control point

N: Number of Bezier control points

with u : $\in [0,1]$

The following part of this section is focused on transforming single curves (straight lines, circles and clothoids) into Bezier curves. That is, a set of specific points will be obtained from each representation.

3.1 Bezier control points of a clothoid

In [4], polynomial coefficients are transformed into Bezier control points using a transformation matrix. For instance, we can also obtain Bezier control points from a third order polynomial s-series with parameters A=50, R=27 as shown in Fig 3.

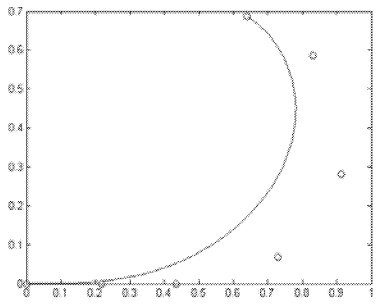


Fig 3 s-series / Bezier points

3.2 Bezier control points of a circle

Approximating circles by Bezier curves is still under research in spite of the fact that many papers have been published on this topic. The main objective of researchers is to obtain a very good approximation with a limited number of Bezier points. In particular, [1] and [2] explains how to approximate a piece of circle by using few Bezier control points. The number of points depends on the angle of the arc and are all of them predefined at a certain distance to the centre of the circle. Fig 4 shows the Bezier approximation of an arc of circle of pi/4 with five control points.

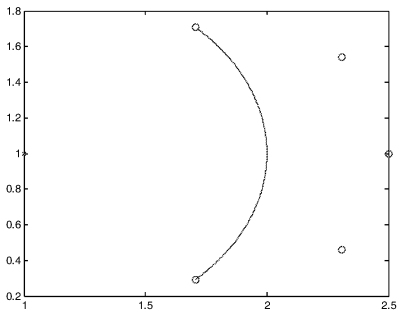


Fig 4 Bezier control points of a circle

3.3 Bezier control points of a straight line

The polynomial approximation of a straight line to a Bezier equation is obvious. Control points are located in the straight line approximately 1/3 of the distance from the start point to the endpoint.

4. Generating three clothoids paths

Based on the combination of three *elementary paths,* as shown in Fig. 5, it is possible to generate trajectories for avoiding obstacles and overtaking vehicles. The fist two clothoids are computed according to the technique explained in [11] and [12], while the third one is computed with a technique named *V-Shape* [13].

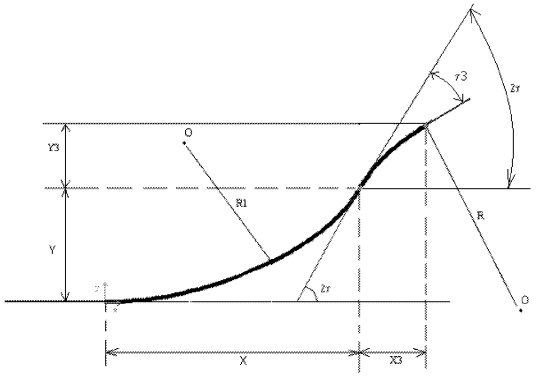


Fig 5 Three clothoids

This algorithm can be used to avoid people or obstacles in general at a security distance. In Fig. 6, sets of three clothoids are computed for avoiding several collisions.

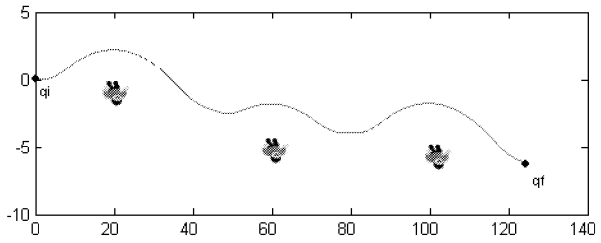


Fig 6 Simulation of collision avoidance

In the case of overtaking other vehicles, our vehicle returns to the original trajectory by using the same three clothoids [15]. That is generating a symmetrical trajectory, as can be seen in Fig 7.

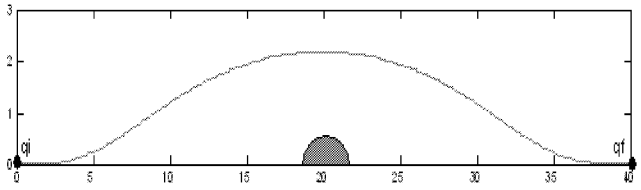


Fig 7 Simulation of overtaking

5. Dual-rate interface

A dual-rate hold is a discrete-time mechanism used to generate output signals at high-frequency sampling rate according to input signals values at low-frequency sampling rate. Therefore, two different sampling periods exist, $\overline{T}$ at low frequency and $\overline{T}$ at high frequency, also named frame-period and base-period respectively. An integer relation of periodicity exists between the sampling rates given by N, that is $\overline{T} = N \cdot T$.

Dual-rate hold circuits can be computed based on primitive functions. The primitive function is used to approximate a continuous function based on the n+1 previous samples (control points) at low frequency [14].

$$f_j(t) = \sum_{l=0}^{n} f_{n,l}(t - t_j) \cdot u(t_{j-l})$$

with $t_j \le t < t_{j+1}$.

This function can be re-written in the following form:

$$f_j(t) = \sum_{l=0}^{n} f_{n,l}\left(\left(\frac{t}{T} - j \cdot N\right) \cdot T\right) \cdot u((j-l) \cdot N \cdot T)$$

and in discrete-time at base-period is:

$$f_j(k) = \sum_{l=0}^{n} f_{n,l}^{*}(k - j \cdot N) \cdot u((j-l) \cdot N)$$

In particular, it is possible to generate dual-rate holds based on Bezier function as primitives. That is for the general, Dual-Rate N-Order Bezier Hold (DR-NOBH)

$$f_{n,l}(t - t_j) = \frac{n!}{l!(n-l)!} \cdot \left(1 + \frac{t - t_j}{t_j - t_{j-n}}\right)^{n-l} \cdot \left(-\frac{t - t_j}{t_j - t_{j-n}}\right)^{l}$$

One of the most important characteristics of this curve is the fact the curve interpolates the first and the last point, meanwhile approximates the rest of points for a given control polynomial. In this sense, we use this curve in order to generate the outputs in $t_j \le t < t_{j+1}$, using the last $n+1$ updated inputs, as shown in Fig. 8:

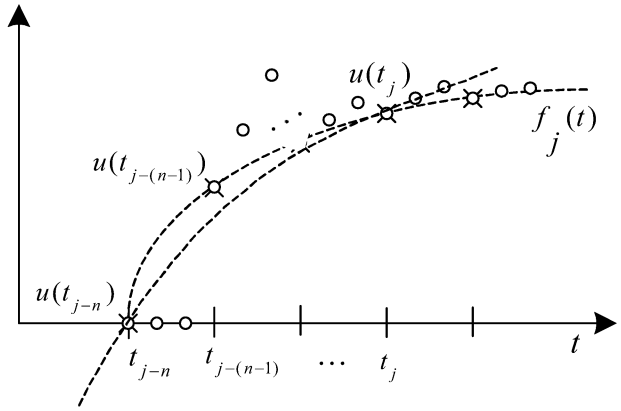


Fig 8 DR-NOBH

The discrete-time function expressed at base-period is:

$$f_j(k) = \sum_{l=0}^{n} \frac{n!}{l!(n-l)!} \cdot \left(1 + \frac{i}{n \cdot N}\right)^{n-l} \cdot \left(-\frac{i}{n \cdot N}\right)^{l} \cdot u(k - i - l \cdot N)$$

and the DR-NOBH transfer function is:

$$\overline{G}(\overline{z}) = \begin{bmatrix} I \\ \dfrac{\overline{z}^n \cdot \left(1 + \dfrac{1}{n \cdot N}\right)^n + \ldots + \left(-\dfrac{1}{n \cdot N}\right)^n}{\overline{z}^n} \cdot I \\ \vdots \\ \dfrac{\overline{z}^n \cdot \left(1 + \dfrac{N-1}{n \cdot N}\right)^n + \ldots + \left(-\dfrac{N-1}{n \cdot N}\right)^n}{\overline{z}^n} \cdot I \end{bmatrix}$$

The following table shows a numerical example for $n = 0,1,2$ and $N = 3$ of the DR-ZOBH, DR-FOBH and DR-SOBH. That is zero-, first- and third-order,

n	DR-NOBH
0	$\overline{G}(\overline{z}) = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$
1	$\overline{G}(\overline{z}) = \begin{bmatrix} 1 \\ 1 + \left(\frac{1}{3}\right) \cdot \frac{\overline{z}-1}{\overline{z}} \\ 1 + \left(\frac{2}{3}\right) \cdot \frac{\overline{z}-1}{\overline{z}} \end{bmatrix}$
2	$\overline{G}(\overline{z}) = \begin{bmatrix} 1 \\ \frac{49 \cdot \overline{z}^2 - 14 \cdot \overline{z} + 1}{36 \cdot \overline{z}^2} \\ \frac{64 \cdot \overline{z}^2 - 16 \cdot \overline{z} + 4}{36 \cdot \overline{z}^2} \end{bmatrix}$

These hold circuits are used to obtain trajectory control points at high frequency using the previous control points and the next destining control points updated at low frequency.

6. Control Architecture

The control architecture is shown in *Fig 9* for industrial vehicle navigation in warehouse [16]. On the one hand, the global planner generates middle points to be achieved according to a predefined road map. On the other hand, the local planner generates trajectories among them. These trajectories are usually straight lines unless some obstacle will have to be avoided. In this case, a set of three clothoids will be calculated. In all cases, Bezier function control points will be finally stored for generating the position reference for the movements. The Dual-rate Interface will be responsible for generating these references based on an interpolation method.

Obstacles are detected and modeled by a convex hull, which will be avoided by the vehicle at a give safety

distance. The avoiding trajectory will converge to the initial trajectory.

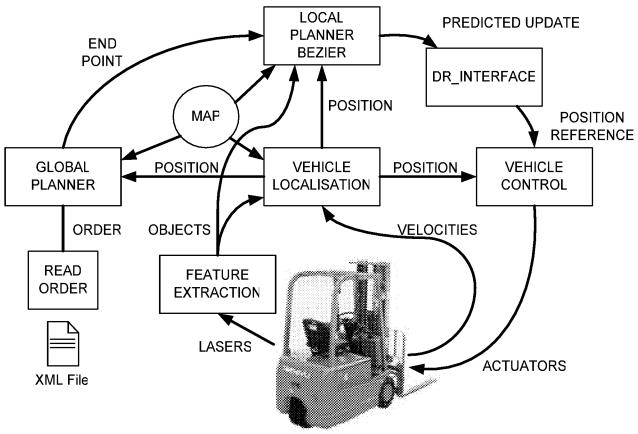


Fig 9: Control Architecture

7. Conclusions

An accurate approximation of clothoids by Bezier equation is explained by the use of s-series. This solves the main problem using clothoids that is the high computational cost. An accurate approximation of circles and straight lines to Bezier equation is also explained. We use these approximations to find free collision continuous path with low computational cost. In order to optimize the path search, the *u* parameter can be adjusted using vehicle dimensions [15].

In addition and in order to adapt sampling frequencies between the planner layer and the vehicle control layer, a dual-rate also based on Bezier functions have been presented.

As further work, it would be interesting to combine several Bezier equations in just one equation.

It is also possible to obtain Bezier control points of clothoids directly from geometrical considerations.

The use of three clothoids to join circles and straight lines has more applications, for instance, it is an ideal solution to collision avoidance vehicle overtaking and lane changing.

References

[1] Chou.J.J.,"Higher order Bezier circles".Computer-Aided design.1995.27(4). 303-309

[2] J.Sanchez Reyes."Higher-order Bézier circles." Computer Aided Design, Vol 29,Nº6, pp.469-472,1997.

[3] J.Sánchez-Reyes,J.M.Chacón."Polynomial approximation to clothoids via s-power series." Computer –Aided Design. December 2003

[4] J.Sánchez-Reyes. "The symmetric Analogue of the Polynomial Power Basis". ACM Transactions on graphics, vol 16, No 3, pages 319-357. July 1997

[5] Papadopoulos,E.,Poulakakis,I.,Papadimitriu,I.,"On Path Planning and Obstacle Avoidance for Nonholonomic Platform with Manipulators: A polynomial Approach", Int. Journal of Robotics Res., Vol 21, No.4,2002,pp.367-383.

[6] P.Jacobs and J.Canny. "Planning Smooth Paths for Mobile Robots." Proc of the IEEE Int.Conf. on Robotics and Automation. April 1989

[7] J.P.Laumond, J.Jacobs, M.Taix and M.R.Murray. "A Motion Planner for Nonholonomic Mobile Robots." IEEE Trans. On Robotics and Automation. October 1994

[8] Scheuer and Fraichard. "Continuous-Curvature Path Planning for multiple Car-Like Vehicles." IEEE Int Conf. on Intelligent Robots and Systems. September 1997

[9] Th.Fraichard, A.Scheuer, and R.Desvigne. "From Reeds and Sheep's to continuous cuvature paths." IEEE Int.Conf. on Advanced Robotics. October 1999

[10] Fraichard and Ahuactzin. "Smooth Path Planning for cars." IEEE. Int Conf on Robotics and Automation. May 2001

[11] *Scheuer and Fraichard ."Planning Continuous-Curvature Paths for Car-Like Robots*."IEEE Int. Conf. on Robotics and Automation. November 1996

[12] *Scheuer and Fraichard ."*Collision-Free and Continuous-Curvature Path Planning for Car-Like Robots." IEEE Int. Conf. on Robotics and Automation. April 1997

[13] Yossawee,W. Tsubouchi, T. Sarata, S. Yuta, S. "Path generation for articulated steering type vehicle using symmetrical clothoids." IEEE. Int. Conf. on Industrial Technology.Vol 1, pp.187-192.2002

[14] L.Armesto and J.Tornero. "Dual-Rate High Order Holds Based on Primitive Functions. "Proceedings of the American control conference. June 2003

[15] N.Montes and J.Tornero. "Lane changing using s-series clothoidal approximation and dual-rate based on Bezier points to controlling vehicle." Int. Conf of system Theory and scientific computation. December 2004

[16] L.Armesto, J.Tornero."Autotrans:Management and transport automation in warehouses" ISC 2005

Genetic Algorithm Based Optimization of Model Reference Adaptive Control Approach for a Vehicle Active Suspension System

Ahmad Bagheri
Assistant Professor, Department of Mechanical Engineering,
Guilan University,
Rasht, P.O.Box 3756, Iran.
bagheri@guilan.ac.ir

Mousa Rezaee
Assistant Professor, Faculty of Mechanical Engineering,
University of Tabriz,
Tabriz, 51736-14766, Iran.
m.rezaee@tabrizu.ac.ir

S. Mehdi Hashemi
Graduate Student, Faculty of Mechanical Engineering,
University of Tabriz,
Tabriz, 51736-33911, Iran.
hashemi81@yahoo.com

Keywords
Active Suspension, Genetic Algorithm, Adaptive Control.

Abstract

A model reference adaptive control technique is applied to a quarter car active suspension system with two degrees of freedom. An ideal mathematical hybrid reference model which represents the response of a desired suspension system is selected and optimized using genetic algorithm. The Lyapunov's direct method is used to design the adaptation laws of adaptive control system which provides a good tracking between the reference model output and active suspension system. Simulation results show good improvement of active suspension performance over the traditional passive suspension systems.

Introduction

Design of a desired suspension system is a compromise between the ride quality and vehicle road holding that indicates the stability of a vehicle. It is not possible to provide both criteria simultaneously in passive systems with fixed physical parameters. Therefore there is inevitable trade off between two design criteria especially for different road inputs [1].
Active and semi-active suspension systems were proposed in 1970s firstly [2] and several researchers have described different approaches to those systems in application to passenger vehicles, military vehicles and trucks. In the active suspensions, a hydraulic actuator is used to provide the needed control force to control the suspension response. But in the semi-active suspensions, the control signal is the variable damping coefficient of a special damper. Although the active suspensions need much more power than the semi-active suspensions, they are more interesting because of their better performance.
Most research has been devoted to consider the control approaches of such systems. Simple On-Off control scheme, optimal control techniques [3], robust control systems [4] and intelligent control systems have been applied and analyzed theoretically and experimentally for active and semi-active suspension systems.
Adaptive control systems can diagnose the input disturbances and structural or parameter variations and adjust the controller parameters via on-line adaptation mechanisms to overcome these uncertainties. Because of wide range of road inputs and uncertainties due to physical parameter changes of vehicle suspension, adaptive control seems to be a suitable control technique for active suspensions.
Sunwoo et al [5] developed the application of model reference adaptive control (MRAC) approach for a quarter car active suspension system with one degree of freedom. In the present paper, the application of an optimized MRAC system for a quarter car active suspension system with two degrees of freedom is presented. All MRAC systems consist of four major parts shown in Figure 1. In this case, the plant is the mechanical implement of active suspension. After modeling the plant, an ideal mathematical reference model is selected and optimized using genetic algorithm. Then control law is chosen and adaptation laws are obtained through Lyapunov's direct method to provide a good tracking between reference model and active suspension.

Plant Mathematical Model

Model of an active suspension system is shown in Figure 2. Equations of motion for this two degree of freedom system are derived and written in state space as:

$$\dot{X} = AX + BU \tag{1}$$

$$\begin{aligned} X_1 &= X_s - X_u \\ X_2 &= \dot{X}_1 \\ X_3 &= X_u - X_{in} \\ X_4 &= \dot{X}_3 \end{aligned} \tag{2}$$

where

$$A=\begin{bmatrix} 0 & 1 & 0 & 0 \\ -K_s(1/M_s+1/M_u) & -C_a(1/M_s+1/M_u) & K_u/M_u & C_u/M_u \\ 0 & 0 & 0 & 1 \\ K_s/M_u & C_a/M_u & -K_u/M_u & -C_u/M_u \end{bmatrix}$$

$$B=\begin{bmatrix} 0 & 0 \\ 0 & 1/M_s+1/M_u \\ 0 & 0 \\ -1 & -1/M_u \end{bmatrix}, \quad U=\begin{bmatrix} \ddot{X}_{in} \\ U_c \end{bmatrix}$$

M_s and M_u represent the quarter car chassis (sprung mass) and one of the tires (unsprung mass) respectively. All of dampers and springs in this model are assumed to be linear and U_c is the control force of actuator. All of these parameters are unknown for controller and they are known only for dynamic simulation. Actuator dynamics can be neglected because of its fast response in comparison with other dynamics of plant.

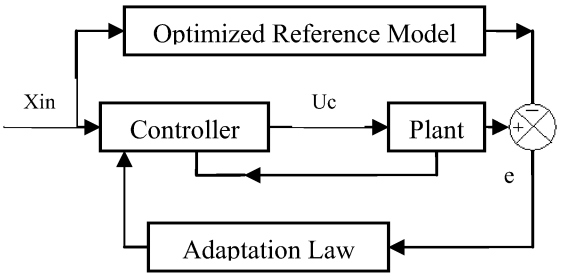


Figure 1 MRAC System Scheme

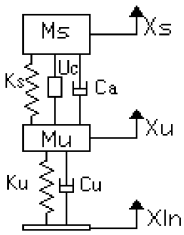


Figure 2 Active Suspension System

Genetic Algorithm Optimized Reference Model

Reference model is a conceptual mathematical model which provides an ideal response to most inputs, which is a sufficient factor. Also the reference model should be in the same order as the plant.

As mentioned before, there are two principal criteria to design and evaluate the performance of suspension systems. RMS of sprung mass acceleration (α) determines the ride quality and maximum of unsprung mass and road surface relative displacement (β) determines the vehicle road holding level which is related to vehicle safety directly.

Ideal reference model should have a low α and β simultaneously. Three different reference models have been proposed: skyhook [2], groundhook [6] and hybrid [7] models. Bagheri et al [8] considered these reference models and analyzed their responses to different road inputs. The skyhook model is the best reference model in the case of ride quality criterion but it doesn't provide an appropriate response for road holding criterion.

Hybrid reference model is the ideal reference model because it has reasonable responses for both criteria. In this model the sprung mass and unsprung mass are attached to the inertial reference via two separate dampers (Figure 3). Equations of motion for hybrid model are derived and state space form of system is written as following:

$$\dot{X}_m = A_m X_m + B_m U_m \tag{3}$$

$$\begin{aligned} X_{1m} &= X_{sm} - X_{um} \\ X_{2m} &= \dot{X}_{1m} \\ X_{3m} &= X_{um} - X_{in} \\ X_{4m} &= \dot{X}_{3m} \end{aligned} \tag{4}$$

where

$$A_m=\begin{bmatrix} 0 & 1 & 0 & 0 \\ -K_s(1/M_s+1/M_u) & -C_3/M_u-(C_1+C_3)1/M_s & K_u/M_u & -C_1/M_s+(C_2+C_4)/M_u \\ 0 & 0 & 0 & 1 \\ K_s/M_u & C_3/M_u & -K_u/M_u & -(C_2+C_4)/M_u \end{bmatrix}$$

$$B_m=\begin{bmatrix} 0 & 0 \\ 0 & -C_1/M_s+C_2/M_u \\ 0 & 0 \\ -1 & -C_2/M_u \end{bmatrix}, \quad U_m=\begin{bmatrix} \ddot{X}_{in} \\ \dot{X}_{in} \end{bmatrix}$$

All values of mass and spring constant of reference model are selected as same as of the plant in Tab. 2, but damping coefficients are selected by an optimization procedure because of their greater influence on transmissibility of road inputs. So design vector of optimization problem becomes:

$$C=[C_1, C_2, C_3, C_4] \tag{5}$$

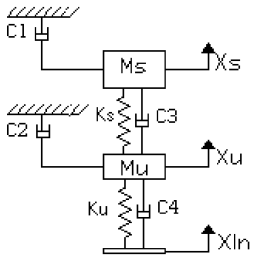


Figure 3 Hybrid Reference Model

α and β are two functions which are requested to be minimized for a sinusoidal road input of Figure 4 in time domain. Thus we have a multiobjective optimization problem. The problem can be simplified to a single

objective problem using weighting factor method. Resulting cost function which should be minimized is:

$$\theta(C_1, C_2, C_3, C_4) = w_1\alpha + w_2\beta \tag{6}$$

Since ride quality has more importance in passenger vehicles, weighting factors are computed as:

$$w_1 = \frac{0.5}{\alpha_{\max}} \ , \ w_2 = \frac{0.5}{\beta_{\max}} \tag{7}$$

Four inequality constraints are applied to optimization problem to limit the values of design variables. Thus the optimum solution will be a normal dynamic system. The cost function is not available analytically, but it can be evaluated in parallel in each step of numerical methods of ordinary differential equation solution method.

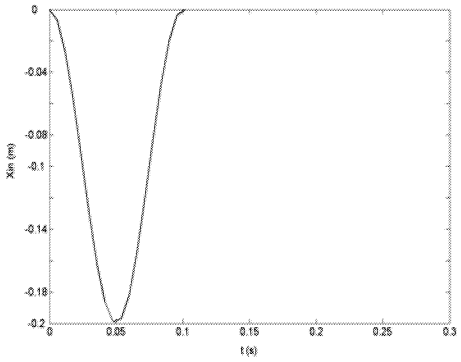


Figure 4 Sinusoidal Road Input

Results of numerical optimization methods highly depend on the initial point and local minimum points may be selected instead of global minimum points in these methods because of periodic properties of cost function. Also such methods need derivative of cost function in optimization procedure which is not available in this case.

Genetic Algorithm as one of stochastic sampling methods is used to optimize the cost function. It is based on maintaining and manipulating a family, or population, of solution and implementing a survival of the fittest strategy in their search for better solutions. In general, the fittest individuals of any population tend to reproduce and survive to the next generation, thus improving successive generations. However inferior individuals can, by chance, survive and also reproduce. Genetic Algorithms is shown to solve linear and non-linear problems by exploring all regions of the state space and exponentially exploiting promising areas through mutation, crossover, and selection operations applied to individuals in the population [9]. Complete steps and procedures of genetic algorithm optimization method are not included here.

In addition to value of cost function, which includes the time response specifications of system, the frequency response specifications are taken into account in the selection of optimum solution throughout the results of several optimization procedure of problem because of their great importance.

Final optimum solution is given in Tab. 1 and they are used in subsequent sections. It should be noted that the solution is a sub optimal solution and maybe better solutions exist.

Table 1 Optimum Parameters of Reference Model

Parameter	Optimum solution (N/m/s)
C1	1458.085
C2	1495.23
C3	925.75
C4	710.7

Control Law and Adaptation Laws

It is evident that although the selected reference model has an ideal response, it cannot be used directly in suspension systems because of its physical limitations. Using active suspension systems, we can adjust the control force online in such a way that active suspension can follow the output of reference model by means of a suitable control law and adaptation mechanisms.

The state equations of plant and reference model (Eqs. (1) and (3)) can be written in alternative forms:

$$\dot{X} = AX + B_1 U_c + B_2 \ddot{X}_{in} \tag{8}$$

$$B_1 = \begin{bmatrix} 0 \\ 1/M_s + 1/M_u \\ 0 \\ -1/M_u \end{bmatrix} \quad , \quad B_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \end{bmatrix}$$

$$\dot{X}_m = A_m X_m + B_{m1}\dot{X}_{in} + B_{m2}\ddot{X}_{in} \tag{9}$$

$$B_{m1} = \begin{bmatrix} 0 \\ -C_1/M_s + C_2/M_u \\ 0 \\ -C_2/M_u \end{bmatrix} \quad , \quad B_{m2} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \end{bmatrix}$$

It is appropriate to suppose a control law which contains state feedback of system as :

$$U_c = M\dot{X}_{in} + LX \tag{10}$$

which M is a scalar adjustable gain and L is an (1×4) adjustable gain matrix.

Substituting Eq. (10) into Eq. (8) yields:

$$\dot{X} = (A + B_1 L)X + B_1 M\dot{X}_{in} + B_2\ddot{X}_{in} \tag{11}$$

Comparing the Eqs. (9) and (11), a set of M^* and L^* for adjustable gains which guarantee the strong tracking of plant and reference model are obtained as:

$$A + B_1 L^* = A_m \tag{12}$$

$$B_1 M^* = B_{m1} \tag{13}$$

Using these parameters in control law of Eq. (10), solution will depend on plant parameters which were assumed to be unknown. Thus the following approach is used.
Using Eqs. (9)-(11), the error between reference model and plant can be written as:

$$e = X - X_m \tag{14}$$

$$\dot{e} = A_m e + B_{m1}(M^{-1}L - M^{*-1}L^*)X + B_{m1}(M^{*-1} - M^{-1})U_c \tag{15}$$

Defining new parameters for error equation:

$$\dot{e} = A_m e + B_{m1}\varphi X + B_{m1}\psi U_c \tag{16}$$

$$\varphi = M^{-1}L - M^{*-1}L^* \tag{17}$$

$$\psi = M^{*-1} - M^{-1} \tag{18}$$

Eq. (16) is similar to the state equation of a linear time invariant (LTI) system. Lyapunov's direct method can be used to have asymptotical stability for error equation and converge it to origin. Positive definite Lyapunov function for this LTI system can be written as [5,10]:

$$V = e'Pe + \varphi F\varphi' + \psi G\psi' \tag{19}$$

where F and G are symmetric positive definite weighting matrices and P is determined by Lyapunov Equation as:

$$Am'P + PAm = -Q \tag{20}$$

Lyapunov's method states that necessary and sufficient condition for a LTI system stability is that for any symmetric positive definite matrix Q, the unique matrix P for solution of Eq. (20), should be symmetric positive definite [11].
Derivative of Lyapunov function which should be negative definite for stable systems becomes as following using Eq. (16) to (21):

$$\dot{V} = -e'Qe + 2e'p(B_{m1}\varphi X + B_{m1}\psi U_c) + 2(\varphi F\dot{\varphi}' + \psi G\dot{\psi}') \tag{21}$$

As A_m is selected to be a stable matrix, the first term of Eq. (21) will be negative via Lyapunov's stability method. Two last terms of Eq. (21) will converge to zero if we adjust two parameters as:

$$\dot{\varphi} = -B'_{m1}PeX'F^{-1} \tag{22}$$

$$\dot{\psi} = -B'_{m1}PeU_cG^{-1} \tag{23}$$

Thus, derivative of Lyapunov function will be negative definite and selection of Eq. (19) is a valid selection as Lyapunov function and error will converges to zero:

$$\dot{V} = -e'Qe \tag{24}$$

Using Eq. (17) and (18) yields adaptation laws as a system of equations:

$$\dot{M}^{-1}L + M^{-1}\dot{L} = -B'm1PeX'F^{-1} \tag{25}$$

$$-\dot{M}^{-1} = -B'm1PeUcG^{-1} \tag{26}$$

Proper selection of weighting matrices F, G and Q and initial values for M and L plays a significant role in stability and tracking of MRAC system.

Simulation and Results

Simulation of designed active suspension with MRAC approach was done in Matlab/Simulink software.
Nominal parameters of commercial passenger vehicle, which is listed in Tab.2 and a sinusoidal road input of Figure 4 were used in simulation.
Figure 5 shows the acceleration of sprung mass and Figure 6 shows the relative displacement of unsprung mass and road surface for active suspension, passive suspension and ideal reference model. Both RMS and maximum of different results are shown in column diagrams. Figure 7 illustrates the control signal (actuator force), which has been computed by MRAC system during the simulation.

Table 2 Parameters of Active Suspension System

Parameter	Value
Ms (Kg)	310
Mu (Kg)	45
Ks (N/m)	40000
Ku (N/m)	190000
Ca (N/m/s)	1500
Cu (N/m/s)	100

Results show that a good improvement can be achieved by the active suspension over passive suspension in the form of 13% reduction in RMS of sprung mass acceleration and 21% reduction in maximum of unsprung mass relative displacement which are the most important design criteria of suspension systems. Also, 10% reduction in maximum of sprung mass acceleration and 23% reduction in RMS of unsprung mass relative displacement can be seen.
In the MARC approach, a persistent input is required to have a complete tracking between plant and reference model [11], but the usual road input of Figure 4 is not such a input. Although complete tracking hasn't happened in the system, the active suspension improved the design criteria in comparison with the passive system.

Conclusions

An ideal reference model is optimized via a multiobjective optimization method. The active suspension system with MRAC approach is able to overcome the trade off between design criteria of passive suspensions and it improves the ride quality and vehicle road holding simultaneously. Lyapunov's stability method was used to derive the adaptation laws of MRAC system which controls the force of active suspension system by means of a hydraulic actuator. State feedback is needed to calculate the control input, so they were selected as the relative velocities and displacements, which are easier to measure.

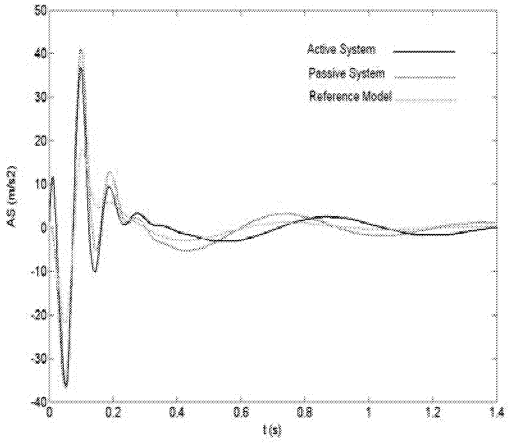


Figure 5a Acceleration of Sprung Mass for Active, Passive and Reference Models

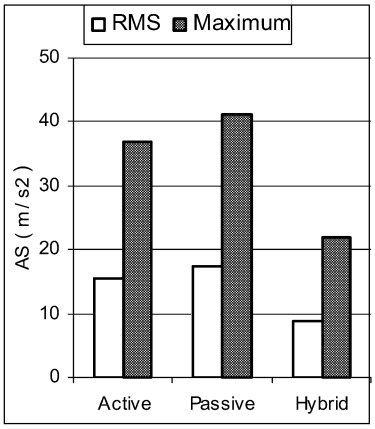


Figure 5b RMS and Maximum of Sprung Mass Acceleration

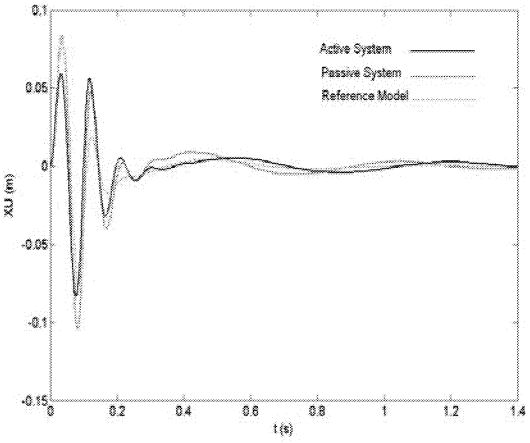


Figure 6a Relative Displacement of Unsprung Mass for Active, Passive and Reference Models

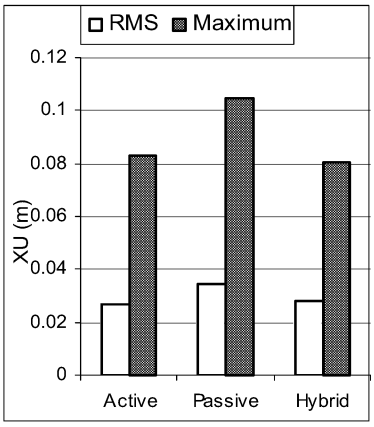


Figure 6b RMS and Maximum of Unsprung Mass Relative Displacement

References

[1] Wong, J.Y., 2001, Theory of Ground Vehicles, John Wiley & Sons Inc.

[2] Crosby, M.J, and Karnopp, D.C, 1973, "The Active Damper," Shock and Vibration Bulletin, 43[4] Kransnicki, E.J., 1982 "The Experimental Performance of an On-Off Active Damper," Shock&Vibration Bulletin, vol.51, part1.

[3] Hrovat, D., 1992, "Application of Optimal Control to Advanced Automotive Suspension Design," ASME J. of Dyn. Syst., Meas., Control, Vol. 114, 328-335.

[4] Ray, L.R., 1992, "Robust Linear-Optimal Control Laws for Active Suspension Systems," ASME J. of Dyn. Syst., Meas., Control, Vol. 114, 592-598.

[5] Sunwoo, M., Cheok K.C, 1991, "Model Reference Adaptive Control of Vehicle Active Suspension System," IEEE Trans. on Industrial Electronics, vol.38, 217-222.

[6] Valasek, M., Novak, M., 1997, "Extended Groundhook-New Concept of Semi-active Control of Truck's Suspension," Vehicle System Dynamics, Vol. 27, 289-303.

[7] Ahmadian, M., 1997, "A Hybrid Semi-active Control for Secondary Suspension Applications," Proceedings of the 6th ASME Symposium on Advanced Automotive.

[8] Bagheri, A., Hashemi, S.M., 2003, "Selection of an Appropriate Reference Model for Active Suspension of Vehicles," 7th ISME International Conference, 1890-1898.

[9] Houck, C. R., Joines, J. A. and Kay, M. G., " A Genetic Algorithm for Function Optimization:A Matlab Implementation", Mathworks corporation.

[10] Astrom, K.J., and Wittenmark, B., 1995, Adaptive Control, Addison Wesley Pub.

[11] Slotine., E., and Li, W., 1991, Applied Nonlinear Control, Prentice Hall.

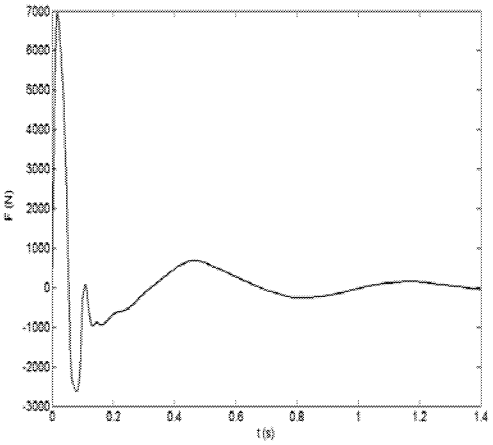


Figure 7 Control Force

USING MODELLING TECHNIQUES FOR FAULT PREDICTION IN HYDRAULIC DRIVE SYSTEMS

C. Angeli
Department of Mathematics and Computer Science
Technological Institute of Piraeus
Konstantinoupoleos 38, N. Smirni
GR-171 21 Athens, Greece
E-mail: c_angeli@otenet.gr
c.angeli@computer.org

A. Chatzinikolaou
S. Patsi 62
GR-118 55 Athens,
Greece
E-mail: achatzi@otenet.gr

KEYWORDS

Model based fault diagnosis, fault prediction, hydraulic systems, real time process.

ABSTRACT

Dynamic modelling techniques involved in the structure of a diagnostic system offer the opportunity to predict faults at an early stage. This can be achieved by developing a suitable mathematical model able to simulate the dynamic behaviour of a system and compare this behaviour to the real-time measurements. In this paper, the characteristics of hydraulic systems are studied and used for the prediction of incipient faults. Both simulation results and experimental results show that the proposed method is capable of predicting faults and can be responded successfully to the requirements of real-time performance.

INTRODUCTION.

Early fault detection is a critical issue for the industrial systems and is considered particularly beneficial for the production technology. The main reasons are the safety and the economical benefits offered to the production procedure by preventing extensive process failures. Recent attempts in developing fault prediction systems have been reported among others researchers by (Isermann 1984; Li et al. 1998; Karri 2000; Zhou and Paik 2000; Angeli and Chatzinikolaou 1999).

Hydraulic drive systems are used in a wide variety of applications from precision control systems in robotics and aerospace to heavy industrial systems as rolling machines and defence applications and tend to find more applicable fields due to their main advantages to transmit high forces in long distances with very little loss as well as their compactness. The development of advanced diagnostic procedures that permit early fault detection for these systems will affect the reliability and safety of the total system and in consequence will be of great interest for the industry.

Although there is a lot of recent valuable research work in performance analysis, condition monitoring and fault diagnosis of hydraulic components (Kwan et al. 2003; Edwards and Friswell 1998; Gao and Patton 2003) there are only few attempts in developing diagnostic methods for the total hydraulic system (Mundy and Stammen 2002; Chen and Lu 1995; Angeli and Chatzinikolaou 1995). This paper presents the use of dynamic modelling information in comparison to the actual process offers the opportunity of an early fault detection. The influence of an insipient leakage on the dynamic response of the hydraulic system is the main feature that is used for the development of the fault prediction procedure.

MODELLING PROCESS

For experimentation purposes a typical hydraulic system was used. This system consists mainly of a hydraulic motor that rotates a disk. Figure 1 presents this part of the hydraulic system.

The hydraulic motor is controlled by a proportional valve. An input current I controls a proportional 4-way valve and thus the flow rate to the hydraulic motor. The rotation speed of the hydraulic motor depends on the actual flow rate through the motor ($q_{mb} = q_{ma} \cdot \eta_v$) and on its displacement volume V_m. The major components of concern for this presentation are the hydraulic motor (2.1), the proportional 4-way valve (1.3) and the pipe connections between them.

The main considerations for the description of the dynamic behaviour of hydraulic systems refered to the fact that the most components in a hydraulic system can be regarded as quasi-steady, because the volume of the incorporated hydraulic fluid is usually very small, so it can be regarded as incompressible, and in addition their moving parts are so small that their inertia is unimportant. Hence, the dynamic state of the components can be regarded as a sequence of stationary states while the whole system is dynamic.

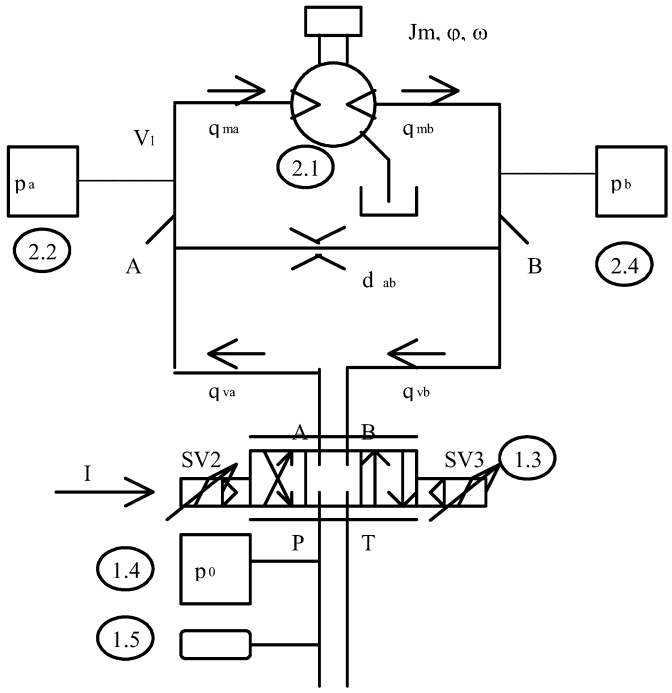


Figure 1: The actual hydraulic drive system

The developed mathematical model takes into account the non-linear character of hydraulic systems and the incompressibility of the hydraulic fluid in the pipes as well as the special characteristics of the used hydraulic elements. The mathematical equations that describe the behaviour of the hydraulic elements are based on their technical specifications, function curves and parameter values given by the manufacturer and tested in the laboratory so that the model represents the behaviour of the system elements more accurately.

For those components which contain large quantities of oil, as pipes and hydraulic actuators, it may be assumed that the change in pressure in the dynamic state is proportional to the net inflow of oil, that is:

$$dp / dt = (E / Vo) \cdot \Sigma Q \qquad (1)$$

where $\Sigma Q = Qin - Qout$,

Qin is the incoming flow to the volume V_l of a connecting pipe

and Qout the outgoing flow from V_l.

That means that the pressure increase in a pipe element at a junction is proportional to the algebraic sum of the incoming and outgoing flows and inverse proportional to the included oil volume.

The functions of the flow rates Qin and Qout through the elements at a junction depend on the nature of the elements and can be determined easily. Important is to put the right parameter values in the equations.

The use of the relation (1) in every junction of the circuit and the equations that express the flow rates, build the mathematical model of the hydraulic system. For example, for the connecting pipes the relationships describing the increase in pressure at the junctions A and B are:

$$\frac{dp_a}{dt} = \dot{p}_a = \frac{E}{\frac{V_m}{2} + V_l} \cdot \left(q_{va} - q_{ma}\right) = C_p\left(q_{va} - q_{ma}\right)$$

$$\frac{dp_b}{dt} = \dot{p}_b = \frac{E}{\frac{V_m}{2} + V_l} \cdot \left(q_{mb} - q_{vb}\right) = C_p\left(q_{mb} - q_{vb}\right)$$

$V_m/2$ is half the displacement volume of the hydraulic motor. This value must be added, because the hydraulic motor includes a considerable oil quantity that has to be added to the oil quantity of the connecting pipes at both sides.

The above equations together with the mathematical expressions of the flows through the motor and the proportional valve as well as the Newton relation of the rotated mass represent the mathematical model of the system.

SIMULATION OF THE SYSTEM

The modelling of the hydraulic elements with the coupled moved masses leads to a non-linear system of equations. This system can be written in the following state vector form:

$$\dot{\bar{x}} = f\left(\bar{x}\right) \qquad \text{where} \qquad \bar{x} = \begin{bmatrix} p_a \\ p_b \\ \dot{\varphi} \\ \varphi \end{bmatrix}$$

Thus:

$$\begin{bmatrix} \dot{p}_a \\ \dot{p}_b \\ \ddot{\varphi} \\ \dot{\varphi} \end{bmatrix} = f \begin{bmatrix} p_a \\ p_b \\ \dot{\varphi} \\ \varphi \end{bmatrix}$$

The system of the equations was solved using the standard fourth-order Runge-Kutta method. The results of the simulation process for the pressures p_a and p_b at the ports of the hydraulic motor are presented in Figure 2.

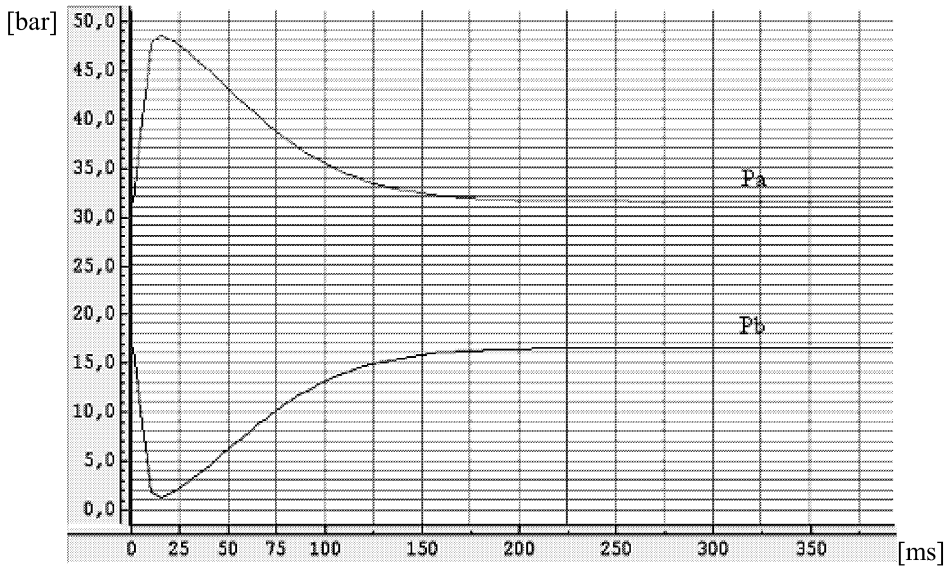


Figure 2: Simulation results for the pressures p_a and p_b

These results represent the behaviour of the fault free system and were used for the fault prediction process. The simulation results were validated after a comparison with the actual measurements on the fault free hydraulic system.

4. USE OF MODELLING INFORMATION FOR THE DETECTION OF INCIPIENT FAULTS

The results of simulation for the pressure p_a for various leakage situations are presented in Figures 3. In this Figure the parameter is the diameter in mm of a bypass orifice placed between the two motor ports that simulates a small additional motor leakage. The curves in Figure 3 represent the pressure p_a of the dynamic response of the system. In this Figure the influence of the leakage on the pressure p_a can be observed. The two curves represent the pressure p_a in case that the additional motor leakage is zero (dab = 0) and in case of a small leakage between A and B (dab = 1 mm).

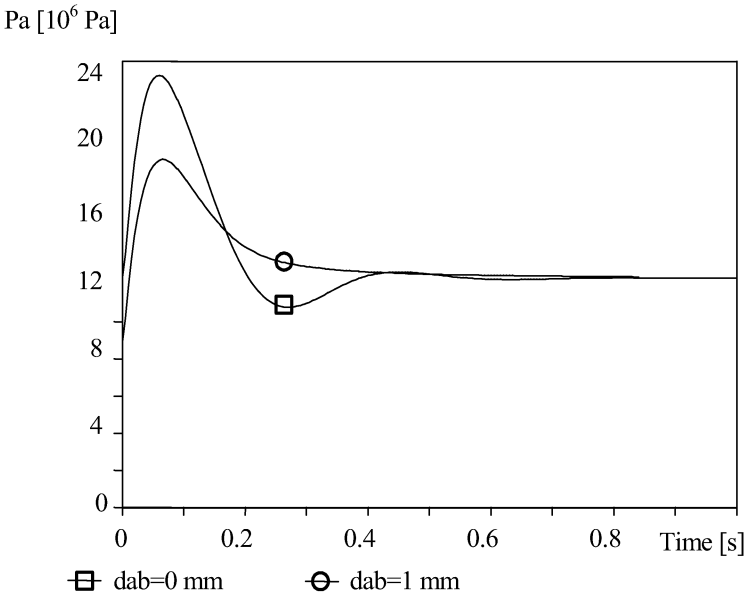


Figure 3: Pressure p_a for various orifice combinations.

It is obvious that the dynamic response changes dramatically in comparison to the static behaviour in case of a small leakage.

IMPLEMENTATION OF THE FAULT PREDICTION

The prediction of faults was implemented as part of a diagnostic expert system. Sensors were placed at suitable positions for the measurement of the system's variables and the acquired signals were sent to the computer through an I/O card. The simulation program was executed in parallel to the run of the real system so that the values of the measured variables could be compared with the calculated values from the simulation program. The deviation of the curves from model and system in the dynamic range was used for the prediction of a fault.

The structure of the fault prediction process is presented in Figure 4. The input to the actual system is the control signal I from the control system and the outputs which are fed to the expert system are the angular velocity ω, the pressures p_a, p_b, and the state signals from the devices of the power unit.

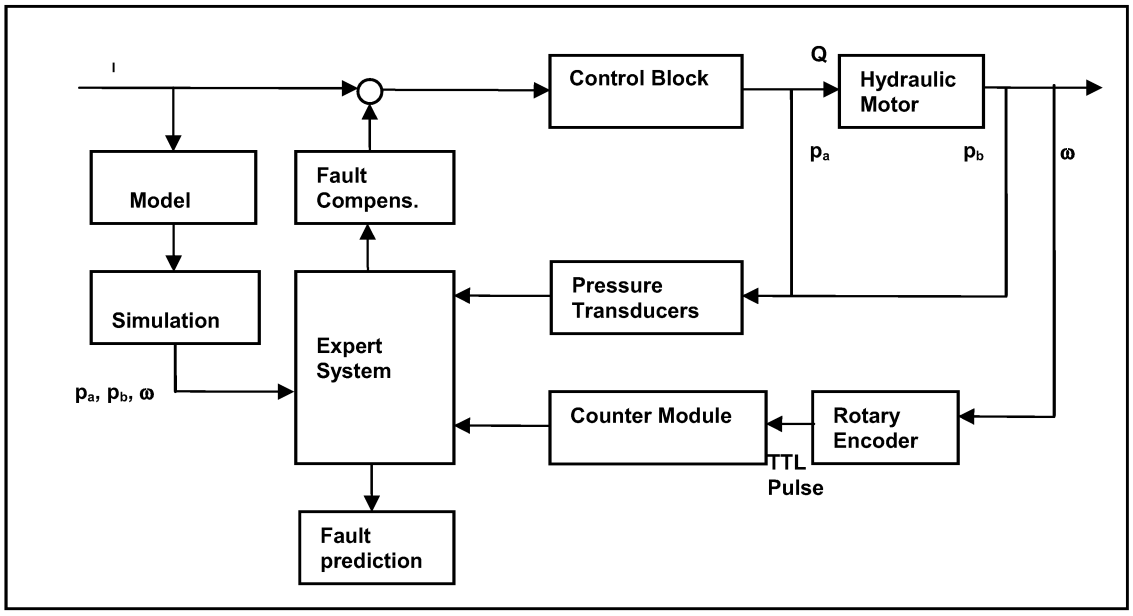


Figure 4: Fault prediction process structure

The data acquisition system acquires data for the angular velocity and the pressures while the speed is changing. The simulation program is running simultaneously and calculates the values of the angular velocity and the pressures for the same period of time.

The comparison of the measured and the calculated values is then performed and the difference is written to an output file. A measurable leakage of the hydraulic motor has as result a decrease of the speed ω. The difference $|\Delta\omega|$ between measurement and calculation (model) is a criterion for that decrease and can be used by the expert system for leakage detection. If the leakage is very small it is not yet measurable but it has a considerable influence on the pressure curves p_a (t) and p_b (t) in the acceleration phase of the hydraulic motor. By comparing the measured curves with the results of the simulation this small leakage can be visible and evaluated. The differences Δp_a and Δp_b between measured and calculated curves can be measured and by exceeding of a given threshold a beginning motor leakage can be detected that otherwise could not be measured at this early phase.

The conclusion whether these differences are acceptable or not is written to text files. The final conclusion about the prediction of a fault is performed by the knowledge base of the system that is also responsible for the diagnosis of faults.

CONCLUSION

In this paper, the use of modelling techniques for the prediction of incipient faults is presented. The influence of the leakage on the dynamic behaviour of a hydraulic system was successfully used for the establishment of a fault prediction system. Dynamic modelling information involved in the fault detection system offered the capability of predicting real time faults.

The main advantage of this method is that the simulation model is realistic and can respond to the requirements for real-time performance. The experimentation results show that the fault prediction and diagnosis procedure is reliable and the method is applicable to real world system. The importance of developing fault prediction functions lies in the higher degree of safety and reliability offered for the industrial systems.

REFERENCES

Angeli C. and A. Chatzinikolaou. 1999. “Fault Prediction and Compensation Functions in a diagnostic knowledge-based system for hydraulic systems”. *International Journal of Intelligent and Robotic systems,* Vol. 25, No.2, 153-165.

Angeli C. and A. Chatzinikolaou. 1995. “An expert system approach to hydraulic systems”. *Expert Systems*, Vol. 12, 323-329

Chen Z. and Y. Lu. 1995. “State Monitoring and Fault Diagnosis for Hydraulic systems”, *Pneumatics and Hydraulics*, 2, 3-7.

Edwards S., A. Less and M. Friswell. 1998. “Fault diagnosis of rotating machinery”, Shock and Vibration Digest, 30, 4-13.

Gao Y. and R. Patton. 2003. “Application of wavelet analysis for performance monitoring and diagnosis of hydraulic pump”, In Proceedings Safeprocess 2003, Washington, U.S.A. 333-338.

Isermann, R. 1984. “Process Fault Detection Based on Modelling and Estimation Methods - A Survey”. *Automatica*, Vol 20, N. 4. 387-404.

Karri, V. 2000. “Drilling Performance Prediction Using General Regression neural Networks”, *In 13 International Conference on Indusrial and Engineering Applications of Artificial Intelligence and Expert Systems, IEA/AIE*, New Orleans, Louisiana, USA, June 19-22, 67-72.

Kwan C., R. Xu and X. Zhang. 2003. “Fault detection and Identification in Aircraft Hydraulic Pumps Using MCA”, In Proceedings Safeprocess 03, Washington, U.S.A. 1137-1143.

Li I.S., K. Kwok and J.Zurcher. 1998. “Prediction and Prevention of Sheetbreak Using PLS and an Expert System”. In IFAC Symposium, *Fault Detection, Supervision and Safety for Technical Processes*, Kingston Upon Hull, UK, 1159-1164

Mundy S. and C. Stammen. 2002. “Condition Monitoring for fluid technology”, o+p Oelhydraulic and Pneumatic, Vol. 46, Nr.2.

Zhou M. and J. Paik. 2000. “Integrating Neural Network and Symbolic Inference for Prediction in Food Extruction Process”. *In 13 International Conference on Indusrial and Engineering Applications of Artificial Intelligence and Expert Systems, IEA/AIE, New Orleans*, Louisiana, USA, June 19-22, 567-572.

DESIGN OF A CUSTOM PLOWING PROCESS USING SIMULATION

Allan Burke
Víctor Acosta
José Carlos Miranda
Juan de Dios Calderón

Automotive Mechatronics Research Center, ITESM Campus Toluca
San Antonio Buenavista, Toluca, Mexico 50110
Email: {allan.burke, a01102240, jmiranda, jcalderon} @itesm.mx

KEYWORDS
Computer-Aided Engineering CAE, Eulerian, LS DYNA, Plowing simualtion

ABSTRACT

Copper commutator manufacturing for direct current motors requires a special machining process called plowing. This process, which forms the material without chip removal, is carried out to generate small fins used to fix a dielectrical resin to the copper core. These fins are of extreme importance since the strength of the union between resin and copper is the most relevant mechanical property of commutators. For the generation of fins of optimal geometry, it is necessary to understand this particular machining process in order to set critical angles, radius and lengths in the cutting tool. For this purpose, a simplified analytical model was elaborated and an Eulerian dynamic Finite Element model based on that of Raczy, Altenhof and Alpas [1] was develop to simulate material flow around the tool. From the results obtained, tool geometry could be established with the objective of maximizing the strength of the union between copper and resin.

INTRODUCTION

The manufacturing of copper commutators for electrical motors requires a special process known as plowing or scarfing which forms small fins inside them as shown in figure 1. This process, which forms the material without chip removal, is required to generate small fins used to fix a dielectrical resin which isolates the copper from the shaft. These fins are of extreme importance since they dictate the strength of the union between resin and copper. This union is most relevant since at high rpms the resin tends to separate from the copper which results in the malfunction or failure of the motor. Hence, the design of a plowing tool must have as objective to generate the fin geometry that maximizes the strength of the union between copper and resin. In the present work analytical and numerical models used to simulate the plowing process are presented together with the steps followed to design the cutting tool. Some data has been intentionally excluded to maintain project confidentiality with the industrial partner.

PROCESS DESCRIPTION

Process description

Machining processes are generally characterized by the removal of the material from the surface of a work zone, generating chip formation by deformation and continuous shearing of the material. In the case of the plowing process, material is not removed from the work zone; it is repositioned as the tool moves axially creating fins and channels. This tool is fastened to a press by means of thread in its inside and tightly held in the outside.

Tool description

The fins shown in figure 1 are generated by a tool with a geometry defined by three main features: the rake angle (α), the incidence angle (λ) and the actual body of the tool (see figure 2). Rake and incident angles are responsible for the initial deformation and the tool body is responsible for the final shape of the fin. Like in conventional cutting processes, changes in tool geometry and cutting parameters alter directly tool performance and forces needed to make the repositioning of the material. The way these variables affect the behavior of the tool can be studied in analytical and numerical models, as shown next.

ANALYTICAL MODEL

Based on theoretical foundations, it is possible to adapt the Merchant model for orthogonal cutting [4], to estimate cutting forces during the plowing process. The variables included in the analytical model developed are shown in table 1. The main responses of interest are the forces that act over the tool: the cutting force *(Fc)* and the thrust force *(Ft)*.

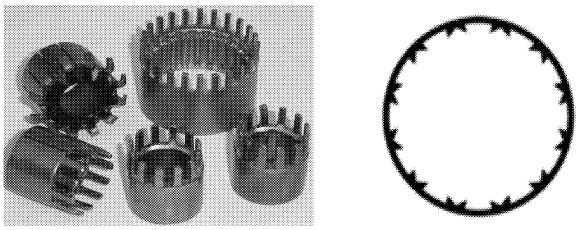


Figure 1: Commutators and Fins Profile

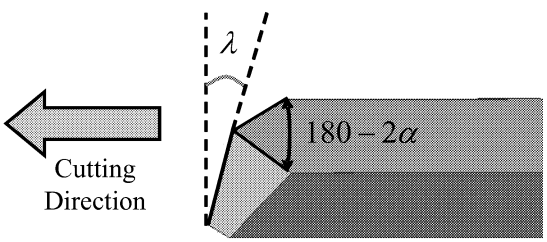


Figure 2: Tool Geometry

Table 1: Variables and Factors Used in the Model

Input	Cutting speed of (V)
	Friction coefficient (u)
	Rake angle (α)
	Friction angle $(\beta) = \tan^{-1} u$
	Depth of cut (a_p)
	Yield stress (σ)
	Feed (f)
Output	Shear angle (φ)
	Thrust force (Ft)
	Cutting force (Fc)
	Shearing force (Fs)
	Resulting force (R)
	Normal force (N)
	Friction force (F)

The thrust and the cutting forces create a resultant force *(R)* that is decomposed in two parts, a friction force *(F)* and a normal force *(N)*. The form in which the forces are applied in the tool is shown in figure 3. From the model and using the variables listed in table 1, the following relations arise:

$$\varphi \approx 45 + \frac{\alpha}{2} - \frac{\beta}{2} \quad (1)$$

$$Ft \approx \frac{\sigma \times \sin(\beta - \alpha) \times a_p \times f}{\sin\varphi \times \cos(\varphi + \beta - \alpha)} \quad (2)$$

$$Fc \approx \frac{\sigma \times \cos(\beta - \alpha) \times a_p \times f}{\sin\varphi \times \cos(\varphi + \beta - \alpha)} \quad (3)$$

$$F \approx R \times \sin\beta \quad (4)$$

$$N \approx R \times \cos\beta \quad (5)$$

Equation (1) shows that a relation exists between the tool geometry and the shear angle *(φ)*. As this angle increases, the cutting pressure decreases and the material undergoes a smaller deformation.
Figure 4 shows the behavior of the forces when the rake angle (α) is varied between 15° and 80°. From the figure it is possible to infer that:

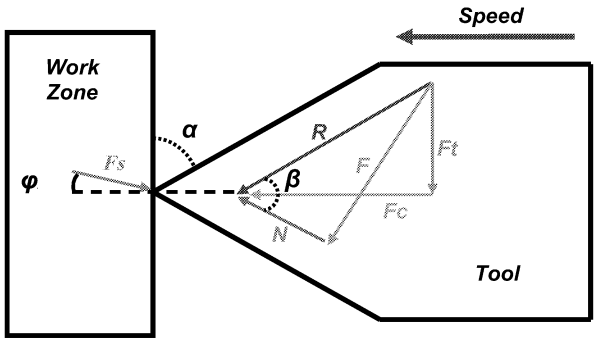


Figure 3: Analytical Model

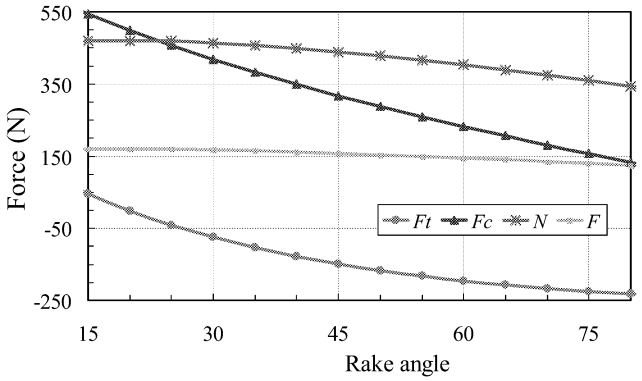


Figure 4: Behavior of the Resulting Forces with the Change of the Rake Angle

1. The negative values of the thrust force *(Ft)* indicate that the magnitude of the friction angle (β) is smaller than the rake angle (α), given that the magnitude of *Ft* is assumed positive.
2. As the rake angle is reduced, the resulting force *(N)* shows an increment in its magnitude and reaches its maximum values between 20° and 30° of the rake angle.
3. The behavior of the cutting force *(Fc)*, shows an inverse relationship to the rake angle as it is with to thrust force *(Ft)*.
4. A stable interaction of the forces over the tool can be found when *Fc* is greater than *Ft*. *Ft* has a great influence over the tool due to the low stiffness of it in this direction.
Hence, it is recommended that (α) is between 55° and 65° where forces are equilibrated.

NUMERICAL MODEL

Static analysis of the tool respect to the geometry variation

Once cutting forces from the analytical model were obtained, they were applied in different models. These models tested different geometries, varying the tool body and the incidence and rake angles to determine best structural performance.
The first feature analyzed was the rake angle, leaving fixed the other features. Two rake angles were used, 55° and 65°, in order to observe its performance. Two load cases were considered: balanced and unbalanced forces. The balanced load case establishes that tool is perfectly aligned with the workpiece, seen as an ideal situation. The unbalanced load case establishes a lack of alignment, causing unequal loads in both sides of the tool. The balanced load case shows similar stress magnitudes regardless of the value of the rake angle, as shown in figure 5. On the other hand, the unbalanced load case presents an stress increment of 175% of the 65° over the 55° angle.
The structural behavior of the tool body geometry was analyzed in a balanced load case through the four different geometries shown in figure 6. These geometries were selected by the fin profile they produced. All the tested tools showed stresses 50% under the material yielding stress.
The effect of the incidence angle in this process was investigated also through a balanced load case. Six models were investigated where the incidence angle took values

from -5° to 5°. From the experiments it was observed that the maximum stress stayed almost unchanged as the angle was varied having smaller values for negative angles.

Dynamic Model

The objective of this simulation was to compare and review the final geometry of the fins using the four different tool shapes and at the same time compare the forces obtained with the analytical method. The rake angle was set to 55° due to its better performance against unbalanced loads.
The model developed for LS-DYNA [3] is composed by three parts: the tool, the copper and the air mesh shown in figure 7. Symmetry was used to reduce computation time and keep a detailed geometry of the model. To work with the Eulerian formulation is necessary for the air space where the simulation is done to be bound. This space has a Lagrangian formulation and is defined as a void space. This space allows the copper to flow, determining the behavior of the metal following the track of each copper element as deformation takes place.
This model is based on previous work by Raczy, Altenhof and Alpas [1] who developed an Eulerian Finite Element Model for orthogonal cutting. Due to its Eulerian formulation, the copper part geometry is defined by the cross section of the workpiece. This characteristic reduces the amount of nodes and elements needed to describe the part.
The tool is located inside the air mesh and has a lagrangian Bellytschko-Tsay shell formulation. The tool is modeled as a rigid body to reduce the solution time.
The material assigned to the copper part uses an elastic plastic hydrodynamic constitutive material model where the material behavior is described by means of a function of the equivalent plastic strain. For the problem at hand, the copper material model was taken from [2]. To describe the pressure-volume relationship of the copper specimen considering impact, the Gruneisen equation was used [5].

Three boundary conditions were specified in the model:
1. A constant velocity of 600 mm/s in the *x* direction was applied to the copper part. This constant velocity describes the route of the elements until they starts deforming against the tool.
2. The bottom of the copper and the air mesh is restricted to move only in the *x* degree of freedom.
3. The tool is fixed in all directions.

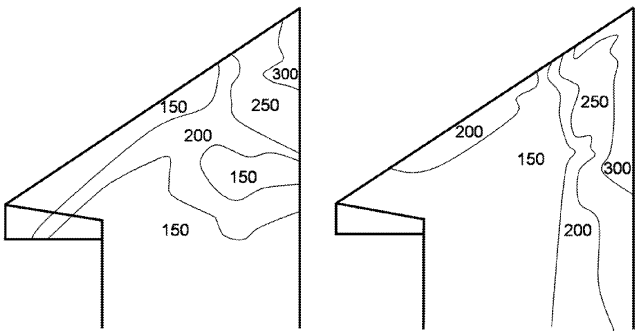


Figure 5: Distribution of Stresses (MPa) at the Tip of the Tool for a Rake Angle of 65° (left) and 55° (right) for the Balanced Loadcase

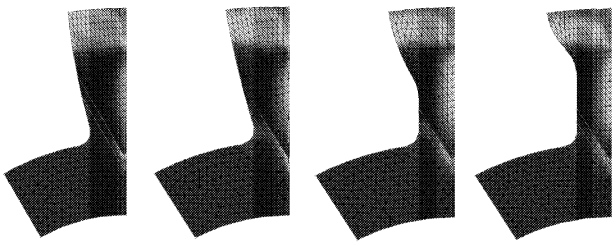


Figure 6: Stress Distribution on Tools' Bodies. From Left to Right: Tool #1, Tool #2, Tool #3 and Tool #4

The model makes the following assumptions and simplifications:
1. The force needed to divide the flow of the copper at the tool tip is neglected for simplicity.
2. The tool is undeformable assuming that the hardness difference with the copper part is significant enough to consider the tool deformation as negligible.
3. Geometrical simplifications were made to ease the meshing process.
4. The copper part displacement is completely restricted. In the real process the radial pressure applied to the commutator is not always enough to avoid movement.
5. The contact between the tool and the copper is modeled through an algorithm that couples Eulerian and Lagrangian elements. The friction coefficient was neglected assuming that is already included in the material formulation.

The termination time of the analysis is set at 5 milliseconds. This time is though to be sufficient to evaluate the final form of the workpiece. In average, an analysis took about 150 hours to run in computer equipped with a 3 GHz Xeon processor and 3 GB in RAM.

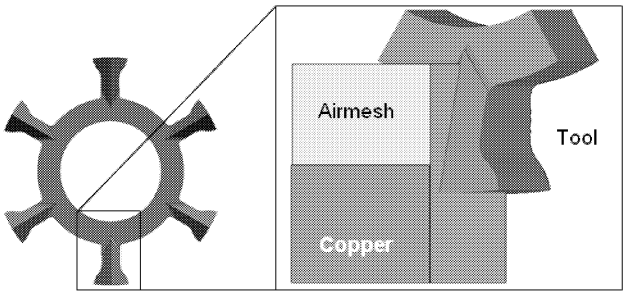


Figure 7: Tool, Airmesh and Copper Part

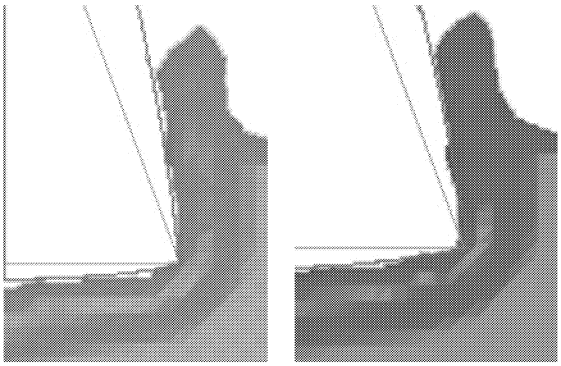


Figure 8: Final Shape of the Fin Due to Tool #1 (left) and #2 (right)

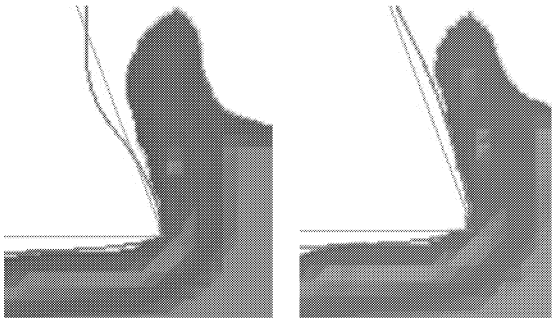


Figure 9: Final Shape of the Fin Due to Tool #3 (left) and #4 (right)

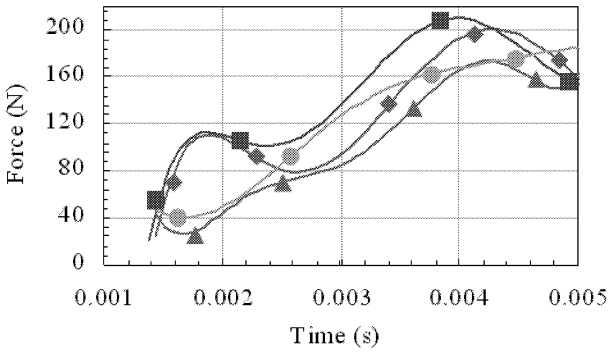


Figure 10: Cutting Force of the Tools #1 (Triangle), #2 (Square), #3 (Rhombus) and #4 (Circle)

The final profiles of the workpieces due to the four different tools are shown in the figures 8 and 9. Each profile is compared against a blue reference line thought to be the ideal fin angle. From the figure, it can be observed that all the analyzed tools create a final geometry that facilitates the resin to adhere. Nevertheless, the final geometry of tool #3 creates a smaller gap to the reference line, whereas tool #1 shows the biggest gap among the analyzed tools.

The cutting force needed to reposition the material with the analyzed tool geometries is shown in figure 10. From the plot, the maximum cutting force is obtained using tool #2 at 212 N while the minimum cutting force in the is found using tool # 1 at 178 N.

CONCLUSIONS

If modified properly, the fundamental theory used by Merchant for orthogonal metal cutting appears to be a sound option for a first approach in a cutting process as different as plowing. The forces obtained through the developed model differed about a 31 % at most with the forces obtained from the dynamic analysis carried out with LS-DYNA. This result, despite that Merchant theory does not consider cutting velocity as a variable.

Since the difference in stresses between tools with rake angles of 55 and 65 degrees was not significant for the balance load case, for the final design the rake angle was set at 55 degrees to provide a low sensitivity against unbalanced forces that may be present at the production press.

The incidence angle was set as negative as possible as the trend of the analysis indicated that negative values resulted in slightly lower stresses. The only limitation of this angle is manufacture feasibility as the cutting edge of these tools is made by grinding and therefore is limited by grinding wheel geometry. As static structural analyses over the different tool bodies revealed no significant differences in stress magnitude and distribution, the final shape of the tool body was selected solely on the results of the dynamic analysis. From these results, tool body #3 generates the best final fin shape. From the process followed and the results obtained, it is clear that Eulerian models are suitable to model different metal cutting processes in order to simulate chip formation despite the large computation time that they require.

ACKNOWLEDGMENTS

The authors wish to thank Dr. William J. Altenhof of the University of Windsor, Canada for the information and support provided for the realization of this work. Also, the support of Industrias Kirkwood de Mexico in the form of access to information is hereby gratefully acknowledged. The LS-DYNA and Hypermesh software used for the realization of this research project was donated to ITESM Toluca by Livermore Software Technology Corporation and Altair Engineering respectively, as part of the Partners for the Advancement of Collaborative Engineering Education (PACE) program.

REFERENCES

[1] Altenhof, W. 2004. "An eulerian finite element model of the metal cutting process". *8th LS-DYNA Users Conference.*

[2] Elmadagli, M. and Alpas, A.T. 2003. "Metallographic analysis of the deformation microstructure of copper subjected to orthogonal cutting". *Materials Science and Engineering Analysis*. p. 249-259

[3] Livermore Sofware Technology Corporation. 2003. "LS-DYNA 970 Keword User´s Manual Volume I". Livermore, CA.

[4] Merchant, E. 1945. Mechanics of the metal cutting process. I Orthogonal cutting and type 2 chip. *Journal of applied physics*. 16. p 267-275.

[5] Steinberg, K., 1996. "Equation of state and strength properties of selected materials". *Lawrence Livermore National Laboratory.* Livermore, CA

BIOGRAPHY

ALLAN BURKE and **VÍCTOR ACOSTA** are mechanical engineers currently enrolled in the M.Sc. in Automotive Engineering at ITESM Campus Toluca and are research assistants at the Automotive Mecatronics Research Center.

JOSÉ C. MIRANDA has a Ph.D. in Computational Mechanics and is director of the Automotive Mecatronics Research Center at ITESM Campus Toluca. His areas of expertise include simulation, design and optimization.

JUAN DE DIOS CALDERON has a Ph.D. in Mechanical Engineering and is head of M.Sc. in Automotive Engineering program of the ITESM. His areas of expertise include the optimization of manufacturing processes, especially metal cutting.

FACTORY WAREHOUSE AND STORAGE MANAGEMENT

GREENHOUSE AUTOMATION
A Model for Design and Evaluation

Jaap A. Ottjes, Joan C. Rijsenbrij and Auke Groenenboom
Delft University of Technology, the Netherlands, Faculty of Mechanical, Maritime and Materials Engineering
j.a.ottjes@wbmt.tudelft.nl

KEYWORDS

Transport, Greenhouse-automation, mobile robot, process oriented simulation

ABSTRACT

A model structure is proposed for the design and evaluation of greenhouse factories equipped with an automated system for transportation of half products. The model is set up according the process interaction method. The system control decisions are made decentralized by a number of system elements and are modeled and clearly identifiable at several levels in the model. A series of experiments with a prototype model is carried out and presented. The possibilities of re-using the control part of the model in a real system as well as the potential of the design tool are discussed.

INTRODUCTION

Mass production of plants and flowers is performed in two steps. The first step concerns the growth from seed or cutting to relatively small plants in specialized greenhouse factories and the second step is the full development of these half products in greenhouses specialized in a small number of product types, (F+H 2004 and Florensis 2005). This paper concerns the first step, the growing of half products. In this branch only a few enterprises take care of the total market demand. Some of these enterprises have a broad assortment of products. Each product may have its own ideal grow path in terms of time, temperature and moisture rate.

There are two ways to approximate the ideal grow path of a plant. The first way is to keep the plant at a fixed location and to adapt the temperature and moisture rate as a function of time according the ideal grow path. The second way is to establish a limited number of climate zones with a stabilized temperature and moisture rate and to move the plant from zone to zone in such a way that the climate fits best to the plant ideal grow path. The first way has the advantage that the ideal grow path is approximated well, but in case of a large assortment with many different grow paths this will demand a lot of space. The second way requires less space but puts high demands on the timely transportation of plants from one zone to another. Moreover the fit of the realized grow path on the ideal grow path is worse. In the paper we focus on the second way of production, using climate zones with transportation of products between the zones. Next we will describe a generic model of a greenhouse factory where the transportation of products between climate zones is automated using mobile robots. The focus in this paper lies on the modeling. First the modeling technique used is explained, then the model is presented in pseudo code and finally a series of experiments is shown.

MODELING APPROACH

We use the process interaction method for modeling, (Zeigler 2000). The modeling approach can be summarized in three steps:

- Decompose the system to be studied into relevant classes of elements or functions, preferably patterned on the real-world elements of the system.
- Identify the properties and methods of each class. A class is characterized by its properties and methods. An instance of a class is called an element.
- Distinguish the "living" element classes and provide their process descriptions. A process governs the dynamic behavior of each element. A process is a method and has special features with respect to so called 'time consuming commands' that may occur in its body

Advantages of process interaction modeling are that it allows communication with non simulation experts and it offers a natural opportunity to model agent based systems because active elements are autonomous and capable of interacting.
We apply a simple vocabulary expressing the typical commands needed for process interaction modeling as well as a simple pseudo code (Ottjes and Veeke , 2002). All commands of our vocabulary will be written in *italic* style. We will use the common object oriented 'dot' notation for qualifying methods and properties
The clock time in the modeled system is called *now.*

A distinction can be made between time_consuming commands that govern the time consumption in a process and interaction commands that affect a state change of elements.

Time consuming commands

We have three forms of time-consuming commands to be used in a process description of an element class.

– *advance* (t) where t is a time interval generating a scheduled element state lasting that interval. t is determined by the element itself.

- *advance* (*while* condition) where the element state becomes scheduled as long as the condition applies. The condition is determined by the element itself.
- *advance* without any argument. Here the element state becomes suspended. A command (*resume*) originating from the process of another element is needed to continue the process.
- *finish* has the same effect as *advance* but skips to the end of the process.

Aliases for *advance* are used to increase the readability of the model. Here we use *wait* and *drive.*

Interaction commands

Interaction commands can be given by some element to influence the process of another element. We distinguish *interrupt* causing the targeted element to change from scheduled into suspended state and *resume* causing the targeted element to continue its process.

Use of sets

A set is used as a collection of elements. A set may represent a physical waiting queue but is also used to establish a 'one to many' relationship. The latter concept is often used control functions. An element may reside in more than one set at a time but not twice in the same set. Both an element and a set own methods for set operations. An element itself may *enter* and *leave* a set and a set may *add* and *remove* an element to/from its content.

Each element in a set knows its *predecessor* and its *successor* and each set owns the up to date values of its *mean length, maximum length*, *mean waiting time* and *maximum waiting time.* If there is no *predecessor* or *successor* the value *none* is assigned.

Other Set related commands are:
copy (Set1,Set2) All elements of Set1 are copied into Set2
move (Set1,Set2) All elements of Set1 are copied into Set2 and removed from Set1
empty (Set1) All elements in Set1 are removed from Set1

In order to look up elements with special characteristics we use compound commands like:
anElement = *first element* in aSet (*with* condition)
Here aSet is an arbitrary Set

The *loop* command is used for repeating actions.. There are three ways to use *loop*:
- *loop*: repeat indefinitely
- *loop* (n): repeat n times
- *loop* (*while* condition): repeat while condition

If a command acts on a command block, the commands in that block should be indented

All identifiers referring to a set will contain the word 'Set'. If it is necessary to add any explanation to a command this is preceded by //
All method names will be written in capitals.

THE MODEL

Now we will apply the process interaction modeling method on the greenhouse system

We distinguish three sub systems:
- The products
- The infrastructure with climate zones
- The mobile robots

Applying the modeling recipe we first decide on the element classes and relevant properties and methods.

The products

A product (plant or flower) is defined by its name and its set of production steps. Each production step owns a reference to the climate it has to be grown in and the duration of this step: the GrowTime.
The basic product carrier is a rack. A rack is a plate, carrying a number of products of the same product type. A rack is the standard transportation unit.
A lot is the production unit. It consists of one or more racks. A lot contains aggregated individual client orders but is produced as one single production order. After production completion a lot has to be split up into the originating client orders and distributed to the final customer. We restrict ourselves to the lot production part of the production process.

In our model a lot acts as an 'agent' taking care of timely presenting itself for transportation. Consequently a lot, in process interaction terms, is an active element and owns a process. This process will be explained later in Table 2. The product related element classes are defined as follows:

Product
- Name
- ProductionStepSet //contains all specific steps

ProductionStep
- ClimateZone
- GrowTime

Rack
- MyLot
- CurrentPosition // position in climate zone
- NewPosition

Lot
- Product
- RackSet // contains all racks of the Lot
- RacksLeftToMoveSet
- CurrentClimateZone
- CurrentStep
- NextClimateZone
- NextStep
- PROCESS

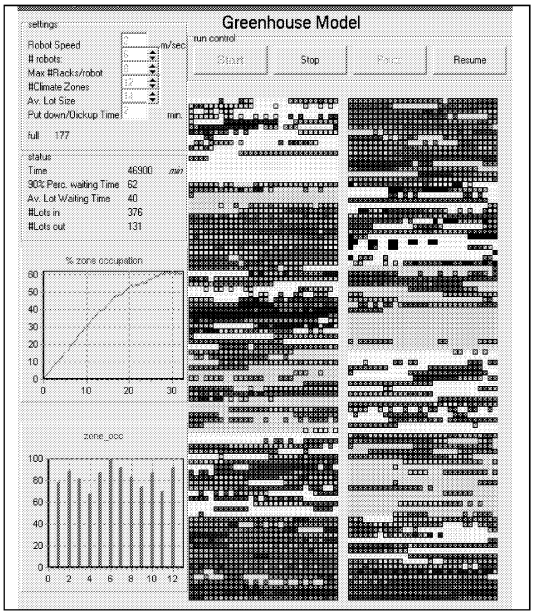


Figure 1. Snapshot of the simulation screen. In the left part the overall occupation of all zones is shown as a function of time and the maximum occupation for each zone.

The infrastructure with climate zones

In Figure 1 a snapshot of the model screen shows an arbitrary configuration of 2 x 6 climate zones, separated by a middle path.

In the greenhouse model a climate zone has its own typical climate and its set of rack positions. Each rack position has a reference to its Rack and knows its own location. Each climate zone has a DestinationSet with Destination elements each representing one of the other climate zones. The idea is that all racks to be moved from some zone A to zone B are sorted in the RacksToMoveSet of the Destination representing zone B.
The classes and Sets in this subsystem are:

```
ZonesToServeSet   //contains the zones with transport
                  //demand sorted according criteria
ClimateZone
- Climate
- RackPositionSet
- FreeRackPositionSet
- DestinationSet
- Location        // coordinates of entrance
Destination
- Zone
- RacksToMoveSet
RackPosition
- ClimateZone
- Location        // coordinates
- MyRack          // reference to a Rack or None
```

The Mobile Robot system

Figure 2a shows the principle of the mobile robot and figure 2b a real one. A robot can one by one load racks up to a maximum number and after transportation put them down one by one according the rule 'first in last out'. Robots take care of the transportation of lots between and inside climate zones. The task of the robots is to assure that all racks arrive at the right time in the right climate zone with a minimum of transport delay. In the configuration of Figure 1 the middle path is used by the robots to drive from one zone to another. Inside a zone a robot drives over the racks both in x and y direction. It is necessary that there is enough room between the racks for the wheels of the robot. A robot can load racks one by one up to a maximum number of MaxRacks and transport these racks to another zone. The class robot owns a Speed and a RackSet in which all racks that physically are picked up by the robot (are onboard) are listed. All control decisions are taken locally and are explained in the robot process description given in Table 1

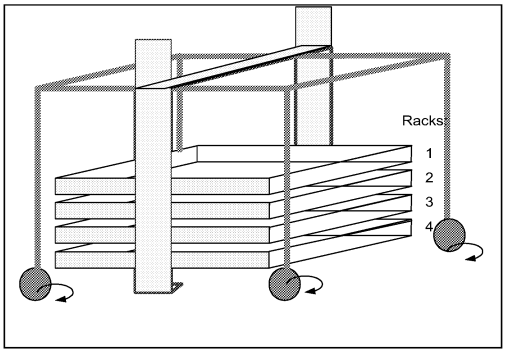


Figure 2a. Sketch of the automated (un)loading mobile robot.

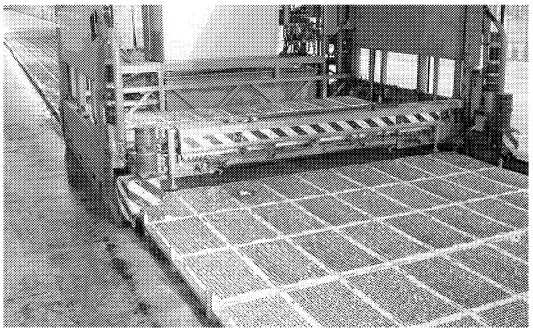


Figure 2b. A real (un)loading mobile robot, driving over a lane with racks containing trays with plants. (Florensis 2005)

The classes and sets in the robot subsystem are:

```
RobotIdleSet  //this set contains idle robots.

Robot
- Speed
- MyLot
- MyRack
- RacksToPickSet // racks to be loaded
```

- RackSet // loaded racks ´on board´
- MyLocation
- MyZone
- MaxRacks // maximum #racks possible
- PutDownTime
- PickUpTime
- SELECTDESTINATION
- PROCESS

The Processes

The robots act like agents, autonomously following their process description. The processes of the robot and lot classes are interrelated. The pseudo code of the robot process is given in Table 1 and the code of the lot in Table 2. Lines preceded by '//' contain additional comment. A short explanation will be given of both processes.

The Robot stays idle as long as there are no racks to be transported (ZonesToServeSet is empty). If there is work to do the robot selects a new destination (=climate zone) to go to. The method SELECTDESTINATION is used for that. A simple implementation of that method could be to just take the first climate zone in the ZonesToServeSet. If the destination is known the robot drives to the entrance of that zone and reserves the racks to be picked.

Table 1. Process of the Robot class

```
Robot PROCESS
Loop // this process is repeated infinitely
 enter(RobotIdleSet)
 wait while ZonesToServeSet Empty   //wait for work
 leave(RobotIdleSet)
 //Determine and drive to the right zone
 MyDestination =  SELECTDESTINATION
 MyZone = MyDestination.Zone
 Drive (Time to MyZone.Location)
 n=MAX(MyDestination.RacksToMoveSet.Length,
   MaxRacks)
 Move first n Racks from MyDestination. RacksToMoveSet
   into RacksToPickSet
 //Collect the racks to be transported
 Loop while(RacksToPickSet.NotEmpty)
   MyRack = RacksToPickSet.First
   MyRack.Leave(RacksToPickSet)
   Drive(Time to MyRack.Currentposition.Location)
   Wait(PickUpTime)
   MyRack.Currentposition.Enter(
     MyZone.FreeRackLocationSet)
   MyRack.Enter(Rackset)
 //Drive to next Climate Zone and distribute
 MyZone = MyDestionation.MyZone
 Drive (Time to MyZone.Location)
 Loop while(RackSet Not Empty
   MyRack=RackSet.Last
   MyRack.NewPosition = MyZone.FreeLocationSet.First
   MyRack.NewPosition.Leave(MyZone.FreeLocationSet)
   Drive (time to MyRack.NewPosition.Location)
   MyRack.leave(RackSet)
   Wait(Put_down_time)
   MyRack.Leave(MyRack.MyLot.RacksLeftToMoveSet)
   If MyRack.MyLot.RacksLeftToMoveSet Empty
      MyRack.MyLot.Resume
```

After that the robot loads these racks in a sequence of driving and picking and then drives to the entrance of the destination zone. After arrival it distributes the loaded racks in the zone by occupying free locations from the FreeLocationSet of the zone. These free locations are sorted in a proper way, for example fitting together. After each put down action the robot tests whether or not a lot-transportation-job is completed. If that is the case the lot is resumed and will continue its own process

The lot process is described in Table 2. After having determined the next climate zone to go to, the lot identifies the right destination in the destination set of its current climate zone. Then it makes its racks available for transportation by copying them into the right RacksToMoveSet according a sorting criterion. Sorting criteria are free to be chosen, for example according the physical location of the racks in the zone in order to minimize robot pick-up times. The lot further copies all racks into its own RacksToMoveSet, enabling the robots to keep track of the progress of transportation of the lot. Finally it sorts its CurrentClimateZone into ZonesToServeSet thus indicating that there is work to do in that zone. After having done so the lot *waits:* It becomes *suspended.* Only after all its racks are transported to the proper zone, the lot process is resumed by the robot that has placed the last rack. The delay that is caused in that way is the key performance indicator of the model. After having been resumed the lot waits during the grow time of its next step and after that repeats its process until all steps are done.

Table 2. Process of the Lot class

```
Lot PROCESS
NextStep =  Product.ProductionStepSet.First
Loop While (NextStep Not None)
   NextClimateZone = NextStep.ClimateZone
   NewDestination = First Destination in
     CurrentClimateZone.DestinationSet with
     DestinationZone = NextClimateZone
   Copy (RackSet, NewDestination.RacksToMoveSet)
   Copy (RackSet, RacksLeftToMoveSet)
   (re)sort CurrentClimateZone in ZonesToServeS
   Wait  // wait until the all racks are transported
   Wait NextStep.GrowTime
   CurrentClimateZone= NextClimateZone
   NextStep = NextStep.Successor
   (Product.ProductionStepSet)
Finish
```

The assignment of robots to climate zones is controlled via the ZonesToServeSet. The zones that need robots are sorted according criteria that are also free to choose. Examples of criteria are the distances to the robots, the total number of racks to be moved from or to the zone or the cumulative delay of the waiting lots or combinations of these criteria. In this case a robot only needs to choose the first zone in the ZonesToServeSet.

Apart from the mobile robot class and the lot class we need an element to take care of the generation of lots to be produced by the system: The lot generator. This generator may use historical input, for example all lots produced in the past year, or just generate input with the help of statistical distributions. If the model is detailed to a certain extend it would be possible to feed the model with real time data and run it real time, for example to use it as a shadow system for planning and control.
The lot generator will not be worked out further here.

Using the model in practice may demand that some additional functionality have to be added.
- Equipment failures
- Using other transport equipment, such as overhead cranes
- Reshuffling of free rack positions, comparable with the compression function for a computer hard disk
- Flexible input generator
- A distributed implementation of the model (Duinkerken et al. 2002)

The model is implemented in the simulation package Tomas (Veeke and Ottjes, 2000 and 2002)

EXPERIMENTS

In this section some test results are shown as an example how to use the model for design and evaluation purpose. The main parameter settings are listed in Table 3. The lot-stream input used is characterized in Figure 3 and Figure 4, showing the distribution of total growing times and the distribution of the lot sizes in numbers of racks respectively.

Table 3. main run parameters	
Run length	90 days
Mobile robot speed	0.5 m/s
#Mobile robots	4
Pick-up & put down times	2 min.
Max. # racks per robot	8
# climate zones	12
Dimensions of climate zones	30 x 16 racks
Rack dimensions	2 x 2 m. incl paths
#product types	1000
#steps per product type	Uniform[3,4,5]
#Racks per lot	Exp[8] with max. 24
Input location	31,1
Output location	32, 96

The green house configuration applied, is according Figure 1: there are 12 climate zones, each having 16 x 30 = 480 rack positions. The rack dimension is taken 2 x 2 m. including the robot tracks. The robot speed was set to 0.5 m/sec for all runs. The first series of experiments concern the number of robots needed. Figure 5 shows the results. The transport delay is the performance criterion. We conclude that 5 robots are sufficient. Next the influence of the maximum number of racks per robot is investigated. From Figure 6 we conclude that the lot delay tends to dramatically increase below 3 racks per robot. Finally the role of pick up and put down times of racks is studied. The results are put in Figure 7. We conclude that in this configuration the rack handling times have a very strong influence on systems performance. In practice numerous experiments varying much more variables are necessary.

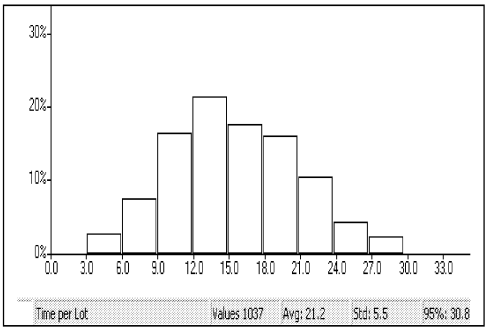


Figure 3. Distribution of total grow times per lot (days)

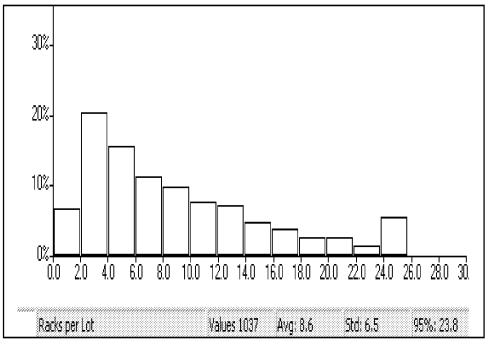


Figure 4. Distribution of the number of racks per lot

Next to the data incorporated in the graphs, all kinds of other information is available in the model such as robot travel distances, robot trip distances, occupation rates, rack filling rates, climate zone occupation etc.

The question arises if it would be possible to reuse a control system, developed and tuned with simulation, to a real system. To answer that question we should realize the practical consequences of the approach. The key issues here are to keep the necessary information up-to-date and to make the control distributed. The information part relates to all data about the lots and the racks. The type of information needed can be derived from the properties of the element classes and from the processes. For example the free rack position status should be updated every time a rack is handled and each robot should have access to a correct set of destination zones. Technologies that are very promising

in that respect are radio frequency identification (Ni L.M et al 2004 and Stefansson G. and Tilanus B, 2000) and multi agent control (Bussmann et al., 2004)

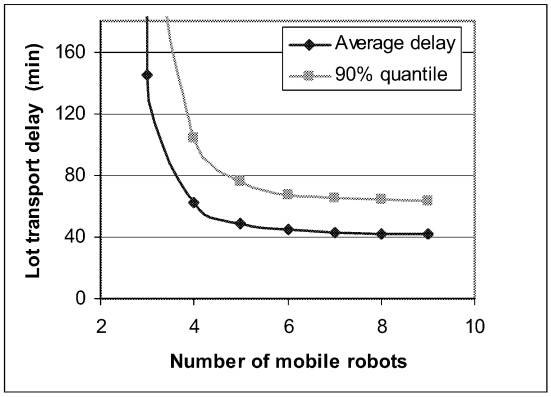


Figure 5. Transport delay as a function of the number of mobile robots

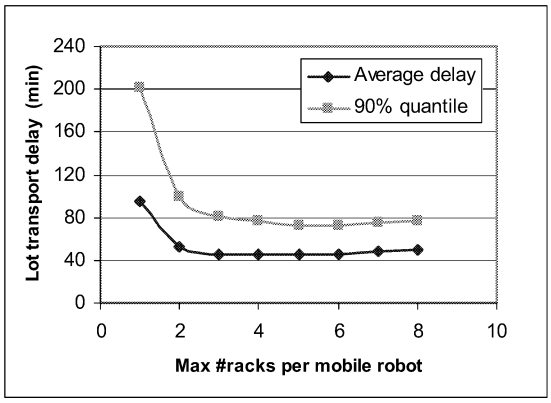


Figure 6. Transport delay as a function of the rack load capacity of a robot with 5 robots.

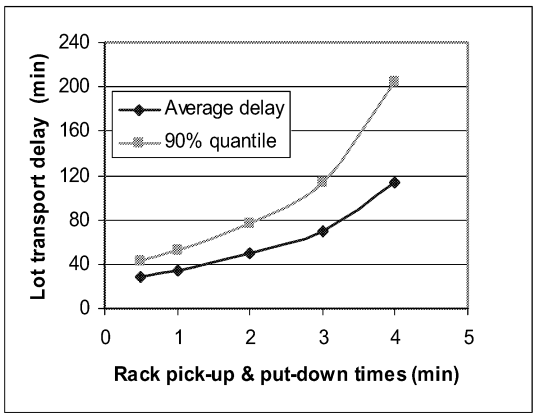


Figure 7. Transport delay as a function pick-up and put-down time per the rack with 5 robots

CONCLUSIONS

A model is developed for the design and evaluation of facilities for mass production of half products for the horticulture industry.. The model focuses on a production method in which production lots have to be transported frequently between different locations having different temperature and humidity settings. The focus was put on the modeling. The model is described in pseudo code and experiments have been carried out to illustrate the use of the model. Future model extensions will relate to the concept evaluation and analysis of other rack handling systems in the greenhouse such as overhead handling cranes and the analysis of climate zone arrangements and interchange buffers.

REFERENCES

Bussmann S, N.R. Jennings and M. Wooldridge,2004. "*Multiagent Systems for Manufacturing Control*", Sringer-Verlag Berlin Heidelberg 2004

Duinkerken, M.B, J.A. Ottjes, G. Lodewijks, 2002; *The Application of Distributed Simulation in TOMAS: Redesigning a Complex Transportation Model.* Proceedings of the 2002 Winter Simulation Conference (WSC 2002). December 2002. San Diego. ISBN 0-7803-7615-3

F+H , 2004 Jungpflanzen automatisch auf Tour. (Fordern und Heben) 54 (2004) Nr5. pp 280-281.

Florensis 2005. http://www.florensis.nl/

Fishman, G.S. 2001. *Discrete Event Simulation. Modeling, Programming, and Analysis.* @001 Springer-Verlag New York, Inc. ISBN: 0-387-95160-1, 52-59

Ottjes J. A. , Veeke, H. P.M., , 2002 "*Prototyping In Process Oriented Modeling And Simulation*", Proceedings 16th European Simulation Multiconference ESM 2002, pp 20-26, ISBN 90-77039-07-04, Darmstadt, Germany

Ni L.M., Liu Y, Lau Y.C. and Zhang X, 2004. LANDMARC: Indoor Location sensing Using Active RFID. Wireless Networks, Vol 10 Issue 6, Nov 2004, pp 701-710

Stefansson G. and Tilanus B, 2000. *Tracking and tracing:principles and practice*. Int. J. Technology Management, Vol 20, Nos ¾, 2000

Veeke, Hans P.M., Jaap A. Ottjes, 2000. Tomas: Tool for Object-oriented Modeling And Simulation. *In proceedings of Advanced Simulation Technology Conference (ASTC2000).* April 16-20, 2000, Washington, D.C. pp. 76-81. The Society for Computer Simulation International (SCS), ISBN: 1-56555-199-0

Veeke, Hans P.M., Jaap A. Ottjes, 2002. TomasWeb: web site: www.tomasweb.com

Zeigler B.P., Praehofer H and Kim T.G., 2000. "*Theory of Modeling and Simulation* 2nd Ed. Academic Press, San Diego.

PERVASIVE ITEMS FOR FACTORY SECURITY

Alessandro Genco, Gaspare Scherma, Salvatore Sorce
Dipartimento di Ingegneria Informatica
Università degli studi di Palermo
Viale delle Scienze, edificio 6, 90128 Palermo- Italy
genco@unipa.it, gaspare.scherma@libero.it, sorce@unipa.it

KEYWORDS
Pervasive Systems, Factory Management, Security, Multi-agent Systems, Augmented Reality

ABSTRACT

Pervasive systems need security items to be implemented for wide and complex environment control. This paper discusses a pervasive security system which works according to a model for hybrid entity exhibition and interaction in augmented reality. The highlights of that model are first discussed. Then, a security subsystem is proposed which was designed to be managed by entity cooperation. Environment safety control and worker monitoring are the main goals of the system which is being tested by means of a simulation tool finally discussed.

INTRODUCTION

Pervasive systems use a variety of smart devices to be spread on real-life environments. They play a role of interaction media in providing people with context-aware useful services which take into account context factors such as: who is where, why, and when. For instance, a pervasive system which is designed to run in a factory, has to provide workmen with services compliant to their role, skill, task, and current position. A pervasive system should not be a hindrance for workers, and it should be perceived right at the time someone needs to interact with it.
Interaction between humans and Augmented Reality (AR) environments need to be safe both as for people and system. Actually, there are many security issues which need to be taken into account, including communication data protection against malicious intrusions as well as environmental parameters to be monitored. Here we deal with safety aspects involving workers and objects in a plant. Besides service provision tasks, a pervasive system is therefore charged to perform security controls by processing parameters and values, thus pointing out occurrences which are likely to be dangerous either for system, workers, and environment. To this end, the security system has to collect and manage a lot of information, including worker current position and task, and sensor data, as for instance room temperature.

THE PERVASIVE APPROACH

Pervasive systems are the operating core of the Ubiquitous Computing vision (Weiser 1991; Abowd and Mynatt 2000), with the end of enabling people to be provided with any information service from anywhere at any time. A current trend of these systems is leading us to a new life style in an information society which will be free from aware interactions with computers. These will be hidden in the environment (Wont et al. 2002) and we will use PC's only for daily office tasks, to write a letter or to make computations. Our reality will become augmented and we will keep on perceiving it like the old natural one (Streitz 2001; Starner et al. 1997).
In such a kind of information system we need to solve some security problems which are more complex than in typical distributed systems (Kindberg and Fox 2002; Al-Muhtadi et al. 2002). AR is populated by hybrid entities which wrap real and virtual resources, and interact according to a distributed paradigm (Genco 2004a). Security therefore needs to be brought up to the hybrid entity level and thus enabled to check hybrid entities and their peer to peer communication.
An object in AR is not only our sensorial perception of it; an object can become what we want or we need to perceive of it, including semantic contents and actions not naturally belonging to that object. A model for hybrid entity exhibition in AR is proposed in (Genco 2004a). In more detail, the so-called HAREM model deals with AR entity exposition from an ontological point of view.
We implement HAREM entities as FIPA agents with some application specific operating roles in addiction to the basic roles which were thought for any kind of entity in AR.
In particular, *consistency, mission*, *vocation, and ability* roles are devoted to manage entity operating structure.
The *survival* role performs safety reactions to dangerous requests, and also manages fault tolerance strategies; the survival role implements security for the entity it belongs to, and eventually interacts with a security entity which might be running in a given application environment.
In this paper we use that model to describe an entity-based pervasive security system.

PERVASIVE SECURITY IN FACTORIES

The pervasive system main task is to detect dangerous occurrences, and to warn factory security staff consequently. To this end, our system works in AR environments by deploying a set of hybrid entities which cooperate and exchange ontology data, so as to keep workers and plants safe.
All involved entities are designed according to the HAREM model. Security issues are mainly processed by the roles defined in the *EnvironmentCheck* entity, which includes the main logic for site safety management. The *EnvironmentCheck* entity aims at inferring dangerous or situations in a room, and performs some checks: *proximity detection, crowding detection, temperature check, noise check.*
Positions of workers within factory are detected and managed by a *Site Positioning System (SPS)* entity, which is based on two cooperating technologies: Bluetooth for coarse positioning, and image recognition for worker compass detection (Genco 2004b).

The *EnvironmentCheck* entity receives position data from the *SPS* entity, and use them to perform all checks on worker position. Among these, *proximity detection* alerts the system about a potential dangerous position of a worker who is standing too close to a tool range of action. *Crowding detection* aims at pointing out unexpected crowds of workers which were classified as not belonging to any registered workgroup, or belonging to different workgroups. This can be the case of a worker in danger or injured and other workers come in help.
Position data can be also compared with path routing information which can be inferred from workers tasks as described in their profile ontology. This kind of investigation can be useful in pointing out some unexpected worker behavior. This task is accomplished by the *Investigate* entity, which is devoted to control the coherence of paths taken by workers.
Temperature and noise values are also processed to infer and prevent critical situations.
All events in a specific room are managed by the *Room* entity.

Abilities and Knowledge

The *EnvironmentCheck* entity, as well as other involved entities in the security system, owns a minimal knowledge base which enables it to detect critical situations occurring in the site. Fig. 1 gives an overview of the *vocation* role in the *EnvironmentCheck* entity. As it can be noticed, there is a link to the *ability* role which is devoted to resources lookup when these are requested by other roles.

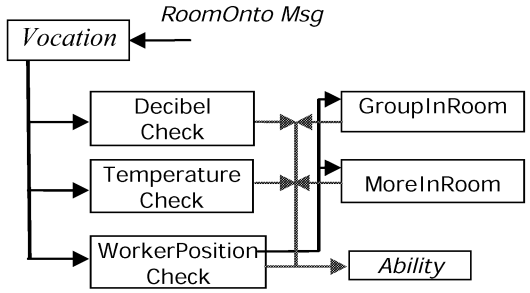


Figure 1: *Vocation* Role of *EnvironmentCheck* entity

Room data are checked by the *TemperatureCheck*, *DecibelCheck*, and *WorkerPositionCheck* roles, in order to detect and signal dangerous situations.
An *Ability* role is implemented in each entity and allows the system to conveniently share hardware and software resources in the environment, as for instance, multimedia devices, common methods, libraries and so on.
Dangerous occurrences are detected by the *EnvironmentCheck* entity by processing data supplied by the *Room* entity, which receives positioning data from the *SPS* entity (fig. 2).

Ontology data

Each entity in the system behaves according to a specific ontology. Ontologies can be extremely simple or extremely complex depending on entity type. In some cases, as for instance the *WorkerPositionCheck* role, ontology is reduced to a simple data structure. Opposite the case of the *UserProfile* ontology, which can be very complex so as to conveniently describe worker tasks, skill, age, behaviors and so on.

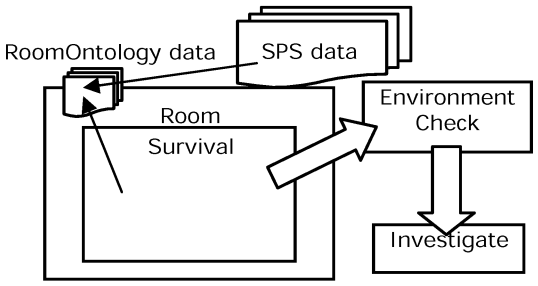


Figure 2: Ontology Processing

A context depending room ontology section (Fig. 3) is periodically supplied by the *Room* entity to the *EnvironmentCheck* entity. This includes all data of workers who are currently in a room and in the nearest ones, along with temperature and noise values.

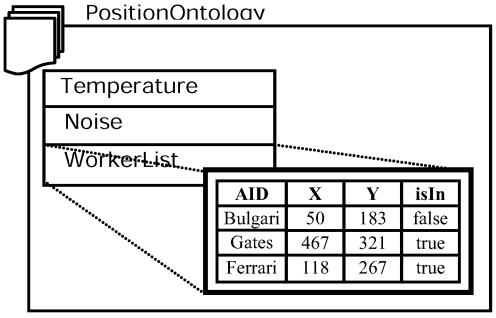


Figure 3: Room Ontology

Entities and Roles Details

The pervasive security system is designed to make three Harem entities cooperate in a wider pervasive system for workers service provision in a factory. Here we give some further details about the security entities and their cooperation logic.
A *Room* entity instance is created for each room or area of the factory. It uses its *Ability* role for resource lookup and provides other entities with information about the room status. In particular, the *EnvironmentCheck* entity queries the *Room* entity for workers data second by second.
The *EnvironmentCheck* entity is created by the *Survival* role in the Room entity (fig. 2), and it can be considered the system core component.
The entity creation mechanism is something featuring the pervasive behavior of HAREM entities. Generally speaking, a new entity can be created when an entity cannot directly cooperate with another entity. The new entity is provided with the methods and resources requested for cooperation. The *EnvironmentCheck* entity can be created by the system administrator or it can be created by other entities which need security control to be externally performed. In our case, the *EnvironmentCheck* entity comes into the system on *Room* entity demand, after a resource gathering process which takes into account the security needs of a room. Once created, an *EnvironmentCheck* entity is charged with room security control by continuous interaction with the *Survival* role of the *Room* entity.
Security check starts from a request for *PositionOntology* which is made by the *Room* entity to the *SPS* entity. Once the request is received, *Vocation* role forwards incoming data to the *DecibelCheck*, the *TemperatureCheck* and the *WorkerPositionCheck* sub-roles, which extract selected data

from the ontology. Results are then forwarded to the *MoreInRoom* sub-role in order to check congestion, and to the *GroupInRoom* sub-role in order to detect any group inside the room.
Finally, the *EnvironmentCheck* entity queries the *Investigate* entity to check the coherence of the room status, by means of workers profile and task analysis (fig. 2).
The *Investigate* entity was implemented to collect information about workers and their current task related to their current position, in order to infer and notify any potential dangerous situation. This entity is called by the *EnvironmentCheck* entity when all information are collected from the *Room* entity ontology.

SYSTEM SIMULATION TOOL

The simulation tool we now discuss was implemented to test the system accuracy when detecting dangerous situations. Furthermore, the tool can be used to set up or test safety plans which have to be taken in case of an emergency within the factory. In more detail, the tool allows us to sketch the environment and to set some parameters (e.g. temperature, noise, number of workers in a room). It returns a visual report of the detected situation along with information for the security staff.
The simulation tool, as well as the whole project, was implemented by means of the FIPA compliant and Java-based JADE agent development platform (Bellifemine et al. 2004), and most of methods the tool includes are the same of the actual security system. In particular, entities of our system was implemented as JADE mobile agents, whose behaviors implement HAREM roles.
Fig. 4 shows the graphical user interface we implemented for simulation. The main component of the window is an interactive area where a factory map is sketched. This area shows the real-time workers simulated position, thus allowing us to check the system reaction correspondingly to a given workers layout. Furthermore, there are some visual tool (buttons, radio-buttons, check-boxes, text areas) which allow users of the tool to set temperature and noise values, to insert a new worker in the work area, to delete all workers, or to check the status of each worker.

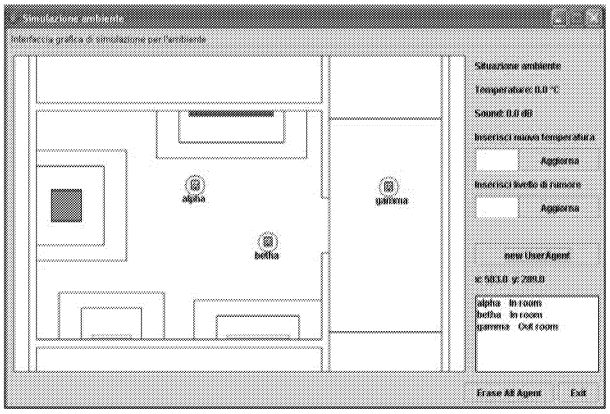


Figure 4: Simulation Tool Window

When a worker is created, an icon shows his position on the map. Information about the worker can be visualized by clicking on the icon, such as name, coordinates, and time elapsed since his entrance in the area.

The simulation tool uses the *WorkerPositionCheck* role which is performed by the *EnvironmentCheck* entity. As above mentioned, this role is the most important in processing workers position inside a room, both for simulation and for actual system running.
Position data are then processed and results are shown by means of a string whose content indicates whether the worker is in an area he is not supposed to stay in. The supposed factory security staff can be provided with some useful information on how to operate in an emergency occurrence.

CONCLUSIONS

An entity-based pervasive security system has been discussed for factory safety monitoring which was arranged according to the HAREM model for hybrid entity exhibition in AR. A simulation tool has been also presented which allow us to tune and evaluate system reactions to occurrences which are perceived by the system as likely to be dangerous. This feature can be used in the actual system to directly control environment parameters, without human intervention, for instance by means of modern Java-enabled appliances., which can execute remote commands or host small Java classes (like Jade agents).

ACKNOWLEDGEMENTS

This work has been partly supported by the CNR (Council for National Research of Italy) and partly by MIUR (Ministry of the Instruction, University and Research of Italy).

REFERENCES

Abowd, G. and E. Mynatt. 2000. “Charting past, present and future research in UC”, ACM Transactions on Computer-Human Interaction, Vol. 7, No. 1, March 2000

Al-Muhtadi, J.; Ranganathan, A.; Campbell, R.; Mickunas, M.D. 2002. “A flexible, privacy-preserving authentication framework for ubiquitous computing environments”, proc. of the 22nd Int. Conf. on Distributed Computing Systems Workshop, 2-5 July 2002 pp: 771 – 776 (IEEE)

Bellifemine F., Caire G., Poggi A., Rimassa G. 2004. “JADE: a White Paper”, EXP, Vol. 3, n. 3, http://exp. telecomitalialab.com

FIPA, Foundation for Intelligent Physical Agents, http://www.fipa.org, Specifications # SC00001L, SC00037J, SC00023J, SC00061G, SC00029H, SC00026H *Abstract Architecture Specification*, Specification SC00037J, 00086D,

Genco, A. 2004. “Augmented Reality for Factory Monitoring”, proc. of Int. Conf. ISC’2004 – Industrial Simulation Conference, June 7-9, 2004, Malaga, Spain

Genco, A. 2004. “HAREM: The Hybrid Augmented Reality Exhibition Model”, WSEAS Trans. On Computers, Issue 1, Vol. 3, Jan. 2004, ISSN 1109-2750

Kindberg, T., A. Fox. 2002. “System software for UC”, IEEE Pervasive Computing, vol. 1, issue 1, Jan–Mar 2002, pp.70-81

Streitz, N. 2001. “The role of UC and the disappearing computer for Computer-Supported Cooperative Work”, proc. of 7th Int. Workshop on Groupware, 6-8 Sept. 2001, pp: 2 - 3 (IEEE)

Starner, T., Mann, S., Rhodes, B., Levine, J., Healey, J., Kirsch, D., Picard, R.W., Pentland, A. 1997. “Augmented Reality Through Wearable Computing”, Presence, Vol. 6, No. 4, August 1997, pp. 386-398

Weiser, M. 1991. “The Computer for the 21st Century”, Scientific American, Vol. 265, No. 3, pp: 94 - 104, September 1991

Wont, R., T. Pering, G. Borriello, K. Farkas. 2002. “Disappearing Hardware”, IEEE Pervasive Computing, Jan.- Mar. 2002

AutoTrans: Management and transport automation in warehouses*

Leopoldo Armesto
Dept. of Systems Engineering and Control,
Technical University of Valencia
Camino de Vera, s/n 46022, Valencia, Spain
leoaran@isa.upv.es@isa.upv.es

Josep Tornero
Dept. of Systems Engineering and Control,
Technical University of Valencia
Camino de Vera, s/n 46022, Valencia, Spain
jtornero@isa.upv.es@isa.upv.es

***Abstract*— In this paper, an integrated solution for management and transport automation is presented. The proposed solution considers different applications: an optimal location assignment application, taken information form an ERP and determining products allocation in dynamic warehouses; transport reports generation on electronic devices such as PDAs; a SCADA application, with remote supervision and control over the whole warehouse; a vehicle navigation system, which controls AGVs moving products from one location to another; a teleoperation applications to remotely control the vehicle from an industrial cabin; and a vision-based line tracking application to automatically guide an industrial vehicle for tracking a line painted on the floor. These applications are part of the result of AutoTrans, a research project involving the University Team and four enterprises. Some simulation and experimental results are provided at the end of this paper.**

***Index Terms*— Warehouse automation, vehicle navigation, AGV, SCADA, industrial teleoperation.**

I. Introduction

Nowadays, companies need to become more and more competitive. In this context, companies must adapt their structures to market changes, technological innovations, etc. The key for survival is to assume a position close to client requirements, reducing delays on product deliveries and services. However, Integral Logistics continues being a pending challenge for many companies. For effective logistics, we shall organize the merchandize and information flows efficiently (requested orders, services, product reception, product returns, storage management, transportation, etc.).

Most of the small and medium enterprizes use conventional industrial forklifts for the transport of pallets driven by operators. The daily routes of these industrial forklifts are repetitive and therefore drivers make a singsong and boring work. In addition, these routes become a mess due to irregular industrial lay-outs. All these factors lead to a waste of human time which can be cut down by introducing Auto Guided Vehicles (AGVs).

In this paper, we present some of the final results of the research project AutoTrans, which main purpose consists of providing an integrated solution for automated production plans, warehouse management and transport processes using a fleet of industrial forklifts. In this sense, we describe several applications in which the information system of a warehouse, based on an ERP (Enterprize Resources Planning), has been integrated with the management and transport processes, as shown in figure 1. The components are:

- **Optimal location management**: Consist of determining the best location for picking and delivering pallets in warehouses in order to reduce transportation times. It generates movement orders to the vehicles.
- **Electronic Transport Reports**: Movement orders are generated as electronic reports supported by electronic devices such as PDAs, on-board vehicle HMI, external terminals, etc.
- **SCADA**: Consists of an application for the real-time monitoring vehicle positions in warehouses. It also performs data-logging and alarm handling.
- **Vehicle Navigation System**: A navigation system that uses a laser ranger for mobile robot self-localisation, object detection and avoidance particularized for warehouse environments.
- **Teleoperation**: It uses an industrial cabin with an steering wheel and pedals. The application combines real and virtual images. It considers protection and warning areas for the vehicle.
- **Vision-based Line Tracking**: It is an industrial forklift automated guidance for line tracking. The line is painted on the floor, detected with a vision system.

Commercial AGVs such as those mentioned in [1], are typical based on magnet, wire or laser beacon guidance, placed on strategic points of the warehouse. The drawbacks of these solutions are, in our opinion, lack of flexibility and the expensive installation required. Another interesting solution for industrial forklift automation are given in [2]. This application named as ROBOLIFT© uses a vision system to detect artificial landmarks on the floor to guide the vehicle. [3] and [4] solve the problem of loading and unloading maneuvers with pallets, also based on a vision system. These applications are usually based on a standard on-board PC, which may not be robust enough for industrial requirements.

In this paper, a proposal for warehouse management and transport automation is described in detail. This proposal is based on PLCs and industrial communications, where

*This work has been supported by the Spanish Goverment (Ministerio de Ciencia y Tecnologia), research project PDI2000-0362-P4-05.

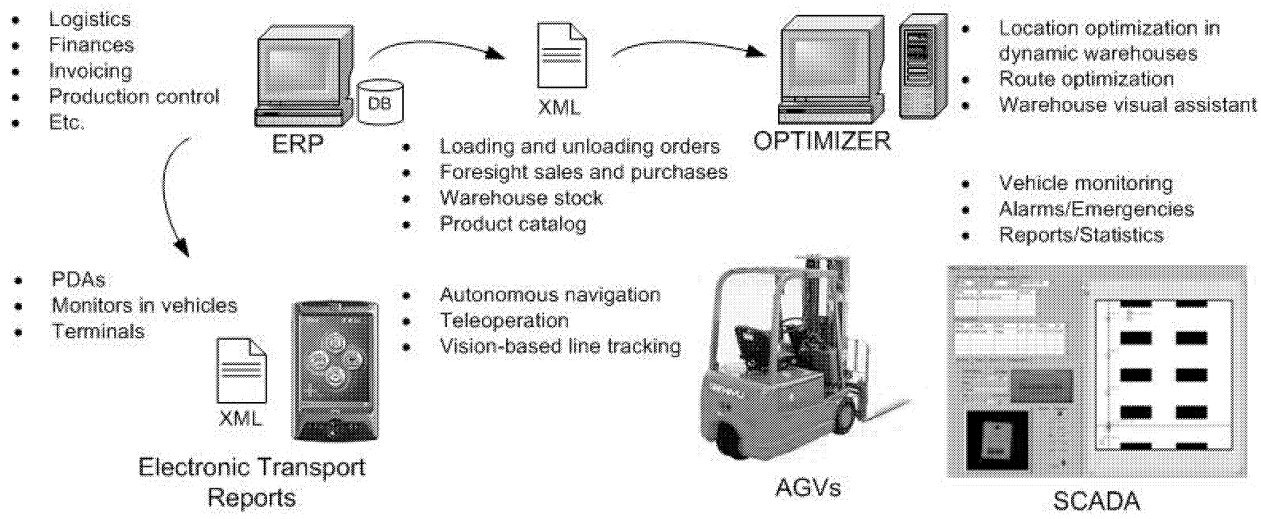


Fig. 1. AutoTrans general overview.

preliminary results of the overall application, including the warehouse management, vehicle navigation system, warehouse supervision and teleoperation have been presented in [5], [6].

II. Warehouse management and Optimal Location Assignment

Information systems such as an ERP (Enterprize Resource Planning) are designed to improve the management of a company, integrating all applications of its departments under the same software. The installation of an ERP is important since it introduces crucial changes on management philosophy.

In the present solution, the software KEWAN [7] has been selected and, in particular, the module related to production management is used. This module has been designed to manage material movements in warehouses, from one location to another, considering among others modules the selling and supply chain, production orders, product catalog, etc.

Commonly, product locations are fixed (static), that is, a zone of the warehouse is commonly assigned to a type of products according to the warehouse layout. However, it is well known that dynamic warehouses, where material locations can change, introduces significant improvements on storage capability and transportation cost reduction. Therefore, the main issue is to determine, in an efficient proposal, where to locate products. Traditionally, the manager performs this allocation based on its own experience without considering the whole possible alternatives. The search of the optimal solution among all possible would consider foresight sales and purchases as well as the present stock.

In this paper, it has been developed an optimal location assignment (OLA) application, based on binary linear-programming (LP) and dynamic programming (DP) techniques. The solution of the problem is a combination among well-known transport and assignment approaches in LP and route map optimization in DP. It can be summarized as follows:

- The system works with palletized units.
- All pallets form inlet docks must be stored to one of existing possible locations (loading orders).
- All material at outlet docks must be delivered from all possible locations (unloading orders).
- Loading orders on already occupied locations, may be still executed if exist any previous unloading order on the same location. Otherwise, the loading orders for these locations are not accepted.
- Unloading orders on void locations may be still executed provided that a previous loading order exists at this location.
- If a location is initially void any loading order on it can be accepted.
- If a location is initially occupied any unloading order related with the same product can be accepted.

At present, the application has been designed for warehouses with one pallet per location, being all locations always accessible. Other type of warehouses, such as those with FIFO and LIFO storage cues will be considered on further versions.

In addition, the application has been integrated with the ERP, providing the necessary information to perform the optimization task. This information includes the present stock in the warehouse, the foresight sales and purchases, and the product catalog.

The data flow between both applications is represented in figure 2. The ERP generates an XML file in "XMLOptimizer" folder, containing all required information sent to the optimizer. OLA application computes the optimal movements to be performed and generates a XML file for each movement containing all required information such as the type of product, origin and destiny locations, date, quantity, vehicle, etc. During execution, each order can be in one of three following states: pending, in process and processed. The file associated to each order is moved form



Fig. 3. Optimal Location Assignment Interface.

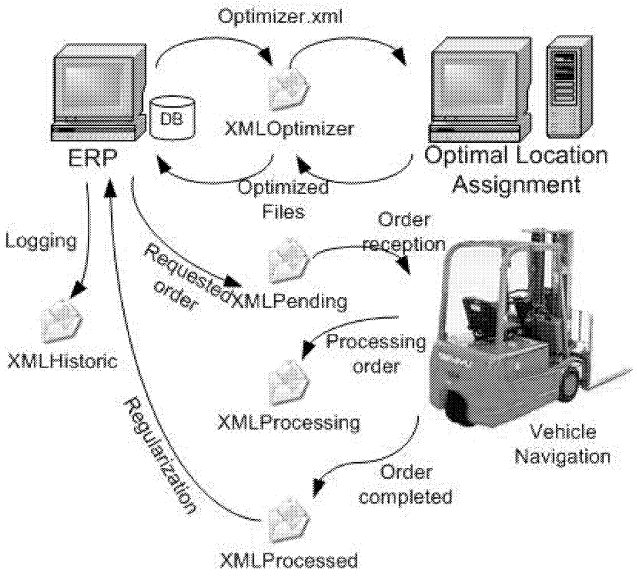


Fig. 2. Data flow between applications through XML files.



Fig. 4. PDA interface reporting movement orders

one folder to another according to its state.

In addition, and for statistic reports a copy of the order file is stored in "XMLHistoric folder at the end of the movement. A simulator has been implemented in order to validate the present application. Figure 3 shows the user-interface for the OLA application, where occupied locations are represented with a small colored square boxes according to the product family code. The simulator depicts in a window the proposed movements of the layout of the warehouse before execution. This will help the driver, on manual operations, the proposed trajectory to be followed.

III. Electronic Report Generation

The ERP data can be accessed form different terminals distributed along the warehouse. Data for each movement is transmitted via XML files containing: product code and description, required quantity of units, origin and destination locations, date, expected stock, driver assigned id. number and comments. This XML file is also used by automated vehicles unmanned for the navigation application. PDAs may improve the communication flow between the warehouse manager and operators, reporting movement orders on a mobile support, as shown in figure 4. Warehouse inventory can automatically be updated by using these electronic devices. In addition, on-board HMI terminals have been installed on vehicles, in order to access

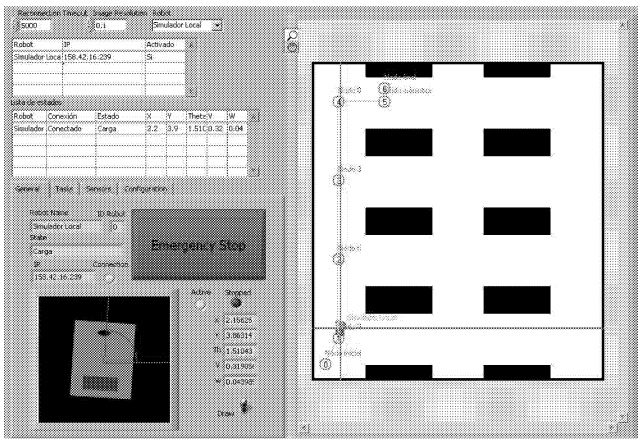


Fig. 5. SCADA Interface.

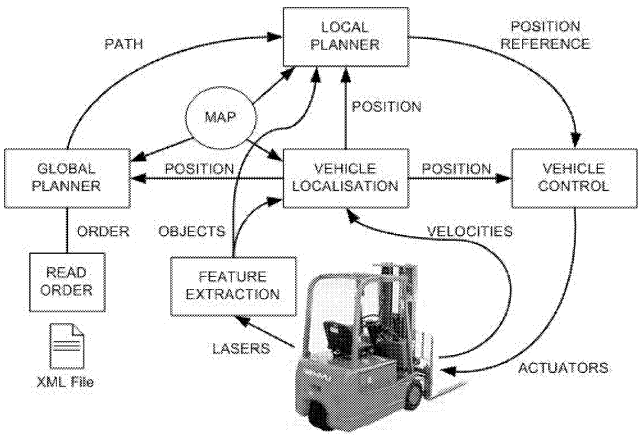


Fig. 6. Control architecture for the vehicle navigation system

to orders if necessary.

IV. SCADA

The SCADA application performs the remote supervision and control of vehicles in a warehouse. This application communicates wireless with each of the vehicles, receiving vehicle positions and velocities, activation flags, etc. From the user point-of-view, the main purpose of the SCADA application is to represent vehicles in the warehouse layout as they move. In addition, it represents some interesting statistics for data-logging purposes as well as connectivity tests, in order to detect communication failures. Warning and protection zones of vehicles as well as emergency stops have been implemented in this application.

Figure 5 shows the SCADA interface created for this application, where the warehouse layout and vehicle position are represented in the right-hand side image, together with the order which is being performed by the vehicle. On the left-hand side of the interface, there is a navigation menu, where the user can appreciate in detail all monitored variables for each vehicle.

V. Industrial Vehicle Automation

Hardware and software architectures for automating industrial vehicles have been previously proposed in [6], [8]. The hardware architecture is based on a PLC from Omron for the low-level vehicle control and a RT-PXI from National Instruments, implementing the navigation system. The software architecture has industrial and real-time characteristics, supporting the RT hardware installed on the vehicle and data communications through a client/server application. It is possible that in the future, the RT-PXI is planned to be replaced by an RT-Embedded PC (Compact RIO), also supported by the software architecture with higher robustness and lower cost but with less computational resources.

In the automated industrial vehicle, the following sensors and actuators have been installed:

- Two incremental encoders to compute wheel speed.
- An absolute encoder to computer the steering wheel angle.
- Two analog outputs for controlling two DC motors at the front wheels.
- Two lasers rangers form SICK providing a 180º scan with 0.5º angular precision. The depth map is sent via RS232 to the on-board PC for robot self-localisation as well as object detection and avoidance. Warning and protection zones are configured on the laser ranger, so that if an object invades this zone, a digital output is activated. These outputs have been connected directly to the PLC for security purposes.

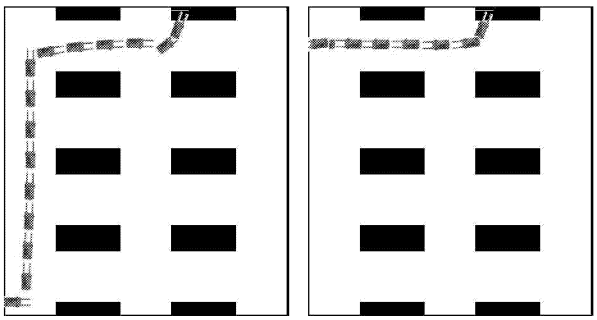
A. Vehicle Navigation System

The navigation system is a complex application where its main goal is to guide vehicles inside the warehouse. In this sense, five different modules have been implemented, as shown in figure 6. That is: global and local planner, vehicle localization and control and feature extraction.

The global planner module receives the transportation order to execute and establishes the shortest route to move the robot from the current position to the picking location and then to the delivery location. This is performed by using a known road map consisting on intersections of corridors (nodes). The local planner module has been implemented using the Artificial Potential Fields (APF) technique [9], considering each node as a partial destination, where the robot is attracted. Objects detected along the path generate an artificial rejection force for collision avoidance. However, in order to reach goals which are quite close to obstacles such as the final destiny located on a shelf, a modified solution for APF is used [10].

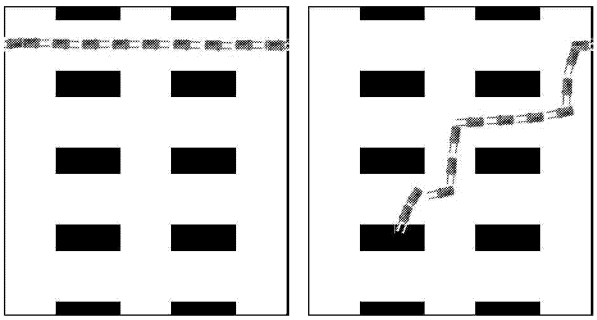
The local planner generates the trajectory points to be tracked by the robot, which has been implemented as a simple kinematic control described in [11], determining wheel reference velocities to apply.

The vehicle localization module is based on Extended Kalman Filter and uses the sensor data to estimate the robot Cartesian positions and orientation, as well as linear and angular velocities. Firstly, a feature extraction module, based on Split & Merge algorithm [12], is used in order



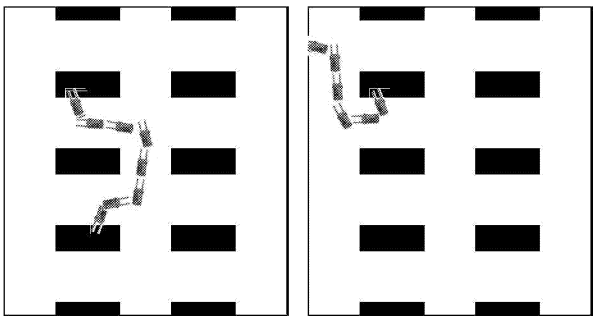
(a) Loading operation (b) Unloading operation

Fig. 7. Movement sequence for order #1



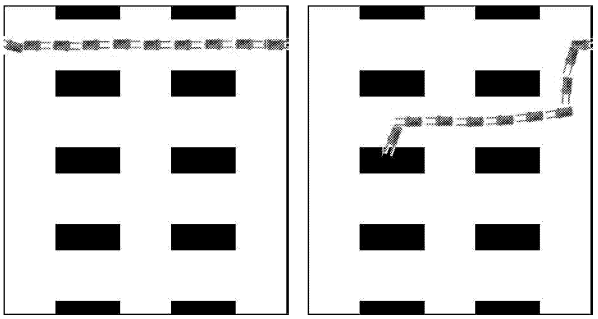
(a) Loading operation (b) Unloading operation

Fig. 8. Movement sequence for order #2



(a) Loading operation (b) Unloading operation

Fig. 9. Movement sequence for order #3



(a) Loading operation (b) Unloading operation

Fig. 10. Movement sequence for order #4

to detect lines representing walls of the environment. Secondly, a data association is performed to determine the correspondence of the map and detected lines [13], [14]. The vehicle uses a known map in order to fuse detected lines by laser ranger and wheels velocities. The Kalman filter used to compute the vehicle position is described in [15], [16], extending standard Kalman filter algorithms (Extended and Unscented Kalman filters) to fuse sensors in a multi-rate sense, that is, at different sampling rates.

B. Simulations and Experiments

The navigation system has been validated through simulation and also with real experimental data. For instance, figures 7(a) to 10(b) show an example of the navigation procedure for a sequence of loading and unloading operations. The loading dock is on the right-hand side of the figures, while the unloading dock is on the left-hand side of the figures.

In addition to this, an experiment has been performed on the parking place at the CPI of the Technical University of Valencia, where our lab is. Figure 11 shows robot trace for this experiment where it firstly moves to Goal 1 from its initial position and later on moves to Goal 2. It is also shows the behavior of the robot when avoiding obstacles which have not previously included in the map, such as persons.

VI. Industrial Vehicle Teleoperation

A teleoperation application has been implemented using industrial vehicles. The hardware architecture shown in figure 12 is based on an industrial vehicle automated with a PLC as described in section V. In addition, for this application, a wireless camera has been installed in order to display an image of the real environment at the teleoperation cabin as shown in figure 13. A DeviceNet (with RF antennas) module (slave) has been installed on the vehicle PLC to communicate data to the teleoperation cabin PLC (master). The teleoperation cabin has properly sensorized in order to transmit steering wheel and speed references to the vehicle and an emergency light indicates any type of problems that may occur during the teleoperation task.

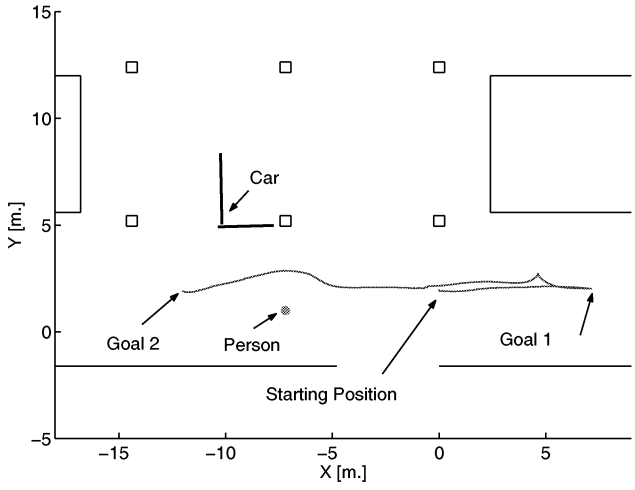


Fig. 11. Experiment of robot navigation in a parking place.

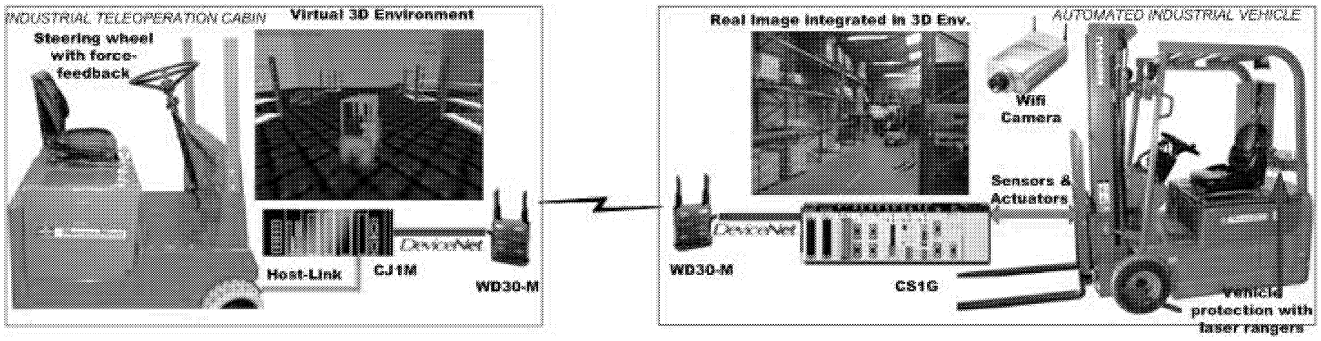


Fig. 12. Hardware architecture for teleoperation

In addition to this, a 3D virtual environment representing the industrial warehouse has been designed in order to integrate the real image with the virtual ones, as shown in figure 14. This application has been designed to monitor the vehicle pose and laser measurements, transmitted from the vehicle to the cabin through DeviceNet and then to the 3D application using host-link commands.

The application automatically reconfigures the represented 3D environment according to the warehouse layout, using a XML file containing the map, which is also used by the vehicle navigation system and the optimizer (OLA). Moreover, warning and protection areas, provided by the laser are also represented in case of intrusion. In addition, the operator can also perform emergency stops by pushing an emergency button installed on the cabin. Due to deterministic properties of DeciveNet network, we can assure that the vehicle will stop with a very short time delay. Network disconnections are also detected by the PLC at the vehicle, causing emergency stops.

Control software architecture for vehicle teleoperation is depicted in figure 15, including: motor control, communication tasks, feature extraction, vehicle localization and a simple reactive object avoidance based on Potential Fields. In fact, due to the modularity of the global application the vehicle localisation and feature extraction are the same than the ones used in the navigation system.

In ftp://158.42.164.64/Videos there is a clip showing the teleoperation of an industrial forklift at Valencia International Fair [17]. Despite of the fact that we were not dealing with an industrial environment, the the validation was good in the sense that it was crowded of people and objects, stopping the vehicle each time anything was detected.

VII. Vision-based Line Tracking

Vision-based line tracking has been successfully performed on several applications such as agriculture applications [11], [18] and road applications [19], [20]. All of them use complex image processing techniques in order to detect lines on the image and consequently guide the vehicle. All of them are based on a PC platform which are not valid for industrial environments. In our application, the hardware architecture is based on the F150 vision controller from Omron connected to the PLC which controls the vehicle, as shown in figure 16. For simplicity, in this applications sonar sensors are used instead of laser rangers. In this sense, a ring of sensors surrounding the vehicle is used to detect objects on the environment. These sensors have been integrated on a CAN node, which consist of a μC PIC connected to the Polariod 6500 ranging module.

The vision controller can be configured to robustly detect border line points on the image. These points are transferred to the PLC and a simple feature extraction based



Fig. 13. Real image of the warehouse

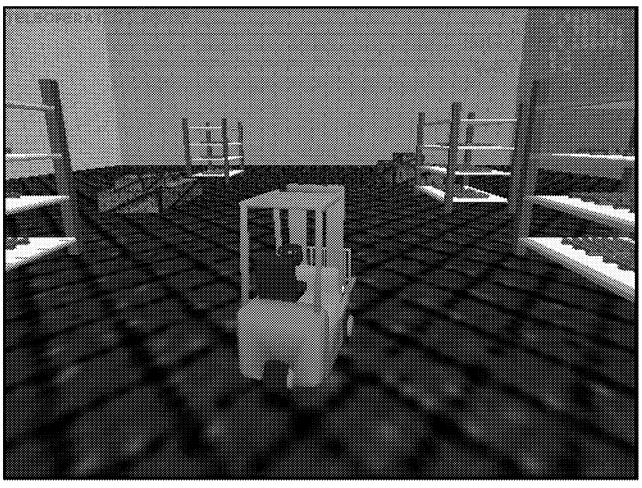


Fig. 14. Virtual image of the warehouse

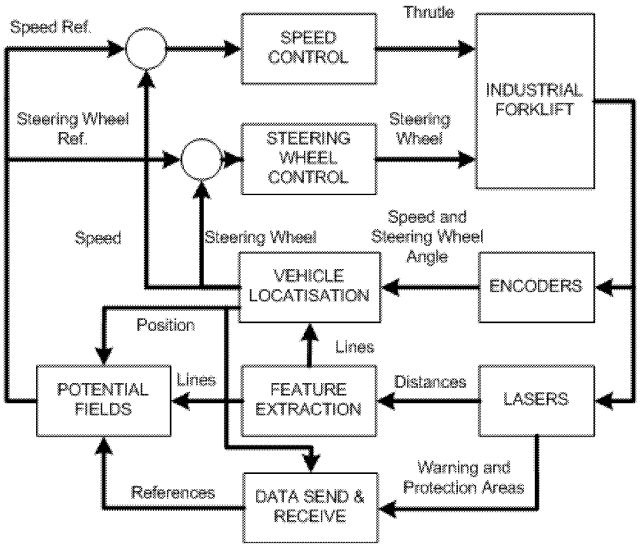


Fig. 15. Software architecture for teleoperation

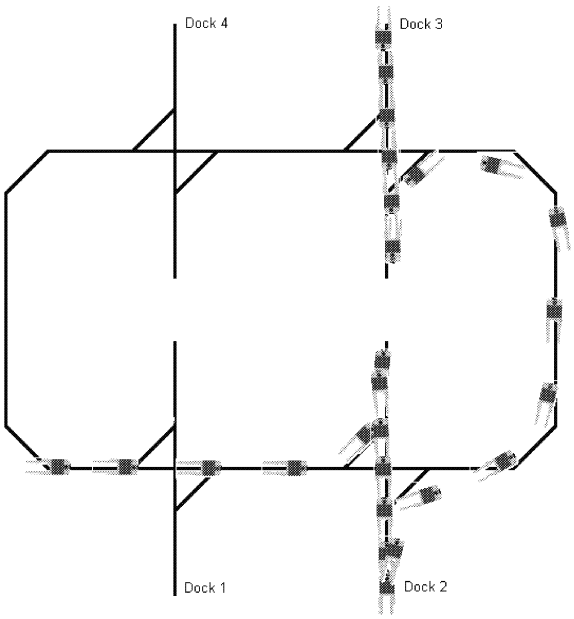


Fig. 17. Example of line tracking

on least-squares fitting algorithm is performed in order to compute the relative position and orientation of the robot with respect to the line. The vehicle kinematic control is based on the Pure Persecution control law [21].

This application has been designed to guide vehicles on closed-circuits, as shown in figure 17. In fact, only two line are detected at each step by the vision system, one on the left and another on the right of the image. The angle between these lines is used to detect and count way-outs. When a way-out is detected and belongs to the destiny dock, the vehicle tracks the line detected on the left to exit from the circuit. Then, a sequence of predefined operations based on the vision system completes the maneuvering operation. The operator performs manually the loading/unloading operation when the vehicle stops and afterwards indicates to the vehicle the next destination dock. Figure 18 shows the control architecture for the vision-based line tracking application.

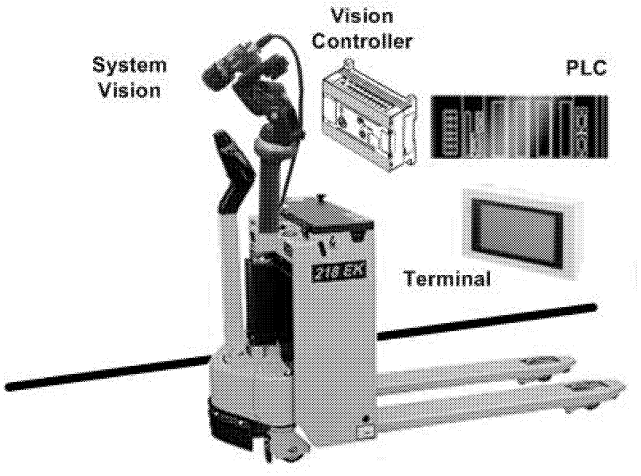


Fig. 16. Hardware architecture for line tracking

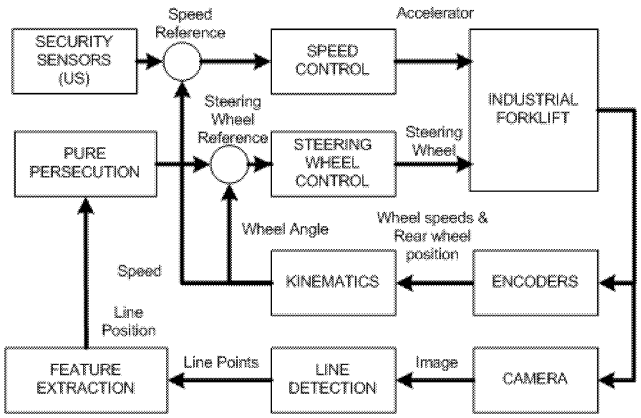


Fig. 18. Software architecture for line tracking

Figure 19 shows an experiment, where it can be appreciated that the vehicle tracks a single line at different velocities.

VIII. Conclusions

In this paper, most of the results of AutoTrans, a research project involving an university team and four enterprises, have been included. In this sense, the main purpose of the project consists of providing an integrated solution for management and transport automation using a fleet of industrial forklifts. The proposed solution is completely modular and several inter-connected applications have been presented.

Management automation module consists of an ERP acting as information system which covers from accounting to production control. An application for dynamically allocating products has been implemented, based on a combination of well-known transportation and assignment solutions in linear and dynamic programming. This ap-

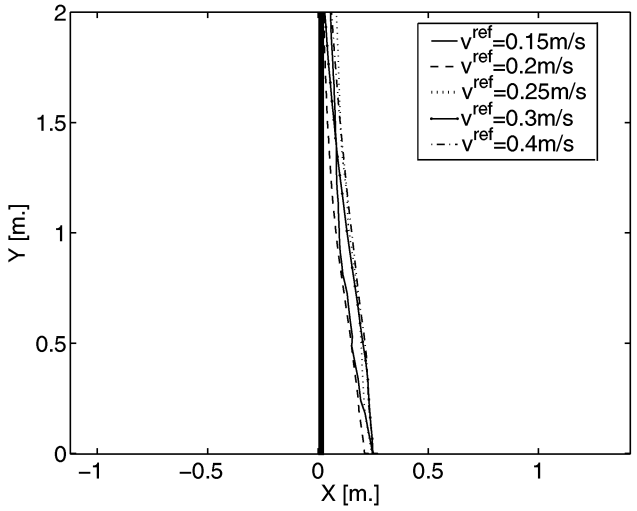


Fig. 19. Vision-based line-tracking experiment

plication uses income and outcome orders together with foresight sales and purchases, production plans, etc. in order to perform the optimization task. Therefore, the optimizer generates optimal allocations to reduce transportation costs associated with delivery times. In addition, electronic transport reports on portable devices such as PDAs will improve the communication flow between the manager and the operators at the warehouse (including forklift drivers).

Warehouse supervision is achieved by means of a SCADA application located on a dedicated computer. This application performs the vehicle monitoring and also data-logging and alarm handling. The transport automation module consists of a complete application set for industrial vehicle control and automation. In particular three different applications are included: robot navigation, industrial teleoperation and vision-based line tracking. All these applications have been particularized for warehouse environments and they are mainly based on industrial PLCs, according to robust requirements of this application.

However, it is interesting to remark that a complete automation of a warehouse is still a difficult task due to its high cost for automating vehicles and also for legal aspects such as security and confinement, which have not been deeply treated.

All described modules have been implemented and tested. In particular, the navigation system and vision-based line tracking are still under development in order to constitute perform commercial modules, that could be implemented in real industrial applications.

In this sense, a modular proposal is the only solution for introducing automation in warehouses, where modules will be implemented separately depending on the needs of the enterprise and the benefits obtained from automation.

ACKNOWLEDGMENTS

Authors would like to thank to Marta Mora, Vicente Franch, Alvaro Herráez, Andreas Hartmann, Guillem Garí, Luís Domenech and Pau Ribes for their support in the implementation on different modules of the final solution.

REFERENCES

[1] C. Lafuente, "Agv: automatizar el transporte y los flujos internos," *Manutención y almacenaje*, no. 401, pp. 71–77, 2005.
[2] G. Garibotto, S. Masciangelo, P. Bassino, C. Coelho, A. Pavan, and M. Marson, "Industrial exploitation of computer vision in logistic automation: autonomous control of an intelligent forklift truck," in *Proceedings of the Int. Conf, on Robotics and Automation*, vol. 2, 1998, pp. 1459 –1464.
[3] J. Pagés, X. Armangué, J. Salvi, J. Freixenet, and J. Martí, "A computer vision system for autonomous forklift vehicles in industrial environments," in *The 9th. Mediterranean Conference on Control and Automation*, 2001, pp. 379–384.
[4] M. S. J. Yoder, "Automatic pallet engagment by a vision guided forklift," in *IEEE Conference on Robotics and Automation*, 2005.
[5] L. Armesto, M. Mora, and J. Tornero, "Supervisión, teleoperación y navegación de vehículos industriales y su integración en el sistema de gestión," *Revista Iberoamericana de Automática e Informática Industrial*, vol. 2, pp. 55–63, 2005.
[6] M. Mora, V. Suesta, L. Armesto, and J. Tornero, "Factory management and transport automation," in *IEEE Conference on Emerging Technologies and Factory Automation*, vol. 2, 2003, pp. 508–515.
[7] "Dimension informatica, http://www.dimension-informatica.es/."
[8] L. Armesto, M. Mora, A. Herraez, and J. Tornero, "A client server teleoperation application of industrial forklifts," in *Symposium on Robots and Control*, 2003, pp. 613–618.
[9] J. C. Latombe, *Robot Motion Planning*. Kluwer Academic Press, 1991.
[10] S. Ge and Y. Cui, "Path planning for mobile robots using new potential functions," in *Asian Control Conference*, 2000, pp. 2011–2016.
[11] L. Gracia and J. Tornero, "Kinematic control system for car-like vehicles," in *Advances in Artificial Intelligence*, Iberamia. Springer-Verlag, 2002, pp. 882–892.
[12] R. Duda and P. Hart, *Classification and Scene Analysis*. New York: John Wiley and Sons, 2001.
[13] Y. Bar-Shalom and T. Fortman, *Tracking and Data Association*. New York: Academic Press, 1988.
[14] J. Neira and J. Tardós, "Data association in stochastic mapping using the joint compatibility test," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 6, pp. 890–897, 2001.
[15] L. Armesto and J. Tornero, "Slam based on kalman filter for multi-rate fusion of laser and encoder measurements," in *IEEE International Conference on Intelligent Robots and Systems*, 2004, pp. 1860–1865.
[16] L. Armesto, S. Chroust, M. Vincze, and J. Tornero, "Multi-rate fusion with vision and inertial sensors," in *Int. Conf. on Robotics and Automation*, 2004, pp. 193–199.
[17] "Protodesign, http://protodesign.feriavalencia.com/."
[18] G. Jahns, "Automatic guidance in agriculture," *ASAE paper NCR*, vol. 83, no. 404, 1983.
[19] E. Dickmanns, "Dynamic computer vision for mobile robot control," in *Int. Symp. Expos. Robots*, 1988, pp. 314–327.
[20] M. Bertozzi, A. Broggi, and A. Fascioli, "Vision-based intelligent vehicles: State of the art and perspectives," *Robotics and Autonomous Systems*, vol. 32, pp. 1–16, 2000.
[21] A. Ollero, *Robtica. Manipuladores y robots móviles*. Barcelona: Marcombo, S.A., 2001.

VIRTAINER: GRAPHICAL SIMULATION OF CONTAINER STORAGE YARD

Miguel Escrivá
Marcos Martí
José Manuel Sánchez
Emilio Camahort
Javier Lluch
Roberto Vivó

Computer Graphics Section
Department of Computer Science
Polytechnic University of Valencia, Spain
E-mail:{mescriva|mmarti|josanchez|jlluch|camahort|rvivo}@dsic.upv.es

KEYWORDS

Stacked object, container block, scene graph, cull, vertex array, device tracking.

ABSTRACT

Currently, 3d graphics are being applied to an increasingly wide range of applications. Industrial processes monitoring is an area where this type of simulation is rarely used, being much more common to represent the information in traditional 2d interfaces.

We introduce here an application of data visualization for maritime container terminals. This system is modular and adaptable to any stacked objects problem. It is architected on a real data base and on the operation processes of a logistic industry. In this paper we describe the architecture of this system, as well as the graphic acceleration techniques that allow us to maintain the frame rate independently of the data size.

INTRODUCTION

Stacked objects are very common in the industry, especially in logistic enterprises. Maritime container terminals are a particular case of this statement. The management of this processes information is usually done by systems with poor alphabetical interfaces which lack the ability to adapt to changes. Advanced graphical simulations model the real process with high fidelity and show changes in a much more convenient manner.

This kind of applications require the access to the system information, usually stored in a database, and the ability to receive the changes occurred in that information through a communications layer. We will show the problems found in the non-graphic parts of these systems, and the approaches we have taken to solve them.

Large data-sets, characteristic of these industries, pose a challenge to the development of graphical simulation systems. The techniques we describe here solve two main problems: real-time visualization of very large data-sets and the modification of this data with an event oriented model. Visibility culling techniques (Cohen-Or et al., 2001), such as geometric simplification, occlusion (El-Sana, 1997) or impostors, have been developed to reduce the amount of data employed to render the scene while maintaining the image quality.

Another indispensable part in any Virtual Reality application is the tracking of input devices. We have implemented a tracking application, based on *OpenTracker* (Reitmayr and Schmalstieg, 2001), which can be configured using standard XML files.

Existing commercial applications, such as those described in (Gambardella and Rizzoli, 2000), offer text or 2d graphics interfaces, and are unable to respond to real time events. Our system offers a richer interface in a virtual representation of the maritime terminal. It is a multi-platform system with integrated data services, XML configuration files, standard database access, and a geometric layer based on an open source scene graph engine. It integrates several graphic accelerations which exploit the spacial coherence of stacked objects to speed up the rendering system. Virtual reality is completely supported, as our system works with any configuration of processing machines, screens and devices. Frame synchronization is done at swap buffers level, allowing the simulation to run simultaneously in any number of computers.

The structure of this paper is described now. First we begin explaining the antecedents and main goals of the project. Then we give some details about the graphic acceleration techniques we have used. A description of devices and interaction in *Virtainer* follows. The next section details *Virtainer's* architecture, its constituting modules and how they work. Finally we present some conclusions and the main directions for future works.

BACKGROUND

This project was conceived as an extension of a previous 2d monitoring application, *Visual Gama* (Jorquera et al., 2003), adopted in the enterprise Marítima Valenciana, one of the most important container terminals of the Mediterranean sea. We felt that the 2d interface wasn't rich enough, and we thought that the natural extension was adding real-time 3d graphics to create a more realistic representation of the terminal with smooth animations to offer higher fidelity and more up-to-date information.

Virtainer is built on top of completely standard open source projects which deliver compatibility and great performance. *OpenSceneGraph* (Burns and Ostfield, 2003) is employed as our graphics engine, providing Virtainer with object management, standard devices handling and other facilities. *OpenProducer* gives us a rendering context and camera control, and *OpenThreads* (Haines and Lagendoen, 1997) provides thread management. The communication layer can employ any protocol, but we are currently using *SCore* (Poza et al., 2002; Posadas et al., 2002), because it is the system adopted in the container terminal to send and receive event messages. Finally, *OpenTracker* encapsulates the devices events in a message which is received and parsed by *TrackerManipulator*. In *Gama* we dispose of data access and communication layers, so *Virtainer* is on the top of *Gama* and fully integrated with it.

The goals of our system are real-time navigation with sustained frame-rate, realism and virtual reality support. An event oriented system is required to maintain the real state of the yard, and multi-view and multimode inspection (fly, walkthrough, cab views, etc) are also desired. The system should offer interaction facilities for selection and camera control.

ARCHITECTURE

Usually container terminals store their information system in a centralized server, using different technologies like ODBC, OLE/DB or web services based on SOAP or XML-RPC. This database stores at any moment the state of the terminal, so the only information needed by our system resides here. Containers, trucks and cranes information is stored in a certain relational scheme that is accessed to represent these elements graphically. It is generally possible to associate a trigger to each interesting event of the database. When any change in the information system occurs, this trigger can broadcast the modification to any application or software agent.

Virtainer is a system based on standard C++ libraries and constructed over open source standard projects. It consists of a set of seven modules, implemented as dynamic libraries, each one handling a different aspect of the simulation.

vrtUtil provides logging facilities and a connection to the Xerces-XML library used to parse configuration files. *vrtDB* encapsulates the code required to connect to a database and load information from its relational schemes. All the logical information of the terminal is stored in *vrtData*. Up-to-date logical positions, destination, vendor and other container data are maintained here. The same applies to trucks, cranes and other machinery. *vrtGeo* manages the geometric aspect of the simulation. It loads the machine models, handles its animations, and is responsible of the creation, destruction and manipulation of the containers geometry. The communication layer resides in *vrtComm*, which is the module employed to receive terminal events and to interchange synchronization messages between the elements of a cluster in a mutiple machines configuration. *vrtGUIHandler*. Keyboard, mouse events and related aspects are managed here, including key bindings and objects selection. *virtainer* serves as an interface to the library set, providing methods to initiate and finish the execution of the library modules, as well as a Viewer class to control the visualization.

When the application begins, the data is loaded from the server database using the *vrtDB* module. This information flows to *vrtData*, and then to *vrtGeo*, where it is used to create the graphic simulation. The changes in this information are received via *vrtComm*, are passed to *vrtData*, and then to *vrtGeo*, where the simulation is updated according to the event occurred. Figure 1 shows this scheme.

Each module has one or more associated configuration files in standard XML format where the relevant parameters for that module can be specified.

GRAPHIC TECHNIQUES

Frustum culling, explained in (Hadwiger and Varga, 1997), is the most common method of culling used in computer graphics. Objects outside the camera frustum are removed from the rendering pipeline. The scene graph structure makes this technique very efficient, as complete object groups can be removed with a simple comparison between the frustum and the bounding volume of the group. *OpenSceneGraph* employs two object lists populated by the culling process. The

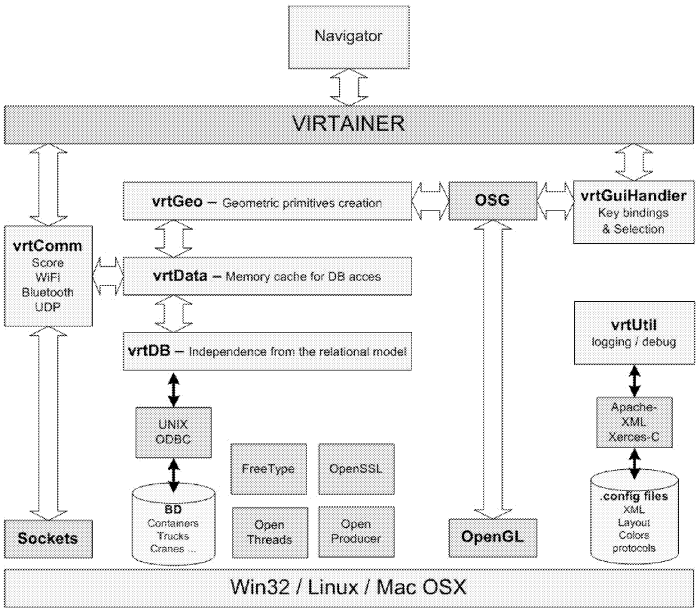


Figure 1: Virtainer modular structure

first list stores the opaque objects ordered by state (material, texture, vertex programs, etc) and the second the transparent objects, ordered by depth. Different bounding volumes are used to improve performance: boxes for geometry and spheres for object groups.

Multi-resolution objects are the next technique implemented. Objects in *Virtainer* have static pre-computed levels of detail, to reduce the number of polygons rendered without reducing noticeably the quality of the resulting images.

Impostors, as explained in (Schaufler, 1995), are a technique which consists of rendering to a texture the objects which are far from the observer, and substituting the real objects with billboards containing that texture. As long as the camera does not move we can achieve very high frame-rates. We have also developed a specific method to create impostors for the container blocks, which consists of using the faces of the bounding box of each individual block as pre-rendered impostors. This method is used when the container block is far from the observer, reducing the geometry of a whole container block to only five quads.

Non-visible geometric culling is other method applied. Container faces which are not visible from any view are removed from the scene graph before visualization. This technique consists of rendering only the surface of each container block, and leads to a noteworthy reduction of the polygon number and an increase of the frame-rate.

Using *Back-face culling* (Kumar et al., 1996; Hultquist, 1990) we can avoid rendering the back parts of the containers. It is easy to see that of the six faces of a container we can only see three at the same time, so with this technique we can reduce the number of polygons to be rendered in a 50%.

Our system can use portal rendering (Elmqvist, 1999a,b) to accelerate the visualization in walkthrough mode. When the application starts, the information about containers is loaded from the database, and it is then used to create cells and portals automatically following the *Breaking the walls* (Lerner et al., 2003) algorithm. Our method, based on (Luebke and Georges, 1995), introduces portal reconstruction to adapt to the dynamic geometry of the container terminal.

Vertex arrays is the last acceleration technique used in *Virtainer*, and consists of grouping the geometry of the container blocks in a certain way which is better handled by the graphics card.

Multiple machines using *Virtainer* can be interconnected to increase the computer power when rendering to more than one display device. We have chosen a master-slave architecture where the master computer controls the camera and sends the model view matrix to all the slaves. Then the camera can be modified to adapt the view to the physical display configuration. Our election of this architecture does not mean we are limited to connecting the devices directly to the master: any machine can receive this kind of events and send them to the master, which will then distribute the changes through the cluster.

INTERACTION

We have developed a utility to manage input devices which offers support for any controller supported by the operating

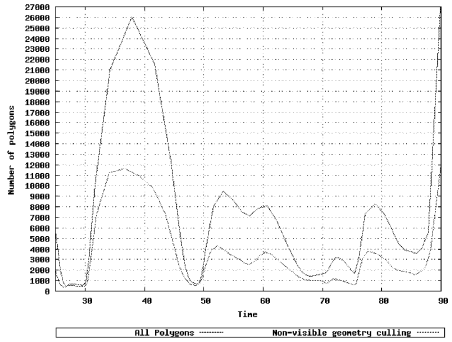


Figure 2: Number of polygons rendered using or not Non-visible Geometry Culling

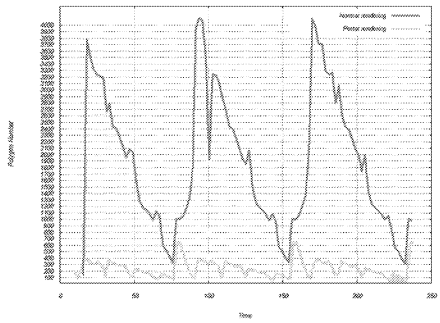


Figure 3: Number of polygons rendered with and without Portal Rendering

system. It is based on *OpenTracker*, benefiting of its advantages: configuration based on XML files, device abstraction and transparent network access. Our tracking system can be launched in a different host that the main application, allowing the device management to run in a dedicated host. Finally, we use XML configuration files to map tracking data to movements and select the interaction type.

RESULTS

We present here the impact in the frame-rate of the graphic techniques implemented. Removing non-visible container faces deliver a noteworthy Figure 2, but the frame-rate is still low; some more improvement is required.

We have reduced considerably the number of polygons using portals, as can bee seen in figure 3. This results in approximately a 75% increase in the frame-rate.

Finally, as you can see in Figure 4, employing vertex array lists delivers a great increase of the frame-rate. Modern graphics cards are capable of dealing with a very large number of polygons, provided that they are organized in a compact structure.

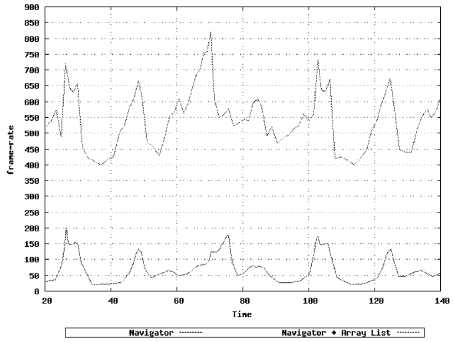


Figure 4: Frame-rate comparison using or not Vertex Arrays



Figure 5: Virtainer running in a cockpit view configuration

We have used joysticks, game-pads and a wheel to control *Virtainer*, and we have also a tracker which follows the user head movements, enabling the user to look around the terminal. The screen configurations tested are jumbo-box, a cockpit view with three monitors and a head mounted display. An example can be seen in figure 5.

Movements of the terminal machinery are reflected by generating smooth animations of the models. When a machine starts to move, an event message is received, and its representation reacts accordingly. All the machines send a message with its position regularly, thus the error of the positions is minimal.

CONCLUSION AND FUTURE WORK

We have presented here a multi-platform simulation system capable of representing a maritime container terminal in 3d with a sustained frame-rate. It offers support for virtual reality configurations and any device can be used to control the navigation. It is possible to select relevant elements of the terminal to show related information. We have created a web site where we publish project related information. The address is http://www.virtainer.org.

We are currently working in improving the quality of the image even more using shader based techniques, like those in



Figure 6: New water shader and Displacement Mapping in containers

(Dumont et al., 2003; Wang et al., 2003), and in designing an event simulator to produce animations outside the maritime terminal network. Figure 6 shows some of these improvements.

Other interesting areas are the improvement of the portal culling and the use of *occluders* (Coorg and Teller, 1997) for systems with lower specifications or even larger scenes. There is also some room for improvement in the synchronization method used; a peer-to-peer scheme may be better suited for our system than the current master-slave configuration.

ACKNOWLEDGMENTS

The *Virtainer* project is being supported by grant TIC2002-04166-C03-01 from the Spanish Ministry of Science and Technology and IST-2-004363-STP from the GameTools project. We thank *Virtainer*, *Gama* teams for their work and collaboration. We are also grateful to Marítima Valenciana for their support of our research efforts.

REFERENCES

Burns, D. and R. Ostfield. Open Scene Graph - An Open Source Solution for Real Time Image Generation. 2003. IMAGE Society.

Cohen-Or, D., Y. Chrysanthou and C. Silva. A survey of visibility for walkthrough applications. 2001. IEEE TVCG.

Coorg, S. and S. Teller. Real-time Occlusion Culling for Models with Large Occluders. 1997. IEEE TVCG.

Dumont, R., F. Pelacini and J. A. Ferwerda. Perceptually-Driven Decision Theory for Interactive Realistic Rendering. 2003. In *ACM Transactions on Graphics*.

El-Sana, J. Integrating Occlusion Culling with View-Dependent Rendering. 1997. IEEE TVCG.

Elmqvist, N. Axis-Aligned Bounding Box Culling For Portal Rendering. 1999a.

Elmqvist, N. Introduction to Portal Rendering. 1999b .

Gambardella, L. M. and A. E. Rizzoli. The role of simulation and optimisation in intermodal container terminals. 2000. European Simulation Symposium.

Hadwiger, M. and A. Varga. Visibility culling. 1997. The Institute of Computer Graphics and Algorithms, Vienna University of Technology.

Haines, M. and K. Lagendoen. Platform-independent runtime optimizations using OpenThreads. 1997. In *11th International Parallel Processing Symposium.*

Hultquist, J. Backface culling. 1990. In *Graphics Gems*, 346–347.

Jorquera, P., T. Barella, D. Llobregat and R. Vivó. Visual GAMA. 2003. EUROGRAPHICS.

Kumar, S., D. Manocha, W. Garrett and M. Lin. Hierarchical back-face computation. 1996. In *Eurographics Rendering Workshop*, 235–244.

Lerner, A., Y. Chrysanthou and D. Cohen-Or. Breaking the Walls: Scene Partitioning and Portal Creation. 2003. In *Pacific Graphics 2003*.

Luebke, D. P. and C. Georges. Portals and Mirrors: Simple, Fast Evaluation of Potentially Visible Sets. 1995. In *Proceedings 1995 Symposium on Interactive Computer Graphics*, 105–106, Department of Computer Science, University of North Carolina.

Posadas, J.L., P. Pérez, J.E. Simó, G. Benet and F. Blanes. Communications structure for sensory data in mobile robots. 2002. *Engineering Applications of Artificial Intelligence* 15, 341–350.

Poza, J.L., J.L. Posadas, J.E.Simó and A.Crespo. Data and Event Management in a Maritime Terminal of Containers. 2002. *New Technologies For Computer Control* 269 – 274.

Reitmayr, G. and D. Schmalstieg. An Open Software Architecture for Virtual Reality Interaction. 2001. ACM.

Schaufler, G. Dynamically generated impostors. 1995. In *D. W. Fellner, editor, Modeling Virtual Worlds - Distributed Graphics*, 129–136.

Wang, L., X. Wang, X. Tong, S. Lin, S. Hu, B. Guo, and H. Shum. View-Dependent Displacement Mapping. 2003. In *ACM Transactions on graphics, 22.*

SUPPLY CHAIN SIMULATION

A METHODOLOGY FOR SUPPLY CHAIN SYSTEM MODELING AND PROCESS EVALUATION

M.Chabrol[2], M.Comelli[1], P.Féniès[1]and N.Tchernev[1]
LIMOS UMR CNRS 6158
[1]IUP Management et Gestion des Entreprises Pole Tertiaire
26 Avenue Léon Blum - 63000 Clermont-Ferrand –France
[2]ISIMA - Campus Scientifique des Cézeaux
63173 Aubière Cedex – France
Email {chabrol, comelli, fenies, tchernev}@isima.fr

KEYWORDS
Modeling Methodology, Supply Chain, Performance Evaluation, Activity Based Costing.

ABSTRACT

The aim of this paper is to present a modeling methodology in order to evaluate financial and physical performance in Supply Chain domain. This methodology is applied to a real industrial case study in a Supply Chain company. In this case study, we compare two production strategies, a push and pull strategy by combining heuristic and simulation.

INTRODUCTION

The purpose of this article is to propose an approach to evaluate performance in a Supply Chain by discussing connections among two Supply Chain flows concerned with management science and computer science. These flows are the physical flow and the financial flow. Judging from the literature, these flows do not always overlap in a Supply Chain. Nevertheless, we believe there are important synergies worth exploring (Féniès et *al.,* 2004). Implementing a modeling environment in a Supply Chain gives the possibility to evaluate industrial performance of the Supply Chain's physical and financial flows. As shown by Vickery et *al.,* (2003), integrating financial flow and physical flow in Supply Chain management is essential to optimize financial flow. Nevertheless, almost all scientific literature concerning physical flow impact on financial flow deals with Supply Chain network configuration which is a strategic problem. There is not, in the literature an approach which proposes to evaluate thanks to ABC and cash flow level the impact of physical flow on financial flow in operational and tactical problems. In a literature review about Supply Chain modeling, Hongwei et *al*., (2005) show a state of the art about global methodology and that many models are only a conceptual network without implementation, therefore very abstract to realize or only designed and simulated for a specific problem and hence lack general applicability: there is a need for a general approach for both Supply Chain modeling and its evaluation which combines physical and financial flows
This paper is organized as follows: in next section, a modeling methodology is presented in section 1. Section 2 is dedicated to a case study in automotive industry, where the modeling environment for Supply Chain (SC) is implemented by using simulation. The case study is introduced to illustrate the modeling and simulation processes by use of the modeling environment. Finally, conclusion and future research work are presented.

A MODELING METHODOLOGY FOR SUPPLY CHAIN EVALUATION

Gourgand and Kellert (1991) have proposed a methodology called ASDI (Analysis, Specification, Design, Implementation) dedicated to the design and the implementation of modelling and simulation software environments for a domain. In this section we propose to use ASDI in order to conceive a modeling methodology for Supply Chain evaluation. In the first paragraph, we show links between systemic approach, knowledge model and PREVA in the methodology process. In paragraph two, we give a generic knowledge model for Supply Chain which integrates physical and financial flows. A model for PRocess EValuation and Analysis (PREVA) is given in paragraph three.

Methodology Process

Modeling Supply Chain behavior may as well be a complex task. That's why we propose an approach dedicated to the study of such systems.

The first part of the methodology is dedicated to the Supply Chain class system. The Generic Domain Knowledge Model is dedicated to the domain. This model is built during Analysis and Specification steps, by identifying the common points of all the systems of the domain according to the entities they contain and their functioning. During the Design step, the Generic Design Model is built by discovering and integrating thanks to the Generic Knowledge Model mathematical or simulation models. The

component coding is done during Implementation step in order to build the Software Libraries.

The second part of the methodology is dedicated to the study of a system of the domain, by proposing an iterative modeling process, and presented in Figure 1. The first part considers the construction of the knowledge model for the studied system: the system is firstly described globally and then some or all points can be detailed. Then, the adapted modeling approach is chosen according to the performance criteria it provides and the study objectives. The associated action model (a simulation model, or an analytical model) is then generated using the software components belonging to the libraries. The system performance is then evaluated with PREVA.

This modeling approach constructs a model according to an iterative process. This process is divided into four phases (figure 1): *(i)* Analysis and formalizing of data; this system specification leads to the design of the knowledge model; *(ii)* translation of the knowledge model into an action model using a formalism allowing its exploitation to provide performance criteria; *(iii)* exploitation of the action model to provide performance criteria; *(iv)* interpretation of results and decisions about actions to perform on the system.

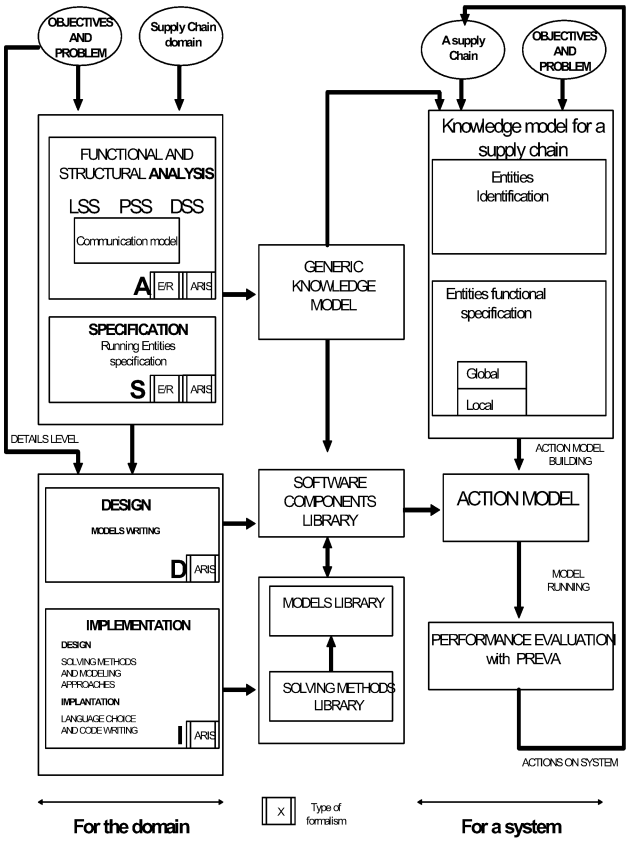


Figure 1 The proposed methodology.

Analysis and specification for a generic Supply Chain case: an object approach

The development of a global knowledge model proposing a holistic view of a studied domain seems interesting before building a model of a system, whatever its type. By considering that Supply Chain is a set of complex systems, we use ASDI to define the stages of modeling for the system of the Supply Chain class. The used systemic decomposition arises in the following form, for a complex system which is integrated in the Supply Chain, as for the Supply Chain system: *(i)* The logical subsystem (LSS) contains the transactions flows treated by the system. *(ii)* the physical subsystem (PSS) is structured in small units containing physical entities needed for elementary operations; *(iii)* the decision subsystem (DSS) is structured in decision center which contains the management rules, resource allocation and system functioning.

The three subsystems are thus complementary and communicating two to two. Supply Chain is considered as a finished number of complex systems which are structured with the systemic decomposition suggested by ASDI.

The knowledge model of a system is a formalization in a natural or graphic language of the system operation and structure. In order to formalize knowledge on a Supply Chain, ARIS (which means Architecture of Integrated Information systems) suggested by Scheer (1991) is used because it is a tool which can be used to model processes (Green and *al.,* 2000), as well as the model Entity Relationship (E/R) suggested by Chen (1976) and which is integrated in ARIS.

The system class analysis enables us to distinguish three families' flows in LSS: the principal flow or main flow (customer flow), the information flow which includes any type of flow providing with the information on the entities circulating in the system, and the financial flow. The establishment of financial flow as a partial monetary counterpart of physical flow makes it possible to have a completely total and integrated Supply Chain vision.

To be able to control these flows, it is essential to describe them: the flow circulation network seems the material support of all the possible movements. Supply Chain is represented like a set of entities or nodes which are interrelated and its activities consist to treat, transform and store the flow elements. The PSS includes Unit Value for Customer (a sort of business unit), which are consisted of the main network, while DSS includes the decision-making centers and the management units which are consisted of the financial and informational networks. The global model of a Supply Chain, presented in Figure 2 characterizes the links between the three sub systems and gives an operational and a functional view.

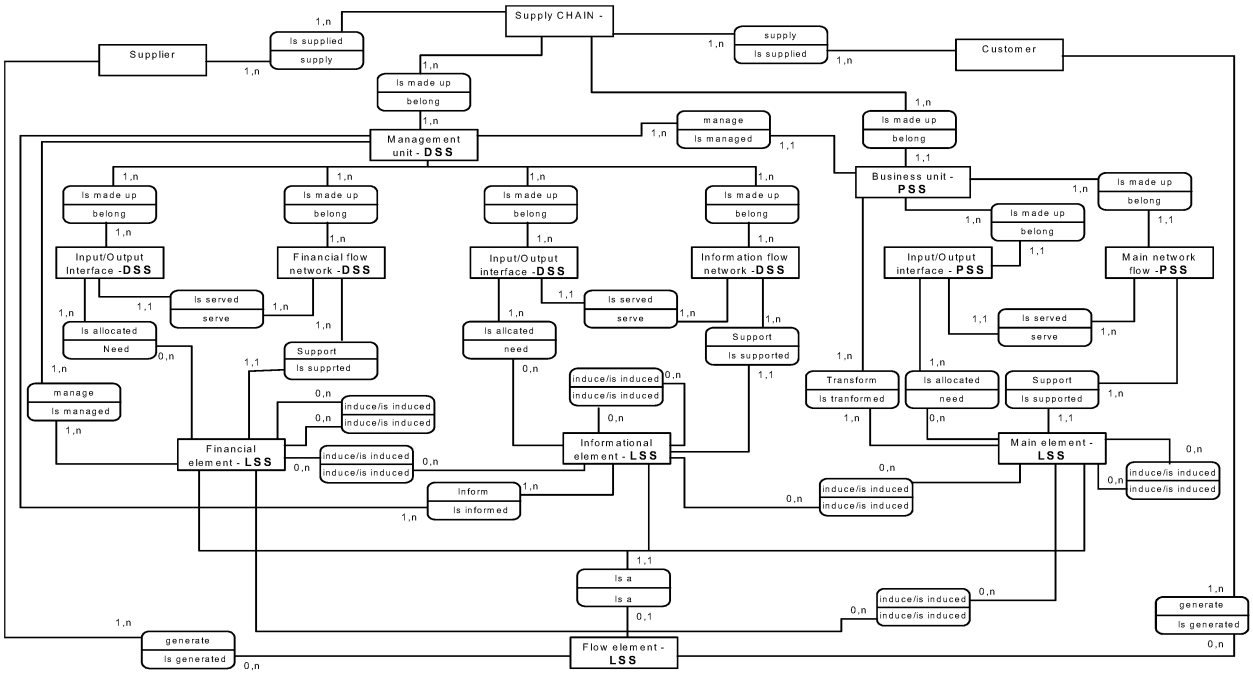


Figure 2 Extract of the generic domain knowledge model for Supply Chain

PREVA: a generic approach for process evaluation

In this paragraph, the conceptual principle of the generic evaluation model, called PRocess EValuation and Analysis is given. This evaluation model is conceived thanks to generic knowledge model. In a Supply Chain company, tens of activities are carried out during the process of getting raw materials, transforming them into finished products and delivering them to customers. Our approach gives Supply Chain manager the possibility to evaluate the impact on Supply Chain financial flow of different Supply Chain physical flow decisions, and to choose between strategies the best one. A synoptic model of the Supply Chain has been defined based on the Knowledge model. Action model is derived from the synoptic of the Supply Chain.

The evaluation is done in three steps:

• Step 1 Physical Process Evaluation: Mathematical approach or simulation can be used. Heuristics with financial and physical aspects are built in order to improve Supply Chain planning thanks to knowledge model and Supply Chain objectives.

• Step 2 ABC evaluation: Our generic approach is able to determine indirect costs consumption thanks to process cost evaluation on each business unit and for the global complex system. In this context, the cost of a business unit is the summation of the costs of activities and processes in each entity with the costs of services and products. Stock value can be evaluated in one entity and for the global complex system. The potential of value creation is also evaluated by combining the difference between demand and quantity sold by an entity or the global system with contribution margin. Between each business unit, transactions are evaluated in transfer price or in market price. Differences between ABC cost and transfer price /or market price give managers the possibility to evaluate value creation (a sort of profit level) in each entity of the system.

• Step 3 Financial process evaluation: thanks to ABC evaluation, direct and indirect resources consumptions and net sales are determined. Because of the nature of the cost (calculated cost and real cost) and because of the payment term which is different between each type of resources and each type of customers, there is a big difference, in medium term, or in short term between profit level and cash flow in the same period. Our model proposes to integrate cash flow thanks to resources consumption evaluation in ABC: we are able to evaluate cash position and cash flow in each entity of the chain and for the global complex system.

A formalization of PREVA for Supply Chain evaluation is given in (Comelli et *al.,* 2005). Figure 3 shows a synthesis of our work which gives Supply Chain manager the possibility to evaluate value creation, cash flow, cash position and potential of value creation in the whole Supply Chain, in each business unit and for each master unit. Thanks to resources consumption in ABC and resources payment, a bridge is done between physical flows planning impact on financial flows.

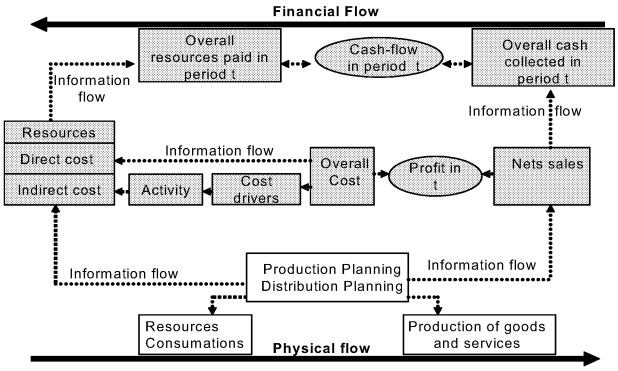


Figure 3 PREVA, a bridge between physical and financial flow in Supply Chain

In next section, we give an example of the methodology implementation for a Supply Chain in a case study and we present an action model.

METHODOLOGY IMPLEMENTATION: ACTION MODEL AND CASE STUDY

Case study is presented in paragraph 1. The action model built thanks to methodology is given in paragraph 2. Simulation results are given in paragraph 3.

Case study presentation

The automative company Supply Chain is comprised of 6 factory, called F1, F2, F3, F4, F5, F6. The system is a company Supply Chain where twelve month demand is known. Six types of product, called P1, P2, P3, P4, P5, P6 are manufactured in the system. The chain is shared in four steps. After each factory, products are stocked. In first step, there is one factory (F1). In step 2, F2, F3, F4 are working both together and have the same structure. In step 3, product from F2, F3, F4 are manufactured on a special plateform.This type of plateform gives the product its name and its quality. In step 4, F6 products are prepared for european, asian, american markets.

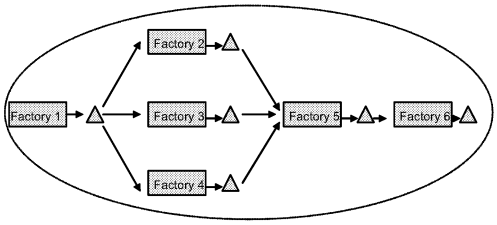


Figure 4 Supply Chain structure

The horizon level is 12 months, and the planning horizon level is the week. This organization runs under either push or pull system. Therefore, 2 types of collaborative production strategy are evaluated: a push and a pull strategy. Push strategy is called strategy 1, pull strategy is called strategy 2. The company is organized in business unit. The capacity of each plant is limited, but the biggest problem is step 2.
The main objective of this study consists in financially evaluating two strategies on a whole Supply Chain. Such a study does not need as highly detailed model as for instance operational control or manufacturing system optimization. The distribution planning is not an input variable but just a parameter. It has been built from commercial forecasts and Supply Chain objectives. Output variables to observe are the monthly stock level of each type of product in each factory, the monthly number of campaigns of each factory, and the number of products processed monthly in each factory.

Action model selection and implementation

Discrete event simulation was preferred to mathematical model for many reasons such as modeling constraints and computation time. Specific modeling was done to take into account particular constraints caused by horizon level and planning horizon.
In this case study, thanks to the specific knowledge model, we choose to simulate families of products rather than singular products. That is why products getting similar process characteristics are grouped together in a same family. Six families have been defined by industrial experts. Furthermore, regards to the depth of planning horizon, it seems to be judicious to simulate a batch of products rather than a singular product. In our study, an entity represents a batch of 1,000 products.
Hence, a resource and six upstream stocks model a factory. Each factory is characterized exclusively by a monthly capacity. Routings between factories are similar to those described in Figure 4. At the beginning of the simulation, stock levels of each factory are average levels frequently observed in each factory. Resources are idle. Hence, in order to minimize the impact of initial state, the two first months are devoted to the warm up of the system.
In this Supply Chain, the difference between pull and push strategy depends on storage policy. This event occurs after each end of production campaign. Discrete event simulation is based on a scheduler, which means that each event occurred creates on the scheduler other events. In our model, each "end of campaign" event involves the choice of a stock. Unlike push strategy consists in choosing the highest upstream stock level, pull strategy choose the type of products, which gets the least autonomy. The modeling of the two strategies Push and Pull is similar to intelligent inventory and production management. More precisely, the modeling of a strategy reproduces the decision-making. Decision-making is thus modeled by a dedicated heuristic.
State of system is used as input data by two heuristics (one for push, one for pull strategy). The state of system taken into account is composed only of upstream stock levels ($upstreamstock_i$) for the Push strategy. On the contrary, Pull heuristic takes into account finished products stock levels since these data allow estimating Supply Chain autonomy for each factory ($autonomy_i$). The two outlines are given in Figure 5.

Push Strategy	Pull Strategy
productType=0 upstockmax=0 i = family of products $i \in \{1,\ldots n\}$ For i =1 to n If $upstreamstock_i$ > upstockmax Upstockmax= $upstreamstock_i$ productType = i End if i=i+1	productType=0 autonomin =A i = family of products $i \in \{1,\ldots n\}$ For i =1 to n If autonomy i < autonomin autonomin = autonomy i productType = i End if i=i+1

Figure 5 Push and pull strategy heuristic

The pull strategy modeled in our study is based for each product on the calculation of a parameter called "Autonomy". This one corresponds to the number of periods where customer demand can be satisfied with the actual stock level of finished products. Regards to these data, pull strategy consist in prioritizing in each factory the family of products, which get the least autonomy. A full strategy is thus reproduced along the Supply Chain.

Results

The simulation is deterministic and represents one year and two months. This 2 Mo model has been developed with Arena 7.0 and takes 1 minute to run with a PC equipped with 1.8 GHz processor and 256 Mo of RAM.
Two strategies are evaluated thanks to PREVA. In Table 1, global physical flow metrics are given. It is quite difficult for supply chain manager to choose between strategy 1 or 2 just with physical flows metrics: the demand satisfaction level is the same and the difference in stock level in the end of one year is under 1% between the two strategies. An evaluation of cash flow level and ABC level will help supply chain manager. Therefore, that is very interesting to evaluate (as shown in section 1) the value creation (figure 6), cash position (figure 7). Value creation evaluation gives manager the possibility to know which products create value and where (in which plant) value is created and where value is destroyed because of collaborative production planning. To conclude this section, table 2 gives final results and strategy selection.

	Strategy 1	Strategy 2
Customer satisfaction	88	88
Stock level (quantity)	135480	133680

Table 1 Physical flow planning evaluation thanks to simulation

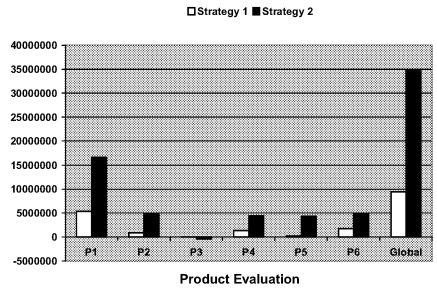


Figure 6 Value creation for each product in strategy 1 and strategy 2

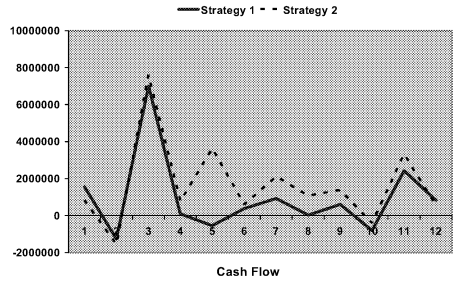


Figure 7 Cash flow levels in strategy 1 and strategy 2

	Push Strategy (1)	Pull Strategy (2)
Customer satisfaction	=	=
Stock level in quantity	=	=
Value Creation	-	+
Cash position	-	+
Cash flow	-	+
Potential of value Creation	-	+
Closing stock level value	-	+
Planning selection	-	+

Table 2 Evaluation final result and strategy selection

CONCLUSION

In this paper, we have proposed a methodology for Supply Chain modeling and process evaluation. This paper has described how object-oriented modeling can provide a new way of managing and designing complex systems with financial, physical and informational flow. The global evaluation model (PREVA) is tested in an internal Supply Chain (a multinational company) and Supply Chain tactical planning are selected thanks to it. The proposed approach improves the visibility of cost but the most important point in our approach is to show how physical flow impact is passed down to financial flow.
In this case study, our evaluation model is coupled with simulation model. This choice, explained in section 2, will allow integrating stochastic phenomenon (such as demand variability, uncertain breakdown,…) in order to study strategies robustness. Nevertheless, the approach is generic and leaves the possibility to be used with mathematic models.
In further research we want to integrate ABC and cash flow in action model in order to improve Supply Chain working. We want to do an extension of this approach in a global Supply Chain to improve its working and give Supply Chain manager the possibility to share value creation which is done by collaborative planning.

REFERENCES

Chen P. 1976. The entity relationship model – toward an unified view of data. ACM transaction on database system.

Comelli M. Fenies P. Gourgand M. and Tchernev N. 2005. A generic evaluation model for cash flow and activity based costing in a company Supply Chain. Accepted in International Conference on Industrial Engineering and Systems Management. IESM, Marrakech, Moroco.

Fenies P. Gourgand M. Tchernev N. 2004. A framework for Supply Chain Performance Evaluation. 5e RIRL, Fortaleza, Brazil.

Gourgand M. et Kellert P. 1991. Conception d'un environnement de modélisation des systèmes de production. 3ème congrès international de génie industriel, Tours.

Green P. Roseman M. 2000. Integrated Process Modelling: an ontological evaluation . In Information systems, Vol. 25.

Hongwei D. Benyoucef L. Xie. 2005. A modelling and simulation framework for Supply Chain design in Supply Chain Optimisation. Ed. A.Dolgui, J. Soldek et O. Zaikin, Editions Springer.

Scheer A.W. 2002. ARIS– Business Process Modelling. Springer.

Vickery S.K. Jayaram J. Droge C., Calantone R. 2003. The effects of an integrative Supply Chain strategy on customer service and financial performance: an analysis of direct versus indirect relationships, Journal of Operations Management.

A MODELLING FRAMEWORK FOR SUPPLY CHAIN PLANNING SIMULATION UNDER UNCERTAINTY

François Galasso
LAAS-CNRS
7, Avenue du Colonel Roche,
31077 Toulouse, France
LGP-ENIT
47, Avenue d'Azereix-BP1629
65016 Tarbes Cedex
E-mail: fgalasso@laas.fr

Colette Mercé
LAAS-CNRS
7, Avenue du Colonel Roche
31077 Toulouse, France
E-mail: merce@laas.fr

Bernard Grabot
LGP-ENIT
47, Avenue d'Azereix-BP1629
65016 Tarbes Cedex
E-mail: bernard@enit.fr

KEYWORDS
Supply Chain, planning, time horizon, decision support system, uncertainty.

ABSTRACT

The purpose of this paper is to present an analytical model of a supply chain element aiming at providing decision support for the manager. The suggested framework is based on literature and case studies and includes the main decision making activities which are the core of the supply chain problems, i.e. adapt the capacity and the requirement planning according to the forecasts and firm orders known on a given period of time. The capacity adaptation may consist in the use of extra-hours, subcontracting, 2 or 3-shifts-works etc. and is constrained by the necessary delay for the set-up of each action. The suggested simulation framework takes into account the points of decision to provide a feasible and possibly optimal plan and evaluates its cost. In that way, it offers simulation facilities for supporting the decision making process at an enterprise level in a supply chain context.

INTRODUCTION

In an industrial context where markets are evermore competitive, the companies need to be flexible and adaptive. In order to reach these goals, the relationships between the entities of a Supply Chain (SC) have to be improved, which requires to address problems regarding information sharing, confidentiality and management of uncertainty. Many IT techniques and methods, such as Electronic Data Interchange, Efficient Customer Response, Vendor Management Inventories, Business-to-Business, Business-to-Consumer, enterprise portals… are dedicated to the SC development and were suggested as the Information Technology grew-up. Advanced Planning Systems also provide the company with an effective tool for the optimisation of the SC, but their use is limited by the availability of accurate information from the suppliers and customers. In the daily life, this information is often pervaded by uncertainty, which must be considered in the sales process in order to allow a better understanding of the customer's requirements (Bartezzaghi and Verganti 1995). Depending on the necessary supply lead-times embedded in the production process, this uncertainty might be taken in account for ordering components and making decisions in the workshop before the customer sends firm orders.
In this paper, we suggest an analytical framework for the modelling of the standard production constraints such as production lead-times, capacity and associated costs. We introduce the concept of a*nticipation delay* of a decision, which allows on one hand to formalise the fact that some decisions have to be anticipated, and on the other hand that the first's periods of a planning are constrained by previous decisions. This framework is based on observations made in various companies and on literature studies and aims at providing a generic decision support that a manager can fit to the characteristics of its enterprise in order to have a better view on the choices available for him during a period of time. Typical choices are for instance: make or buy; subcontracting or changing to 2 or 3-shifts-works, use overtimes or produce in advance and create stocks.

This paper focuses on the most usual way to process uncertainty, i.e. through periodic refresh of the planning. As a second step, we intend to explicitly quantify the demand uncertainty in the model. This framework has been tested on an example of supply chain structure specially designed in order to be representative from real problems.

The paper is organised as follows: we firstly position our approach regarding current academic researches and case studies on uncertainty management in a supply chain. Then, we describe a typical case of SC on which our modelling framework is applied in the following section. Finally, we discuss some preliminary results and assess the applicability of the modelling framework.

CASE STUDIES AND LITERATURE REVIEW

Taking into account the demand uncertainty in Production Planning and Scheduling has been the purpose of a huge literature, e.g. through improvements in the dimensioning of safety lead times (Whybark and Williams 1976), safety stocks (Grasso and Taylor 1984), (Krupp 1997), safety stocks and lot-sizing policy (Bodt and Van Wassenhove 1983), (Ho and Ireland 1998), (Brennan and Gupta 1993), or by using only lot-sizing policies (Johansen 1999). Management of uncertainty can be viewed as a comprehensive process in which the primary important phase is to better understand what the final demand will be

(Bartezzaghi and Verganti 1995). They also showed that the over-costs due to the variability of the demand can be self-regulated by the reuse of overstocks. A state of the art on the dimensioning of safety stocks in a MRP (Material Requirement Planning) context can for instance be found in (Guide and Srivastava 2000), concluding that no method clearly outperforms the others.

Another way to model an uncertain demand is to propagate uncertainty down to the final planning. Possibility theory is used in that purpose in (Grabot et al. 2005), uncertainty being propagated all along the Material Requirement Planning process. The results are a fuzzy requirement planning and a load planning that can help the manager to view the possible charge he will have to deal with. In (Fargier and Thierry 2000), a representation of the imprecision on the ordered quantities is propagated in order to determine the global demand corresponding to a set of orders. An uncertain demand can also be addressed by periodic re-planning, after including the forecasts in the master production planning process. In (Zhao and Xie 1998) is developed in that purpose a model integrating a variable demand in the Master Production Schedule. The planning process suggested in this study is dependant on lead times and on a rolling time principle that introduce a frozen horizon. They show that the length of this frozen horizon has an impact on both the overall costs and stability of the SC. Van Landeghem and Vanmaele (2002) provide an overview of the sources of uncertainty that can be handled by a risk assessment, supposed to lead to a better management of uncertainty, reducing then the number of re-planning periodicity. Similarly, Génin (2003) developed a tactical planning method based on a reference plan which defines the re-planning periodicity, based on the frozen horizon induced by the production process and lead times. He shows that this method has an important impact on the limitation of the bullwhip effect and the stability of SC.

In order to check the consistence of these academics researches with the real industrial practices, we have conducted a series of interviews with supply chain managers of large enterprises, from the automotive, aeronautic and air-conditioning sectors. These interviews showed that the actors of the SC usually have a poor visibility on the internal processes of their partners, leading to difficulties for a good coordination. As a consequence, supply chain management is often reduced to a peer-to-peer relationship, each partner taking into account the demand of his customers and redirecting this demand on its suppliers. In order to speed up the information flow and improve coordination, the companies are trying to develop data sharing by mean of common information systems based on internet, using ERP facilities or enterprise portals, but their main difficulty is to handle the demand uncertainty, most of the time addressed through slack times and safety stocks which increase the costs and decrease flexibility. In a context of more and more specific products, safety stocks are not always possible and very often, managers have to improvise solutions based on their experience to cope with disturbances like unexpected increase of the demand or tardy supplies. Being able to better anticipate these disruptions and to better quantify their consequences is in our opinion a key issue for being able to prepare a set of efficient decisions in advance.

This study also confirmed that the planning periodicity and the length of the frozen horizon in which the possibilities of available decisions is reduced have an important impact on the global performance of the SC. Based on theses hypothesis, our approach aims: 1) at providing the manager with a good vision of the situation of his enterprise facing an uncertain demand and 2) at showing him the decisions which he could make through time in order to address the expected problems. At each step, his choice is then to make an early decision (which can reveal to be sub-optimal if the demand changes) or to take the risk to wait until the demand is more precisely known, with the possibility that his choices are then reduced. In all cases, we do think that illustrating the consequences of a decision making at each decision node is a more promising axis of research than building so-called optimal plans which are rapidly set into question and do not allow the decision maker to use his experience.

TYPICAL SUPPLY CHAIN CONFIGURATION

A test case has been designed after considering many examples of supply chains in order to simplify the industrial reality while being nevertheless representative of the problems that may occur. In that purpose, we have defined an example of SC made of four main links, with a possibility of subcontracting for one of the partners. Various suppliers (with different costs and delays) may be used for the same component. Some of the partners belong to different SC, in order to illustrate the possible conflicts between local (related to a partner) and global (related to a Supply Chain) decision makings.

A representative example of Supply Chain

The considered SC structure is shown in Figure 1. This SC, which manufactures pieces of furniture, is defined as follows:

1. Two families of products are produced by two Supply Chains which have common points: SC1 is producing tables and SC2 produces two references of chairs. The suppliers (FO0, 1 and 2), the manufacturer (MAN), the subcontractor (SUB), the painter (PAINT) and the client number 1 (CLT1) constitute SC 1. The suppliers (FO2, 3 and 4), the manufacturer (MAN), and the clients number 2 and 3 (CLT2, 3) constitute the SC 2. So, FO2 and MAN belong to the two SC.

2. For a given SC, the limitation of the production capacity may require that a manufacturer defines a priority policy between its customers. In our structure, MAN will have to choose in some cases between CLT2 and CLT3.

3. Several suppliers are possible for the same components: FO3 and FO4. The characteristics of these two suppliers are different: FO3 is cheaper than FO4 but FO4 is more reactive (its lead time is lower). MAN

will have to choose between those suppliers: if he can anticipate, FO3 is a better choice but he can choose to use FO4 at the last moment if he prefers to increase his cost than to take the risk of over-stocks.

4. MAN has the possibility to use a subcontractor (SUB) for solving capacity problems. The use of this subcontractor is to balance with the use of 2 or 3-shifts-works, or with extra-hours. It is clear that this choice will depend on the respective costs of these solutions but also 1) on the delays required to use these possibilities (making extra-hours takes less time than sub-contracting) 2) on the expected demand on the following periods (extra-hours can be used for absorbing a punctual over-load whereas increasing the number of shifts must be done on a longer horizon).

5. The painter PAINT aims at also giving to MAN a role of supplier. This case allows to deal with a collaborative process and to assess different information sharing modes.

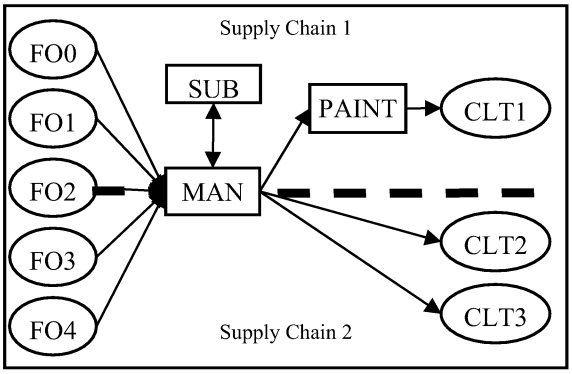


Figure 1: SC Typical Structure

It is clear that the relative values of the parameters have a crucial influence on the interest of a decision compared to another. For instance, the interest of the decision to sub-contract or to make extra-hours depends on the relationship between the cost of sub-contracting, the cost of extra-hours and the carrying costs. What we want to illustrate here is that it also depends on the respective dates when these decisions must be made, and so on their respective "cycle times": making early decisions may allow to minimize the expected cost, but this optimality can be set into question if the real demand differs from the expected one. In that context, we remind that our aim is not here to define an optimal policy on the base of the suggested example, but to define a generic support system which use can be extended to other cases.

Information Flow

In the industrial cases that we studied, the actors do not take into account tier 2 suppliers: their strategy is rather to select their suppliers according to their maturity in production management, allowing them to manage themselves efficiently their own suppliers. This point is an illustration that a centralised management of the SC is seldom possible when the companies are not belonging to the same group, both because of the complexity that it requires and of the autonomy of each entity. As a consequence, we focus here on a peer-to-peer point of view: a company is considered, along with his relationships with his customers and suppliers, on the base of the aggregated information available at a certain period. Thus, each actor acts as an autonomous decision centre and is not aware of the internal processes of the other entities. Therefore, its visibility is limited to the first stage upstream and downstream the material flow. We assume that the demand is known at a reference period and that the planning is actualised following a periodicity which definition is a part of the study.

ANALYTICAL MODEL

In this model, we mainly focus on the manufacturer point of view. The aim of the model is to help the decision maker to choose the best decisions among different possibilities: produce with internal resources or subcontract, adapt capacity by using extra-hours or by increasing the number of shifts, accept backorders, etc.. Each kind of decision is characterized by its anticipation delay and by its cost. The planning process is carried out dynamically, following a rolling-horizon procedure.

We present first the static tactical model that is used at each step of the dynamic planning process and then show how anticipation delays are taken into account through specific frozen sub-horizon within the planning process.

Tactical Planning Model

At each step, this tactical planning model helps deciding what actions can be done in order to satisfy the final product demand implying the purchase of components at a minimal cost. Final products are distinguished by index “p” and components by index “c”. We consider a planning horizon of “T” periods. So, at step τ of the planning process, the planning horizon “HP” covers period τ to period τ+T-1.

The model integrates the internal production cycle time noted “LP” of the manufacturer and the cycle time “LS” of the subcontractor. To be realistic, the planning horizon must be longer than these cycle times.

We introduce the following decision variables:

$X_{p,t}$: internal production of the manufacturer.

$ST_{p,t}$: subcontracted production.

HS_t : extra-hours.

$B_{a,t}$ (binary variable) = 1 if action “a” for increasing capacity is activated, 0 otherwise.

$I^+_{p,t}$; $I^-_{p,t}$: inventory and backorder levels for final products.

$J_{c,t}$: component inventory level.

$L_{p,t}$: effective deliveries of final products

$A_{c,t}$: component’s supplies ordered by the manufacturer

The model is based on the followings data:

CN is the normal internal capacity allowed at every period.
{a} represents the set of actions that may be activated to adjust the capacity of the manufacturer (i.e. 2 or 3-shifts-work).
$D_{p,t}$ indicates the global demand of each final product resulting from all the customers.
R_p represents the unit processing time of each final product.
$\alpha_{p,c}$ models the bill of materials linking final products and component consumption.

The Planning model is as the follows:

$$\max \sum_{t=1}^{T} \left[\begin{array}{l} \sum_p v_p L_{p,t} - \sum_p h_p I^+_{p,t} - \sum_c hc_c J_{c,t} \\ - \sum_p b_p I^-_t - \sum_c f_c A_{c,t} - \sum_p u_p X_{p,t} \\ - \sum_p st_p ST_{p,t} - \sum_a o_a B_{a,t} + sHS_t \end{array} \right] \quad (1)$$

Subject to:

$$I^+_{p,t} - I^-_{p,t} = I^+_{p,t-1} - I^-_{p,t-1} + X_{p,t-LP} + ST_{p,t-LS} - D_{p,t} \quad (2)$$

$$\sum_p R_p X_{p,t} \le CN + \sum_a (B_{a,t} \times SC_a) + HS_t \quad (3)$$

$$J_{c,t} = J_{c,t-1} - \sum_p \alpha_{p,c} (X_{p,t} + ST_{p,t}) + A_{c,t} \quad (4)$$

$$\sum_p \alpha_{p,c} (X_{p,t} + ST_{p,t}) \le J_{c,t-1} \quad (5)$$

$$L_{p,t} = D_{p,t} + I^-_{p,t-1} - I^-_{p,t} \quad (6)$$

$$HS_t \le HSMax \quad (7)$$

The objective function (1) maximises a simplified profit, defined as the cash flow generated by the sale of the final products minus the costs. "v_p" represents the unit selling price of a final product, h_p, hc_c, u_p, st_p, b_p, f_c, o_a, s, are the unit costs associated to the relevant decision variables.

Equation (2) links production quantities (subcontracted or not) and the inventories or backorders. Note that, due to the cycle time, production quantities launched at period t-LP or t-LS are available in period t.

The production allowed at every period is limited by the production capacity. To model this capacity (see Equation (3)), we use a standard capacity "CN" available at each period, to which we can add extra-hours "HS" per period. The actions in {a} can be activated by the binary variables "$B_{a,t}$". This allows an overcapacity "SC_a" specific for each action. This equation reflects the resource sharing between products.

In equation (4), components inventories are calculated according to the production (subcontracted or not) and the bills of materials coefficient "$\alpha_{p,c}$".

Equation (5) introduces a limit on the final product production that cannot exceed the component stock at period "t-1".

Equation (6) permits the calculation of the deliveries "L" of product "p" at each period "t".

Equation (7) limits the use of extra-hours "HS_t" according to a maximum value "HSMax".

All these constraints are defined for any period of the planning horizon i.e. from period t=τ to period t=τ+T-1 (τ is the reference period of the step τ in the planning process)

Anticipation Delay, Frozen Horizon and Dynamic Planning Process

The tactical planning model described hereinbefore provides assistance in the manager's decision making process giving him the best possible actions at a specific planning step. Nevertheless, the need for anticipation of some decisions has to be taken in account. An anticipation delay associated to each decision is marked by freezing a part of the decisions taken in previous planning steps. These temporal concepts must be integrated in the previous planning model in order to obtain a simulation framework that makes this planning model operational in a dynamic context.

For instance, if a subcontractor requires to be notified of a supply order 3 periods in advance, the *Anticipation Delay* associated to subcontracting decisions is 3 periods. So, at step τ of the planning process, the manufacturer cannot change decisions concerning its subcontracted production for, at least, the three first periods of the planning horizon. This sub-horizon is the *Frozen Horizon* for subcontracting decisions. Nevertheless, the manufacturer can freely decide subcontracting quantities on the periods following this *Frozen Horizon*.

Moreover, the dynamic planning process is carried out with a *Planning Periodicity* noted "PP" that implies that a part of the decisions made freely at a specific planning step will be frozen and replicated in the next planning step. An illustration of this Dynamic Planning Process for subcontracting decisions is given by Figure 2. Hence, we see how the first planning periods are frozen and how the decisions made in the previous planning periods influence the new decisions to be made.

In this example, we have a Planning Periodicity "PP" of 2 periods, a length of Planning Horizon T=6 and a *subcontracting Frozen Horizon* FH=3. The first planning step starts at period τ (reference period). The decisions relative to the three following periods are reported from the former planning. The decisions relative to periods τ+3 and τ+4 are decided freely in this planning but will be conserved in the next planning step.

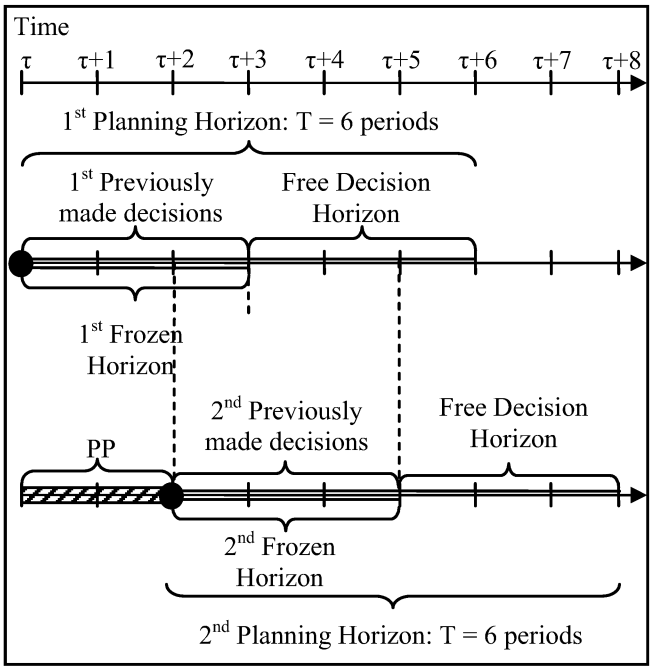


Figure 2: Dynamic Decision Freezing Process

In order to model this dynamical phenomenon, we associate a specific *Frozen Horizon* "FH_k" to each decision k. Additional constraints (8), (9), (10), (11) and (12) force decisions made at the reference period "τ" to be equal to the decisions at the reference period "τ-PP" over the relevant sub-horizon. The upper index refers the reference period (i.e. the planning step). Our model aims at providing a decision support proposing which are the best decisions to made fitting these *Frozen Horizons*.

$$X_{p,t}^{\tau} = X_{p,t}^{\tau-PP} \quad \forall p, \forall t \left| \tau < t < \tau + FH_X \right. \qquad (8)$$

$$ST_{p,t}^{\tau} = ST_{p,t}^{\tau-PP} \quad \forall p, \forall t \left| \tau < t < \tau + FH_{ST} \right. \qquad (9)$$

$$HS_{t}^{\tau} = HS_{t}^{\tau-PP} \quad \forall t \left| \tau < t < \tau + FH_{HS} \right. \qquad (10)$$

$$A_{c,t}^{\tau} = A_{c,t}^{\tau-PP} \quad \forall c, \forall t \left| \tau < t < \tau + FH_{A} \right. \qquad (11)$$

$$B_{a,t}^{\tau} = B_{a,t}^{\tau-PP} \quad \forall a, \forall t \left| \tau \le t \le \tau + FH_{Ba} \right. \qquad (12)$$

RESULTS

Numerical experiments have been done in order to validate the model, the dynamic approach and to assess planning performance. We describe, in this section, three typical simulations characterising the importance of the concepts presented above and integrated in the model. To solve the model, we used the optimisation software Xpress-MP.

These first experiments are based on simple examples with deterministic demand. We firstly run the planning model over a horizon of 20 periods, assuming that demand over the entire horizon is known at the beginning of the horizon. Hence, we obtain a "reference plan". We then apply a dynamic planning process with a planning horizon T of 8 periods and a periodicity of 2 periods. The resulting plan is composed of the successive frozen decisions. In order to highlight the interest of taking in account a specific frozen horizon for each decision variable, we evaluate the resulting plan regarding variation of the frozen horizons on backlogs, subcontracting and use of extra-hours indicators.

In all experiments, cycle time values are constant (LP=1, LS=3). For clarity reasons, we do not show the internal manufacturer's production and the inventory position in the figures, and focus on backorders, subcontracting quantities and Extra-hours. Extra-hours and subcontracting are respectively limited to 10 and 100 in order to be more realistic. Costs have been chosen such as the use of extra-hours is more expensive than subcontracting but, on the opposite, anticipation delay of subcontracting is longer than the one for extra-hours.

Figure 3 shows the results of the first simulation. The visibility on the whole demand allows to anticipate the subcontracting decisions sufficiently in advance to face the demand pick at period 10. Backorders and use of extra-hours are completely avoided as they are the most expensive decisions.

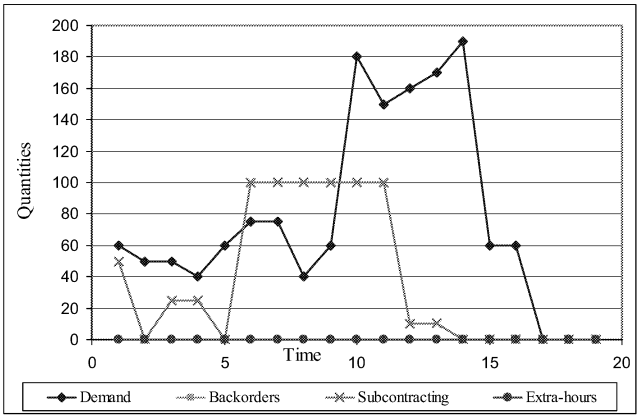


Figure 3: Reference Plan

As a second example we defined the Frozen Horizons as follows: FH_{HS}=1 and FH_{ST}=3. The results of the implemented decisions can be seen in Figure 4. We can see that, compared to Figure 3, extra-hours are used and backorders generated. This is due to the lack of visibility for the manufacturer which cannot anticipate enough with a cheap decision (subcontracting) and use a more reactive solution (extra-hours) to adapt its capacity.

The following simulation points out the impact that could have a Frozen Horizon not correctly adapted to the planning periodicity and horizon. We only changed the frozen horizon of the subcontractor to 4 periods. We can see in Figure 5 a significant increase in backorders, extra-hours used and irregularities in the orders transmitted to the subcontractor. This is due on the one hand to the lack of visibility of the manufacturer on the demand and, on the other hand, to a lack of reactivity in decision making induced by the frozen horizon.

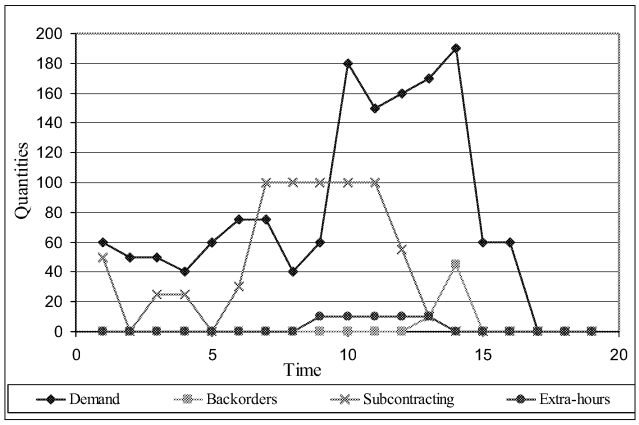


Figure 4: Second Simulation Resulting Plan

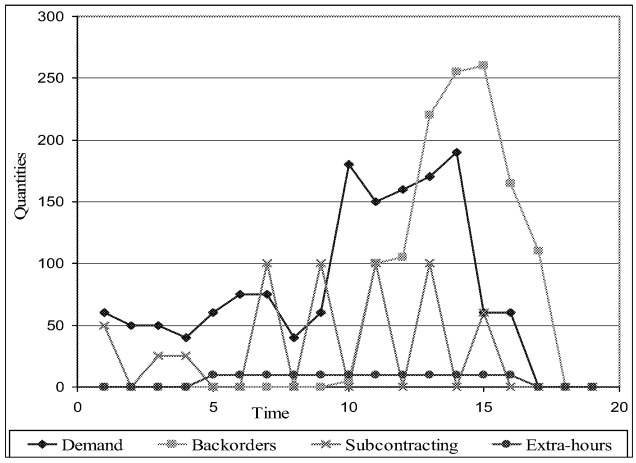


Figure 5: Third Simulation Resulting Plan

In a context in which a rolling horizon planning process aims to balance the lack of visibility on firm orders, these results show that the quality of the implemented production plan can be clearly affected by the relative lengths of the temporal parameters of the model (the *Anticipation Delay*s, and the associated *Frozen Horizons,* the *Planning Periodicity, the planning horizon, etc.)*. In that way, the optimisation model included in the framework provides optimal solutions assisting the manager and giving him guidelines to decide how to elaborate its plan. Moreover, the evaluation of different parameters configurations in order to find which decision is critic can result in improvement actions to gain reactivity if it is impossible to enlarge the planning horizon and the visibility on the demand.

CONCLUSION AND PERSPECTIVES

Efficiently manage the material flow in a Supply Chain requires successive decision makings which are made difficult by the inherent uncertainty on the demand. As it is often done in industrial cases, we suggest here to manage this uncertainty through a periodic refresh of the planning, including the delays required for using the degrees of freedom available for the manager. In order to provide an efficient decision support to the managers, we suggest a simulation framework based on Mixed Integer Linear Programming integrating 0-1 variables enabling the adjustment of production capacity through time.

As a next step, we are now interested in an explicit management of the demand uncertainty using our framework to evaluate a forecasting model based on a flexible horizon in which the demand can vary within an interval agreed between the manufacturer and its customers.

In that way, the framework will serve to evaluate the performance of several planning strategies based either on an "optimistic" approach integrating the maximum values of the possible orders or on a "pessimistic" approach integrating the minimum values. Thus, it will provide a risk assessment that can influence the manager behaviour.

REFERENCES

Bartezzaghi, E. and Verganti, R. 1995. "Managing demand uncertainty through order overplanning". *International Journal of Production Economics*, No. 40, 107-120

Bodt, M. A., and Van Wassenhove, L. N. 1983. "Cost increased due to demand uncertainty in MRP lot sizing". *Decision Science Journal*, No. 25, 345-362.

Brennan, L. and Gupta, S. M. 1993. "A structured analysis of material requirement planning under combined demand and supply uncertainty". *International Journal of Production Research*, No. 31 (7), 1689-1707.

Fargier, H. and Thierry, C. 2000. "The use of possibilistic decision theory in manufacturing planning and control: recent results in Fuzzy Master Production Scheduling". In *Scheduling under fuzziness*, R. Slowinski and M. Hapke (Eds.). Springer-Verlag, 45-59.

Génin, P. 2003. "Planification Tactique Robuste avec Usage d'un APS. Proposition d'un Mode de Gestion par Plan de Référence" Ph.D. Thesis, École des Mines de Paris

Guide Jr, V. D. R. and Srivastava, R. 2000. "A review of techniques for buffering against uncertainty with MRP systems". *Production Planning and Control*, No. 11, 223-233.

Grabot, B. Geneste, L. Reynoso Castillo, G. and Vérot, S. 2005. "Integration of Uncertain and Imprecise Orders in the MRPII method". *International Journal of Intelligent Manufacturing*, 16(2).

Grasso, E. T. and Taylor, B. W. 1984. "A Simulation Based Experimental of Supply/Timing Uncertainty in MRP Systems". *International Journal of Production Research*, No. 22, 485-497.

Ho, C. J. and Ireland, T. C. 1998. "Correlating MRP system nervousness with forecast errors". *International Journal of Production Research*, No. 36 (8), 2289-2299.

Johansen, S. G. 1999 "Lot sizing for varying degrees of demand uncertainty". *International Journal of Production Economics*, NO. 59 (1-3), 405-414.

Krupp, J. A. G. 1997. "Safety Stock Management". *Production and Inventory Journal*, 3rd quarter, 11-18.

Van Landeghem, H. and Vanmaele, H. 2002. "Robust planning: a new paradigm for demand chain planning". *Journal of Operations Management*, No. 20, 769-783

Whybark, D. C. and Williams, J. G. 1976. "Material Requirements Planning Under Uncertainty". *Decision Sciences*, No. 7, 595-606.

Zhao, X. and Xie, J. 1998. "Multilevel lot-sizing heuristics and freezing the master production schedule in material requirements planning systems". *Production Planning and Control*, Vol. 9, No. 4, 371-384.

THE IMPACT OF VARIOUS DELIVERY PRIORITIZATION STRATEGIES IN HETEROGENEOUS SUPPLY CHAIN ENVIRONMENTS

Thomas Thron, Gábor Nagy, Niaz Wassan
Kent Business School, University of Kent
Canterbury, KENT CT2 7PE, UK
Tel.: (+44)1227 827726 Fax.: (+44)1227 761187
Email: T.Thron@kent.ac.uk

KEYWORDS
Supply Chain Management, Delivery Priority, CPFR, discrete-event simulation

ABSTRACT

In this paper, discrete-event simulation is used to point out how varying delivery priority ranking strategies amongst customers within a heterogeneous Supply Chain featuring traditional reorder-point (ROP) and CPFR driven replenishment affect global and individual performance of market participants.

INTRODUCTION

To tackle the matter of supply chain inefficiencies due to a lack of unflawed demand information academic research as well as companies' field studies aim to develop new replenishment approaches which are based on the general idea of information sharing and preferably get along without traditional routine orders. The basic idea is to approach the entire distribution chain as a replenishment-driven process based on reliable forecasts rather than an uninformed carriage system based on rather vague order prediction (Simchi-Levi, 2002). The most common and widespread of such automated replenishment approaches are vendor managed inventory (VMI) and collaborative planning forecasting and replenishment (CPFR). In a VMI partnership the manufacturer (or any other supplier/distributor) makes the main replenishment decisions for the consumer. In practical terms the manufacturer monitors the customer's inventory levels and decides about periodic replenishment actions regarding delivery time and quantity (Waller et al., 1999). A particular advantage is the possibility for the manufacturer to recognise and possibly delay non critical replenishments to smooth demand peaks and rank delivery requests according to actual priority (Kaipia et al., 2002).
CPFR is an advancement of VMI and describes the sharing of forecast and related business information among business partners in the supply chain to enable automatic product replenishment (VICS, 1998).
Collaborative SCM research tries to address the issues arising out of a lack of unflawed demand information but has however mainly focused on either the relationship between one vendor and one customer or the ideal situation of a manufacturer collaborating with all its downstream partners. This evidently provides only limited support to various companies struggling with different more practical problems which are commonly less homogenous. For example, many companies will only selectively chose trustworthy partners to engage in collaboration due to a fear of revealing vital information to too many market participants. The same accounts for a typical medium sized company that is just not in the strategic position to be able to compel its customers into some kind of collaboration, it rather has to adapt to requirements and demand of the bigger players. As such it cannot rely on either a pure ROP or a universal collaborative scenario. In such and many other circumstances a heterogeneous replenishment approach will be necessary. The idea of this research is thus to emphasize on these conditions which constitute an intermediate approach in between traditional isolated echelons and fully transparent data-sharing systems which is a field that has commonly been neglected by SCM research.
This particular paper will take a look on prioritization strategies within the delivery process between a manufacturer and several customers that either share a common demand forecast via CPFR or operate autonomous, thus placing random orders without prior knowledge of the manufacturer. The analysis will evaluate the outcome of several common performance measures to conclude about particular strategies being more or less recommendable.

METHODOLOGY

The scenario that is to be used for several analyses is based on a business framework and thus real demand/delivery data of an actual medium-size food-company and its three major customers (retailers). Primarily it was created to evaluate possible benefits that arise from participation in data-sharing systems, thus using additional demand and inventory data from a varying number of customers. The company the data was obtained from has recently established a VMI/CPFR replenishment system with one of their major customers implying full demand data disclosure on a weekly basis on store level as well as on distribution centre level. So far the benefits of this cooperation could be experienced in a somewhat higher service level outcome as well as less and shorter delivery fulfilment delays. Nevertheless the manufacturer is not too euphoric since the additional cost for entering the retail companies' supplier data system were quite substantial and approaching a replenishment system from a VMI/CPFR perspective made

several changes within production scheduling and further investments into IT equipment necessary.
Considering the ongoing investigation the company was particularly interested as to what extent their actual achievements could be considered as satisfactory compared to a theoretical benefit approach. Hence to point out which particular figures have theoretical potential for further efficiency improvement. Another key issue was to find out to what extent the company could further gain from additionally including their other two key customers into collaborative replenishment approaches.
The simulation model that is about to be used has been laid out in a way to reflect the market conditions of this particular company. Nevertheless it can be considered as general and flexible enough to represent a typical supply chain system and thus should be applicable to model market conditions for a variety of other enterprises of similar size and delivery structure. A graphical representation of the situation can be seen in figure 1.

THE SIMULATION MODEL

The simulation model used for the investigation consists of three suppliers serving the manufacturer which is then serving the distribution centres (DCs) of its three major customers. Moreover, these DCs serve several retail outlets each. Since the three involved retailers account for about 80% of the manufacturer's total production, several less significant customers are omitted from the analysis due to reasons of simplicity and clarity.

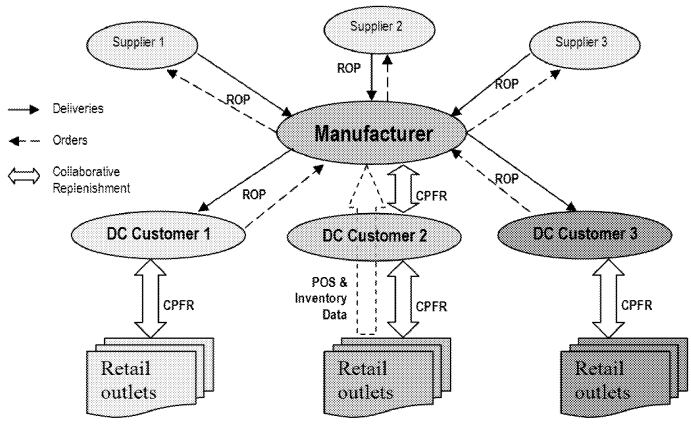


Figure 1: Outline of the modelled Supply Chain (one CPFR-customer scenario)

Even though it is part of the simulation model, the investigation will not discuss any issues arising out of the relationship between the manufacturer and its three suppliers of raw material. Data that is used to run this part of the system is rather taken as a given.
Within the distribution framework between manufacturer and retailers, deliveries for each particular customer are either scheduled via a traditional reorder policy (ROP) or follow a CPFR approach.
The ROP case simulates a scenario where order requests are triggered by the retailer according to inventory level dropping below a particular lower bound without prior knowledge of the manufacturer regarding point of time or quantity. Within this scenario the manufacturer has to rely only on the standard demand as to say average sales throughout the year with some seasonal adjustments to face e.g. somewhat higher demand in winter and lower in summer as it is the case for the investigated product group. Production scheduling thus has to depend on past experience and a certain amount of estimation and does not involve recent sales information.
The CPFR approach simulates a scenario where sales data on store level as well as inventory levels at retail outlets and distribution centres are available to the manufacturer on a weekly updated basis. Furthermore replenishment requests are triggered by DC-inventories reaching a particular reorder point that is determined by a sales forecast based on past demand data, detailed promotion schedules, seasonal factors and most recent available sales data.
The retail outlets in the simulation model are always connected with their DC via a VMI/CPFR replenishment system. The delivery is on a daily request basis with average replenishment approximately once a week. This can be considered as common practice for many kinds of products and within most major retail companies (VICS 1999, Holmström et al. 2002). Thus the replenishment strategy between DC and outlets will not be further analysed.
Within the system described above the focus of the following analysis will lie on how different ways of ranking delivery priority amongst several customers that are replenished via either ROP or CPFR will affect numerous performance measures that investigate vital efficiency targets.
The demand simulation part of the model is based on three years of actual sales data that have been obtained from the three involved retailers. The sales data under consideration is characterised by a certain volatility in demand behaviour, remarkable seasonalities and particularly intensive promotional activities which are mainly initiated by the retail-companies. Based on this market intelligence a sales forecast system was implemented that served the CPFR part as basis to estimate replenishment points and necessary quantities. For further details about the implemented forecasting approach see Thron (2004a).
Due to a generally lean inventory policy and high delivery frequencies between manufacturer and customers there is a severe amount of delivery request clashes where available product supply does not match the total requested quantities. Actual request/delivery records of the company show that more than half of the incoming delivery-requests cannot be fulfilled instantly out of inventory and have to be delayed to some (often very short) extent. The company is thus very much interested in improving its delivery priority ranking system to minimize delays and service level gaps especially in the light of introducing CPFR systems with at least one of their customers. Increasing safety stock is not considered as an option for the time being due to storage capacity restrictions.
The actual priority ranking system classifies delivery requests into several levels of urgency based on outlet & DC inventories and forecasted demand data respectively in case it is available via CPFR. The data are then used to set delivery priorities in case of insufficient inventory to minimize total loss of sales. Rankings vary in between 0 (extremely urgent) to 1 (non-crucial). In case of ROP

replenishment-policy incoming requests are either served on a first come first serve basis (impartial fulfilment) or according to a pre-assigned priority ranking which commonly favours the larger customer for the smaller. Within heterogeneous scenarios (CPFR+ROP mixed) ROP orders are commonly preset to average urgency due to a lack of information about their real importance.

The manufacturing company that provided most of the data for this research is a medium sized enterprise with an annual turnover of fifty million pounds. It manufactures a particular kind of product in various forms and shapes and additionally in various sizes. The product that was chosen for the analysis was taken out of a set of data that contained sales information about the top five selling SKUs of the particular manufacturer. Altogether sales behaviour from all products within this group is almost identical. Replenishments are commonly made ordering the entire set of products in similar amounts and any kind of sales promotion always affects the entire set. Thus the analysis focuses on one particular SKU instead of the whole set since several test runs with each of the other four products resulted in a virtually identical outcome.

The company is using a weekly production cycle which adjusts production to expected forthcoming demand within the following four weeks. Production is capacity constrained since overtime work can only be applied to a limited amount due to union agreements and can commonly only be set up at the beginning of the week. As mentioned before the inventory will be set on a rather lean level as this goes along with the company's long term inventory policy that average stock-level of each inventory class A SKU throughout the year must not exceed an average week's demand. According to several interview sources this apparently is a very typical approach within that industry. Thus, this limitation will hold for the case of ROP and CPFR.

The simulation software used was Simul8 R11 (Simul8 Corp, 2004). The model is prior to each run set up with sufficient startup inventories to be able to fulfil replenishment respectively incoming orders. Minimum fill-loads were set to 70% of originally requested quantities. To reduce most of the initialization-bias and assure an unflawed observation result-collection is initiated after a 13 weeks warm-up period.

SIMULATION - OUTCOME

There can be generally no doubt about that a complete change from ROP to CPFR delivery strategy results in tremendous gains of efficiency, meaning higher service level or less and shorter delivery delays. Just to see how significant the impact is altogether and particularly within each stage of the supply chain from the manufacturer through the distribution centres to the retail outlets we will first evaluate the results taken from the two "pure" scenarios. These are solely ROP replenishment on one hand and pure CPFR on the other.

Table 1: Results for two initial ROP and CPFR delivery

			All ROP+ impartial fulfilment	All ROP+ Prioritized fulfilment	All CPFR
Service Level outcome		*C1*	1.88%	1.34%	0.40%
		C2	6.83%	0.79%	0.64%
		C3	0.88%	6.03%	0.32%
		global	**4.15%**	**2.08%**	**0.50%**
DC-Replenishment	*denied request ratio*		1.55	2.20	0.68
	% non – delayed deliveries	*C1*	45.56%	40.68%	66.67%
		C2	31.71%	32.12%	55.64%
		C3	53.21%	37.69%	73.74%
		global	**43.70%**	**36.81%**	**65.08%**
	% deliveries delayed over 4 days	*C1*	12.23%	15.49%	3.93%
		C2	38.31%	10.92%	9.12%
		C3	6.86%	32.22%	3.50%
		global	**18.58%**	**18.82%**	**5.61%**
Retail-Outlet Replenishment	*% delayed*	*C1*	1.75%	1.81%	0.51%
		C2	8.43%	0.83%	0.69%
		C3	0.87%	5.38%	0.36%
		global	**3.67%**	**2.62%**	**0.52%**
	% delayed over 4 days	*C1*	0.44%	0.31%	0.22%
		C2	2.88%	0.26%	0.26%
		C3	0.31%	2.34%	0.22%
		global	**1.21%**	**0.95%**	**0.23%**

The table contains two columns for ROP, one featuring impartial fulfilment meaning a first come first serve policy in case of request clashes. The second featuring prioritized fulfilment as it was the case at the particular company the analysis was carried out for. Here customers' delivery priority was set according to importance for the manufacturer's business (in particular C2 > C1 > C3). Thus a larger retailer had priority about a smaller one.

From the data we can see that impartial fulfilment seems to disadvantage the largest customer (C2). Switching to prioritized fulfilment which ranks e.g. C2 higher than C1 not surprisingly leads to a major performance increase for C2 and consequently a decline for C1&C3. As a result the prioritized setting has overall a clear advantage regarding service-level outcome whilst the impartial-fulfilment setting results in slightly better delay-figures. A total implementation of CPFR on the other hand reduces delays and service-level gaps for all three retailers absolutely substantially: The global service level gap for example reduces at least by the factor 4. A similar picture arises for the distribution centre replenishment. Global figures show a significant increase in non-delayed deliveries and a reduction of extensively delayed deliveries. The positive impact of CPFR continues once more at the retail level. Due to the fact that all three retailers already manage the distribution between their DCs and the retail outlets via a CPFR approach, differences between ROP and CPFR scenarios are not as substantial as in previous stages.

As it can be further seen from the above table, delivery priority has an immense impact on the performance measures of an individual company. Comparing the outcome of both ROP scenarios further reveals that global figures are not all that different. Thus choosing a particular priority approach does not seem to concern the manufacturer too much. Priority rankings should as a result be of major importance mainly to particular customers since they are much more affected by this decision.

The impact of Delivery Priority Settings

Something that is practically just as important as to demonstrate what an entire CPFR enabled supply chain is capable off is the evaluation of a heterogeneous supply chain framework where CPFR and ROP replenishment strategies coexist.
Since a general analysis on benefits and threats for particular customers under certain ROP/CPFR settings has already been undertaken in a previous paper (Thron, 2004b), the following investigation will concentrate on another issue that arises within heterogeneous supply systems: prioritisation of certain deliveries/customers.
This topic involves for once the delivery preferences a manufacturer may have regarding his several customers. This may be due to particular agreements, traditional habits, size and importance of a customer or simply monetary incentives. Due to any of these reasons, given the case that actual supply cannot fulfil the incoming demand, a particular customer's delivery request may get priority over another.
Supplementary to that there is the case of impartial delivery where distribution of a scarce product is just done on a random basis (i.e. first come first served).
In a ROP driven delivery scenario the manufacturer has commonly no reliable information about the actual importance and urgency of an incoming delivery request. Hence in case of limited inventory and thus request clashes, deliveries are either assigned on a preferential customer basis as explained above or simply on a random basis.
One of the main advantages of a CPFR orientated system is the possibility to assign deliveries not just on a random or prioritized basis but according to predicted demand and actual urgency. This is due to the increased demand visibility that is obtained through data sharing. When engaging with several customers in a CPFR relationship, a manufacturer should be able to gather and evaluate recent demand information from store level together with inventory data from store and distribution centre level. Taking this information into account a production forecast can be set and deliveries can be assigned according to their real level of urgency. In case of delivery request clashes this makes it possible to send scarce products to the place where they are needed the most and thus improve perceived delivery performance and avoid service level gaps as much as possible.

Heterogeneous Systems

Once it comes to heterogeneous delivery structures there is additionally the question of delivery prioritisation between ROP and CPFR customers. Since in this case information regarding the delivery urgency is available from some customers but not from others the question arises which priority strategy is the most successful one to obtain the best possible supply chain performance. Since any particular customer would obviously always interfere to its own best the following analysis is from the viewpoint of the manufacturer that aims to balance between the certain retail companies to achieve best possible supply chain performance.

To evaluate the impact certain preferential combinations have on individual performance measures we created five settings:

1. ROP customers always have priority
2. ROP advantaged – here ROP deliveries are set to be preferred from CPFR unless a CPFR request is classified as particularly urgent
3. Balanced scenario – ROP requests are set to the mean level of urgency of a standard CPFR request
4. CPFR advantaged – here CPFR deliveries are set to be preferred from ROP unless a CPFR request is classified as rather uncritical
5. CPFR customers always have priority

These five settings were tested in a variety of scenarios:

1. Single CPFR customer + 2 ROP customers

Here again there was a distinction made as to how the delivery requests were handled between the two ROP customers.

1.1 Impartial delivery (first come first serve)
1.2 Prioritized delivery (larger customer prioritised)

For both scenarios we analysed all three possible cases (customer 1-CPFR & C2,C3 ROP; C2-CPFR & C1,C3 ROP; C3-CPFR & C1,C2 ROP)

2. Two customers CPFR + 1 ROP

The possible scenarios for this setting were C1,C2-CPFR & C3-ROP; C1,C3-CPFR & C2-ROP and finally C2,C3-CPFR & C1-ROP.

Altogether there were nine different scenarios that were then tested for each of the five above defined priority settings. Together with the three reference scenarios (2x pure ROP, pure CPFR) 48 different simulation setups were created and were run 20 times each to obtain sufficient confidence intervals for the results.

In order to shrink down the fierce amount of output data and to present the obtained outcome in an understandable way a marking system was introduced that awarded +1 point for a particular good outcome compared to the other 4 settings and -1 for a rather inferior outcome. Moreover a mark of -2 is given to settings that resulted in a practically unacceptable outcome and thus complied with certain disqualification rules. These particularly were defined as

a) Substantial drop in one of the major performance measures (service level, non-delayed deliveries and long delayed deliveries over 4 days) for a CPFR customer compared to its prior ROP outcome.
b) Major drop of several performance measures for any ROP customer beyond a reasonable point.

The reason for red-flagging these cases is due to practical business conventions that just make it unacceptable that for example a customer engages into a CPFR relationship and as a result of changed priority ranking gets off significantly worse regarding several major performance measures. Settings containing such cases in any of their underlying performance measures are marked red and should be considered with great caution even though the actual figure will be stated for reasons of comparison.

The following tables are classified into *Service-Level, DC-Delivery* and *Retail-Level* which stand for groups of performance measures. These are evaluated for the five above defined priority-settings and according to the three retail-companies' replenishment methods (ROP or CPFR). Numbers in a particular row/column represent thus a cumulated judgment from several underlying results.

Service-Level was taken as a measure of gaps in inventory at store level and therefore certainly lost sales and customer goodwill. The *Service-Level* figure includes global service level outcome for all three customers combined as well as individual figures for each of them.

DC-Delivery expresses a combination of several delivery performance measures for the distribution between the manufacturer and the customers' distribution centres. Included are global figures as *denied request ratio* that stands for the number of denied requests per delivery that was carried out successfully, or figures for each individual customer including *immediately fulfilled delivery requests* and *long delays of over four days.*

Retail-Level finally cumulates replenishment performance measures at store-level. These are again delivery failures or (long) delayed transfers.

The lower part of each table accumulates the individual scores to give a combined, more general view.

Scenario 1: Single CPFR customer + 2 ROP customers - Impartial delivery among ROP customers

C1 CPFR	ROPprio	ROPadv	balanced	CPFRadv	CPFRprio
Service Level	0	1	1	1	1
DC-Delivery	-3	-4	-2	-1	3
Retail-Level	2	3	1	0	2
C2 CPFR					
Service Level	0	1	1	0	0
DC-Delivery	5	-3	0	-1	-1
Retail-Level	3	0	0	-2	-3
C3 CPFR					
Service Level	1	2	1	1	-1
DC-Delivery	-1	0	0	2	1
Retail-Level	-4	3	0	-1	-1
TOTAL	**ROPprio**	**ROPadv**	**balanced**	**CPFRadv**	**CPFRprio**
ServiceLevel	1	**4**	3	2	0
DC-Delivery	1	-7	-2	0	**3**
Retail-Level	1	**6**	1	-3	-2

Giving priority purely to ROP customers results in this scenario in unacceptable outcome at several occasions once customer 1 engages in CPFR delivery. These cases are consequently marked red. The best possible outcome for each particular customer's scenario is marked light-green. This will indicate the recommended priority-strategy the manufacturer should choose in each particular case to maximise overall performance. If for example the manufacturer engages into CPFR with just customer 3, the altogether best possible option would be to prefer ROP deliveries from CPFR unless a CPFR request is classified as particularly urgent (Setting 2).

For the above case, the recommended setting changes for each customer and thus no particular setting seems to be significantly superior over the others. Overall we can state that the balanced approach results in the most stable outcome as to say reveals no significant weakness in any setting. It is nevertheless inferior to the CPFR priority approach regarding the DC-delivery performance. Setting 2 (ROP advantaged over CPFR) shows some interesting outcome and obtains very good scores for service-level and retail-level but seems to fall behind considering DC-delivery. Since there seem to be some complications in case C1 engages in CPFR it has to be examined carefully before considering any practical adoption.

Scenario 2: Single CPFR customer + 2 ROP customers - Prioritized delivery among ROP customers

C1 CPFR	ROPprio	ROPadv	balanced	CPFRadv	CPFRprio
Service Level	1	1	-1	1	1
DC-Delivery	2	-6	-3	-2	2
Retail-Level	3	-2	-3	-2	-1
C2 CPFR					
Service Level	-1	1	0	0	-1
DC-Delivery	5	-3	-4	-4	0
Retail-Level	3	3	0	0	0
C3 CPFR					
Service Level	-1	1	2	2	-1
DC-Delivery	-2	1	-1	1	-1
Retail-Level	-2	2	6	3	1
TOTAL	**ROPprio**	**ROPadv**	**balanced**	**CPFRadv**	**CPFRprio**
ServiceLevel	-1	**3**	1	**3**	-1
DC-Delivery	**5**	-8	-8	-5	1
Retail-Level	**4**	3	3	1	0

The overall performance in this scenario seems to be rather inferior compared to the other two scenarios. We can see again some major drawbacks in certain circumstances where setting 1&2 result in unacceptable outcome. The recommended settings for C1/C3 engaging in CPFR are settings 1&4 respectively. For the scenario of customer 2 engaging in CPFR no reasonable recommendation can be given since none of the five settings results in an unflawed outcome.

Altogether Setting 5 (CPFR always priority) seems to be the most stable one here but with certainly inferior individual results. Setting 1 even though it has one red flagged case seems to have by far the best overall outcome and makes it thus worth considering as long as the single unacceptable case is not an issue. Altogether every setting has its strengths and weaknesses in different fields. Hence picking one of the settings as favourite should be done according to individual preference as to decide which performance measure group is considered as most important and according to which particular customer engages into CPFR.

Scenario 3: Two CPFR customers + 1 ROP customer

C1&C2 CPFR	ROPprio	ROPadv	balanced	CPFRadv	CPFRprio
Service Level	-4	0	2	1	1
DC-Delivery	-3	0	0	3	3
Retail-Level	-3	1	5	2	3
C1&C3 CPFR					
Service Level	1	1	0	1	-1
DC-Delivery	-1	-1	-4	-1	5
Retail-Level	0	0	4	4	-2
C2&C3 CPFR					
Service Level	-1	2	2	2	0
DC-Delivery	3	-4	1	1	0
Retail-Level	0	0	4	3	2
TOTAL	**ROPprio**	**ROPadv**	**balanced**	**CPFRadv**	**CPFRprio**
ServiceLevel	-4	3	**4**	**4**	0
DC-Delivery	-1	-5	-3	3	**8**
Retail-Level	-3	1	**13**	9	3

In an advanced stage of having two out of three customers engaged in a CPFR process, the CPFR-favouring settings 4&5 outperform by far the ROP-favouring settings 1&2.

Overall best scenario is clearly setting 4 which prefers CPFR deliveries to ROP unless a CPFR request is classified as rather uncritical. This setting is among the recommended settings for all three different customer replenishment scenarios.
Setting 3 and 5 obtain remarkably good results as well but only for specific scenarios.
Altogether we can state that settings 3&4 on one hand and 5 on the other have distinct strengths and weaknesses. The rather balanced approaches (3&4) perform very well for the retail-level considering service-level and delivery performance but fall back slightly on distribution-centre delivery performance. This is the actual strength of Setting 5 (CPFR always priority) which on the other hand performs just about average regarding Service-Level and Retail-Level performance. Once more the analysis reveals individual strengths and weaknesses for particular settings and scenarios but this time there is setting 4 which can without any doubt be globally recommended. Overall the difference between a good and a bad setting seem to increase when moving from one to two CPFR replenished customers. Thus choosing the right strategy is even more important.

GLOBAL RESULTS AND CONSIDERATIONS

Table 5: Overall accumulated figures

TOTAL	ROPprio	ROPadv	balanced	CPFRadv	CPFRprio
ServiceLevel	-4	**10**	8	9	-1
DC-Delivery	5	-20	-13	-2	**12**
Retail-Level	2	10	**17**	7	1
Sum	**3**	**0**	**12**	**14**	**12**

Accumulating all the above figures in one table may distort the underlying facts to some extent. Nevertheless it can be useful to point out some overall performance conclusions.
Looking at the figures we can definitely state that there are distinct strengths and weaknesses between the pure (1&5) and balanced (2,3,4) settings. The balanced clearly outperform the pure settings regarding service-level and retail-level figures. Particularly service-level measures are much better which does not come as a surprise since balancing demand fulfilment should usually lead to lower supply gaps. The particularly good outcome for retail-level performance goes along with the high service-level since a retail outlet that is delivered on time will most likely always have enough inventory to be able to cope with demand. The pure settings on the other hand reveal much better DC delivery performance such as shorter and fewer delivery delays due to the fact that some customers are always delivered instantly whilst others have to stand back and try to adjust to the lower priority. The balanced settings rather lead to a situation in which incoming customer requests somehow "block" each other which leads to fewer immediate fulfilled delivery requests and lower fill-loads which could be observed while running the simulations.
Accumulating the figures makes it seem that overall settings 3, 4 and 5 are the recommendable ones with setting 4 achieving the highest overall score. Nevertheless, the above investigation showed that each scenario has indeed very distinct characteristics respectively strengths and weaknesses.

Setting 1 in which ROP customers had absolute delivery priority proved to be the most complicated since in all three scenarios there were situations arising in which at least some of the obtained performance measures were simply at an unacceptable level. This was due to the nature of the approach to purposefully neglect CPFR demands. Keeping that in mind it was somewhat surprising that in some cases like for *single CPFR customer + 2 ROP customers and priority delivery* this setting turned out to be the overall most successful one. Altogether the performance was nevertheless often inferior in numerous scenarios, lacking a proper consideration of all supply chain performance aspects.
Setting 2 where ROP deliveries were set to be preferred from CPFR unless a CPFR request is classified as exceptionally urgent achieved remarkably good results particularly regarding service-level. Overall it revealed reasonably good results regarding retail-level delivery as well and in one scenario could be recommended as the setting of choice. However, at a global level it shows inferior performance compared to setting 3, 4 & 5 due to mostly very poor DC-delivery performance.
Setting 3 which served as a balancing scenario whereas ROP requests were always set to the mean level of urgency of a standard CPFR request and thus treated equally obtains the best overall score regarding retail-level delivery performance. In 2 out of 9 scenarios it could be recommended as the method of choice for implementation. This setting is most likely the most used in practice. Its major weakness is a rather poor DC-delivery performance which makes an unrestricted global recommendation difficult to support.
Setting 4 where CPFR deliveries were set to be preferred from ROP unless a CPFR request is classified as rather uncritical turns out to be the winner of the above analysis. For once it achieved the highest accumulated score of all five settings. Furthermore it was recommended as the method of choice in 4 out of 9 scenarios. A further strength of this setting is the homogeneity of the achieved scores which do not polarise that much towards either good service-level or good delivery performance as in the other settings.
Setting 5 in which CPFR customers generally had delivery priority turns out to perform very solid with no major shortcoming in any of the particular scenarios. It could be recommended 2 out of 9 times for possible implementation. This setting reveals the highest outcome considering DC-delivery performance thus it should be the strategy of choice for a manufacturer that values timely delivery to the customers distribution centres the most. The setting however suffers from rather poor service-level performance. It is nevertheless not unusual in practice that a buyer-supplier relationship is rather more affected negatively by a few severe delivery delay incidents and not so much by globally achieved service-level. Hence a manufacturer has to evaluate for each particular circumstance which figure is the more important one to pursue.
It seems altogether that the balancing settings 2,3 & 4 which do not setup clear priorities and rather try to balance demand requests lead in the end to higher service levels and thus less supply gaps at store level compared to the non-balancing settings. From a global point of view it is setting

4 that can most certainly be recommended for practical implementation since it leads to a very high possible service-level and thus product availability at store level which should be of most interest for manufacturer and retail companies. At the same time it results in solid performance considering DC-delivery measures.
Considering all the above cases a possible way for further improvement of actual delivery priority ranking approaches in heterogeneous ROP/CPFR systems is to take a balancing strategy but lean it somewhat towards CPFR prioritising. Thus it would not only be the particularly urgent requests that get priority over standard ROP deliveries but also averagely, rather less-critical cases. This could apparently lead to a significant improvement regarding DC-delivery performance measures combining the strengths from setting 3 and 5.
Contrary to these suggestions, particular delivery agreements with specific customers or certain issues between retailers procurement and manufacturers delivery departments can lead to a behaviour that rather focuses on maximum delivery performance between manufacturer and distribution centres instead of achieving global goals. In such circumstances setting 5 would be a very reasonable scenario.
As it could be seen from above, the discrepancies between optimising for either DC-delivery targets or service-level diminishes when moving from an all-ROP to a 1,2 or all CPFR customer scenario due to a global forecast taking over to setup delivery quantities and timings.

CONCLUSION

It has to be stated that even though some settings altogether look more stable or result in superior outcome, the actual results can very much vary depending on the particular case. Even though the above analysis resulted in the suggestion of setting 4 being the best choice overall there are plenty of individual cases which would show a superior outcome if they would use one of the other strategies.
Hence the best solution would be using a flexible approach that would first pick the optimal strategy for given market circumstances but is then also capable of adapting to different situations like adding or releasing customers or switching from ROP to engage in CPFR. Delivery priority settings could thus be dynamic to obtain the maximum benefit which would certainly be higher than just taking a standard approach like the practically most often used balanced scenario (setting 3). Especially in heterogeneous supply chain environments that are in a transition process from traditional ROP to Vendor Managed Inventory or CPFR driven systems this could be a worthwhile consideration. Hence this article shall serve as encouragement to investigate a particular market change beforehand and adjust priority strategies afterwards. The actual gains from these adoptions can and should be remarkable as the above proceedings show. On the other hand it is necessary to evaluate possible problems that hasty changes and adoptions can cause. Changing delivery priorities spontaneously and inconsiderately can cause turmoil amongst customers and lead to major inefficiencies and substantial interruptions of supply at store level. This will not just jeopardise the relationship between a manufacturer and its customers but will also cause heavy damage to the reputation of all involved trade partners.

REFERENCES

Holmström, J., Fraemling, K., Kaipa, R. and Saranen, J. (2002), "Collaborative planning forecasting and replenishment: new solutions needed for mass collaboration", *Supply Chain Management: An international Journal*, Vol. 7 No. 3, pp. 136-145

Kaipia, R., Holmström, J. and Tanskanen, K. (2002), "What are you losing if you let your customer place orders?", *Production Planning & Control*, Vol. 13 No. 1, pp. 17-25

Simchi-Levi, D., Kaminsky, P. (2002), "Designing and Managing the Supply Chain", McGraw-Hill Higher Education, 2002

SIMUL8 Corp, Boston, USA, Simul8 R11 Process Simulation Software Package (2004) – The leading PC simulation software, Webpage: http://www.simul8.com

Thron, T. (2004a), "Opportunities and Vulnerabilities of Demand Forecasting Approaches prior to and after Implementation into Computer Simulation Models", Working Paper Series, University of Kent, 2004

Thron, T. (2004b), "The impact of increasing demand visibility on order fulfilment and service level in heterogeneous ROP/CPFR environments", submitted for publication in *International Journal of Physical Distribution & Logistics Management*

VICS, Voluntary Interindustry Commerce Standards Association (1998), "Collaborative planning forecasting and replenishment (CPFR)", available at: http://www.cpfr.org

VICS, Voluntary Interindustry Commerce Standards Association (1999), "The Roadmap to CPFR: The Case Studies", available at: http://www.cpfr.org

Waller, M., Johnson, M.E. and Davis, T. (1999), "Vendor-managed inventory in the retail supply chain", *Journal of Business Logistics*, Vol. 20 No. 1, pp. 183-203

AUTHOR BIOGRAPHY

THOMAS THRON was born in Riesa, Germany and went to the Philipps-Universität in Marburg, Germany and the University of Kent, UK where he studied Management Science and Operational Research and obtained degrees as MSc and MA. Since 2003 he is pursuing his PhD at the Kent Business School. Since then he has already published several papers in journals, conference proceedings and working-paper series. His research is concerned with logistics and particularly efficiency advances through computer-simulation of supply chains.

DECISION SUPPORT SYSTEM FOR PLANT LOCATION ANALYSIS BASED ON FUZZY MODEL

Francesco Costantino
Giulio Di Gravio
Massimo Tronci
Department of Mechanical and Aeronautical Engineering
University of Rome "La Sapienza"
Via Eudossiana 18, 00184 - Rome, Italy
E-mail: giulio.digravio@uniroma1.it

KEYWORDS
Industrial plant location, territorial planning, multiple criteria model, fuzzy logic, decision support system (DSS)

ABSTRACT
Localization analysis of an industrial plant is one of the most important decisional steps of design process. In fact, even if this process is developed only once during the life cycle of the plant, it plays a fundamental role on the survival possibility of the company in the long term. The choice for a new factory positioning, on the basis of a choice of the industrial activity to realize, is generally preceded by four design phases:

- definition of plant layout and its development plan;
- definition of needing and requirements for communication accesses and connection with service plants;
- determination of energy resources quantities;
- check of safety and environmental condition of the industrial processes in the community, to best avoid the significant impacts.

The article presents the implementation of a decision support system to locate a general industrial plant, when the common language in many cases seems sufficient to describe the information that derives from the observation of the physical world and social or economic situation. The objective is to introduce fuzzy logic in territorial multiple criteria models, integrating representations of complexity and inaccuracy of available data and information.

INTRODUCTION

An industrial plant is one of the most advanced structures of technological processes, autonomous and connected, with a common scope: it transforms raw materials or derivatives in finished products, acting in a more complex system, the enterprise, with economic and financial objectives. So, a new plant to build cannot be designed without considering the whole social, political and economic environment where it has to live. In fact, industries survives only where certain advantageous conditions are set and that is the result of accurate decisions relative to particular places and time. It has to:

- be inserted harmonically in an organic development plan of its country and the available resources;
- be convenient and grant stakeholder interests (investors and management, customers, working people, suppliers, community);
- produce the best quality results;
- work in safe and guaranteed conditions.

In other words, an industrial plant has to:

- produce the quantity of market required product at possible lowest costs;
- need affordable investment and general and maintenance costs, in relation to their incidence on production costs;
- allow easy expansion and adaptation to possible increase or development of production;
- ensure optimal working condition to the employees;
- not cause damages to the environment.

LOCATION ANALYSIS

A plant location study is a strategic decisional process made of two different but parallel activities. The first (*geographic location*) is the choice of country and region where to set up the productive system; the second (*topographic location*) restricts the field to single out an area and then a precise site.
In particular, the choice of the region where to build the factory is effectuated on the basis of the characteristic political, social and economic lines. Otherwise, the purchase of the site can be effectuated when a general or executive design is ready and the shape or extension of the needed surface has been calculated. Only at this moment, in the region previously individuated, a punctual location and an appropriate and convenient plot is defined. In the end, the primary target of location analysis is to select the optimum place to minimize the sum of related costs: setting (soil purchase, factory building, connection with infrastructure), resource acquisition, product distribution, services.
Localization choices of industrial plants, warehouse or any other enterprise activity are one of the main strategic management decision. In fact, the number of variables (*location factors*) to consider is very high and concerns all the different company functions.

Furthermore, the timeline to refer and evaluate the consequences of the decision (both for return on investment and in general on the mission and vision of the organization) is wide and critical as any long term action, mainly for the cost and relevance related to an eventual relocation problem. To build an evaluation model, that could work as a decision support system, is necessary to consider all the different aspects on the ground. A first proposal is to divide the parameters individuating four categories of location factors:

- environmental factors – describe the relation between company and natural environment, as impact and influences caused and suffered by the production system;
- social and political factors – define the legislative texture as bounds and opportunities;
- economical factors – group the main costs and revenues of the organization;
- technical factors – consist of available resources and technical frame.

FUZZY IMPLEMENTATION

In this model, the construction of a fuzzy model fits the necessity of combining quantitative evaluations (as costs and revenues) that can result from detailed analysis or simple assessments with qualitative ones expressed by opinions of experts. The frame proposed defines as input a set of variable purely linguistic (i.e. "far, medium, low") and compute MBF (membership function), L-shaped, from two to five elements, as shown in figure 1.

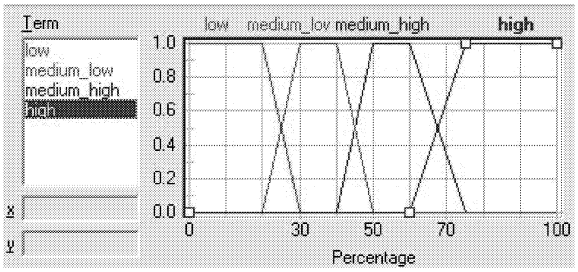


Figure 1 – MBF "Financing" variable

The fuzzy logic based system uses production rules to represent the relationships among the inputs. The rules consist of preconditions (IF-part), linked together by linguistic conjunctions like AND and OR, and a consequence (THEN-part). The influence that each IF statement can have on THENs is managed through linguistic relations that goes from "very negative" to "very positive" (figure 2).

The computation of fuzzy rules is called fuzzy rule inference and is a calculus consisting of the following steps: aggregation, composition and, if necessary, result aggregation. The first step of the fuzzy inference, aggregation, determines the degree to which the complete IF part of the rule is fulfilled. Given the structure of the model, the standard MIN operator is chosen which represents the AND aggregation function. The second calculation step of each production rule (composition) uses the support of the precondition to calculate the support of the consequence. In standard inference methods, the consequence of a rule is considered equally as true as the condition. Finally, if more than one rule produces the same consequence, an operation must aggregate the results of these rules. A result aggregation step determines the maximum degree of support for each consequences which is used for all further processing. In this case, the MAX method selects the rule with the maximum firing degree of all rules matching to the term.

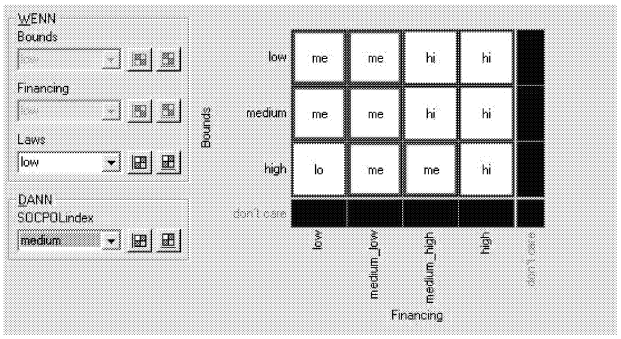


Figure 2 – Rules Matrix

The result produced from the evaluation is, of course, fuzzy. Membership functions are used to translate the fuzzy output into a crisp value. This defuzzification process is performed using the Center-of-Maximum method (CoM). Because more than one output term can be evaluated as valid, the algorithm must compromise between the different results, computing a crisp output as a weighted average of the term membership maxima, by the inference results.

THE MODEL

Input data and information contribute to generate four partial indexes, relative to the factor categories, carrying out a global index that measures the risk and opportunity level associated to the location under exam. The values go from 0 (high risky) to 100 (high opportunity) express a optimum degree of the solution adopted. The four indexes are composed as shown in figure following the structure described below. The global index, as seen in figure 3 is a weighted average of the partial ones using CoM method.

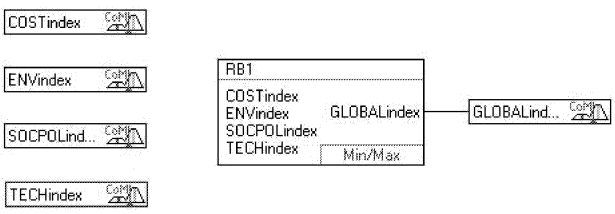


Figure 3 – Global Index structure

SOCIAL – POLITICAL INDEX
The index, shown in figure 4, is composed of the following three influence factors:

- *Bounds*: a location with bounds or servitudes of other public or private site, whose activity are in contrast with the one of the new plant, needs agreement or compromises with generally not predictable expenses;
- *Financing*: national and regional programming has an high weight on favouring industrialization in particular zones, with particular tools as no-refund contribution, facilitated rate financing, fiscal and custom facilitation;
- *Laws*: location choice has to consider actual and active laws, as public regulation on urban and territorial plans.

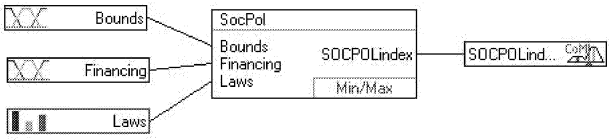


Figure 4 – Social – Political Index structure

ENVIRONMENTAL INDEX
The index, shown in figure 5, is composed of the following three influence factors:

- *Climate*: climatic situation of the area influences the correct development of production processes (first of all humidity, temperature and wind intensity, especially for precise working activities) and affects in an onerous way consumption of energetic resources;
- *Decline*: qualitative factor that defines the level of decline of the site, eventual or potential, identifying possible area of negative impact for environmental risks and safety in the plant;
- *Impact*: qualitative factor that indicates the level of direct environmental impact, generated by the production system on neighbours and community.

COST INDEX
The index, shown in figure 6, is composed of three sub-systems, each one defined by a set of correlated variables.

Budget sub-system gives an estimate of the ratio between allocated budget for location and real associated costs. The variables are relative to:

- *BudgetFix*: represents investment costs, principally site purchasing and settlement and plant and constructions building;
- *BudgetVariable*: variable annual costs, in particular transport costs, depending on the different typologies, service fees and infrastructures

District sub-system describes the development level of the region and its relation with economic texture. The variables are relative to:

- *Factories*: balances the presence of auxiliary and complementary industries in the same region with the concurrent or substitutive ones that can interfere on the market ;
- *Infrastructure* and *Service*: analyse the availability of services and infrastructure, favouring the location where advanced tertiary is more developed.

Web subsystem combines the information relative to supplier and customer market and connection net. The variables are relative to:

- *Communication*: is a weighted evaluation of the level of performance of traffic connections and its different transportation system (interstates, motorways, railways, etc...). A good location can be used also for advertisement;
- *MarketEnd:* defines the distance from the centre of gravity of the customer market, concentrated or distributed;
- *RawMaterials:* analyzes the raw materials market, defining the availability of supplies (duration, efficiency and effectiveness), regularity and continuity, distance from the centre of gravity of the market.

TECHNICAL INDEX
The index, shown in figure 7, is composed of the following three influence factors:

- *Labour*: the working force availability problem has to be considered both at regional and local level. It is also necessary to verify the needed amount of labour and its opportune professional level, skills and experience;
- *Energy*: each energy form (water, electricity, fuels) has to be evaluated in its main parameters that influence the supply yield as closeness to the source, service continuity and amount required. In the model, an *Energy* block is present for each form of energy requested;
- *Terrain*: the composition of the variable comes from a factor related to the extension of the surface to occupy (percent of used surface on surface of possible expansion) and many factors that can describe important technical aspect of the terrain and its physical and technical characteristics.

CONCLUSION

The model solves in an adequate way the problem, so to represent an important decision support system for company location strategy. Its application depends on the availability of input data and becomes more effective when partial analysis are carried out by people expert of the different sectors. Its strength is in combining different and multi cultural issues in a simple and correlated way, summing up in four comparable index both qualitative and quantitative information. Its simple formulation allows an implementation in every industrial initiative, simply tuning the more significant variable for different requirements, "turning on or off" the eventual parameter of influence , comparing in an objective way different choices. Further develops will be conducted on purely linguistic variables, to an eventual MBF conversion, and on some specific factors as impact and decline of the environment, where all the single direct factors (emissions in air and water, waste management, contamination of ground, use of natural resources and raw materials, local and transportation matters, incident risks and damages, effect on biodiversity) and the ones related to the different items of fix and variable budget allocation will be defined.

BIBLIOGRAPHY

1. Wu F.G., Lee Y.J., Lin M.C. (2004), *Using the fuzzy analytic hierarchy process on optimum spatial allocation*, International Journal of Industrial Ergonomics, n.33
2. Kahraman C., Ruan D., Dogan I. (2003), *Fuzzy group decision-making for facility location selection,* Information Sciences, n.157
3. Chen C.T. (2001), *A fuzzy approach to select the location of the distribution center*, Fuzzy Sets and Systems, n.118
4. Badri M.A. (1999), *Combining the analytical hierarchical process and goal programming for global facility location-allocation problem,* International Journal of Production Economics, n.62
5. Hsu H.M., Chen C.T. (1997), *Fuzzy credibility relation method for multiple criteria decision making problems,* Information Sciences, n.96
6. Hsu H.M., Chen C.T. (1996), *Aggregation of fuzzy options under group decision making*, Fuzzy Sets ans Systems, n.79
7. Chase R.B., Aquilano N. (1995), *Production and Operations Management,* Richard D. Irwin Inc.
8. Ross T.J. (1995), *Fuzzy Logic with Engineering Applications,* McGraw-Hill
9. Stearns T.M., Carter N.M., Reynolds P.D., Williams M.L, (1995), *New firm survival: Industry, Strategy, and Location*, Journal of Business Venturing,, n.10
10. Sule D.R. (1994), *Manufacturing facility: location, planning and desing,* PWS Publishing
11. Zimmermann (1994), *Fuzzy Sets and its Applications,* Kluwer Publishing
12. Kosko B. (1993), *Fuzzy Thinking: The New Science of Fuzzy Logic*, Hyperion
13. Moore J.M. (1979), *Plant layout and design,* Macmillan

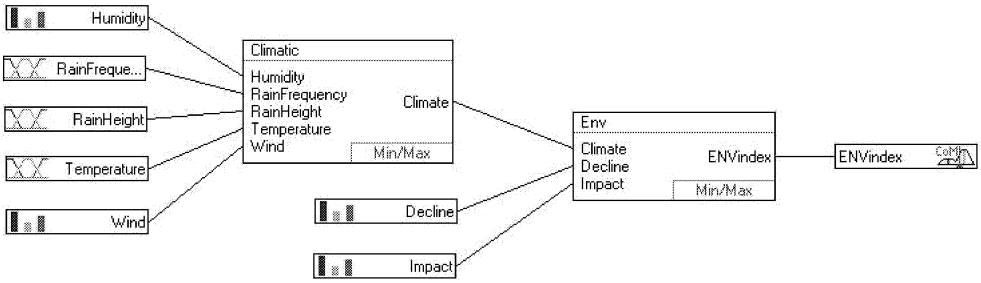


Figure 5 – Environmental Index structure

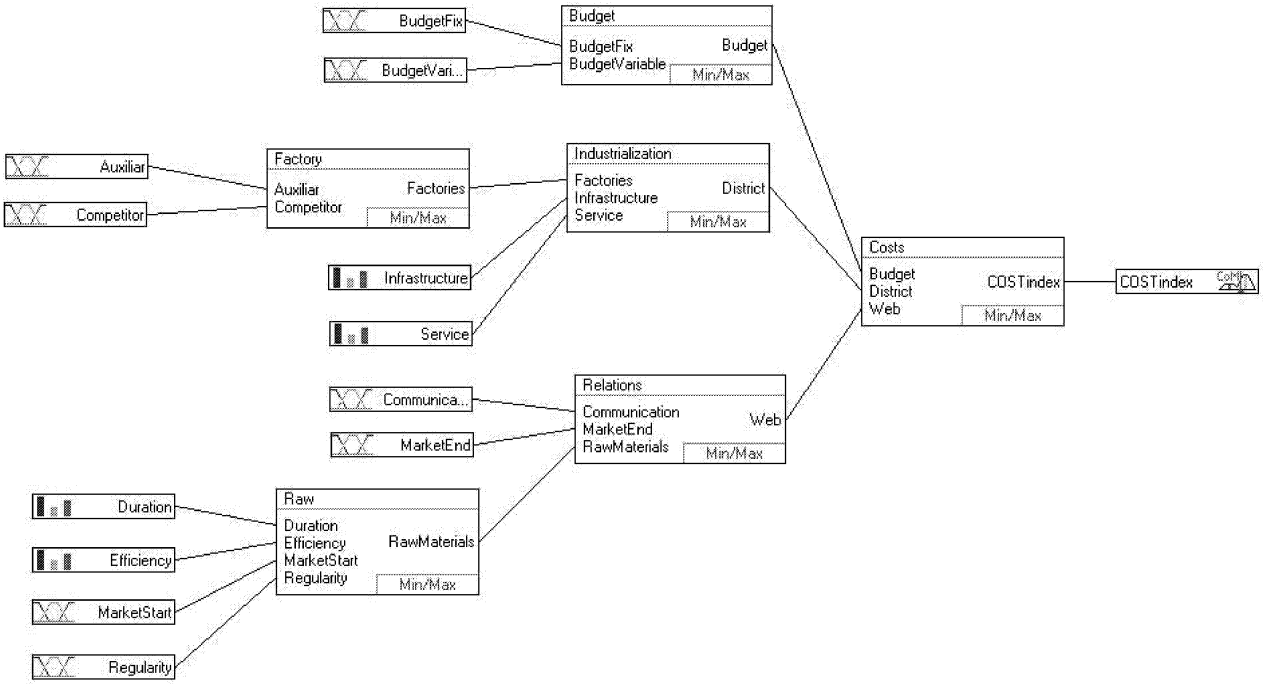


Figure 6 – Cost Index structure

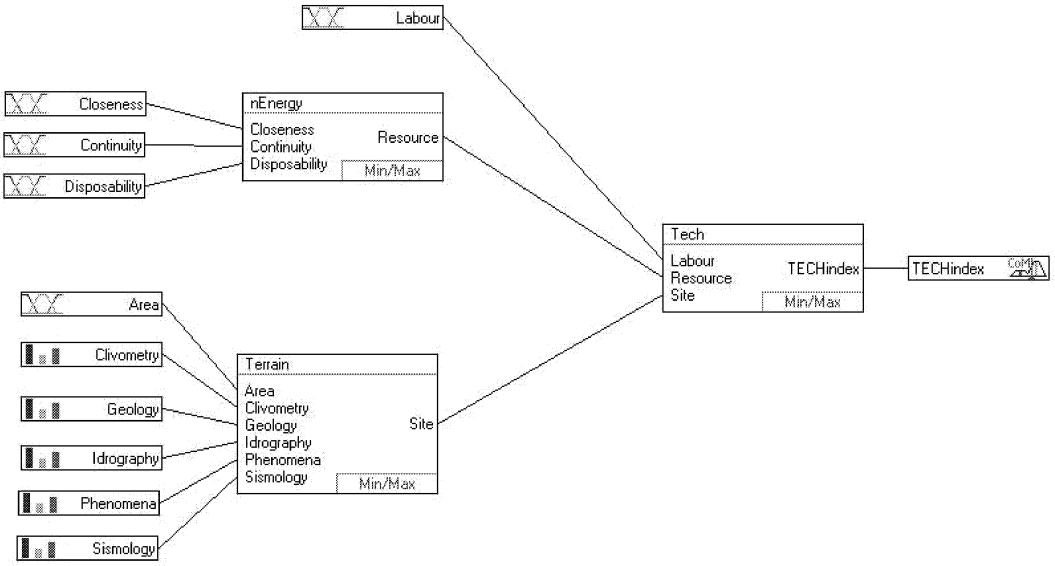


Figure 7 – Technical Index structure

TRAFFIC SIMULATION

BASIC ASPECTS OF TRANSPORT DEMAND MODELLING

Ing. Nina Kudlackova
Ing. Alexander Chlan, Ph.D.
University of Pardubice
Jan Perner transport faculty
Studentska 95, 532 10 Pardubice, Czech Republic
Phone: +420 466 036 384, Fax: +420 466 036 374
Email: Nina.Kudlackova@upce.cz

KEYWORDS

Modelling, The Four Step Model, Activity-based Models, Trip-based Models

ABSTRACT

The aim of this article is to present the transport demand methodology and show possibilities to future improvement. The history of travel demand modeling has been dominated by the four step model. Travel, always viewed in theory as derived from the demand for activity participation, in practice has been modeled with trip-based rather than activity-based methods. As the four step model was very succesfull nowadays there are efforts to use the travel-based models, activity-based models and the landuse together to get better results.

INTRODUCTION

When we talk about the modelling in transport, we have to answer the question: Why to simulate? The answer is, that transport modelling may give us answers on questions like this: What types of transport investment will give maximum improvement of future mobility in region?, Which places are possibly jeopardise by congestions?, How can landuse and developmental scenarios influence future levels of traffic congestions? or How many people will use public transport, car or van for journey to work?

These answers are important for prognoses and future decissionmakeng. We need transport models and prognoses to get the idea of future state of transport. Regional transport (prognostic) model of travel demand is mathematial interpretation of transport suply and demand in urban area. Transport demand is generaly represented by road system and transit network. The road system present all types of main roads in region and the transit network present all public transport services in region. Travel demand is represented by series of mathematic models.

Models

Main elements of transport models are:

1. input data
2. models (mathematic procedures a representations)
3. results (outputs)

Inputs of transport models may be:

- forecastings of future population, sizes of households, employment in region
- informations about future planned changes in transport networks, or potential changes which should be tested on possibility of improvement of current transport system

Outputs of models:

- flows on communications
- speeds on communications
- OD (origin-destination) schemes – are represented by trip tables
- modal split
- emissions (need emission models)

THE FOUR STEP MODEL

The history of demand modeling for person travel has been dominated by the modeling approach which has come to be referred to as the four step model (4SM).

Travel, always viewed in theory as derived from the demand for activity participation, in practice has been modeled with trip-based rather than activity-based methods. Trip origin-destination (OD) rather than activity surveys form the principle database. The influence of activity characteristics decreases, and that of trip characteristics increases, as the conventional

forecasting sequence proceeds. The application of this modeling approach is near universal, as

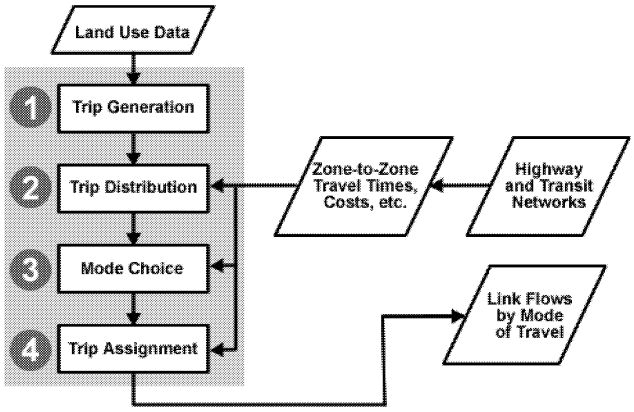
in large measure are its criticisms. The current 4SM might best be viewed in two stages.

In the first stage, various characteristics of the traveler and the land use - activity system are evaluated, calibrated, and validated to produce a nonequilibrated measure of travel demand (or trip tables). In the second stage, this demand is loaded on the transportation network in a process than amounts to formal equilibration of route choice only, not of other choice dimensions such as destination, mode, time-of-day, or whether to travel at all. Although this approach has been moderately successful in the aggregate, it has failed to perform in most relevant policy tests, whether on the demand or supply side.

The four step model is the primary tool for forecasting future demand and performance of a

transportation system, typically defined at a regional or sub-regional scale. The models system was developed for evaluating large scale infrastructure projects and not for more subtle and complex policies involving management and control of existing infrastructure or introduction of policies which directly influence travel behavior. It is called „The four step model“ because it sontains four basic steps:

- trip generation
- trip distribution
- mode choice
- traffic assignment



The four step model has considerable requirements on data, because it must define systems of activities and transport. There is the need of dates which define travel behaviour. These we can get from travel researches and we use them for validation. Befor the realization of FSM we have to get and process:

1. The zone social-economic data as for example the number of households, number of employed or level of population are.
2. Prototypal demographic data of households as the ownership of car or receipt of households are.

These data are used for data validation of sample and for to relevation and determination of trip generation, trip distribution and mode choice models.

Trip generation

The target of this first stage of the FSM process is to define size of total daily trips in model system, by households and on zone levels for different purposes of trips (activities). This first stage also explicitly transfer FSM from the stage based on activities to the stage based on trips and simultaneously divide single trips to production and attraction. At the same time we have to ensure, that the measure of net performance would not influence the number of executed trips.

There are two basic methods of trip generation problem solution:

Direct method

This method is based on statistical research. Direct way of probing of intensities qi from area Gi (respectively into the area) may be accomplished by simple count. For example in personal railway transport we just count the ticked sold at stations in given area Gi to stations out of the area. Sometimes the simple count is not possible, so we estimate the proportional figure.

Indirect method

In case of more complicated direct assessment of trip generation, we can use the indirect method. It is possible to come out from condition:

$$qi = f(x1,\ldots, xm),$$

where x1,…,xm are some parameters describing the Gi area. For the estimate of number of cars coming out of area Gi we can use the theorem:

$$qi = a1x1 + a2x2 + a3x3 + a4x4 + a5x5,$$

where x1 is number of inhabitants in the area Gi ,
x2 is number of households in area Gi,
x3 is number of cars in area Gi,
x4 is average revenue of household in area Gi,
x5 is the density of settlement in area Gi,
a1, …, a5 are coefficients.

Trip distribution

Trip distribution in given period is expressed using OD (Origin-Destination) matrix. For each two junction u, v in given area, we determine the current intensity from the junction u to the junction v. OD matrix is considered to be the most important continuous model of transport demand and it is a tool of manager decisionmaking.

Current steer we can get even from calculation, but we have to know total current intensities qi, outcoming from Gi, or incoming to Gi. On the base of these dates we can construct mathematical model which help us to count values qij from known values qi, or $\overline{q_j}$. Conditions:

$$\sum_{j=1}^{n} q_{ij} = q_i \qquad i = 1, \ldots, n, \qquad (1)$$

$$\sum_{i=1}^{n} q_{ij} = \overline{q}_j \qquad j = 1, \ldots, n, \qquad (2)$$

$$q_{ij} \geq 0 \qquad (3)$$

Gravitational model

We can prove that the value qij is the bigger, the bigger are values qi a $\overline{q_j}$ and the smaller is the price of travelling cij. So there is the analogy with the Newton gravitational principle. Gravitational model help us to calculate intensities before two areas.

Trend methods

Trend method is relatively useful and effective tool for determination of among-area partial intensities qij in case hat we know values qi(0), $\overline{q_j}$ (0) a qij(0) from past. We

find out actual values qi a $\overline{q_j}$ and wether these do not differ from values qi(0) a $\overline{q_j}$ (0) too much, we calculate values of partial intensities qij according to so called trend model:

$$q_{ij} = k_{ij} \cdot q_{ij}^{(0)}$$ (1), where

$$k_{ij} = \frac{k_i + \overline{k}_j}{2}$$ is average among-area intensities trend,

$$k_i = \frac{q_i}{q_i^{(0)}}$$ is trend of traffic streams outcoming from area Gi,

$$\overline{k}_j = \frac{\overline{q}_j}{q_j^{(0)}}$$

is trend of traffic streams incoming to area Gj,

The Detroid method can be also used. This method has the same calculation form as classic trend model. The difference is in the way of assesment of coefficient kij:

$$k_{ij} = \frac{k_i \cdot \overline{k}_j}{k}$$, where $$k = \frac{q}{q^0}$$

Mode Choice

Nowadays is very important to choose methods of modal split reflecting factors of present policy as car-sharing and car-pooling are. We can say that the most important factors influencing the modal split are:

- accessibility of public passenger transport,
- car property,
- existence or nearness of car-pool systems,
- cost of each mode of transport,
- average parking costs,
- time required on using single modes of transport.

The factors of modal split are ordered into theorem, which express possibility of each passenget of choosing existent mode of transport. The principle of modal choice result from mathematic choice theory, which work with utilities of each alternetive and with possibilities of the choices. Most common model for modal choice is the logit model. Using the logit model we work with assumption that alternatives of choices are independent.

Traffic assignment

After mode choice comes the last phase of the FSM-Traffic assignment. The aim of this phase is to find out and compare real streams separation determined in trip distribution (OD matrix) on the net and theoretical streams separation. Real streams separation we get using straight measuring on given sections. The theoretical streams separation comes out of price optimal streams separation on net and assign cars on transport net lines with minimal price.

Conditions:

1. Rated net S = (V, H, d), where d(h) ≥ 0. Assessment of d(h) express for each section h∈H its price
2. Internodal intensities qij from vi to vj (known from trip distribution).

The route choice depends on carrier and customer and their influence is in different modes of transport different. The way the customer choose the mode of transport can be described using logit model-this way we get the best variant.

Software support for traffic assignment

As example ot software tool for traffic assignment in real time we can give british programm Contram (Continuous Traffic Assignment Model). Software tools allow us (thanks to high quality of computer technology and computer grids) to assess aboundance of factors. The Contram software dispose of more than 20 tables with dates for transport model needs. The algorithm of traffic assignment optimalise through minimal total sensed costs as special function:

Total sensed costs = aD + bTT + cDV2 + pTQ → min,

where D = distance (km),
TT = total travel time (hrs),
TQ = total queuing time (hrs),
V = average travel speed (km/hr) (including queuing time),
a = coefficient – unit costs/km,
b = coefficient – unit costs /hr,
c =coefficient – unit costs /(km(km/hr))2,
p = coefficient – unit costs /hod.

Programm allow us arbitrary amendment of actually solved transport scheme, for example addition of new zone, knot, crossing, junction, communication or change of their characteristics.

ACTIVITY BASED APPROACH

The ABA (Aktivity Based Approach) access is the reaction on imperfections of conventional transport models (as for example FSM is) primarily in areas of valid dates representing basic travel behaviour. Abilities of the FSM access are limited on evaluation of consequences of big investment infrastructure actions. The basic ABA philosophy is to substitute conventional access to travel behaviour by more holistic access where the travel behaviour is analysed as daily and multi-daily schemes of behaviour coming out of differences in lifestyles and in participation on activities within population. The heart of the ABA access is travel behaviour based on activities. ABA refer to inability of transport based model reflect on basic behaviour and so inability to react on new developing measures oriented on management.

TRAVEL-BASED MODELS VS ACTIVITY-BASED MODELS

Critics of the FSM point out that conventional travel models ignore reciprocal spatiotemporal interlock related with travel behaviour of households. The prime principal of passanger demand (that the passanger demand is derived from demand on patrticipation on activities) is here absolutely ignored. Travel based models work only with generating trips, activity based models are looking for answer on question: What generated those activities, which made trips? So we can say that with deficit of landuse integration prognostic models into the FSM model the future activities systems will be independent on future networks of communications.

CONCLUSION

The history of public passenger transport demand modelling is commanded by modelling accesses refering to the Four-step model (FSM). Transport, in theory always seen as derived from activities participation demand, is in praxis modelled rather on trip-based methods than on activity-based methods. Main complex of dates is made rather O-D method than on activities research methods. So in traditional prognosis the influence of characteristics descend and the travel charasteristics grow. Application of this modeling access is almost universal and togerher with it goes reviews. We have to look on both methods and find optimal solution if possible. Only time and additional research will give us answers.

REFERENCES

Cerna, A., Cerny, J.: Teorie rizeni a rozhodovane v dopravnich systemech. Institut Jana Pernera o.p.s., Pardubice, 2004, ISBN 80-86530-15-9.

McNally, M. G.: The four step model. In Handbook of transportation modelling, Elsevier Science Ltd, 2000

Working Together On Transportation Challenges. [online]. WWW: < http://www.mwcog.org/uploads/pub-documents/8l5WXA20040409151527.pdf#Forecasting >.

Opportunities for the ITS in the Czech Republic. [online]. WWW: <http://www.its-cz.cz/activities.html>.

Four-Step Model. [online]. WWW: <http://www.mwcog.org/transportation/activities/models/4step/step3.asp>.

Purpose of Modeling. [online]. 2004. [cit. 25.03.2004]. Dostupný na WWW: <http://www.mwcog.org/transportation/activities/models/purpose.asp>.

Transportation Models & Forecast, Dostupný na WWW. [online]. 2004. [cit. 25.03.2004]. <http://www.mwcog.org/transportation/activities/models/default.asp>

RESEARCH PROJECT THEORY OF TRANSPORT SYSTEMS

Prof. Karel Šotek
Alexander Chlaň, Ph.D.

Jan Perner Transport Fakulty
University of Pardubice,
Studentska 95, 532 10 Pardubice
The Czech Republic
E-mail: Karel.Sotek@upce.cz, Tel: +420 466 036 508, Fax: +420 466 036 094,
E-mail: Alexander.Chlan@upce.cz, Tel: +420 466 036 394, Fax: +420 466 036 094,

ABSTRACT

Research project Theory of Transport Systems belongs among main research and development activities at Jan Perner Transport Faculty, University of Pardubice and will be solved in year 2005 – 2011.
The main subject definition of the project can be divided in following themes: Transport system development in new social economic conditions, Optimization of technology processes and controlling in transport systems, Theory of information technologies and optimizing tasks in transport and Management, optimization and diagnostics of complicated dynamic systems.

KEYWORDS

Research project, theory of transport systems, optimization.

INTRODUCTION:

Research and development activities at Jan Perner Transport Faculty, University of Pardubice proceed on account of solution of gained research projects and transformation and development programs within:

a) Grants Agency of the Czech Republic (GA CR),
b) Fund of Development of Universities´ Council (FRVS)
c) Other resort projects (MD CR),
d) Foreign projects,
e) Development Activities Program at University of Pardubice,
f) Institutional research,
g) Transformation and development programs of Ministry of Education of CR.

Research and development activity proceeds in close relation on seven study fields:

- Railway Vehicles
- Transport Infrastructure
- Operational Reliability, Mechanics and Diagnostics in Transport
- Electrical and Electronical Engineering in Transport
- Informatics in Transport
- Transport Management, Marketing and Logistics
- Transport Technology and Control

1. MAIN TOPICS OF RESEARCH PRODUCT

Research project will be guaranteed by:

Department of Transport Management, Marketing and Logistics (DTMML)
Basic research and development activities are specified in these areas:

- Assurance of basic transport services,
- Problematic of intermodal transport and intermodal transport systems,
- Impact of transport quality growth on region development,
- Problematic of mobility in passenger and freight transport,
- Quality of traffic and transportation processes and services,
- Implementation of process management and quality management in traffic and transportation processes,
- Application of reengineering in management of logistics processes in transport.

Department of Transport Technology and Control
Basic orientation at:

- Optimization of public passenger transport services in regions (1997-1999)
- Optimization of public transport in regions (2000-2002)
- Optimalization of interval transport in suburban agglomeration (2001-2002)
- Project of effective cycling transport in urban agglomerations (2003)

Department of Informatics in Transport
Research activity of this department is orientated on:

- Development and application of SW means and information and control systems in transport,
- Generation of simulation models on a transport network, solution of optimizing problems in transport process,
- Solution of location and allocation problems, problematic of theory of schedules
- Rationalization and objectiveness of decision-making in technologic transport processes.

Research team of Jan Perner Transport Faculty is completed by Department of Process Control and Informatics of

Faculty of Chemical Engineering from University of Pardubice:

Department of Process Control and Informatics (DPCI)

Research activity of this department is orientated on:

- Project of multidimensional systems control,
- Research of adaptive multidimensional regulators based on algebraic control theory,
- Problematic of modelling of significant non-linear systems with utilization of neuronal nets,
- Processing of algorithms of hierarchical process control by means of fuzzy logics.

2. ACTUAL RESULTS AIMED TO

Modelling of technological processes in public passenger and freight transport
Basis for organizing of lock-out activities on railways
Arrangement in freight transport logistics
Generation of methods, models, computer support, planning methodology and decision-making in crisis situations aimed to transport issue
Projecting, innovation and support of information systems for generation of basic control plan (GVD) of Czech Railways company (CD a.s.)
Multidimensional systems control.
Research of adaptive regulators
Processing of modelling algorithms of significant non-linear systems with utilization of neuronal nets,
Methodology processing of significant non-linear processes and processes with time variable parameters.

3. DESCRIPTION OF RESEARCH PROJECT

Subject definition of research activity in the frame of partial task:

- „Transport system development in new social economic conditions“

The subject of partial task solution is theoretical problems of transport system development in new social, economic and ecological conditions. The content of solution is not only modelling of influences that impact the development of transport system, but also new restriction of its sustainable development. On account of these changes we will suggest a model of transport system operation from the view of costs and utility. This model will come out from the balance of supply and demand and from internal and external balance of transport system.

- „Optimization of technology processes and controlling in transport systems”

In the field of technology and control in transport we will solve the problematic of development and applications:

- theory of logistical transport,
- theory of jungle operation on a transport net,
- theory of management of intermodal and integrated transport systems,
- optimization of transport processes,
- modelling of technological processes in passenger and freight transport,

- „Theory of information technologies and optimizing tasks in transport“

It is orientated on:

- theoretical aspects of creation of information and control systems in transport
- new components in creation of discrete simulation models on a transport net
- next development of creation and use of simulation models of real environment
- theoretical questions of optimization of management and technological processes in transport

- „Management, optimization and diagnostics of complicated dynamic systems**“**

Research program in the field of management and optimization of complicated dynamic systems will be orientated on common problems of process management, formation of methodology of their management proposal and optimization and verification of suggested procedures and solution.

Scientific and research activity of the solver is orientated on economical, social, managerial, marketing and transport political development problems in transport sector. There are also solved research tasks and projects from the field of forwarding, development and evaluation of economic effectiveness of transport nets, quality of transport services, optimization of transportation systems of chosen substrates, analyses of efficiency and costs in transport and reengineering in management of logistics systems in transport. The above mentioned research orientation is in the conformity with orientation of solution of research project so from the view of the object so of the staff.

In the field of technology and control of transport the relevant part of the solution proposal of the research project is composed with the consideration on experiences of solution team and quality or level of knowledge of problematic. The problematic of research project consequently belongs to the research orientation of solving team for the field of technology and control in transport.
The field of research project is also in the consensus with the long-term outlook of development of research and development in the field of technology and control in transport. The research project alike as the own field of problematic of technology and control in transport will be orientated on three basic directions of cognition that will be developed in the future:

1) Theory of logistics systems,
2) Theory of transport technology,
3) Theory of transport and management.

4. AIMS DEFINITION OF THE RESEARCH PROJECT

Theoretical elaboration of acceptable development of transport system on account of social, economic and ecological conditions.
In conformity with supposed changes in surroundings of a transport system there will be suggested a model of operation of a transport system. Functional model will come out from the model of balance of supply and demand in transport and from internal and external balance of a transport system. On account of this there will be suggested an effective way of realisation of variants of transport system development.

Theory of logistics transport
Special terminology will be specify and elaborate. Criteria for optimal number and placement of logistical centres will be formulated with the requests for operational parameters for all kinds of transport. Theoretical base for suggestion and management of logistical chains and their relations will be created.

Theoretical basics of jungle operation in a transport net
Problems of optimal jungle equipment with technical means will be formulated and solved and suggestion, organization and management of serving processes in these jungles will be improved and précised. Models of processes that run in jungles in continuous and discrete time will be created and questions of suggestions and management of movement of transport means in the district of a jungle will be solved.

General theory of management of intermodal and integrated transport systems
Creation of mathematical models of decision-making and management processes together with the formulation of minor conditions and optimization criteria. Concentration on management processes models in logistics centres and on distance control of transport processes on a net. Optimization of transport flow management and partial components from the view of minimisation of time and energy need.

Optimization of technological processes in transport
General optimisation principals and methods will be developed and applied on problems of organization and management of logistical transport and on optimisation of supply of transport services. Indicators and factors that influence transport processes will be determined and modelled and computing procedures will be suggested.

Modelling of technological processes in personal and freight transport
In passenger transport there will be created models of individual processes and also integrated model. Next integrated models of long distance transport with combination of more kinds of transport will be created. In modelling of freight transport there will be created single and group models of technological processes, combined models and finally an integrated model.

Field of management and optimisation of difficult dynamic systems
Management of multidimensional systems – research of methods based on easy models gained from investigation of FM characteristics.
Research of adapted multidimensional regulators on algebraic theory of management.
Elaboration of algorithms of modelling significantly non-linear systems with the use of neuron nets.
Processing of algorithms of hierarchical management of strong non-linear processes and processes with time variable parameters.

5. STRATEGIES AND METHODS OF SOLUTION OF THE RESEARCH PROJECT

Strategy of task solution will come out from the use of existing results of research to partial problems of the research project and from conception materials, e.g. Conception of transport infrastructure development in the Czech Republic till year 2010, Transport policy of the Czech Republic after the entrance into the EU, Sector and operational program Transport and environment, Conception of logistics centres construction etc. and from own accesses and methods of solution.

Statistical analyses and suitable models as a tool of prognoses creation will be used to mathematical description of development tendencies at modelling of indicators that influence the sustainable development of a transport system.
Functional model of a transport system from the view of costs and utility will come out from the analyses of all indicators and from conditions of their balance. Research of supply and demand on a transport market will come out from the macroeconomic point of view. At modelling of factors influencing on supply and demand there will be used the concept of elasticity.
At solution of internal and external balance of a transport system we will come out from modelling of supplies and consumption among regions, from the model of modal split or from the model connected with supply and demand or balance of a transport net.
Verification of results will be realised upon contacts and collaboration with transport companies and organizations. Gained results will be groundwork for improvement of teaching process of bachelor, master and doctoral students and chosen partial problems will be solved as topics of projects, diploma thesis and dissertations.

Strategy of solution of partial tasks in the field of Management, optimisation and diagnostics of difficult systems will be based on:
- Research of methods based on models gained from experimental way (FM characteristics) and in the field of adapted multidimensional regulators on algebraic theory of management.

- Modelling of non-linear systems with the use of artificial neuron nets with use of sigmoidal and RBF transmission functions.
- Processing of algorithms of management of non-linear systems based on using of reference model in the shape of ahead neuron net and with use of inversive model realized by neuron net
- Processing of suggestion methodology of fuzzy regulation for strong non-linear processes or processes with time variable parameters.

CONCLUSION:

This research project sets up new possibilities for enlargement of theoretical and consequential applied research in transport.

REFERENCES:

Chlaň, A., Bucháčková, P.:
Transformation to Process Organization.
Proceedings from 11th international konference in Žilina „Veda, vzdelávanie a spoločnosť", s. 77-80, ŽU v Žilině, 2003,
ISBN 80-8070-120-2.

Melichar,V., Chlaň, A.:
Demand Price Elasticity Modelling in Public Passenger Transport.
Proceedings of „Third Scientific Conference", p. 7-14, University of Pardubice (2003).
ISBN 80-7194-568-8.

Šotek, K.:
Optimisation of Closure of Traffic by Simulation Model.
Scientific Papers of the University of Pardubice Ser.B, 3:147-151,1997.
ISSN 1211-6610.

Šotek, K.:
Applying of Simulation Models in Information and Steering Systems in Railway Transport.
Proceedings of 20th International Workshop, Advanced Simulation of Systems, ASIS 1998, VŠB-TU Ostrava, Krnov 1998,217-222.
ISBN 80-85988-26-7.

Šotek, K.:
Computer Technology and Informatics in Transport.
Monograph, Institut of Jan Perner, University of Pardubice, 1999, 245 pages.
ISBN 80-7194-230-8.

Šotek, K., Bachratý, H., Tavač, V., Kršák, E.:
The Transport Line Control Simulation Model.
MOSMIC ´99 The International Workshop – Modelling and Simulation in Management, Informatics and Control. Sulov, 5.-7. October 1999, 83-88.
ISBN 80-7100-635-1.

KAREL SOTEK was born in Trebic, The Czech Rublic and went to the University of Zilina, in former Czechoslowakia, where he studied Operation and Economics of Railway Transport and obtained his degrees in 1968. After doing his military service, he worked for a couple of years for the Czechoslawakia Railway Company before moving in 1993 to Pardubice and in 2003 he became dean of Jan Perner Transport Faculty, University of Pardubice, where he has been working ever since.

Preparation of Digital Maps for Traffic Simulation; Part 1: Approach and Algorithms

Daniel Krajzewicz, Georg Hertkorn, Julia Ringel, Peter Wagner
German Aerospace Centre, Institute of Transportation Research
Rutherfordstr. 2
12489 Berlin
Germany
E-mail: Daniel.Krajzewicz@dlr.de, Georg.Hertkorn@dlr.de, Julia.Ringel@dlr.de, Peter.Wagner@dlr.de

KEYWORDS

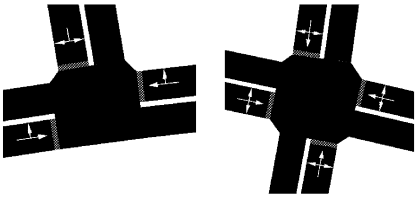
Microscopic traffic simulation, digital road maps, open source, traffic research

ABSTRACT

Traffic simulations are an accepted tool for investigations on road traffic and used widely within the traffic science community. Modern computer systems are fast enough to model and simulate traffic within large areas at a microscopic scale regarding each vehicle, replacing macroscopic simulations in most cases.
Although microscopic traffic simulations offer better quality than macroscopic ones, they also need additional data to describe the modelled road networks. A street's lanes are modelled explicitly within microscopic simulations and in most cases also the connections between their lanes over junctions.
If one wants to model large areas, the best source to get the description about their road network is the usage of digital maps. Unfortunately, most of these are used for routing purposes and do not contain the fine-grained information mentioned above that is needed by microscopic simulations. This document describes an algorithm for the computation of the needed information from simple road networks.

INTRODUCTION

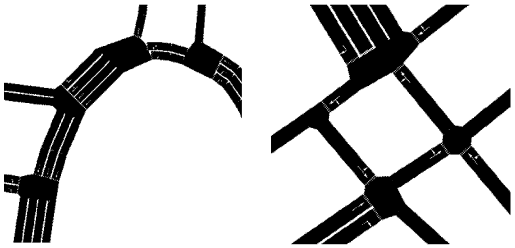
"SUMO" – an acronym for "Simulation of Urban MObility" – is an open source traffic simulation package developed at the Institute of Transportation Research at the German Aerospace Centre. This application is used within some of our institute's projects for simulating impacts of new optical sensors (see [1]) or traffic management strategies (see [2]). Descriptions of the SUMO package may be found in [3] or at the project's webpage [4].
The microscopic view on road networks contains information about the connections between consecutive lanes. When a junction has to be crossed, vehicles have to choose a proper lane in order to get to the desired one. Beside these lane-to-lane connections, one also has to model which of these connections are foes to which other connections in order to make vehicles coming from directions with a smaller priority wait.



Picture 1: Two rather uncomplicated junctions

We deal with large areas – cities or larger parts of these or highway networks – and we wanted to reduce the amount of work a user has to perform before a simulation can be started. Often, one has to edit all information mentioned above by hand. Our approach is an automatic generation of such data. For this purpose, beside the simulation program itself, the SUMO package also includes a program which allows the conversion of networks from other formats. These formats have their origin in other simulations, both macro- and microscopic ones, or in routing systems. While both microscopic simulations SUMO can import, Vissim by ptv (see [5]) and ARTEMIS by Prof. Dr. Peter Hidas ([6]) contain information about the described lane-to-lane connections, this information are not stored within the inputs to macroscopic simulations or routing systems. Macroscopic simulations do not use them as the flow is not modelled using vehicles but in an abstract way (see [7]) and so no lanes, but the streets' capacities are used (see. [8]). Routing systems do not consider the information about lane connections, too.
The needed automatic computation of lane-to-lane connections and of the information about foes was not as trivial as it sounds first. Road networks do hold many special cases to regard. Due to this, we decided to publish the algorithm as it may be interesting to other and also allow other to take a look at it in order to improve it.
Within this publication, we will describe the main paradigm used for modelling road networks and the information made available by the formats we regard. Then, we will describe which information is needed by SUMO and then, how we compute

this information from the available data. Some comparisons to the reality will follow, together with a conclusion and some thoughts about further work.
This publication does not contain the description about how the geometry of junctions or streets is computed, but only how to compute the logical connections.



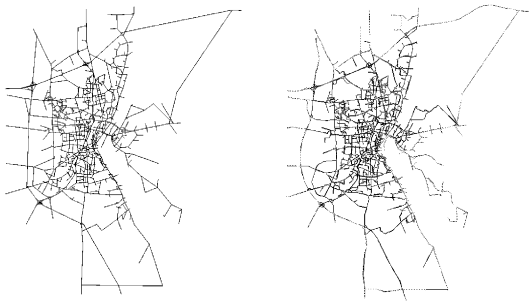
Picture 2: Two sets of rather complicated junctions as computed by the algorithm

ROAD NETWORK DESCRIPTIONS

Road networks are normally stored as a directed graph. Junctions are represented as nodes and streets as edges. Leaving out Vissim which has a different view on road networks, this paradigm is found throughout the formats we import, namely in Visum, ArcView, FastLane, ARTEMIS, or Navteq - networks.
The description of a junction is normally quite minimalist. Beside the position it is located at and an identifier, sometimes the information about its type is found which in most cases distincts between simple priority junctions and junctions controlled by a traffic light. The descriptions of the traffic light plans themselves are not that uniform and we will not discuss them herein. Still, one has to keep in mind that the information whether a link is controlled by a traffic light is necessary during the computation whether a link is foe to another.
An edge's description contains more parameter. At first, the number of lanes the edge consists of should be given. In the case of macroscopic simulation networks, this is sometimes not the case and the edge's capacity has to be used to compute the edges' lane number. Assuming the maximum capacity of a lane as 20000veh/h, the formula below computes the number of lanes if this information is not given. The only problem is that in some cases, edges used in macroscopic traffic simulation networks are using unreal values for the flow in order to fit the simulation to reality or to guarantee a high inflow. The second case can be caught and managed because it mostly occurs on feeding edges only which are mostly marked as such (at least in Visum). Furthermore, a maximum number of lanes can be applied. The first case can only be changed by hand.

$$lanes_{edge}=capacity_{edge}/max_capacity_{lane} \quad (1)$$

Further information optionally stored within an edge's description is the edge's type. It is not found in all inputs and also no standard value sets exist. This is quite unfortunate as we will see later. Other road attributes are either stored directly within the edge or within the type and in the second case they must be retrieved indirectly. These attributes are: the maximum velocity allowed on the edge, possibly the length, information whether overtaking is allowed or not and geometrical information as a single edge may be not a straight connection between two junctions but may possess a curvature (see picture 3).



Picture 3: Two views at the city of Magdeburg; the left network uses straight connections for streets, the right one includes the streets' geometries

NEEDED INFORMATION

When looking at a junction in detail, one will see that some further information is needed. The vehicles using a street which has more than a single lane, have to choose the lane they use properly in order to get to the next street (see picture 2). The wish to model this fact is not only a try to make the simulation appear more realistic, but is also important for realistic results.
A further needed information is also mostly not available: whether a vehicle has to stop before entering a junction when another vehicle approaches from another direction. Of course, this behaviour depends on the junction‘s type. In the case of traffic light controlled junctions, the behaviour changes with the traffic lights‘ state. In the case of uncontrolled junctions, it depends on whether the roads participating in the junction have different priorities or the vehicles have to interrogate which of them may pass the junction first. In Germany, for example, many junctions use the right-before-left – rule. In all cases of uncontrolled junctions, vehicles have to wait only if their trajectories cross. Stop signs by now remain unregarded.

INFORMATION COMPUTATION

Some Remarks to the Import Procedure

To avoid duplicate methods, the import is divided into two parts. At first, the input is read using a reader capable to parse the format of the imported file and converted into the net converter's internal structures. Within the second step, the network converter uses the information stored within these internal structures to compute all values the simulation needs. If any of the needed information was already available within the input file, the computation is skipped and the read values are used. This method allows an easy implementation of import modules and guarantees the same results independent of the import format itself.

We will not go into depth about parsing of input files. Most of it is straightforward because the available information is the same for most cases and the algorithm described herein is executed within the second step.

Lane-to-Lane Relationship Computation

After loading, all loaded junctions (called 'nodes' from now on) and streets ('edges') are stored in two containers – one for the junctions and one for the edges. The order of those objects within these containers is not defined. All operations are executed on every junction and on every edge within the according container. The following operations are done in the given order:

1. for each edge: compute turnaround edges
2. for each node: sort each node's edges
3. for each node: compute each node's type
4. for each node: set edge priorities
5. for each edge: compute edge-to-edge connections
6. for each edge: compute lanes-to-edge connections
7. for each node: compute lane-to-lane connections
8. for each edge: recheck lanes
9. for each edge: append turnarounds

We will now describe what those functions do exactly do and why they are needed.

1. Compute Turnaround Edges

Among a street's successors, the turnaround direction is a special case as the number of lanes used to reach it is always one and because this direction is not regarded as an explicit direction: it is uncommon to have a lane which is only used to turn around as the wish to do so only seldom occurs. A further reason for computing the turnarounds is due to needing this information for the computation of the edges' clockwise order within step 2. Due to his peculiarities the backward edge is computed for each edge first and this information is stored to allow neglecting the backward direction on further processing (steps 2-8).

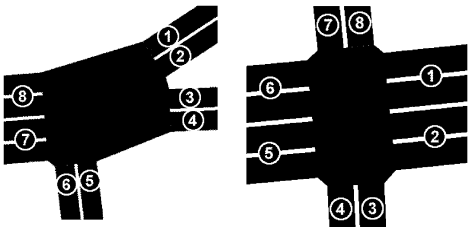
To compute the turning directions, we assume that an edge is the backward edge if the absolute value of the difference between the current edge's direction and the edge considered to be the backward direction is larger than 160°. The angle is measured at the current junction as each edge is a list of straights, and due to this a street's angle may differ along it.

This computation holds a trap: there may be more than a single edge that fits to the 160°-rule. In this case, the edge with the largest absolute direction difference is used. Still, to check for 160° is necessary as the usage of the edge with the highest rotation difference only would be false in cases where an edge has no backward direction at all.

2. Sort each Node's Edges

Within the second step, each of a junction's streets – both incoming and outgoing – are sorted clockwise by their direction. "Direction" means here the angle at the currently regarded junction. After this, the edges are sorted by their logical direction – whether they are incoming or outgoing. This is done by going through the sorted list of edges and checking whether the next edge (the one at the current position + 1) is the backward direction of the current one and is an incoming edge. In this case, the outgoing edge was sorted to lie left to the incoming one, but should be on the right side. In this case, the edges are swapped within the list.

The result of this operation is a clockwise sorted list of edges that participate in the junction and the incoming lanes lay before the outgoing ones as shown in picture 4.



Picture 4: The resulting order of edges after sorting

3. Compute Nodes' Types

We regard three different types of junctions: "no junction", "priority junction", and "right before left junction". The first case, "no junction" is a special case where no right of way rules are applied. A normal network should not contain such junctions, but if for example two highways cross, this type may be used. Within "priority junction" an incoming street and its consecution

has a higher priority than other incoming streets. Within "right before left junctions", all directions have the same priority and as common in Germany, vehicles which come from the right road may pass, the others will have to wait.
The resulting junction type is computed for each pair of edges which income to the junction. If one of the combinations yields in the type "priority junction", this type is used. Otherwise, the type "right before left" is used.
To determine the junction type for a single combination of two incoming edges, a two dimensional matrix is used. As indices, the velocities of the regarded edges are used; the values are the junction types. Table 1 shows the used matrix.

	-10	10-30	30-50	50-70	70-100	100-
-10	r	p	p	p	p	x
10-30		r	p	p	p	x
30-50			r	p	p	x
50-70				t	p	p
70-100					t	p
100-						x

Table 1: Junction types in dependence to two crossing streets; r: right-before-left, p: priority, x: no junction

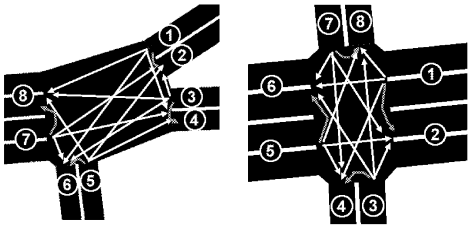
4. Set Edge Priorities

Now for each edge, we set his priority within a junction. This information will be needed later to determine how many lanes shall approach this edge – in the case of an outgoing edge – from an incoming edge. The priority of an edge within a junction may differ from the edge's overall priority because edges of different types may cross within the junction. The edge's overall priority is also not always available within the network description. If not, the edge's speed may be used.
The edge's priorities for a certain junction they participate within are computed as following:

- highest_incoming_edge2 = undefined
- highest_incoming_edge1 = get the incoming edge with the highest priority
- if further incoming edges exist:
  - highest_incoming_edge2 = get the incoming edge with the second highest priority
- if at least one outgoing edge exist:
  - highest_outgoing_edge1 = get the outgoing edge which is almost in the same direction as highest_incoming_edge1
- if further outgoing edges exist and highest_incoming_edge2 is not undefined:
  - highest_outgoing_edge2 = get the outgoing edge which is almost in the same direction as highest_incoming_edge2

5. Compute Edge-to-Edge Connections

Herein, the list of the edges that may be reached from the currently seen edge is computed. This is simply done by connecting all outgoing edges to the currently regarded incoming edge. The connected edges retrieved from the junction are already sorted as described in step 2. Picture 6 shows the generated edge-to-edge connections for our example.



Picture 6: The resulting connections between edges (grey: turnaround computed in step 1)

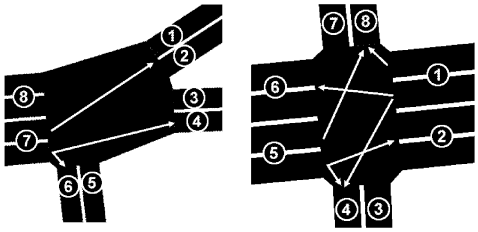
6. Compute Lane-to-Edge Connections

What we have to do herein is to fan out the lanes onto the outgoing edges, in order to compute which edges can be reached from a certain incoming lane of the currently regarded edge.
This computation is quite tricky and requires some heuristics. The problem is that the number of lanes used to reach a certain following edge is depending on the available number of edges and on the amount of traffic flow that will use this connection. For most cases, the flows can be determined using the number of lanes the following edge has. As the comparison at the end of this report show, this is not always true. The flows themselves are not available within the process of network conversion.
The following computation is done for each edge:

- get the list of connected edges beside the turnaround
- sort them by their angle
- for each edge in this list, compute its priority for the current edge:
  - priority = (connected edge's junction_priority + 1) * 2
- if one of the lower priorised outgoing roads goes to the right:
  - divide his importance by 2 as vehicles using it can leave the junction faster
- if there are no major roads at this junctions:
  - multiply the outgoing road that goes straight by 2, making it more important then the others
- compute the number of lanes that shall approach each of the connected edges:
  - sum up all priorities
  - for each outgoing (connected) edge:
    - number of lanes to use to reach this edge = this edge's priority for the current edge / priority sum
    - if number > number of current edge's lanes:
      - number = number of current edge's lanes

We now know how many lanes shall approach each of the connected edges. We still have to compute which lanes are used for which edge or

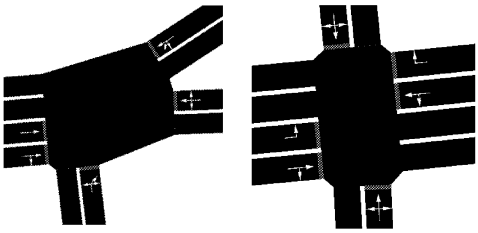
better to say from which lane the next edge may be approached.
This is done using the Bresenham algorithm for line computation assuming the list of lanes of the current edge to be one dimension and the (sorted) list of edges to approach the other one



Picture 7: Connections from lanes to edges; Shown are only connections from edges which have more than one lane – in the other case the single lane is connected to all connected edges

7. Compute Lane-to-Lane Connections

For each of a junction's outgoing edges, we now have to determine, which approaching lanes will yield on which of this outgoing edge's lanes. This is done using the Bresenham algorithm, too. The dimensions the algorithm goes by are the number of edges connected to the currently regarded outgoing edge and the number of lanes the currently regarded outgoing edge has.



Picture 8: After step7, the junctions are completely built for the most cases

8. Recheck Lanes

In step 6, we have computed from which lane which edge may be reached. In some cases, mainly if the connected edges have more lanes than the current edge, not all lanes were filled, yet. Within step 8, such mistakes are being searched and corrected. The algorithm simply checks whether a lane has no connected edge and if a neighbour lane has more than one connected. In this case the neighbour lane's connections to a directionally matching edge are moved to the side.

9. Append Turnarounds

Now, after all normal connections between all edge's lanes are set, the turnarounds may be added to the leftmost lane.

Foe Computation

The computation of which streams are foes to which requires only two steps. The wished output is the information about which connections from one incoming lane to an outgoing lane have to wait while other may drive. The problem is that not only one has to take into account right-before-left relationships, but also compute whether two streams are crossing at all.
Within the first step, for each connection the information is set whether it is controlled by a traffic light or not.
The second step is done for each junction. It assumes that the edges are sorted clockwise as described in "2. Sort each Node's Edges". The algorithm is as following:

- for each incoming edge i:
  - for each outgoing edge o:
    - compute right-hand link crossings(i, o)
    - compute left-hand link crossings(i, o)

where:
compute right-hand link crossings(i, o):
- ip = position of i in the sorted list of edges
- while ip!=position of o in the sorted list of edges:
  - move ip counter clockwise (decrement, wrapping)
  - i2 = edge at position ip
  - if i2 is an incoming edge:
    - op = position of o in the sorted list of edges
    - while op!=position of i in the sorted list of edges:
      - o2 = edge at position op
      - if o2 is an outgoing edge:
        - connection i to o crosses the connection i2 to o2
        - check priority(i, o, i2, o2)
      - move op counter clockwise (decrement, wrapping)

compute left-hand link crossings(i, o):
same as "compute right-hand link crossings(i, o)", but the pointers are moved clockwise (increment, wrapping).

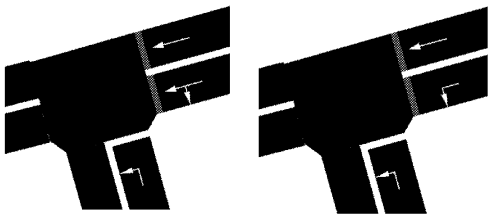
check priority(i, o, i2, o2):
- if already checked:
  - return
- mark as checked
- if connection i to o is a turnaround:
  - i to o is lesser priorised than i2 to o2
- same for i2 and o2
- if i has a higher junction priority than i2:
  - i2 to o2 is lesser priorised than i to o
- if i has a lesser junction priority than i2:
  - i to o is lesser priorised than i2 to o2
- if i and i2 have a high junction priority:
  - if o has a higher junction priority than o2:
    - i2 to o2 is lesser priorised than i to o
  - if o has a lesser junction priority than o2:
    - i to o is lesser priorised than i2 to o2
- if counter clockwise distance between i and i2 is smaller than between i2 and i:
  - i to o is lesser priorised than i2 to o2

- if counter clockwise distance between i and i2 is larger than between i2 and i:
  - i2 to o2 is lesser priorised than i to o
- if counter clockwise distance between i and o is larger than between i2 and o2:
  - i to o is lesser priorised than i2 to o2
- else:
  - i2 to o2 is lesser priorised than i to o

This algorithm has the benefit to compute all needed information without the need to know a junction's geometry. Only logical information is used.

COMPARISONS TO THE REALITY

Most of the evaluations we have made so far is based on evaluating the generated networks by hand and determining whether the generated junctions look like they should or not. The algorithm described in here is about 3 years old and worked well during such tests so far.
The only methodological analysis was done for a small area which real-life counterpart is located around our institute's site. For this case, the algorithm returns the correct result for all (177) but four junctions, being to almost 98% correct. One of the junctions which do not match is shown in picture 9. The reason for the mismatch is due to the fact, that in real life, most of the traffic is turning left – something the algorithm is not able to forecast. A further reason for false computation was the lack of information about additional left-turn lanes within the original network description.



Picture 9: False (left) and corrected (right) junction within the evaluated scenario

CONCLUSIONS

The comparisons to reality show that the error one should expect is quite small. When taking into account that if not using these methods one would have to edit every junction by hand, this approach seems to be very useful: generating the network for a city with about 10.000 junctions takes less than a minute using the SUMO net converting tool. Assuming you would do this by hand, you would have to touch every junction several times – if each of such actions would cost you only 1 second, you still would have to spend about three weeks working on it.

FUTURE WORK

The heuristics used within step 7 of lane-2-lane computation are something that should be revalidated. Although the results are fine for the networks we have used so far, such methods seem quite error-pruned and may cause problems in the future. As reported, only one methodological analysis was done so far. This disallows to predict the algorithm's performance for a broader set of situations than inner-city areas in Germany. Some tests should be done using comparisons of SUMO-generated networks and networks having the complete topology.
A special case for such world-covering evaluations would be networks with left-handed traffic, which so far are not capable to be generated at all.
There are also at least two further important topics which should be covered by additional heuristics. The first is the need to guess for each junction whether a traffic light is positioned on it or not. The second is determining whether a junction is located on a highway, because in some cases some of the connections must not be inserted.

REFERENCES

[1] IVF/DLR. OIS project pages at: http://www.dlr.de/vf/forschung/projekte/ois

[2] INVENT. Invent-Homepage at http://www.invent-online.de

[3] D. Krajzewicz, G. Hertkorn, C. Rössel, P. Wagner. 2002. "SUMO (Simulation of Urban MObility); An open-Source Traffic Simulation" Proceedings of the 4th Middle East Symposium on Simulation and Modelling (MESM2002), Edited by: A.~Al-Akaidi, pp. 183 - 187, SCS European Publishing House, ISBN 90-77039-09-0

[4] D. Krajzewicz, G. Hertkorn, C. Rössel, P. Wagner. 2002-2005. SUMO Homepage. http://sumo.sourceforge.net

[5] ptv. ptv Internet pages at: http://www.ptv.de/

[6] P. Hidas. 2004. "Modelling Vehicle Interactions in Microscopic Simulation of Merging and Weaving". In: Transportation Research Part C: Emerging Technologies, Elsevier Science Ltd. ISSN: 0968-090X

[7] D. Chowdhury, L. Santen, A. Schadschneider. 2000. "Statistical Physics of Vehicular Traffic and Some Related Systems"; Physics Reports 329, (4-6), S. 199-329, Elsevier, 2000, ISSN: 0370-1573; also available at: http://www.arxiv.org (cond-mat/0007053)

[8] C. Gawron. 1998. "Simulation-Based Traffic Assignment: Computing User Equilibria in Large Street Networks", Dissertation, Köln 1998

URBAN TRAFFIC SIMULATION MODELING FOR RECONFIGURABLE HARDWARE

Anders Å. Hansson, Henning S. Mortveit, Justin L. Tripp, and Maya B. Gokhale
Los Alamos National Laboratory
P.O. Box 1663, MS M997
Los Alamos, NM 87545, USA
E-mail: {maya, hansson, henning, jtripp}@lanl.gov

KEYWORDS

Traffic modeling, Reconfigurable hardware, Cellular Automata, FPGA.

ABSTRACT

TRANSIMS (Barrett and et al, 2002) is a large-scale simulation system capable of modeling traffic of entire metropolitan areas such as Portland and Chicago. In this paper we present a traffic model based on the TRANSIMS micro-simulator that targets reconfigurable hardware. We give the general model, and also present the modeling aspects that are specific to our implementation target which is Cray XD1. The prototype implementation, although in its early stages and not optimized, is able to achieve $34\times$ speed up over software alone. This work demonstrates that partitioning and system integration issues play a crucial role in realizable speed up for hybrid simulation systems.

INTRODUCTION

In this paper we present our modeling/adaption of the TRANSIMS micro-simulator that targets reconfigurable hardware. The development platform we have been using is the Cray XD1 (Cray Inc., 2004a), but the principles presented here should be applicable to other types of reconfigurable hardware as well. Our model is based on TRANSIMS which is a highly configurable, large-scale simulator for traffic and transportation systems developed at Los Alamos National Laboratory. This system has been used to model and analyze the traffic of entire cites such as Dallas, Portland, and Chicago. Local traffic departments and the U.S. Department of Transportation are incorporating TRANSIMS in their planning and analysis.

A basic understanding of how TRANSIMS operates will help put our work into context, so in what follows we describe how a TRANSIMS simulation typically is conducted.

First a synthetic population is generated based on survey data and statistics for the particular urban region, e.g. Chicago. This population is partitioned into households with given locations, and the people are assigned workplaces, schools, and so on. Next, a twenty-four hour *activity plan* is generated for each individual of the population. An example of a plan is: (a) drive kids to school, (b) go to work, (c) pick up kids at school, (d) go home. The plans are generated so as to be consistent.

The set of activity plans is used as input to the TRANSIMS *router* module. This module's responsibility is to find a path through the city network for each traveler such that the actual travel times along these paths match real measured travel times. TRANSIMS can handle multi-modal traffic, so a path on the network graph can look like (a) walk from node A to B, (b) go by bus from B to C, (c) continue by car from C to D to E and so on. For a given set of plans it is the TRANSIMS *micro-simulator* that actually moves people, cars and buses around. The two modules operate together in an iterative feedback loop to arrive at the final routes.

The micro-simulator is a cellular automaton (CA) based model of roadway traffic (Nagel and Schreckenberg, 1992; Nagel et al., 1995; Rickert et al., 1996) that handles single- and multi-lane roads, intersections, synchronized traffic signals and more. For these large studies the micro-simulator is typically run as a parallel computation on cluster architectures. However, the underlying CA rules are well suited for specialized hardware. Our work described below shows how to off-load large parts of the micro-simulator computations to field programmable gate arrays (FPGAs). As will become clear a streaming approach is natural for this setup. Previously bandwidth and communication costs would make this unfavorable, but on architectures like the Cray XD1 this is no longer so, and our non-optimized prototype already shows a speedup when applied to, e.g. the Portland road network.

Our work presented here is only concerned with the micro-simulator module, and it is assumed that we already have a set of travel routes. Moreover, we only focus on car transportation.

To give an idea about the scale of the problem we are targeting consider the Portland road network (Barrett and et al, 2002) which we have used for our prototype and software reference model. The Portland road network representation has roughly $124,000$ road segments with average length about 250 meters. Each road segment is split into cells in a process described later. The resulting CA model has more than 6.6 million cells and additionally there are about $100,000$ intersections. For cities like Chicago and Los Angeles these number are an order of magnitude larger.

Previously, two different approaches for using FPGAs in traffic simulations have been proposed. The first approach by George Milne (Milne et al., 1993) simulated road networks by directly implementing their behavior in hardware. Milne's direct implementation used Algotronix's CAL FPGAs to create a long single-lane road of traffic. The cars can be placed on the road and their behavior with respect to each other can then be simulated. Results of the simulation are obtained reading the entire state from each of the chips used in the simulation. Cars were able to have two speeds (stop/go) and their behavior was determined based on the presence of their nearest neighbor. Naturally, the direct implementation approach has a very high degree of concurrency. Its scalability is limited by the amount of hardware available and the level of data visibility (i.e. the CA neighborhood size) that is required by the simulation.

The second approach by Marc Bumble (Bumble, 2001) implements a generalized system for parallel event-driven simulation. His system consists of an event generator, an event queue, a scheduler, and a unifying communications network in each processing element. Each of these elements are built in reconfigurable hardware at a cost of 30–34 Altera Apex FPGAs. The traffic simulation is calculated by streaming data into processing elements. Each processing element is capable of simulating one source, intersection, or destination node with the associated outbound roads of traffic. Bumble states that a system composed of 8000 processing elements could simulate a large traffic network (at a cost of 24,000 FPGAs).

Bumble does not address the scalability of his approach, the visibility of the simulated traffic, or how data is transfered in and out of the system. Also, his road models are limited to single-lanes with simple four-way intersections. This approach does not easily lend itself to realistic simulation of metropolitan areas. For this we require that the system can handle realistic road traffic configurations (e.g. multiple lanes, merge and turn lanes, all common types of intersections), that it scales well, and that the cost of retrieving/capturing the system state is small.

The work presented here differs from previous approaches in three ways. First, we are using simulation models which are currently in production use. Our model include acceleration, stochastic slow-down, different velocities, and cars with routes. Second, our target is metropolitan areas rather than specialized configuration with a small number of roads and intersections. Previously, the cost of FPGA traffic simulations at the metropolitan scale was too expensive, so as a third point we address the cost of transferring data between the host and the FPGA for analysis or monitoring.

The paper is organized as follows: We first present the basic CA traffic model with its driving rules and intersection logic, as well as aspects of how the computation is divided between the host CPU and the FPGA. We then discuss the performance of our streaming approach. Also, the performance of one of our earlier direct implementations (which is somewhat similar to Bumble's model) is presented along with its obvious limitations.

TRAFFIC MODELING AND THE MICRO-SIMULATOR

The TRANSIMS road network simulator, which is based on (Nagel and Schreckenberg, 1992; Nagel et al., 1995; Rickert et al., 1996), can best be characterized as a cellular automaton computation on a semi-regular grid or cell network. The modeling process starts from a more or less complete city network description in terms of links and nodes. A network link corresponds to a road segment with one or more lanes all of which share the same driving direction. (In TRANSIMS links can have traffic in both directions. Our representation splits such TRANSIMS links in two.) A link may also include merge and turn lanes. Network nodes correspond to locations where there is a change in the road network such as an intersection or a lane merging point.

The CA representation is obtained by splitting each lane into 7.5 meter long cells. In general the neighborhood of a cell includes the cell itself, its immediate left and right neighbor cells plus all cells in front and behind these three cells that are up to five cells away, where five is the maximum car velocity. Of course, the neighborhood is adjusted when a cell does not have a left or right neighbor cell.

For the car dynamics the following conventions are used: A cell can hold at most one car, and a car can travel with velocity $v \in \{0, 1, 2, 3, 4, 5\}$ cells per iteration step. The positions of the cars are updated exactly once per iteration step using a synchronous update, and each iteration step advances the global simulation time by one second.

The basic driving rules for multi-lane traffic in TRANSIMS are given by a four-step algorithm. In each step we consider a single cell i in a given lane and link. Note that our model allows passing on the left and the right. To avoid cars merging into the same lane, cars may only change lane to the left on odd-numbered time steps and only change lane to the right on even-numbered time steps. This convention, along with the four algorithm steps described below, produces realistic traffic flows as demonstrated by TRANSIMS.

Local driving rules

We let $\Delta(i)$ and $\delta(i)$ denote the cell gap in front of cell i and behind cell i, respectively, and we let $v_{\max}(i)$ be the maximum speed of cell i.

1. **Lane Change Decision:** Odd time step t: If cell i has a car and a left lane change is *desirable* (car can go faster in target lane as a result of its Δ being larger) and *permissible* (there is sufficient space for a safe lane change, i.e. δ is large enough) flag the car/cell for a left lane change, see Figure 1. The case of even-numbered time steps is analogous. If the cell is empty nothing is done.

2. **Lane Change:** Odd time step t: If there is a car in cell i, and this car is flagged for a left lane-change then clear cell i. Otherwise, if there is no car in cell i and if the right neighbor of cell i is flagged for a left lane change

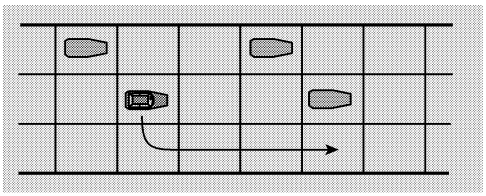


Figure 1: CA Traffic in TRANSIMS

then with probability α move the car from the neighbor cell to cell i. The case of even time steps t is analogous.

3. **Velocity Update:** Each cell i that has a car updates that car's velocity using the two-step sequence:
   - $v := \min(v+1, v_{\max}(i), \Delta(i))$ (acceleration)
   - If $[\text{UniformRandom}() < p_{\text{break}}]$ and$[v > 0]$ then $v := v - 1$ (stochastic deceleration).
4. **Position Update:** If there is a car in cell i with velocity $v = 0$, do nothing. If cell i has a car with $v > 0$ then clear cell i. Else, if there is a car $\delta(i) + 1$ cells behind cell i and the velocity of this car is $\delta(i) + 1$ then move this car to cell i. The nature of the previous velocity update pass guarantees that there will be no collisions.

All cells in a road network are updated simultaneously. The steps 1– 4 are performed for each road cell in the sequence they appear. Each step above is thus a classical cellular automaton Φ_i, and the complete update pass is a product CA, that is, a function composition of classical CAs:

$$\Phi = \Phi_4 \circ \Phi_3 \circ \Phi_2 \circ \Phi_1 \quad (1)$$

Note that the CAs used for the lane change and the velocity update are stochastic CAs. The rationale for having stochastic braking is that it produces more realistic traffic. The fact that lane changes are done with a certain probability avoids slamming behavior where whole rows of cars change lanes in complete synchrony. Note also that some of these passes can and typically will be combined in an implementation which amounts to combining some of the factors in the product (1).

Intersections and Global Behavior

The four basic rules handle the case of straight roadways. TRANSIMS uses travel routes to generate realistic traffic from a global point of view. Each traveler or car is assigned a route that it has to follow. Routes mainly affect the dynamics near turn-lanes and around intersections as cars need to get into a lane that will allow them to perform the desired turn or sequence of turns.

To incorporate travel routes the road links need to have IDs assigned. Moreover, to keep computations local the cells need to hold information about the IDs of upcoming left and right turns. The following describes the extension of the four basic driver rules to handle turn-lanes and intersections.

Modification of the lane change rule

Consider a car in cell i. As before, lane changes to the left/right are only permissible on odd/even numbered time steps. We refer to the adjacent candidate cell as the target cell.

1. If the link ID of the target cell matches the next leg of the travel route and differs from the current link ID a lane change is desirable (desirable turn-lane).
2. Else, if the target cell has a link ID that does not match the next leg of the route and it differs from the current link ID of the route, a lane change is not desirable (wrong turn).
3. Else, apply the basic lane changing rule described above.

Note that this handles lane changing prior to turn-lanes as well as intersections.

Intersection Logic

The TRANSIMS road network data includes a lane connectivity table for each intersection. The table specifies the link and lane targets of every lane in every incoming link. The goal of the additional lane-changing rules is to help cars get into a correct lane so that it can make the turns that are permissible by this table and such that it proceeds along its route.

Each intersection has a set of immediate adjacent road cells. We refer to these as the *intersection road cells*. The intersections operate by dynamically assigning the front and back neighbor cell IDs of the intersection road cells. This allows us to naturally extend the driving rules for multi-lane traffic to intersections without any modifications. The subset of the intersection road cells that come from incoming links have their front neighbor cell set to zero by default. The same holds for the back neighbor of the intersection cells belonging to outgoing links. The intersections operate by establishing front/back pairs between cells to accommodate the routes. Stop intersections and traffic signal intersections impose additional constraints on which cars are allowed to drive at what times by controlling the corresponding connections. Note that our prototype is designed to incorporate travel routes, but this feature is not currently implemented—currently each car's next route segment is set stochastically.

The Computation Model: Software/Hardware Interaction

There are two basic approaches to hardware implementation: a *direct implementation* and a *streaming implementation*. In a direct implementation, each road cell in the whole road network is physically instantiated in an FPGA. The implementation is straightforward and we omit any details: each cell provides its current state to its neighbors so that all the cells in that local neighborhood can calculate their next state. As will be demonstrated in the next section, this leads to computation speeds several orders of magnitude faster than a single-processor software implementation. However, the

results in the next section also indicate that a single FPGA can only hold a few hundred road cells, so the direct implementation requires several thousand FPGAs to simulate even a moderate-sized city (recall that the Portland road network representation has about 6.6 million road cells).

In contrast, the streaming implementation lets a computational unit, an *update engine*, process a *stream* of road data and subsequently output a *stream* of updated data. In this way, the update engine sweeps across the road data, and the number of road cells is no longer limited to available FPGA area. Instead, the only limiting factors are the size of the memory to hold the state of the road cells and the associated memory access time. Thus, a streaming hardware design becomes inherently scalable and can handle large-scale road networks.

To simplify the streaming design, it is desirable that the hardware processing is governed by a simple, homogeneous set of traffic rules. For this reason, we chose to partition the road network in such a way that straight road segments are processed by the hardware, while intersections and merging nodes are updated by the host software module. This hybrid hardware/software strategy means that all road plan decisions are handled by software. Yet, the hardware significantly relieves the computational burden from the software, which is the strength of our hardware/software co-design.

Data representing straight road segments is fed to the hardware update engine against the flow of traffic, starting from the end of each segment. A straight road segment will henceforth be referred to as a road link (or a link). Since the engine lacks knowledge about the network topology and the road plans, it cannot move cars from one road link to another. Specifically, if a car exists in one of the last $v_{\max}$ cells of a road link, it could potentially move to another link in a single update pass, and for this reason, we define the last $v_{\max}$ cells of each link as *overlap region I* (see Fig. 2). (Because of the processing direction of the update engine, cells in overlap region I are processed before other cells.) Cars that are inside overlap region I at the beginning of the hardware pass are not updated by the engine. (Note that the engine can move *other* cars *into* overlap region I.) Instead, the software module is responsible for moving cars inside overlap region I.

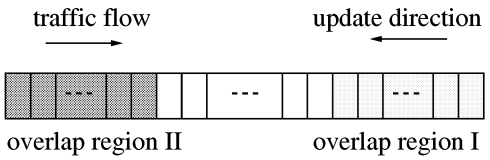


Figure 2: Road Link Structure

The software is also responsible for writing information to the first $v_{\max}$ cells of each road link, which corresponds to new cars moving into a link (arriving from other links, either through an intersection or by merging). However, computing velocities and positions of these new cars requires knowledge of the first $v_{\max}$ cells of each link. The first $v_{\max}$ cells of each link then constitute another region that needs to be overlapped between hardware and software, and this region will be called *overlap region II* (see Fig. 2). Note that the hardware can only move cars *from* these first cells (due to the traffic direction). Thus, if the cells were empty at the beginning of the hardware update pass, the software module does not need to read back the updated status.

The observant reader may have noticed that our discussion on overlap regions was confined to single-lane traffic. Clearly, multiple-lane traffic requires overlap region I to be longer than $v_{\max}$ cells since the lane changing rules assume knowledge of preceding cells (and not only succeeding cells). In order to minimize the need for memory synchronization, it is desirable to process only single-lane traffic in hardware. In fact, 90% of all roads in Portland are one-lane roads, which means that most road cells would still be updated in hardware. The hardware memory constraint (16 MB) allows us to off-load 40% of all straight road segments. We are currently investigating more compact cell data structures in order to off-load 100% of the straight road segments.

Inside the compute engine, a car's velocity is updated first, and then its position. Of course, if no car exists in the incoming road cell, or if the incoming road cell belongs to an overlap region, the incoming velocity is passed out unchanged. In all other cases, the engine initially tests whether acceleration is permissible or not, and this is facilitated by shifting incoming cars through a register with $v_{\max} + 1$ cells. The maximum number of cells a car can move ahead is readily found by scanning the shift register. As explained before, each car is also subject to stochastic slowdown.

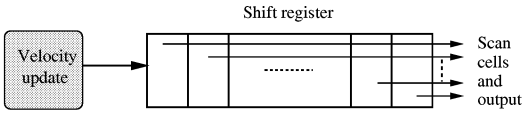


Figure 3: The Position Update

Having calculated the new velocity of a car, the streaming engine calculates the new position by shifting the car one cell every clock cycle (see Fig. 3) until its newly calculated velocity matches the distance from the end of the shift register. At the point when there is a match, the car information is exported to the destination road cell. This pipelined design makes it possible for the update engine to read and write one word of road data every clock cycle. If we have access to N concurrent memories, it is advantageous to instantiate N parallel replicas of the compute engine.

COMPUTATIONAL PLATFORM

The Cray XD1 supercomputer combines high performance microprocessors, a high speed, low latency interconnect network, and reconfigurable computing elements. This provides an environment where data transfer latencies are reduced and bandwidth associated with I/O busses is greatly increased. The tight integration of processors and reconfigurable computing changes the meaning of reconfigurable supercomputing. A supercomputing problem can be split between the

CPU and FPGA with close synchronization and fast communication between software and hardware.

Cray XD1

A single chassis of the Cray XD1, consists of 12 AMD Opteron 200 series processors, with up to 8 Gigabytes of memory per processor. The processors are paired into a SMP processor module as shown in Figure 4. Each processor has 1 or 2 Cray RapidArray links that connect to a fully non-blocking Cray RapidArray fabric switch. The switch provides either 48 GB/s or 96 GB/s total bandwidth between the SMP pairs. The RapidArray fabric switch is able to achieve 1.8 μs MPI latency between SMPs.

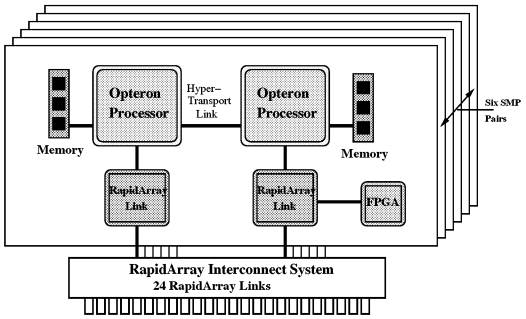


Figure 4: XD1 Processor Module

As shown in Figure 5, one Xilinx Virtex-II Pro (30 or 50) FPGA is available for each processor module from the RapidArray fabric. An FPGA has 3.2 GB/s link to the RapidArray fabric, which connects to the local processors or to other processors on the fabric. The FPGA also has dedicated 2 GB/s RocketIO links to the neighboring SMP module in the same chassis. Four QDR SRAMs are connected to each FPGA providing 3.2 GB/s of bandwidth at 200 MHz (Cray Inc., 2004a).

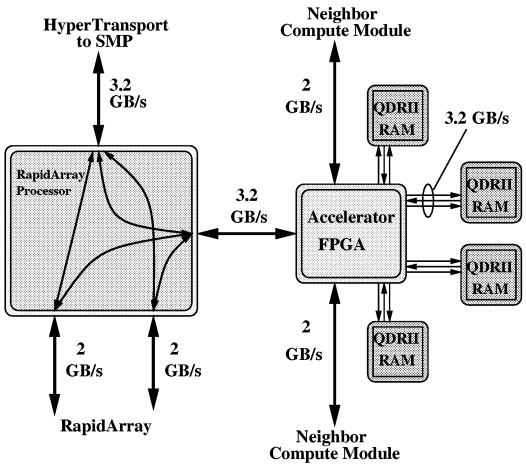


Figure 5: FPGA Expansion Module

System development is performed by first manually partitioning the software and hardware for the reconfigurable system. The FPGA design is written in a Hardware Description Language (HDL) like VHDL or Verilog. This design is combined with behavioral system cores that help simulate the entire system operation. Once the system has been verified in simulation, the design is then synthesized into form that can be used to program the FPGAs. With the FPGA programmed, the software can then be integrated with the FPGA hardware.

On the Cray XD1, the Opteron SMP modules run Linux, and MPI is provided to communicate to other Opteron modules. The FPGAs are accessed under Linux using device drivers provided by Cray. Cray's FPGA API provides functions for loading, resetting, and executing FPGA designs. Functions are also provided for mapping the FPGA's memory into the operating system's memory space and for accessing registers defined on the FPGA.

TRANSIMS on the XD1

Since the FPGAs and processors have tighter integration than most FPGA board systems (e.g., PCI boards), we have partitioned the road traffic simulation between the FPGA and CPUs available in the system. The FPGAs are tailored to process single-lane traffic. The CPUs in the XD1 are responsible for data synchronization between the hardware and the software, and simulating multiple lanes and intersections. Based on the size of the data required, two FPGA nodes are needed for simulation so that all of Portland can fit in the memories available.

The implementation on the XD1 is an improvement over our earlier Osiris based design (Tripp et al., 2004) due in part to better bandwidth between the memories and the FPGA. The QDR SRAMs on the XD1 are fully dual ported and allow for simultaneous reads and writes to any memory location. This provides a large amount of external memory bandwidth to the FPGA.

Despite the large amount of bandwidth on the RapidArray network between the FPGA and the Opteron CPUs, the simulation attempts to reduce the required amount of data traffic. Synchronization of data only occurs if there are cars on a particular road segment and only in the overlap (shared) data regions. This allows for better trade-off between calculation and available bandwidth.

XD1 Communication

As described in the *XD1 FPGA Development* manual (Cray Inc., 2004b), there are asymmetric costs associated with reads and writes between the host and the FPGA's external QDR SRAMs. Writes can take advantage of write combining in the Linux kernel and are non-blocking. On the other hand, reads are blocking and cannot be combined. This creates an asymmetric cost which must be overcome.

The method for accessing the FPGA's SRAMs is shown in Figure 6. The multiplexor (mux) between the traffic engine and the RapidArray Transport (RT) core allows both the host and the traffic engine to write to the SRAMs. Reads do not

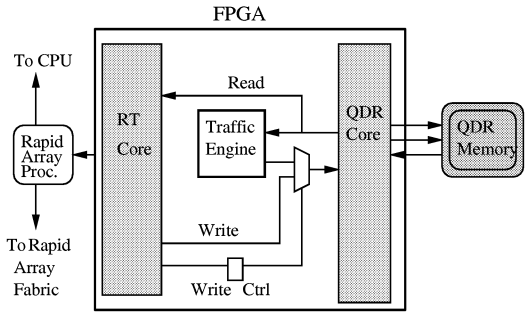


Figure 6: Default FPGA QDR SRAM access (only one memory shown)

require a mux for the data, but both read and write require muxes for the address lines (not shown). Using this setup we benchmarked reads and writes from the host to the FPGA.

Table 1: Read and Write Bandwidth (MB/s) for Hosts to QDR SRAMS

	Array	Pointer	Memcpy
Read	5.94	5.95	6.01
Write	1260	1320	1320

Table 1 shows the bandwidth results for reads and writes accomplished using three different approaches. The FPGA SRAMs are mapped into the hosts local memory and can be accessed using arrays with indexes, using pointers, or by using the *memcpy* function call. Array accesses were found to be slight slower, but the bandwidth is roughly equivalent. However, the difference in cost between reads and writes is a factor of 200. This asymmetric cost is overcome by using a "write-only" architecture as suggested by Cray. The host must write data to the FPGA memory and the FPGA must write the data that the host needs to read.

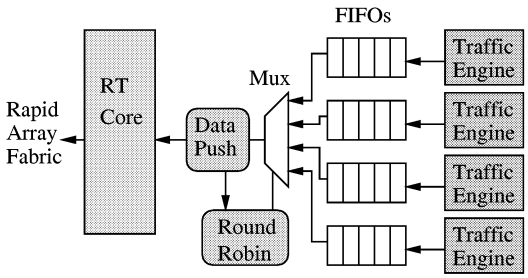


Figure 7: FPGA Push to write data to the Opteron host

Figure 7 details the hardware added to the traffic design to make it a write-only architecture. Only cars in overlap areas are transferred to the host. These overlap cars are stored in FIFOs since they can arrive in bursts of up to ten cars. These bursts are caused because overlap region II of one link is adjacent to overlap region I of the next link. The number of cars in overlap areas is dependent on the traffic data in the memories, so a data push process watches each of the FIFOs through a mux that is rotating through the FIFOs in a round-robin fashion. When the data push process sees a FIFO with data in, it stops the round-robin control and sends up to four data words to the host (due to HyperTransport requirements). Finally, after sending the data the round robin control moves the mux to the next FIFO. Independent of the data push, if a FIFO becomes full it will disable its traffic engine until the FIFO is empty again.

Use of the FPGA data push process allows the traffic design to reduce the time required to transmit the overlap data updates between the FPGA and the host. This allows more time on the processor and FPGA to be spent processing data instead of communicating.

RESULTS AND PERFORMANCE

We now present results for the direct implementation as well as the streaming implementation. The two approaches represent two extremes in terms of concurrency and scalability.

Direct Implementation

The direct implementation was written in VHDL and synthesized to EDIF using Synplify v7.6. The EDIF description was subsequently passed into Xilinx ISE v6.2, targeting a Xilinx XC2VP100 (V2p100) FPGA.

Table 2 compares the results of a hardware implementation of two-lane circular traffic using the Xilinx V2p100 and a software implementation of two-lane circular traffic running on a 2.2 GHz AMD Opteron processor. The V2p100 simulates traffic at a rate $1175.4\times$ the Opteron. This speedup comes primarily from the fact that the FPGA implementation is updating all cells concurrently, and the software implementation, which may have instruction level parallelism, updates each cell individually.

Table 2: Direct Implementation Results Comparison for Two Lanes

	V2p100	Opteron
Cells	640	2 Million
Cells/sec	6.70×10^9	5.7×10^6
Speedup	1175.4	1.0

Clearly, the direct implementation offers very high speedup, but how about the area requirements? As Table 3 indicates, the FPGA chip can only hold at most 650 road cells. Since the Portland road network representation has about 6.6 million road cells, it would require about 10,000 FPGAs to simulate the entire city. (In fact, 10,000 is a modest estimate since the intersection logic would have to be more complex than the logic needed for processing straight road segments, thus requiring more area.)

Table 3: Direct Implementation Design Results

	One-lane	Two-lane
Cells	650	640
LUTs/Cell	97	128
Clock(MHz)	64.17	62.8
(% of FPGA Area)	(71%)	(92%)

Streaming Implementation

The streaming implementation was written in VHDL and placed in VHDL interfaces provided by Cray for the RapidArray and QDR SRAMS (release 1.1). All of this was synthesized using Xilinx XST and the bitstream generated by Xilinx ISE v6.2.

The results shown in Table 4 were timed using a timer register, called the Time Stamp Counter (TSC), which measures processor ticks at the processor clock rate. The 64-bit read-only counter is extremely accurate, as it is implemented as a Model-Specific Register inside the CPU. The overhead of using this register is extremely low and the TSC register on the 2.2 GHz Opteron has a resolution of 450 picoseconds. Now, Table 4 shows that an FPGA design with four update engines (limited by the number of available memories from/to which data can be streamed) operates at a rate $126.3\times$ the speed of a comparable software version running on a 2.2 GHz Opteron.

Table 4: Comparison of Streaming with Software Simulation

	V2p50	Opteron
Clock(MHz)	180	2199
Cells/sec	7.2×10^8	5.7×10^6
Speedup	126.3	1.0

The results in Table 4 are somewhat unfair as they do not account for the cost of transferring data from and to the memories. Instead, Table 5 shows a more accurate speedup which includes the cost of transferring data to and from the FPGAs. If results calculated by the FPGA are read directly by the host, the speed up is only $4.5\times$. This is due to the blocking nature of reads from the host. Pushing the results up to the host results in a more impressive speedup of $34.4\times$.

Although the streaming implementation is a factor of 30 slower than the direct approach, it is still enough of an improvement to provide significant overall speedup. Additional speedup is still possible with more FPGA boards. The most crucial limiting factor in this implementation is the number of memory banks on each board; additional banks would allow us to increase the number of simultaneous data streams. In fact, with the current design, one compute engine requires less than 5% of chip area. Since each compute node has four concurrent memories, it is advantageous to instantiate four parallel engines, but already at this moderate level of parallelism, we run into a bandwidth bottleneck.

Table 5: Comparison of Streaming Including Communication Costs

	V2p50		
	w/o Push	Push	Opteron
Cells/sec	2.56×10^7	1.96×10^8	5.7×10^6
Speedup	4.5	34.4	1.0

The hardware performs extremely well with the straight lane segments, which make up 70–90% of the road segments in a given simulation. FPGA aided simulation done in the scalable, streaming approach may be the fastest way to do extremely large metro-area traffic simulations, especially in light of the advances being made in combined microprocessor/FPGA computing systems. The cellular nature of the road segments meshes well with hardware, and a combined hardware/software approach for the full-fledged simulation fits each of their computational strengths.

System Integration Issues

We will now discuss costs associated with integrating the FPGA computation engine with the microprocessor and its software environment, and how a realistic road network description of Portland affects this integration.

A profile of the software execution time is shown in Table 6. The transaction cost of transferring data from the FPGA hardware to the software by having the microprocessor read the FPGA memory proved to be unacceptably long—12.3% of the overall execution time of one simulation step. As a remedy for this problem, we implemented a push back mechanism in which the FPGA writes back data to a memory space that is local to the host. This solution led to a hybrid design where the communication time for data transfer is negligible (see Table 6).

Table 6: Percentage of Overall Execution Time of one Simulation Step

	Without Push	With Push
Network state update	8.3%	9.7%
Software to hardware	0.2%	0.18%
Intersections	14.0%	13.8%
Lane change update	13.3%	16.8%
Velocity update	21.9%	25.5%
Position update	30.0%	33.8%
Hardware to software	**12.3%**	**0.14%**

Another bottleneck is the size of the FPGA memory. Out of 6.6 million road cells, the 16 MB available FPGA memory can only hold 2 million cells, assuming that 64 bits are used to describe the status of each road cell. If the memories were deeper, it would be possible to off-load significantly more of the overall computation to the FPGA.

We are currently investigating two alternatives to getting more road cells processed by hardware: (i) a more succinct representation using only 32 bits instead of 64, and (ii) on-line FPGA data compression. As already explained, the software is responsible for putting new cars in each single-lane road segment (corresponding to cars coming from adjacent segments). By also keeping track of the order in which these new cars enter each segment, we would not have to encode and send along any car IDs, since cars obviously have to leave a single-lane road segment in the order in which they enter. This implicit road ID method would reduce the size of a road cell to 32 bits.

To do compression, the FPGA could compress the data written by the microprocessor using a simple on-line scheme such as run-length encoding. This is possible because the data written by the microprocessor must first go through the FPGA. We could also use more aggressive compression techniques by exploiting the semantics of the traffic data.

SUMMARY AND FUTURE WORK

FPGA-aided simulation using a scalable, streaming implementation has the potential of being the most effective way to do large metro-area traffic simulations, especially in light of the advances being made in combined microprocessor/FPGA computing systems. The cellular nature of the road segments matches well with hardware, and a combined hardware/software approach for the full-fledged simulation fits each of their computational strengths. Apart from being useful in the context of traffic simulation, our study also provides guidelines for implementing hybrid designs for other types of simulations (Atkins and et al, 2003; Eubank et al., 2004).

The TRANSIMS hybrid simulation demonstrates the high cost of communicating data between different processing nodes. Considering the communication cost, the speedup is $34\times$. However, $34\times$ only refers to the speedup of single-lane road sections. The overall speedup is limited by Amdahl's law, and a factor of 5 would be possible if the entire dataset could fit in the SRAM memory banks. We discussed alternative methods to achieving the potential $5\times$ speedup such as data compression, increasing the number of FPGA nodes used, and increasing the size of the memories. Our work demonstrates that partitioning and system integration issues play a crucial role in realizable speed up for hybrid simulation systems.

We are currently only using one compute blade of the Cray XD1 and only one of the two Opteron processors on that blade. With this design we estimate the speedup relative to a single processor software implementation to be of the order 50–100%. The Cray XD1 has six blades, so we are considering a parallel implementation where each process would be a copy of our existing design potentially yielding speedup factors of 5–10.

ACKNOWLEDGMENTS

This research was supported in parts by Laboratory Directed Research and Development under DOE contract W-7405-ENG-36 to the University of California for the operation of the Los Alamos National Laboratory.

REFERENCES

Atkins, K. A. and et al (2003). ADHOPNET: Integrated tools for end-to-end analysis of extremely large next generation telecommunication networks. Technical report, Los Alamos National Laboratory, Los Alamos, NM.

Barrett, C. L. and et al (2002). TRansportation ANalysis SIMulation system (TRANSIMS) Portland study reports.

Bumble, M. D. (2001). *A Parallel Architecture for Non-deterministic Discrete Event Simulation*. PhD thesis, Pennsylvania State University.

Cray Inc. (2004a). *Cray XD1 Datasheet*. Cray Incorporated, Seattle, WA USA.

Cray Inc. (2004b). *Cray XD1 FPGA Development*. Cray Incorporated, Seattle, WA USA.

Eubank, S., Guclu, H., Kumar, V. S. A., Madhav, M. V., Srinivasan, A., Toroczkai, Z., and Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184.

Milne, G., Cockshott, P., McCaskill, G., and Barrie, P. (1993). Realising massively concurrent systems on the space machine. In Pocek, K. and Buell, D., editors, *FPGAs for Custom Computing Machines*, pages 26–32, Napa, CA USA. IEEE Computer Society, IEEE Computer Society Press. Inspec 4630521.

Nagel, K. and Schreckenberg, M. (1992). A cellular automaton model for freeway traffic. *Journal de physique I*, 2:2221–2229.

Nagel, K., Schreckenberg, M., Schadschneider, A., and Ito, N. (1995). Discrete stochastic models for traffic flow. *Physical Review E*, 51:2939–2949.

Rickert, M., Nagel, K., Schreckenberg, M., and Latour, A. (1996). Two lane traffic simulations using cellular automata. *Physica A*, 231:534–550.

Tripp, J. L., Mortveit, H. S., Nassr, M. S., Hansson, A. A., and Gokhale, M. (2004). Acceleration of traffic simulation on reconfigurable hardware. Technical Report LA-UR 04-2795, Los Alamos National Laboratory, Los Alamos, NM USA.

Simulation of modern Traffic Lights Control Systems using the open source Traffic Simulation SUMO

Daniel Krajzewicz, Elmar Brockfeld, Jürgen Mikat, Julia Ringel, Christian Rössel, Wolfram Tuchscheerer, Peter Wagner, Richard Wösler
German Aerospace Centre, Institute for Transportation Research
Rutherfordstr. 2
12489 Berlin
Germany
E-mail: Daniel.Krajzewicz@dlr.de, Elmar.Brockfeld@dlr.de, Jurgen.Mikat@dlr.de, Julia.Ringel@dlr.de, christian.roessel@gmx.de, Wolfram.Tuchscheerer@dlr.de, Peter.Wagner@dlr.de, Richard.Woesler@dlr.de

KEYWORDS

Microscopic traffic simulation, open source, traffic lights, traffic research

ABSTRACT

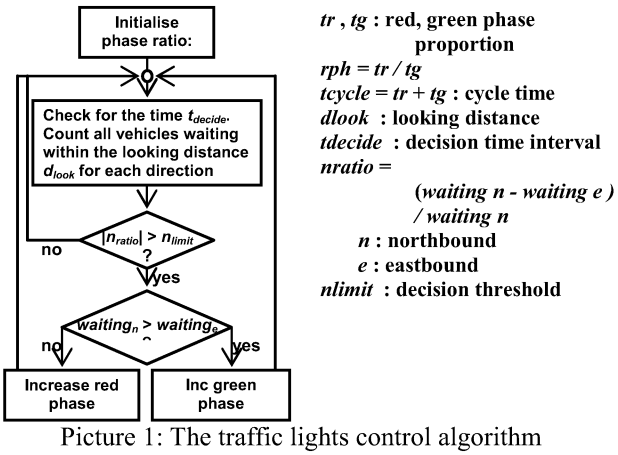
Within the project "OIS" (optical information systems) new traffic control mechanisms had to be invented and tested. One of the most important topics was to optimize the flow over a junction using information from the OIS sensors which can not be measured using normal sensors such as induct loops. For this purpose, an "agentbased" traffic lights logic algorithm was used, which uses the length of a jam in front of a traffic light as input. As we had no possibility to test the traffic lights control within the reality, the improvement of the flow throughput of such junctions was shown using the open source traffic Simulation "SUMO" (Simulation of Urban MObility) [1, 2].
This publication describes the algorithm itself and how it was embedded within the simulation. Furthermore, the simulation results are given.

INTRODUCTION

Simulations are often used to test new systems before implementing them into the real world. This also counts for traffic simulations. Within a project which has been done in a co-operation with universities and a number of companies from Berlin and Brandenburg, new optical sensors have been developed (see [3, 4]). Using digital cameras, we are able to observe traffic and gain areal information, including the length of jams on a street or trajectories of vehicles. Besides developing these systems themselves, another goal of the project was to invent mechanisms which use such information for traffic optimization.
To show the capability to improve traffic flow, we decided to implement our agentbased algorithm (see [5] for a description) for traffic lights controls within the microscopic road traffic simulation "SUMO" ([1, 2]). SUMO is an open source traffic simulation developed at our institute. As the OIS-system has been tested within a certain area around our institute's place, we have used a digital network representing it as our simulation scenario.
The OIS sensors themselves will be not discussed, herein, but only their simulation and the simulation of the traffic lights control algorithm. At first, this algorithm will be presented, followed by a description of how the simulation was prepared. We will then give the simulation results, followed by some conclusions and a discussion of open questions.

THE TRAFFIC LIGHTS CONTROL ALGORITHM

The agentbased traffic lights control algorithm was firstly presented in [5]. The main idea is that each traffic light is trying to solve the jams in his front by itself. To achieve this, he looks into the incoming lanes and measures the jam lengths on these lanes. If at one of these lanes the jam gets longer, this lane gets green for a longer time. Beside these assumptions, several parameters prevent the system from oscillating and from adapting too fast or too strong. This is done by increasing a green phase's duration only if a jam is longer than a threshold. Furthermore, the jam has to occur for a certain amount of time. There are further boundaries for the duration of a phase – beside the standard value given at the begin, a phase must not be longer or shorter than predefined thresholds. The whole algorithm is shown within picture 1.
Beside the advantage to be very simple, the agentbased traffic lights logic can be set on top of existing traffic lights and tries to adapt them to the current traffic amount.



Picture 1: The traffic lights control algorithm

SIMULATION PREPARATION

Setting Up The Scenario

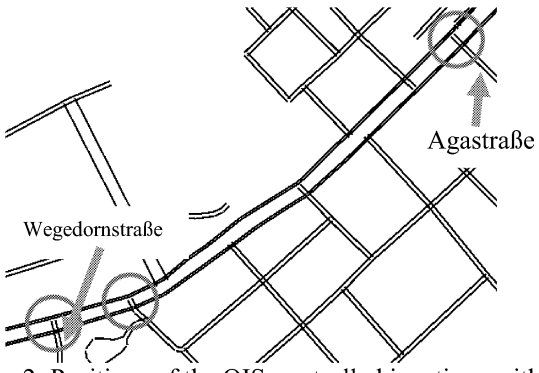
The description we used was a database containing the road network of Berlin, converted from a digital map supplied by NavTech. To extract the area we wanted to use, we had to extract all nodes (junctions) and edges (streets) within a certain rectangle. This was done straightforward using simple SQL-queries.

For the final presentation, two situations had to be shown simultaneously, one showing the scenario running with normal traffic lights as they are implemented in reality, and one using the OIS-detectors. To achieve this, we have duplicated the lists of edges and nodes, shifted them by several hundred meters – the width of the scenario plus an offset – and included this second list within the first one. These lists were then converted into XML files which the SUMO network converter is capable to read. Doing this, we gained two networks combined into a single which could be used for simulating the scenario.

Furthermore, traffic lights definitions computed by the "NETCONVERTER" had been replaced by definitions that exactly match the real world traffic light logics. This has been done for both parts of the network – the one using normal traffic lights logics and the one using agentbased traffic lights logics.

The routes were retrieved from measures of the real world scenario area. Picture 2 shows the junctions the traffic was counted at. From these measures, junction turning percentages for each of the participating, incoming edges were computed. These were then fed, together with the traffic amount to one of SUMO's routing modules for generating single vehicle trips.

Of course, the routes had to be duplicated for the second network, too, in order to gain exactly the same flows for both parts of the network.



Picture 2: Positions of the OIS-controlled junctions within the simulation (marked by circles)

Implementing The OIS Sensors

The OIS sensors were simulated using areal detectors (see [2] for a further description) which look at all lanes in front of the junctions which are meant to be equipped with OIS sensors. Beside the jam length in meters, other values are collected, such as the jam length in vehicles, the number of vehicles, the occupancy degree, the vehicles' mean speed, the halting durations, and several more.

Implementing The Traffic Lights Control

The traffic light control was implemented by extending SUMO's representation of normal traffic lights logics (done using derivation in C++). The application's interface was changed in a way which allows the traffic light to extend its phases. Besides implementing the traffic lights control themselves this way, some further changes had to be done. Among them was the addition of the ability to change the traffic lights parameter on application start and to read additional information about the phases' minimum and maximum lengths from the network description.

Below, the algorithm that is executed every time the light switches is given:

- find the maximum queue length for the lanes that have green light during the current phase
- save this queue length into a list A which holds such values for the last l_h cycles (remove past values if needed)
- if the time after the last decision > t_decide:
  - compute the mean queue length for all phases by averaging the values within list A with the length of list A
  - compute the phase with the largest queue in front
  - compute the phase with the smallest queue in front
  - quit computation if either the phase with the largest queue in front may not be lengthen or the phase with the shortest queue in front may not be shorten
  - compute n_ratio:
    - n_ratio = (max queue length – min queue length) / max queue length
  - if n_ratio > n_limit:
    - increment phase length of the phase with the largest queue length by one
    - decrement phase length of the phase with the shortest queue length by one

With:

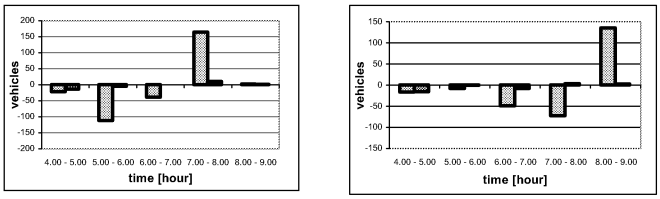
l_h:	learn horizon
t_decide:	decision time interval (see Pic. 1)
n_ratio:	normalised queue length delta between the longest and the shortest queue
n_limit:	decision threshold (see Pic. 1)

SIMULATION RESULTS

The comparisons were created by writing detector measures into files and evaluating them after the simulation has ended. For this purpose, areal detectors, the same as used by the agentbased traffic lights logics, were laid in front of the junction. Two measures were used to compare the traffic lights controls: a) the jam in front of them and b) the throughput of the junctions. The second was computed by summing up the number of vehicles that leave the junction and subtracting the number of vehicles that approach the junction.

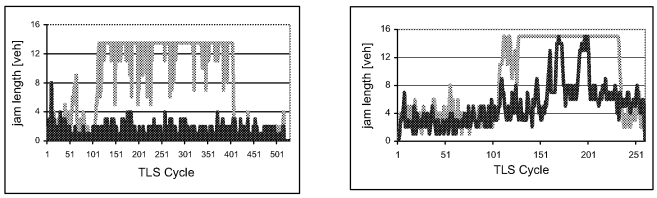
While using the original traffic amount, almost no difference between the OIS controlled and normal traffic lights controls could be observed. For this reason a further scenario has been implemented where the vehicle flow on one of the incoming edges was incremented for the time between 5am and 5.30am. After doing this, the

configurations showed large differences, which prove the improvement of the junctions' phases when areal sensors combined with the agentbased traffic lights control are used.
Below, comparisons for two of the three regarded junctions are given, using the junctions' throughput (picture 3) and the jam lengths in front of the junctions (picture 4). We will now discuss it, briefly.
As the amount of approaching vehicles increases, a normal junction (shown in light grey within pictures 4 and 5) is not "prepared" to solve all incoming vehicles. Most of them get stucked in front of the junction. Due to this, there are more approaching than leaving vehicles and the "throughput" gets negative. After the additional flow has been inserted into the network, the number of vehicles that try to pass the junction decreases and the vehicles that were waiting in front of it are leaving the junction consecutively. This is the reason for the high positive peak in the normal junctions' throughput. As one can see, there are neither positive nor negative peaks within the agentbased junctions (shown in dark grey), what shows that such junctions are capable to solve the additional demand.



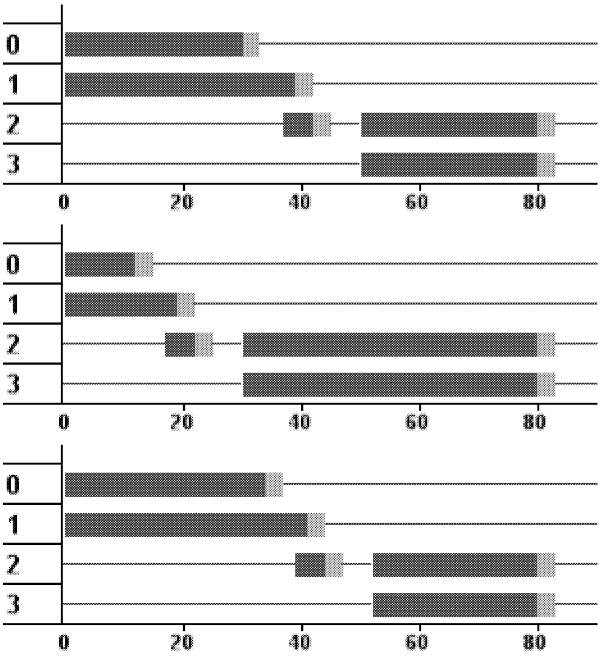
Picture 3: Comparison of the throughput of fixed and agentbased traffic lights logics; light: throughput of fixed tls, dark: throughput of agentbased tls; left: junction Agastraße, right: junction Wegedornstraße

This ability is even more visible when the jams in front of the traffic lights are examined (picture 4). Here, one can see that jams are much longer in front of normal traffic lights when compared with agentbased traffic lights. The constant maximum in front of normal traffic lights is due to a certain length of the detectors that were used to compute the jam lengths.



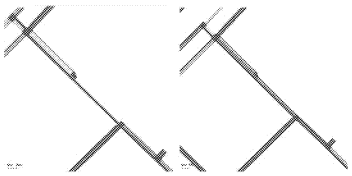
Picture 4: Comparison of the jam lengths in front of fixed and agentbased traffic lights logics for a certain lane; light: throughput of fixed tls, dark: throughput of agentbased tls; left: junction Agastraße, right: junction Wegedornstraße

Picture 5 shows the difference in the traffic lights logic located at the bottom of the rightmost junction from picture 2. One can see here how agentbased traffic light logics adapt to an increased flow and how they readapt to a normal situation as soon as the queues in front of it get similar long.



Picture 5: Comparison of the traffic lights logic before (top) and after the optimization (middle); the display at the bottom shows the phases after the increased flow has passed

Besides collecting the values as shown above, a visualisation component was developed which allows viewing the aggregated values over all lanes. Picture 6 shows the situation at about 6am comparing a normal adaptive traffic lights logic and the agentbased logic equipped with OIS sensors. One can see that the jam, displayed by the luminance is much longer when the normal traffic lights logic is used.



Picture 6: Comparison of the density on the streets when using normal (left) and agentbased traffic lights logics (right); the lighter the street, the higher the density

This increased throughput is what is wanted, but it also contains traps. The simulation shows that the succeeding traffic lights may not be able to solve the greater incoming amount of vehicles. If this happens, the system's jam length may even increase, although locally the throughput of a junction is improved.

CONCLUSIONS

Although the agentbased traffic lights logics show clearly their benefits at certain circumstances, no improvement could be achieved as long as the flows were small and balanced between the different directions of a traffic light.

This means that the best place for such controls in real life would be junctions where the flow is changing tremendously from time to time, for example in places near to venues.
A second result is that we have to investigate the interrelationship between consecutive traffic lights to avoid generation of larger jams in front of succeeding traffic lights. Without such coordination mechanisms, agentbased traffic lights are most effective when used solely, in a larger distance to other traffic lights.
At last, we want to point out that this research would cost a much greater effort when a commercial simulation application had to be used. Especially the implementation of new on-road systems would not be possible or at least very time consuming as long as the simulation's source code is not open. That's why we want to encourage you to take a look at SUMO's project pages located at http://sumo.sourceforge.net/ and try out the software.
Unfortunately, we are not able to make the scenario downloadable, because it is based on commercial NavTech data.

FUTURE WORK

We have seen that some further research has to be done on coordinating the lights. A mechanism for this is not yet designed and should be done as next.
There are also further systems to be evaluated. Currently, we develop a system which reidentifies vehicles that pass a set of camera-equipped junctions (see [6] for a description). Among other things, travel times are computed from these reidentified cars. These travel times are valuable data, e.g. for further optimization of traffic light control. Such traffic improvement methods, could be simulated, too, and may also be one of the further steps.

REFERENCES

[1] D. Krajzewicz, G. Hertkorn, C. Rössel, P. Wagner. 2002. "SUMO (Simulation of Urban MObility); An open-Source Traffic Simulation" Proceedings of the 4th Middle East Symposium on Simulation and Modelling (MESM2002), Edited by: A.~Al-Akaidi, pp. 183 - 187, SCS European Publishing House, ISBN 90-77039-09-0

[2] D. Krajzewicz, G. Hertkorn, C. Rössel, P. Wagner. 2002-2005. SUMO Homepage. http://sumo.sourceforge.net

[3] http://www.dlr.de/vf/forschung/projekte/ois

[4] http://www.fav.de

[5] J. Mikat, E. Brockfeld, P. Wagner. 2003. "Agent Based Traffic Signals on a basic grid". Proceedings of the 4th Workshop on Agent-Based Simulation. Edited by: Jean-Pierre Müller, Martina-Maria Seidel. SCS European Publishing House.

[6] R. Woesler. 2004. Real-time recognition and reidentification of vehicles from video data with high reidentification rate, Proceedings of the 8th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2004), Orlando, Florida, USA. Edited by: Nagib Callaos, William Lesso, Ashraf Ahmad, Vol. VI, pp. 347-352, International Institute of Informatics and Systemics, ISBN 980-6560-13-2

BIOGRAPHY

Born in Bydgoszcz, Poland, 1972, **Daniel Krajzewicz** has finished his study of computer science at the Technical University in Berlin by the middle of the year 2000 with artificial intelligence and computer graphics as main topics. After work on text classification he changed to the Institute for Transportation Research of the German Aerospace Centre in 2001 where he now works on a cognitive driver model and the open-source urban traffic simulation "SUMO".

ENERGY SYSTEMS SIMULATION

INTEGRATION OF A HYDROGEN ECONOMY INTO THE GERMAN ENERGY SYSTEM: AN OPTIMISING MODELLING APPROACH

Michael Ball and Otto Rentz
French-German Institute for
Environmental Research (DFIU/IFARE)
University of Karlsruhe
Hertzstrasse 16
D-76187 Karlsruhe
e-mail: Michael.Ball@wiwi.uni-karlsruhe.de

Martin Wietschel
Fraunhofer Institute for
Systems and Innovation Research (FhG ISI)
Breslauer Straße 48
D-76139 Karlsruhe
email: Martin.Wietschel@isi.fhg.de

KEYWORDS
Hydrogen energy system, energy system model *MOREHyS*, hydrogen infrastructure, GIS, alternative fuels.

ABSTRACT

Possible limitations of oil supply are increasingly triggering the search for alternative automotive fuels. One such fuel, that is recently attracting a lot of attention from both industry and research communities, is hydrogen. The following paper describes a novel modelling approach developed to identify optimal ways for implementing a supply infrastructure for a hydrogen-based transport system in Germany and to assess the resulting effects on the national energy system. Some first results of this model-based system analysis are presented.

1. INTRODUCTION

A projected increase in global energy demand, the economic and geopolitical implications of future shortcomings in the supply of oil, and consequently concerns about energy supply security have brought the discussion about hydrogen as future energy carrier back to the agenda. Notably vulnerable to shortcomings is the transport sector, which still depends worldwide to about 95 % on oil (IEA 2004). Of all the alternative fuels being under discussion for meeting the future demand for automotive fuels, hydrogen seems particularly promising, as it can contribute to the reduction of urban air pollution and as it offers the long-term possibility to be solely produced from renewable energies.

As there is no clearly outstanding hydrogen pathway in terms of economics, CO_2 emissions and primary energy consumption, and as the implementation of an operational infrastructure for the production and distribution of hydrogen as well as the cost-effective production of hydrogen vehicles will require considerable investments over several decades, it is important for all stakeholders to start defining a strategic orientation already by today.

The application of a model-based system analysis to compare and evaluate different hydrogen pathways and their integration into national energy systems, taking into account different energy-political and country-specific framework conditions, is quite new. In addition, modelling approaches (e.g. Tseng and Friley 2003) generally lack an appropriate geographical representation of a hydrogen transport infrastructure, such as location and distribution of demand centres or production sites and related transport distances, modes and costs. Therefore, the *MOREHyS* (Model for Optimisation of Regional Hydrogen Supply) model has been developed, allowing to identify the cost-optimal way for building up and implementing a hydrogen supply infrastructure within a country-specific context (Germany), for an *exogenously given hydrogen demand.*

The objective of the modelling approach is to assess by means of an energy system analysis the economic and environmental effects of using hydrogen as an energy carrier in mobile applications in Germany. The question to be answered within this context is how it could practically be realized to substitute a certain percentage of vehicle fuel by hydrogen, considering country-specific conditions, such as future development of energy and fuel demand, or availability and price development of fossil and renewable energy carriers. The analysis covers a time horizon until 2030.

2. THE MOREHyS MODEL

2.1. General Modelling Approach

MOREHyS, having been developed by *DFIU* in cooperation with *FhG ISI*, is based on the open-source *BALMOREL* model, which has initially been developed to support analyses of the energy sector in the Baltic sea region, with emphasis on the electricity and combined heat and power sec-tors (see also http://www.balmorel.com). The model is implemented in the algebraic modelling language *GAMS*.

MOREHyS is a bottom-up, mixed-integer, linear optimisation model with exogenous demand for electricity, heat and hydrogen. The model determines the minimal cost of production (as well as transport/transmission) and new investments, given some constraints for each year such as maxi-mum emission levels or regional limitations for generation capacities and fuel availability, which might effect the cur-rent energy mix. As at the same time not only the demand for hydrogen, but also for electricity and heat needs to be satisfied in the analysed energy system, hydrogen production has to compete for the same primary energy carriers (fossil and renewable) as conventional electricity and heat generation (see Figure 1). Further interconnections with the conventional energy system concern CO_2 emissions as well as hydrogen production via electrolysis.

The objective function used for the optimisation, which is sequentially carried out on a year-by-year basis, is cost minimisation for the whole region each year. In the optimi-

sation procedure existing technologies compete with their variable and fixed costs against new technologies with their additional annualised investments. Dynamics between the years are introduced by transferring the results of the optimisation (i.e. endogenously found capacities of production and transport/transmission) in the previous year to the beginning of the subsequent year. Thus, the single optimisation periods are interlinked through capacity accumulations and the yearly decommissioning of old capacities. Since the model has *perfect foresight* only within the year, but not be-yond, decision-making can be described as *myopic*.

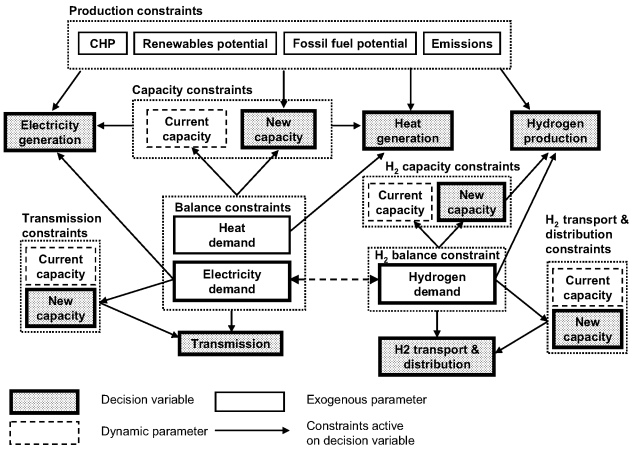


Figure 1: Simplified structure of MOREHyS

Optimisation variables of the model are level of electricity, heat and hydrogen production per time period, technology type and area; investment in new generation capacity per technology type and area; electricity transmission and new investments in transmission capacities; and amount of hydrogen transported between and within all hydrogen areas. At country level, the model further derives resulting CO_2 emissions or total primary energy consumption.

2.2. Integration of Technology Learning

Few energy system models include rudimentary endogenous mechanisms of technological change, such as learning curves, which describe how cost decline with cumulative production. Since learning curves exhibit increasing economies of scale, their implementation results in non-convex non-linear mathematical problems if placed insight the optimisation of perfect foresight models. The *myopic approach*, however, allows to integrate the learning curve concept outside the actual optimisation and thus does not cause mathematical difficulties. The economic parameters are recalculated at the end of each optimisation period (based on the cumulatively installed capacity and the endogenously found new capacity of the current year) and transferred as new input to the following period. Technological learning is applied to both renewable energy generation technologies and hydrogen production technologies.

2.3. Modelling of a Hydrogen Infrastructure

MOREHyS permits the specification of geographically distinct entities, which are *countries*, *regions* and *areas*, where each country is constituted of one or more regions, and each region of zero or more areas. The areas are the building blocks with respect to the geographical dimension: all generation technologies and capacities as well as hydrogen demand are described at the level of areas. For the moment, the following hydrogen areas (urban and rural) are distinguished: Hamburg, Berlin, Ruhr area, Stuttgart and Munich. Figure 2 shows schematically how the hydrogen infrastructure options are modelled in *MOREHyS*.

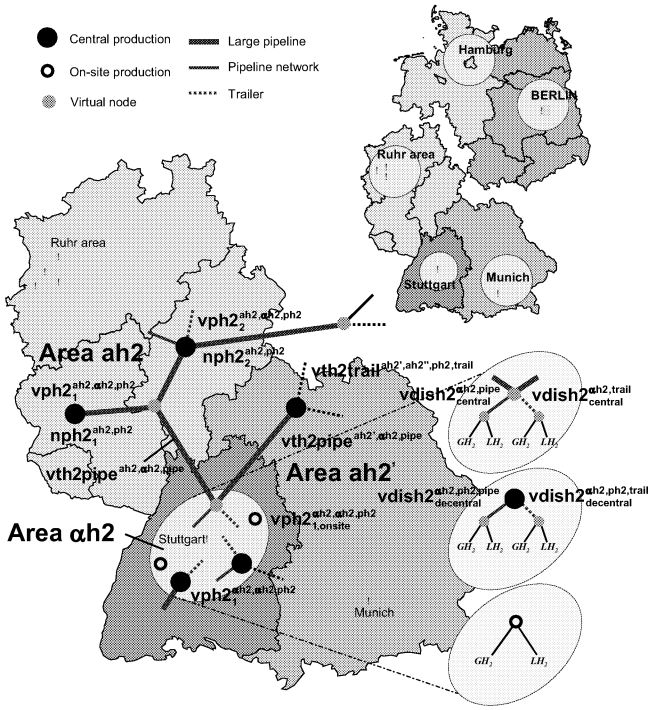


Figure 2: Hydrogen infrastructure options in MOREHyS

$vph2^{ah2,\alpha h2,ph2}$	variable H_2 production on technology ph2 in production area ah2 for demand area αh2
$nph2^{ah2,ph2}$	integer number of production plants ph2 in production area ah2
$vth2pipe^{ah2,\alpha h2,pipe}$	variable H_2 pipeline transport *from* production area ah2 *to* demand area αh2
$vth2trail^{ah2,\alpha h2,ph2,trail}$	variable H_2 trailer transport *from* production plant ph2 in area ah2 *to* demand area αh2
$vdish2^{\alpha h2,pipe}_{central}$	variable H_2 pipeline distribution from central transport node *within* demand area αh2
$vdish2^{\alpha h2,trail}_{central}$	variable H_2 trailer distribution from central transport node *within* demand area αh2
$vdish2^{\alpha h2,ph2,pipe}_{decentral}$	variable H_2 pipeline distribution from production ph2 *within* demand area αh2
$vdish2^{\alpha h2,ph2,trail}_{decentral}$	variable H_2 trailer distribution from production ph2 *within* demand area αh2
GH_2, LH_2	Filling stations for gaseous and liquid H_2 demand

2.4. Application of a Geographic Information System

Infrastructure costs for transport and distribution of hydrogen are partially determined by topological characteristics of the considered regions, while production costs for hydrogen are often site-specific. In order to take these „ real world“ costs into account in the analysis of a hydrogen-based transport system, a *GIS* (Geographic Information System) is applied. In a first step, the GIS has been used for estimating average transport distances between and within

the areas of the model for pipeline and trailer transport of hydrogen. Realistic estimations of transport distances are of major importance as they influence not only the optimal transport modes but also the associated transport costs.

3. DATA BASIS

As it is not possible in the context of this paper to present a comprehensive set of the data used as input to the model, only the major assumptions effecting the introduction of hydrogen as vehicle fuel are discussed. However, the most important references used to derive hydrogen technology data were (Amos 1999), (HyNet) and (HySociety 2004).

Table 1 illustrates the hydrogen consumption of the vehicles distinguished in *MOREHyS* at the beginning of the assumed vehicle uptake in 2010 (DWV 2005, EUCAR 2004). The consumption is differentiated between passenger cars, light duty vehicles (LDV) and buses, as well as between vehicles with fuel cells and internal combustion engines (ICE). The figures for both fuel cell and ICE vehicles are averaged between hybrid and non-hybrid vehicles.

Table 1: Consumption of hydrogen vehicles

Year 2010	[kg/100km]	[l oil equiva-lent/100km]	[kg/a /vehicle]	[Nm^3/a /vehicle]	[MWh/a /vehicle]	Range [km/a]
Car FC	0.78	2.7	87	821	2.9	11,200
Car ICE	1.4	4.9	157	1,474	5.2	11,200
LDV FC	2.4	8.4	269	2,527	8.9	25,000
LDV ICE	4.0	14.1	448	4,211	14.9	25,000
Bus FC	12.0	42.2	1,344	12,633	44.7	55,000
Bus ICE	16.0	56.2	1,792	16,844	59.6	55,000

The demand for hydrogen is exogenously given to the model by yearly penetration rates for hydrogen vehicles, relative to the estimated absolute development of the vehicle fleet in Germany (Hart 2005) (see Figure 3). In 2010, the number of vehicles (in millions) is assumed as follows: passenger cars: 44.63, LDV: 2.06, busses: 0.07. Vehicles are distributed among urban and rural areas according to the number of registered cars per area.

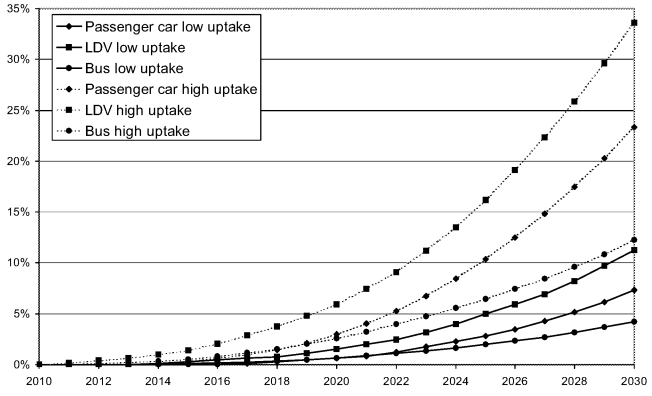


Figure 3: Penetration rates of hydrogen vehicles

A hydrogen infrastructure will be not be fully developed at the start phase of the market penetration of hydrogen vehicles. Based on the assumption, that ICE passenger cars and LDV will be equipped bivalent at the beginning, the share of ICE vehicles is assumed to be higher. For this phase, it is further assumed, that FC cars and LDV are mainly fix located fleet-vehicles (commercial use) and second cars. With the time, FC vehicles will prevail as they have competitive advantages over ICE vehicles due to a better fuel economy. For buses, the fuel cell technology is more reasonable.

In a first step, hydrogen production has been limited to steam reforming of natural gas (SMR), coal gasification and electrolysis, as up to now, these are the only proven production technologies at industrial scale. In addition, the supply of industrial by-product hydrogen (mainly from chlorine-alkaline electrolysis) is taken into account, amounting to around 500 million Nm3/a (DWV 2005). For hydrogen transport and distribution, both compressed gaseous (CGH_2) and liquid (LH_2) hydrogen trailers as well as hydrogen pipelines with different capacities are considered. Liquefaction plants can only be installed at central production plants, as they cannot be operated economically at small capacities.

Today, the majority of hydrogen is produced by steam reforming of natural gas at a rate of 4.7 kg gas/kg H2. A comparison between the consumption of a natural gas passenger car of 5.1 kg gas/100 km (EUCAR 2003) and the equivalent natural gas consumption of a hydrogen car – with the hydrogen produced from steam reforming – shows, that a FC car would consume 3.7 and an ICE car 6.6 kg gas equivalent/100 km. Regarding primary energy consumption, this means, that a hydrogen car is competitive to a natural gas car.

4. SCENARIOS AND FIRST RESULTS

Two scenarios have been defined regarding the uptake of hydrogen vehicles according to Figure 3, and two scenarios regarding the geographic penetration of vehicles, referred to as „urban“ and „urban-rural“ (one with an initial focus on urban areas and one with an immediate uptake in urban and rural areas). Table 2 shows the resulting hydrogen demand for the high and low uptake scenarios.

Table 2: Scenarios for hydrogen consumption

GWh H_2/a	**2010**	**2015**	**2020**	**2025**	**2030**
High uptake	8	1,584	9,609	29,050	59,735
Low uptake	0	188	2,262	8,449	19,920

In order to assess the impact of a sudden increase in hydrogen demand on the production mix, the optimisation has been carried out both on a *year-by-year basis* and by *steps of five years*. As there is not yet a clear strategy of the automotive industry regarding the storage of hydrogen on board, it has further been investigated how the assumption that half of the hydrogen demand will be in the form of gaseous hydrogen and half in the form of liquid hydrogen (Hart 2005) will effect the hydrogen supply infrastructure („urban LH_2“ scenario).

Figure 4 shows the hydrogen production mix for the high uptake case and the scenarios „urban“ and „urban-rural“ between 2015 and 2030. First of all, it can be seen that industrial by-product hydrogen plays a key role for hydrogen supply, particularly during the first years. During the initial phase, onsite steam reforming (SMR) at the filling stations further makes up a large part of hydrogen production, as transport cost of small amounts of hydrogen are more expensive than hydrogen production in small scale units. With increasing demand, however, larger scale production capacities with subsequent hydrogen transport turn out to be more economic. The use of onsite electrolysers can be explained through the surplus electricity generation capacity due to an assumed increase of the installed wind power in Germany, especially offshore. The „urban“ scenario with a concentration of hydrogen demand in urban areas tends more to larger scale production units than the „urban-rural“ scenario, which results in lower production costs (see Figure 5).

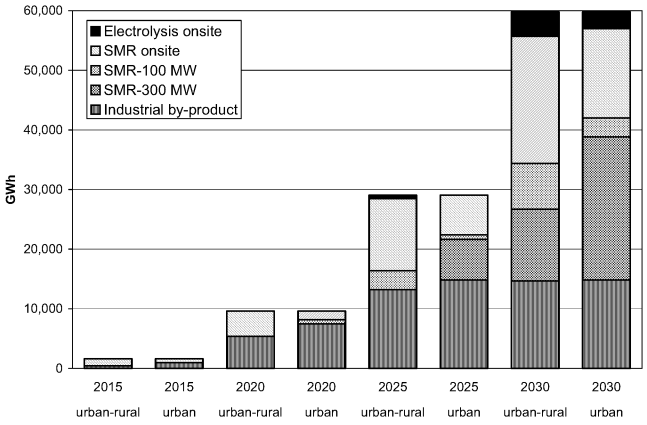


Figure 4: Hydrogen production mix

Figure 5 shows the total costs of hydrogen production in million €, including annualised investments, fixed and variable operating costs. Production costs in the „urban“ scenario are generally lower due to economies of scale resulting from larger production capacities. Figure 5 further illustrates, that the liquefaction of hydrogen increases total supply costs („urban LH_2“ scenario), although production costs in this scenario are lower than in the „urban“ scenario. This is due to the fact that more hydrogen is produced in large plants, which are required for an economic operation of liquefiers; under the scenario assumptions (especially the amount of filling stations) the storage of hydrogen in gaseous form on board would thus be preferable from the perspective of total infrastructure costs.

Finally, the specific supply costs in €/MWh hydrogen are displayed for the „urban“ scenario, which shows the least total costs. The specific costs are relatively high at the beginning of the vehicle uptake, due to the fact that an initial infrastructure (e.g. in the form of filling stations) will be required. They decline in the following years, as a large part of the demand can be met from industrial surplus hydrogen. With increasing hydrogen demand, however, the need for new production capacities, and thus the specific costs, rise again. Nevertheless, with crude oil prices between $50 and $60 per barrel, hydrogen production becomes an economic alternative.

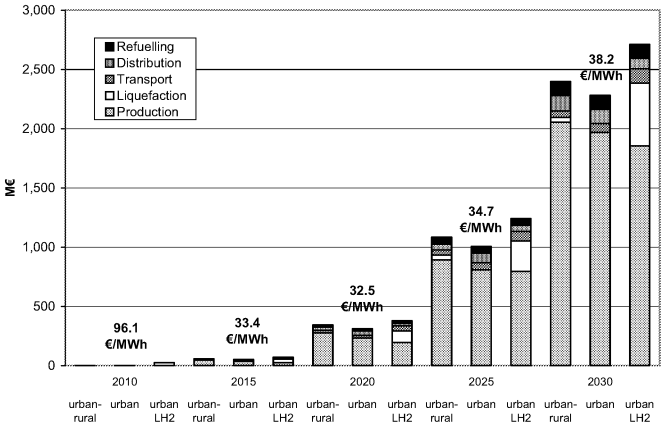


Figure 5: Total cost of hydrogen supply

The high and low uptake scenarios show no major differences regarding the implementation of the hydrogen infrastructure and associated costs, as the low uptake scenario can be considered a shift of the high uptake scenario of some years. The main difference between the scenarios run on the basis of five years respectively one year, is that the latter ones prefer onsite hydrogen production at filling stations during the first years, since the demand is only slightly increasing and transportation costs are generally high.

5. HYDROGEN PRODUCTION AND ELECTRICITY GENERATION

The large scale integration of renewable energies, especially wind energy, into the electricity system represents a technical and economic challenge. This is due to the necessary (local) extension of the electricity grid, the provision of back-up power to adjust to fluctuations in supply (as wind power and grid load are not correlated in time), and the requirements for frequency control and grid stabilisation. If renewable energies are eventually to make up significant shares of total energy supply, it is further crucial that the energy storage problem is solved. Besides compressed air or pumped storage, the production of hydrogen *at off-peak periods* (during night time or at weekends) and at periods with a *surplus of renewable energy* (such as at times with peaks of wind power) is being discussed as a possible option. Since an extension of wind energy (onshore and offshore) will increasingly displace base load power plants, spare capacities will grow, especially at off-peak times, with the consequence of falling electricity prices. The production of hydrogen off-peak would be an option for power plant operators to either produce and sell hydrogen on their own, or to sell their electricity during this time at higher margins to external hydrogen producers. As the need for back up capacities will increase with the installed wind power, these capacities could additionally be used.

During periods of low wind, supplementary or *positive control power* is required to meet the grid load, while *nega-*

tive control power is needed during periods of high wind, when the grid cannot absorb the excessive power; this excess electricity could be buffered and „stored“ by producing hydrogen, especially as it will technically not be feasible to meet the demand for negative control power with a growing share of installed wind power (dena 2005).

Figure 6 shows the load curve of electricity generation and the cost-optimal load distribution of electrolysers during a typical working day and a weekend day for a scenario in which hydrogen demand was solely supplied by electrolysis. It can be seen that electrolysis is mainly run at times of low marginal electricity generation costs, i.e. during off-peak times respectively at weekends. Due to the limited capacity of the electrolysers, however, it is not possible to fully use the available excess electricity generation capacity of the grid, as the electrolysers run already almost entirely at full load.

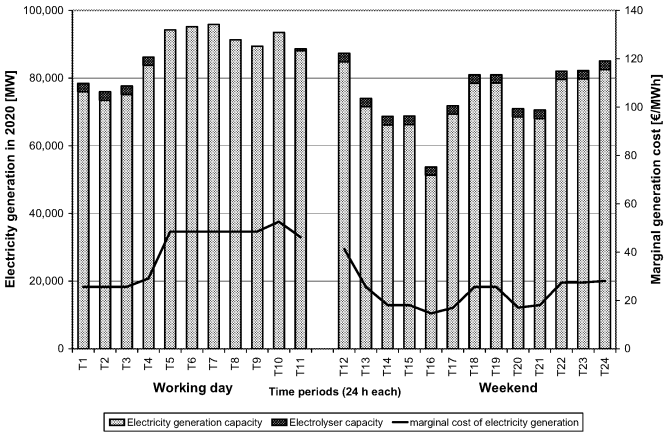


Figure 6: Hydrogen production from electrolysis

The above figure illustrates that buffering wind peaks with hydrogen production from electrolysis may be constrained by the capacities of the electrolysers and the economic requirement to reach a minimum of 2,000 full load hours.

6. CONCLUSIONS AND OUTLOOK

6.1. Conclusions

Some first results of the *MOREHyS* model regarding the implementation of hydrogen-based transport system have been presented. One of the most important aspects regarding a possible introduction of hydrogen vehicles is the fact, that industrial surplus hydrogen could supply as much as 610,000 FC respectively 340,000 ICE passenger cars per year, or provide fuel for all vehicles during the first five years of the high uptake scenario and the first ten years of the low uptake scenario (no need for a model run so far).

The hydrogen transport and distribution mode (LH_2/CGH_2 trailer and pipeline) turned out to be very sensitive to the techno-economic characteristics of the transport technologies and distances, and is strongly influenced by liquefaction costs. The model runs so far have demonstrated, that hydrogen can be an alternative to conventional automotive fuels with crude oil prices between $50 and $60 per barrel.

6.2. Further Research

In a next step, the model will be extended by including more hydrogen technologies, both for production (e.g. electrolysis from wind energy on offshore platforms, biomass gasification, thermo-chemical production in nuclear power plants, etc.) and transportation (e.g. metal hydride trailers or transport of hydrogen/natural gas mixtures in natural gas pipelines). Possible tradeoffs between hydrogen production and electricity generation will be investigated in more detail. To get a more realistic image of a hydrogen supply infrastructure, it might also be necessary to further disaggregate the areas in the model.

The GIS will further be applied for estimating average hydrogen pipeline costs for specific terrain characteristics, such as in urban or mountainous areas. The GIS could further be used for location planning of hydrogen production sites relative to demand centres, e.g. for coal gasification with carbon capture and storage. Thus, transportation costs would become dependent not only on the transport distance, but also on terrain types and site-specific characteristics of production technologies.

In the long run, the model could be extended by integrating conventional road transport as well as the supply chains of both conventional and other alternative fuels, in order to make the penetration rates of alternative fuels become an optimisation variable.

REFERENCES

Amos, W. 1998. „Costs of Storing and Transporting Hydrogen“. National Renewable Energy Laboratory Colorado/U.S. Department of Energy. NREL/TP-570-25106

Deutsche Energie-Agentur GmbH (dena). 2005. „Energiewirtschaftliche Planung für die Netzintegration von Windenergie in Deutschland an Land und Offshore bis zum Jahr 2020“. http://www.deutsche-energie-agentur.de/page/index.php?dena

EUCAR, CONCAWE, Joint Research Center of the European Commission (JRC, Ispra). 2004. “Well-to-wheels analysis of future automotive fuels and powertrains in the European context”. http://ies.jrc.cec.eu.int/Download/eh/31

German Hydrogen Association (Deutscher Wasserstoff Verband, DWV). 2005. “H_2-Roadmap”.

Hart, D. (E4tech). 2005. “The Economics of a European Hydrogen Automotive Infrastructure – A study for Linde AG”. http://www.hydrogenday.de/WGAP/hydrogen/file/rwar-69wdr8.de.1/e4tech%20hydrogen%20study.pdf

HyNet (The European Thematic Network on Hydrogen). http://www.hynet.info

HySociety. 2004. “Technology Database for Hydrogen Pathways, Analysis of Hydrogen Pathways”. Designed by Fraunhofer ISI for the EU HySociety Project. http://www.hysociety.net

International Energy Association (IEA). 2004. “World Energy Outlook 2004”. OECD/IEA, Paris

Tseng, P.; Lee, J.; and Friley, P. 2003. “Hydrogen Economy: Opportunities and challenges”. Office of Energy Efficiency and Renewable Energy, U.S. Department of Energy, Washington, Brookhaven National Laboratory, Upton, NY, USA

BIOGRAPHY

Michael Ball studied Industrial Engineering at the University of Karlsruhe, where he is PhD student at the French-German Institute for Environmental Research since 2002.

Feasibility Study of Static Field Converter

Anil K. Puppala
Mohamed Soliman
Mohammed Safiuddin
University at Buffalo, State University of New York

Andrew Abolafia
Andrew Abolafia Company, New York

Email: {apuppala|soliman3|safium}@eng.buffalo.edu
andrew@inventorone.com

ABSTRACT

The present work is to study, using a Finite Element Modeling (FEM) tool, the feasibility of construction of a device called as Static Field Converter, which is based on the principle of meissner effect of superconductors. It is designed to use a hemispherical superconductive material that rotates around either a permanent magnet or an electro-magnet. Magneto-Static analysis using an FEM tool was basically done for a situation when the superconductive material of the device is "Zero Field Cooled". Voltage picked up by pick-up coil that is wound around a ferromagnetic material (which completes the magnetic path) of the device is calculated according to Faraday's laws.

KEY WORDS

Superconductor (SC) material, meissner effect, Field Cooling (FC), Zero Field Cooling (ZFC), critical field.

INTRODUCTION

Several laboratories around the world have been carrying out the Research and Development (R&D) activities to successfully use superconductors in power generation, power transmission and energy storage. Takeoka et al [1] and Ishikawa et al [2] built a motor that used the principle of meissner effect. They used a heater to make the SC material lose its superconductivity for easing the rotation of SC material into the magnetic field of a Permanent Magnet (PM) and used liquid nitrogen to take it back into the superconductive state so that it may repel the flux lines of the PM when it is in the meissner state.

Camacho[3] opines that the development of applications based on the interaction between a PM and a high temperature superconductor has been limited due to lack of software tools that can simulate a superconductor under a wide range of conditions, in an adequate manner. This probably is because; a valid mathematical expression for phenomenon like flux pinning is still being investigated.

In this paper, a description of the Static Field Converter (SFC) is presented and the principle of operation upon which it has been designed to be working is illustrated. A model was developed to study the SFC dynamic behavior. After presenting the simulation results, the paper ends with the conclusion and suggestions on the future work.

STATIC FIELD CONVERTER

Static Field Converter, which is protected by a U.S. patent is a device that has a superconductive material in the shape of a hemispherical shell, maintained below its critical temperature (T_c), rotating around a PM. Figure. 1. To ease the rotation of the superconducting hemisphere it is attached to another hemisphere of relative permeability of one to form a complete sphere. A pick up coil surrounds the upper hemisphere. The two hemispheres are rotated around the PM by an external mechanical means. The assembly is supported by non-conducting rods on a non conducting base in a glass Dewar (refrigeration

tank), which is held below the critical temperature of the superconducting material. The pick up coil is periodically opened and closed by a switch during the operating cycle. As the hollow hemispheres rotate around the magnet, the flux from the permanent magnet is alternately interrupted by the superconductive material; thereby creating a rate of change in flux, hence inducing a voltage in the pick up coil according to Faraday's law.

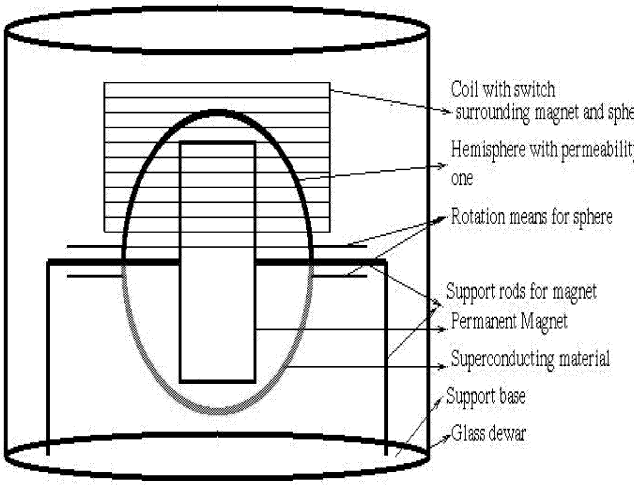


Figure 1. Static Field Converter

MODELING

Some assumptions were made before carrying out the modeling. As the support means, rods, collars and base are proposed to be made of a non-conducting material, it was safely assumed that, these do not interfere with the electric/magnetic circuit and hence are not included in the model. Also, the second half of the hemisphere, which is intended to ease the rotation of the superconductive hemisphere, is not modeled for the same reason. Its relative permeability is assumed to be one, which is same as that of air. Figure 2. shows the simplified form of SFC that was modeled. While modeling the superconductor, its hysteresis effects observed by Hull [4] and also its critical magnetic field were not considered., For a cylinder of radius $r_{m=}$ 5.316 cm, and height $h_{m=}$ 11.14 cm with spherical pole faces, the magnetic properties of commercially available Alnico 5 series permanent magnet were assigned. Coercive force (H_c) of the PM was 50,400 A/m. Around the magnet a hemisphere with inner radius of 8.5 cm and outer radius of 9 cm was modeled to represent the superconducting material. According to Camacho, the meissner state can be constructed in two different ways: imposing a magnetic vector potential constant or canceling the magnetic permeability (μ). The second method was used, with $\mu = 10^{-18}$ H/m assigned to the superconducting entity as assigning zero value was not permitted by the software tool that was used. As the second half of the sphere is not modeled, the superconducting hemisphere is effectively surrounded by the air. To the entity that represented air, $\mu = \mu_0$, permeability of free space was assigned. A 'C'- shaped ferromagnetic flux guide, with $\mu_r = 65000$ was used to complete the magnetic circuit. L_y and H_y of the flux guide are 19.316 cm and 27 cm respectively. The air gap L_g and hence the distance between the PM and the SC is kept to a minimum to minimize the reluctance to the flux lines. The finite element modeling software tool, ANSYS [6] that was used, does not allow the simulation of continuous rotation of the sphere when a permanent magnet is involved in the design. Instead, the SC hemisphere position is changed in steps of 10° each and the simulation is carried out each time to compute the flux density for one complete rotation of 360°.

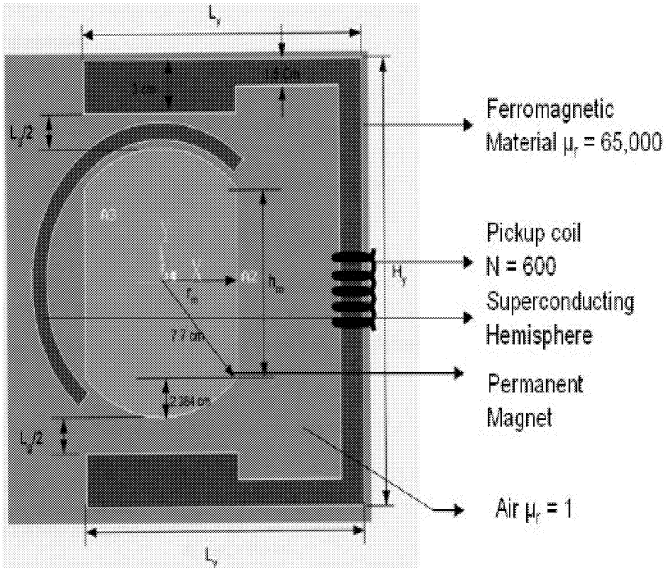


Figure 2. Simplified form of the Static Field Converter that was modeled

SIMULATION RESULTS

Figure 3. shows the pattern of the flux density when the superconductor is at the position shown in Figure 2. Flux densities at the center of the ferromagnetic material and also at a point above the magnet are noted.

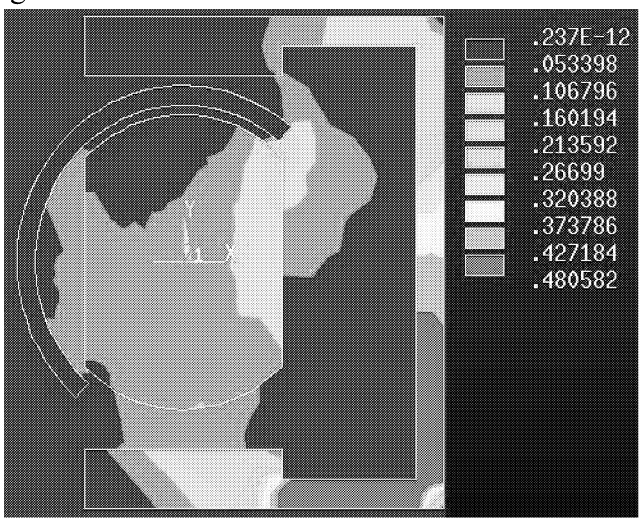


Figure 3. Pattern of flux density for the shown position of SC

Figure 4. gives the plot of the variation of the flux density for one complete rotation of the SC at the center of the ferromagnetic flux guide when superimposed over the straight lines that are obtained using the curve-fitting function of the MATLAB. This is also the shape of the variation of the flux density in a coil when it is wound at the center of the ferromagnetic flux guide. The voltage that would be induced in the pick up coil, based on Faraday's law, is

$$V = AN\frac{dB}{dt}$$

$$= AN\frac{dB}{d\theta}\frac{d\theta}{dt} \qquad (1)$$

A = area of cross section of the coil core
N = number of turns of the pick up coil
B = flux density produced by the PM

If the SC hemisphere was made to revolve at 60 rpm, then $\frac{dB}{d\theta} = 2\pi$ rev/sec. From Figure 4, $\frac{dB}{d\theta}$,(which is the slope of the waveform) was obtained. This again is performed with the help of MATLAB. The pick up coil open circuit voltage is obtained for N = 600 turns using Eq. (1).

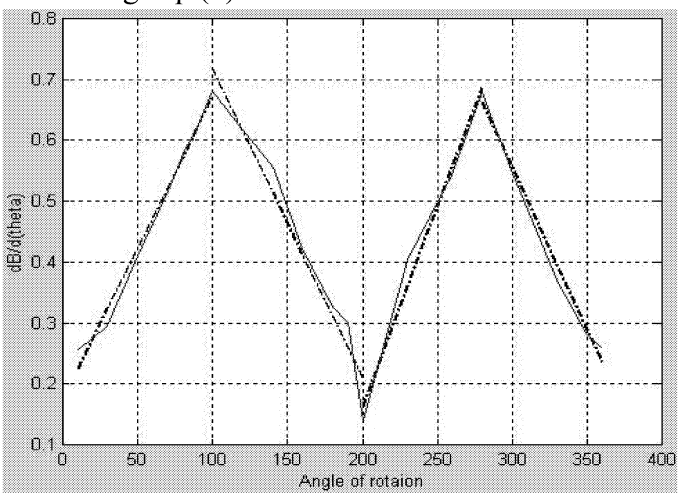


Figure 4: Plot of the flux density with respect to angle of rotation

Figure 5. shows the open circuit voltage plot that would be obtained at the terminals of the pick up coil.

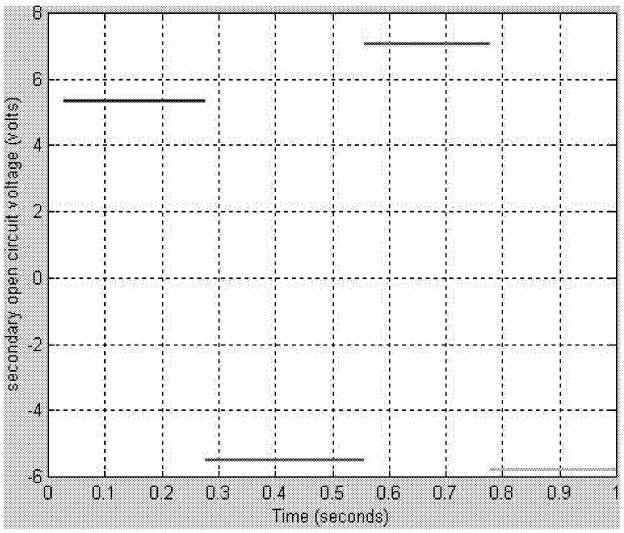


Figure 5. Voltage induced in the secondary coil

DISCUSSION AND FUTURE WORK

In the FC process, the temperature of the SC is brought down below its Tc in the presence of magnetic field. In the case of FC condition, it is assumed that all the flux lines that are inside the SC is perfectly pinned to the material defects in the SC [3]. In such a case, as suggested by Hull [4] and Lee [5] the SC material forms a 'frozen image' of the PM because of the trapped flux inside the SC. Hull observed that force

between the SC and the PM was attractive i.e. the SC tried to pull the PM when it was moved laterally. On the other hand, in the case of ZFC condition, Hull observed that SC behaved as a diamagnet, and the lateral force was repulsive, when it was given a lateral displacement. Hence the SC in the above device (or in any device for a rotational motion around a PM) should be ZFC and the levitation force should be due to meissner effect rather than due to flux pinning. Hence impurities (like silver) that enhance flux pinning should be avoided in the manufacturing process of superconducting materials.

When the switch is opened (to open the pick up circuit), the flux due to the permanent magnet, according to the meissner effect, would repel and hence push the SC hemisphere away, when it is above and below the PM. On the other hand, to ease the rotation of the SC hemisphere into the region above the PM, the switch should be closed so that the counter flux from the pick up coil would decrease the flux from the PM and hence the effective repulsive force due to the magnetic flux lines on the SC material would be least. On the lines that we discussed, when the pick up coil was excited with the voltage (that was calculated before) when the switch is supposed to be kept closed, it was observed that the flux density above and below the PM is lower. The rotation of the SC hemisphere would be easier if the switch is opened and closed at four times the frequency of rotation. When the SC is at the sides of the PM, the circuit of the pickup coil should be closed and when it is above and below the poles of the PM, it should be kept open.

When the number of sectors of the SC material is increased, voltage induced in the secondary coil would increase as $\frac{dB}{d\theta}$ increases. But according to Yang [7], increasing the number of sectors of the SC, would in effect decrease the meissner repulsion force.

The results show that, the construction of the SFC is feasible provided the SC is built by ZFC process.

REFERENCES

[1] Takeoka A., Ishikawa A., Suzuki M., Kishsi Y., Kuwano Y. August 28-31, 1988 Fundamental properties of a New Superconducting Motor" Proceedings of the 1st International symposium on Superconductivity (ISS '88), Nagoya.

[2] Ishikawa A. Takeoka A. Kishi Y Fundamental design for Meissner motor" 11th International Conference on Magnet Technology (MT-11). Elsevier Applied Science. 1990, pp.586-91 Vol.1. London, UK

[3] Camacho D., Mora J., Fontcuberta J. and Obradors X. Aug. 1997, "Calculation of levitation forces in permanent magnet-superconductor systems using finite element analysis", Journal of Applied Physics. **82(3)** pp. 1461-1468.

[4] Hull J.R., Cansiz A. Dec. 1999, A "Vertical and lateral forces between a permanent magnet and a high temperature superconductor", Journal of Applied Physics **86(11)** pp. 6396-6404.

[5] Lee H. Iwasa Y. "Levitation Force in YBCO/Nd-Fe-B Permanent Magnet System" IEEE transactions on Applied Superconductivity, Vol. 11, NO. 1, March 2001

[6] ANSYS Electromagnetic Analysis modeling guide. Available at www.ansys.com

[7] Yang W.M. "Effect of perimeters of induced shielding current loops on levitation force in melt grown single domain YBa2Cu3O7-x bulk." Applied Physics Letters" **79 (13)** pp- 2043-45

SIMULATION OF THE STEAM CONDENSER IN MATLAB ENVIRONMENT

Milan Cepák
Department of Control Engineering
Czech Technical University in Prague
Karlovo náměstí 13
121 35, Prague 2, Czech Republic
E-mail: cepakm@control.felk.cvut.cz

KEYWORDS
Energy, System identification, Model design, Continuous simulation.

ABSTRACT

The steam condensing turbines represent one of the basic units used for the generation of the electrical energy. Contrary to the backpressure turbines whose output pressure does not decrease very low, the condensing ones operate within the range containing the values very close to the vacuum. That is why they are applied as the last equipments in the cascade of the steam turbines in power plants. The indivisible part of the condensing turbine is constituted by the steam condenser where the flowing water cools the steam passing through the cascade and leaving the system. While the modeling of the condensing turbine does not differ too much from the backpressure one, the condensation process seems to be quite complicated event. This paper describes the development of the mathematical model of the condensing unit in MATLAB. The next purpose is the testing of the designed model in the SIMULINK tool. The comparison of the data provided by the model with the real data represents the last goal.

INTRODUCTION

Recently, the thermal power plants contribute to the total production of the electrical energy by the substantial part. Thereby, serious attention is paid to the research and to the development of the steam turbines. The new control algorithms cannot be directly applied to the equipments but they need to be relevantly tested because their possible failure might have undesirable effects. Then, each control design has to be simulated on the system model.
While the condensing turbine model arising from the backpressure turbine model (Cepák 2004) is quite direct and does not contain many crags, the condensing unit represents the complicated thermal system. It is mainly caused by its nonlinear behavior, which in addition emphasizes the fact that two latent components (steam, water) coexist here in the same moment and their ratio continually changes. The description of this process consists in the repetitive solution of the system of partial differential equations, which might be very hard task; certain simplifications need to be taken account during the model design.

NOMENCLATURE

Q_1 [W]	heat power balance at outer side of pipe
Q_2 [W]	heat power balance at inner side of pipe
Q_3 [W]	heat power balance at cooling water
S [m^2]	surface of condenser
L [m]	length of condensing pipe
δ [m]	thickness of condensation pipe
λ_c [W/m ^{o}C]	thermal conductivity of condenser material
α_v [W/m^{2o}C]	heat transfer coefficient
α_p [W/m^{2o}C]	heat transfer coefficient
T_{ps} [^{o}C]	temperature of condensing steam
T_{m1} [^{o}C]	temperature of outer side of pipe
T_{m2} [^{o}C]	temperature of inner side of pipe
T_{vin} [^{o}C]	temperature of influent cooling water
T_{vout} [^{o}C]	temperature of effluent cooling water
λ [W/m^{o}C]	condensing steam thermal conductivity
ρ [kg/m^3]	condensing steam density¨
g [m/s^2]	gravitational acceleration
l [J/kg]	condensing steam latent heat
μ [Pa.s]	condensing steam dynamic viscosity
h_s [J/kg]	condensing steam enthalpy
m_m [kg/s]	condensing steam flow
c_m [J/kg^{o}C]	condensing steam specific heat
m_w [kg/s]	cooling water flow
c_w [J/kg^{o}C]	cooling water specific heat
m_{w0} [kg/s]	the nominal cooling water flow
α_0 [W/m^{2o}C]	heat transfer coefficient for m_{w0}
m_{cs} [kg/s]	condensate flow
h_{cs} [J/kg]	enthalpy of condensate
p_c [MPa]	pressure in condenser
p_{c0} [MPa]	initial pressure in condenser
n [mol]	number of moles
R [J/kgmol]	universal constant of gas
V [m^3]	condensing steam volume

MATHEMATICAL DESCRIPTION

The condensation unit constitutes a certain sort of the thermal exchanger with the task to deprive the steam of its heat content. Consequently, the steam gets cold, changes the state and gets liquid.
The condenser contains the system of pipes installed in several rows. The hot steam coming from the turbine cools down on the pipes and is transformed on the water flowing to the reservoir (fig. 1).

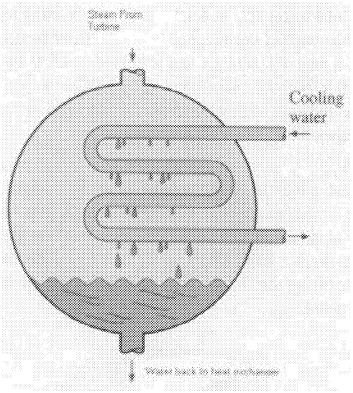


Figure 1: The Condensation Unit

The heat transfer description

The heat transfers taking place in the condenser are described by three thermal differential equations (Cídl 2004). The first equation determining the process of the transition of heat content from the condensing steam to the outer side of pipe has two parts defined as follows (Nožička 2001)

$$\frac{dQ_1}{dt} = \alpha_p S\left(T_{ps} - T_{m1}\right) - \frac{\lambda_c}{\delta} S\left(T_{m1} - T_{m2}\right), \qquad (1)$$

where the first part corresponds to the amount of heat power entering to the outer side of pipe and the second part to the amount of heat power leaving the outer side of pipe and moving to its inner side.
The process of thermal diffusion in the pipe from its outer to its inner side is defined by

$$\frac{dQ_2}{dt} = \frac{\lambda_c}{\delta} S\left(T_{m1} - T_{m2}\right) - \alpha_v S \frac{\Delta T_{vout} - \Delta T_{vin}}{\ln \frac{\Delta T_{vout}}{\Delta T_{vin}}}, \qquad (2)$$

where the first part represents the amount of heat power coming to the inner side of pipe and the second part the amount of heat power leaving it. Except the other quantities, transmitted power depends on the average logarithmic thermal declivity [-] (Malinský 1995)

$$\frac{\Delta T_{vout}}{\Delta T_{vin}} = \frac{T_{m2} - T_{vout}}{T_{m2} - T_{vin}} \qquad (3)$$

The last equation relative to the transition of heat power from the inner side of pipe to the cooling water is given as

$$\frac{dQ_3}{dt} = \alpha_v S \frac{\Delta T_{vout} - \Delta T_{vin}}{\ln \frac{\Delta T_{vout}}{\Delta T_{vin}}} - m_w c_w \left(T_{vout} - T_{vin}\right), \qquad (4)$$

where the first part represents the amount of the heat power coming to the cooling water and the second part the amount of heat power leaving the condenser in warmed water.

The equations (1), (2) and (4) can be further modified and written for the appropriate temperatures T_{m1}, T_{m2}, T_{vout} as

$$\frac{dT_{m1}}{dt} = \frac{1}{\frac{c_m}{2} m_m} \left[\alpha_p S\left(T_{ps} - T_{m1}\right) - \frac{\lambda_c}{\delta} S\left(T_{m1} - T_{m2}\right) \right], \qquad (5)$$

$$\frac{dT_{m2}}{dt} = \frac{1}{\frac{c_m}{2} m_m} \left[\frac{\lambda_c}{\delta} S\left(T_{m1} - T_{m2}\right) - \alpha_v S \frac{\Delta T_{vout} - \Delta T_{vin}}{\ln \frac{\Delta T_{vout}}{\Delta T_{vin}}} \right], \qquad (6)$$

$$\frac{dT_{vout}}{dt} = \frac{1}{c_w m_w} \left[\alpha_v S \frac{\Delta T_{vout} - \Delta T_{vin}}{\ln \frac{\Delta T_{vout}}{\Delta T_{vin}}} - m_w c_w \left(T_{vout} - T_{vin}\right) \right], \qquad (7)$$

The scheme of the condensation is shown on fig. 2.

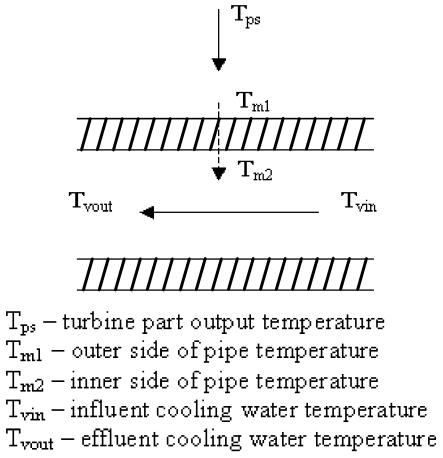


Figure 2: The Condensation Process

The heat transfer coefficients

The heat transfer coefficient α_v for the transition pipe-cooling water depends mainly on the water flow. In literature, e.g. in (Jícha 2001), the formula in form

$$\alpha_v = \alpha_0 \left(\frac{m_w}{m_{w0}} \right)^n, \qquad (8)$$

can be found. The value of the exponent n [-] is derived from the kind of the flux (e.g. for turbulent flux $n = 0.8$).
The determination of the heat transfer coefficient α_p for the transition condensing steam-pipe requires the complex evaluation. The relation

$$\alpha_p = 1.15 \left(\frac{\lambda^3 \rho^2 g l}{\mu L \left(T_{ps} - T_{m1}\right)} \right)^{0.25}, \qquad (9)$$

located e.g. in (Kalčík 1973) can be used.

The pressure in the condenser

The amount of the condensate produced in the condensation unit gives the formula in form

$$m_{cs} = \alpha_p S \frac{T_{ps} - T_{m1}}{h_s - h_{cs}} \qquad (10)$$

The pressure in the condensation unit is derived from the state equation; it is defined by the integration of the difference between condensing steam flow and condensate flow

$$p_c = p_{c0} + \frac{nRT_{ps}}{V} \int (m_m - m_{cs})\, dt \qquad (11)$$

CONDENSER IN THE SIMULINK TOOL

The model in the SIMULINK tool consists of three integrators serving for the computations of the temperatures T_{m1}, T_{m2}, T_{vout} and one integrator for the evaluation of condenser pressure p_c having the appropriate constant ($n.R.T_{ps}/V$) as an integration parameter; the entire condensation unit corresponds to the four-order system. The simulation scheme is composed mainly from the classic blocks executing algebraic operations. The calculation of the heat transfer coefficient α_p is performed through the special bloc where the *Steam Property Package* called as MATLAB function permits to determine the physical properties (conductivity λ, viscosity μ, etc. see (9)) of the condensing steam. The *Steam Property Package* substituting the steam tables contains also the information about the other thermal quantities (enthalpies h_s, h_{cs}, etc.) necessary for the determination of the amount of condensate.

The model of the condenser can be simply implemented as a subsystem in the large-scale system scheme; it possesses five inputs - steam flow, temperature and enthalpy at the turbine output, cooling water flow and temperature - and two outputs - effluent cooling water temperature and pressure in the condenser. This last quantity represents the lower value of the pressure difference arising at the turbine part; hereby, the condenser influences the amount of the steam flow in the system and becomes the indivisible part of the condensing turbine.

The condenser design in SIMULINK was tested on the real data offered by the turbine operator. The turbine steam flow entering to the condensation unit was changed in the whole interval of its feasible values. This approach enabled the gathering of the information sufficient for the evaluation of the model properties and qualities.

However, no all data necessary for the simulation was available, i.e. cooling water flow m_m, the surface of the condenser S, the length of the pipe L and the another ones. These unknown quantities had to be estimated; that is the reason way the real and the modeled data shown on the figures 3 – 5 have not the best coherence.

Nevertheless, in spite of the unavailable measurements, the time developments of the real and modeled quantities differ only in units of percent. Comparing their trends, it can be concluded that the modeled data follows the real one very exactly. Moreover, the certain disproportions between them should be improved by the treatment of the missing quantities.

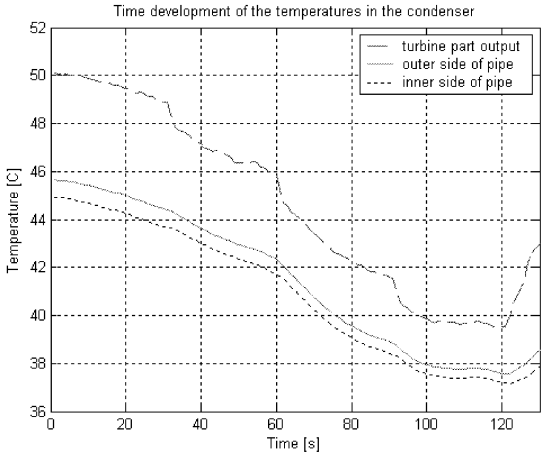


Figure 3: Temperatures in the Condenser

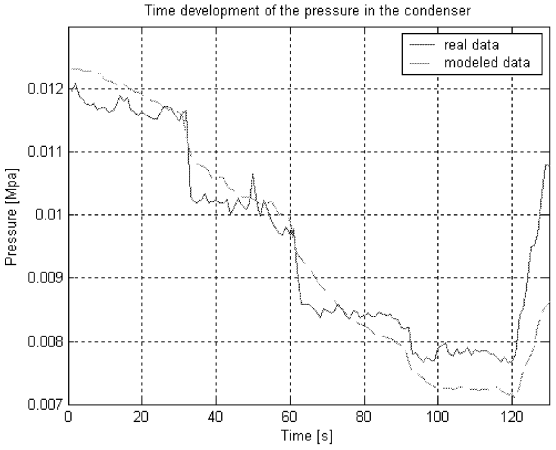


Figure 4: Pressure in the Condenser

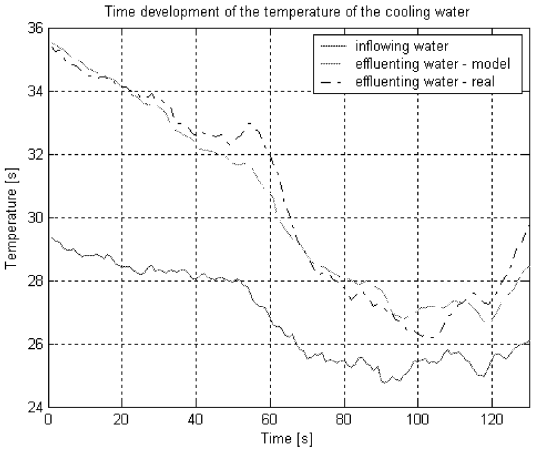


Figure 5: Temperatures of the Cooling Water in the Condenser

The simulation of the condensing turbine

The second phase of the testing of the condenser consists in the link between the steam turbine and the condensation unit. The design of the one-part turbine has been described in (Cepák 2004) and the implementation of the library including it and the other thermal components in (Cepák and Havlena 2004).

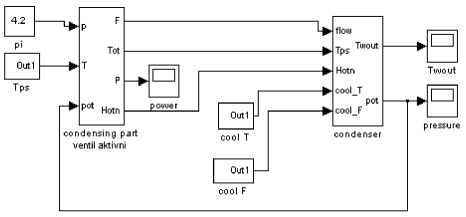


Figure 6: Condensing Turbine SIMULINK scheme

The output quantities of the turbine part (the steam flow, the output temperature and the output enthalpy) represent the initial conditions for the condenser execution. Correct model working demands also the temperature and the flow of the cooling water (for scheme see fig. 6).

The condensing turbine was tested in SIMULINK for the different valve position (0 – closed, 1 – fully open) and at first for the constant values of the temperature and flow of the cooling water (fig. 7 – 9). Further, the simulation is expanded for the situation when the flow and the temperature of the water rapidly move - it can represent the fast changes of the weather or the problem with the station controlling the properties of the cooling (fig. 10 – 12). The valve position of turbine part remains the same as in the case of no problem situation.

The system model was simulated by the solver ode45 (Dormand-Prince) with relative tolerance 1e-3 and 1e-6. The computation of 1200 samples has taken 90 seconds for both cases, but the solution with accuracy 1e-6 was more stable. Then, solvers ode32 (Bogacki-Shampine) and ode113 (Adams) were tested; these methods have reduced computational time about 20%. All simulations have been performed on Pentium 4, 2.5 GHz, 512 MB RAM.

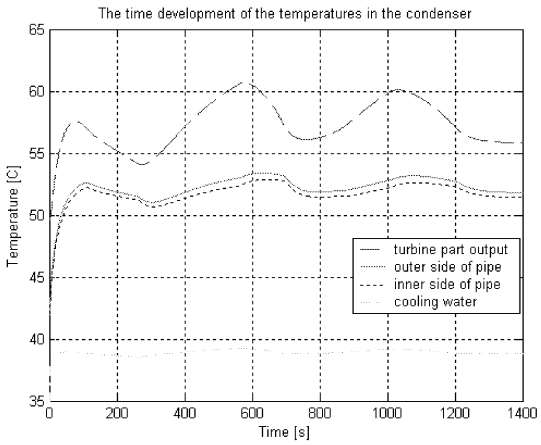


Figure 7: Temperatures in the Condenser

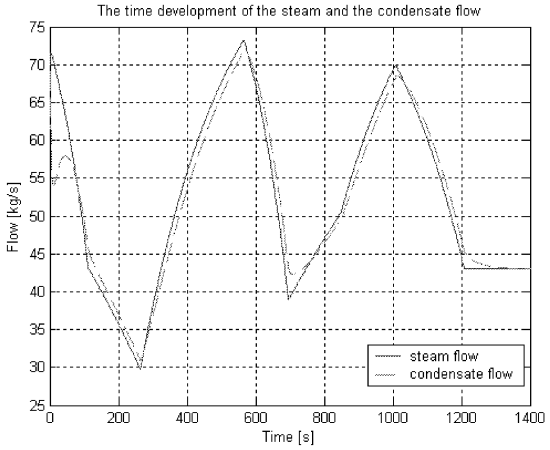


Figure 8: Steam and Condensate Flows

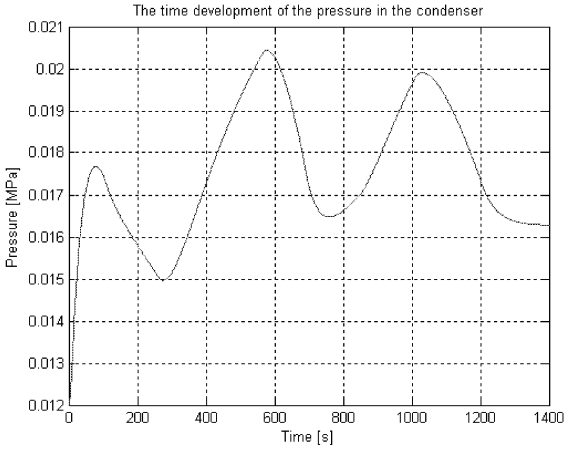


Figure 9: Pressure in the Condenser

In case of slower computer, large approximations of the computations of steam properties instead the direct use of *SPP* package are needed; then, the results become less exact but the condenser can be used in the applications with strong requirement on the model response.

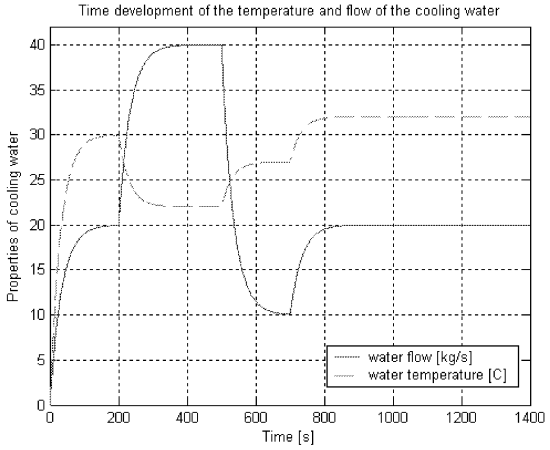


Figure 10: Change of the Properties of the Cooling Water During the Problem Situation

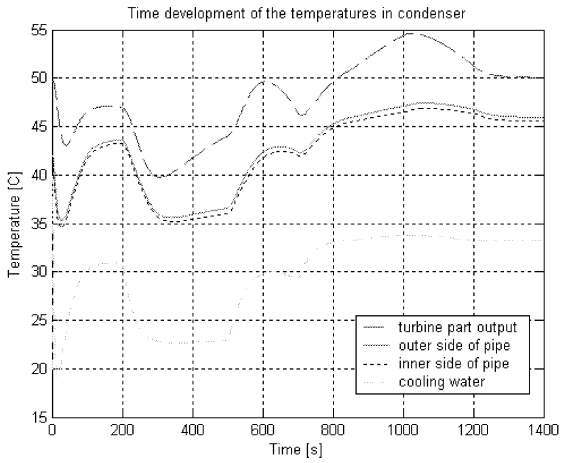


Figure 11: Temperatures in the Condenser During the Problem Situation

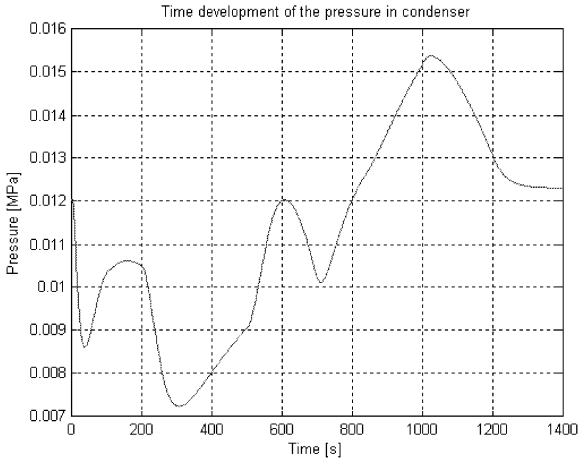


Figure 12: Pressure in the Condenser During the Problem Situation

The implementation of the energetic library

The modeled condensation unit has been built in the special energetic library already containing the steam turbine part (both backpressure and condensing one), the steam header and the injection cooler (fig. 13).

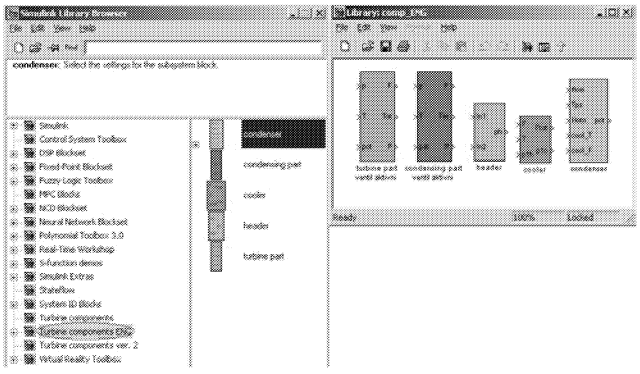


Figure 13: Energetic Library in SIMULINK tool

All these blocks possess the appropriate block parameters. There is certain number of parameters set by the user here. Mainly, the input values necessary for the correct model working are located here; but the other kinds of the form fields (combo box, check box) can be implemented. This approach restricts the direct incidence to the component interior and enables to execute only allowed changes.

CONCLUSION

In this paper, a mathematical model for a steam condenser has been designed. After, some model simulations have been done in SIMULINK and the block representing the system has been implemented in this tool. The designed model follows up the real plant trends but there have been some differences between simulated and true data caused by the unavailability of some necessary quantities. However, the reached accuracy appears as sufficient one and it can be further improved. Therefore, the design provides together with the condensing turbine model strong tool for the simulation of the power plant networks.

REFERENCES

Cepák, M. 2004. „Simulation of the steam turbine in Matlab environment“. In *Industrial Simulation Conference 2004* (Malaga, Spain, June 7-9). Eurosis-ETI, Ghent, p.251-255.

Cepák, M. and V.Havlena. 2004. „The library of the energetic components for the simulation of the steam turbines works.“ In *Matlab 2004* (Prague, Czech rep. Nov. 11). VŠCHT, part 1-2, p. 60-63.

Cídl, J. 2004. „Mathematical model of the steam turbine with suppressed vacuum“. In *Control of Power systems '04* (on CD), Slovak University of Technology in Bratislava, Bratislava, Slovakia.

Jícha, M. 2001. „Transfer of the heat and the material”. CERM Brno.

Kalčík, J. and K. Sýkora 1973. „Technical thermodynamics“. Academia, Prague.

Malinský, K. 1995. „The physics I.“. Faculty of electrical engineering, CTU Prague.

Nožička, J. 2001. „The base of the thermodynamics“. Faculty of Mechanical Engineering, CTU Prague.

ACKNOWLEDGMENTS

This work is supported by the project 1H-PK/22 of Ministry of Industry and Trade of Czech Republic and by grant 102/05/2075 of Grant Agency of Czech Republic. Next, I would like to thank Vladimír Havlena for his help.

AUTOR BIOGRAPHY

MILAN CEPÁK was born in Strakonice, Czech Republic and went to the Czech Technical University in Prague, to the Faculty of Electrical Engineering where he studied the system control and obtained his Master degree in 2003. Now, he is a PhD. student at the Department of Control Engineering at CTU in Prague.

OBJECT-ORIENTED DATABASE MANAGEMENT SYSTEM FOR INDUSTRIAL COMBUSTION PROCESSES

Nuray Kayakol
Instituto Superior Tecnico
Av. Rovisco Pais,
1049-001, Lisbon, Portugal
E-mail: nkaykol@navier.ist.utl.pt

Marcos A. Nogueira
Irradiare,
Oeiras,
Lisbon, Portugal
marcos.nogueira@irradiare.com

KEYWORDS

Object oriented database, data mining, case-based reasoning

ABSTRACT

Object-Oriented Database Management System for Scenario (S_DBMS) processing is developed as part of an Expert System specially designed for industrial combustion systems. The S_DBMS is a hybrid methodology of case-based reasoning, object oriented and data mining. As it can use and/or extract data from both Computational Fluid Dynamics (CFD) simulation results and plant measurements, which are complementing each other, the proposed structure may over come some drawbacks of Expert Systems (ES) such as having a solution domain not covering whole range of operating conditions. The S-DBMS maintains fault rules in a sub-database that can be accessed during a diagnostic search. Diagnostic search is based on identification of faults according to the classification of values of process variables which are either in normal, warning or abnormal data range, and finding the corresponding faults from the pre-defined set of faults (a list of variables having abnormal values) with a matching coefficient

NOMENCLATURE

Abbreviations

CBR	Case-Based Reasoning
CFD	Computational Fluid Dynamics
ES	Expert Systems
OQL	Object Query Language
OO	Object-Oriented
StDev	Standard Deviation
StErr	Standard error
S_DBMS	Object-Oriented Database Management System for Scenario processing

INTRODUCTION

Recent developments in mathematical and artificial intelligence modeling, sensors and computer technology lead to the development of new control strategies for supervision of industrial processes. This paper focuses on the a hybrid methodology of Case-Based Reasoning (CBR), Object-Oriented (OO) methodology and data mining for the use of CFD results and/or historical process data for off-line decision support during supervision of industrial combustion systems.

Due to insufficient and incorrect measurements in industrial processes it is desirable to use CFD models for control purposes. However, the on-line implementation of the transient CFD models on the control systems is limited due to excessive computation time but the ES provides the use of CFD codes at the supervisory level. CFD codes are used to generate a database covering the range of process operations, simplified models or functions for the model library of the ES (Carvalho et al., 2003). For example, The ES allows the use of CFD models for on-line control by employing predefined simulation results which work as reference optimal operating conditions (ACORDE, 1999).

CBR systems have been used in a broad range of domains to capture and organize past experience and to solve new problems using a database of cases which is a list of previous past solutions. A review of some applications of CBR can be found elsewhere (Liao, 2005).

OO programming is one of the best ways to design databases, graphic user interfaces, and a wide variety of other applications (Barry, 1996). OO methodology are becoming popular within last decade and are being used in ES (Liao, 2005) and have been taking part in practical intelligent supervisory control systems for the process industries (Venkat et al., 2003). Data Mining is the process of finding trends and patterns in a large amount of data. It is very useful tool for analyzing CFD simulation results for control purposes and fault detections.

In this study, a hybrid methodology which combines the benefits of methodologies of CBR and OO, and data mining is developed. The hybrid methodology is the main part of an ES, especially designed for industrial combustion systems (Carvalho et al., 2003). It is called as object-oriented database management system for scenario processing, S_DBMS. It uses CFD simulation results and/or historical process data as different scenarios representing operating conditions of the process under consideration. It allows searching single or multivariable data and/or the trend of

variables among stored scenarios, and reports found value(s) with a matching coefficient.
The S-DBMS can also be used for fault detection. It maintains fault rules in a sub-database which can be accessed during a diagnostic search. Diagnostic search is based on identification of faults according to the classification of values of process variables which are either in normal, warning or abnormal data range, and finding the corresponding faults from the pre-defined set of faults (a list of variables having abnormal values) with a matching coefficient.

OBJECT-ORIENTED DATABASE MANAGEMENT SYSTEM

Object-oriented database management system for scenario processing is developed from scratch by using object-oriented programming language C++. The S-DBMS is independent from any commercial CFD or database software and also is process independent. It is designed to be applicable to any industrial combustion system. The S_DBMS provides a user-driven environment incorporating all available data from model-based CFD simulation results and/or historical data from process sensors, which can be accessed by query commands in explaining cause and effect relationship of the process variables and in helping fault detection by a diagnostic search. Different operating conditions of the process under consideration are stored in a scenario database. It is not designed as a real time database.

The single or multivariable search allows the operator to get values of the input variables from output variables, called reverse search, and/or values of the output variables from values of input variables, called direct search, as shown in Fig. 1.

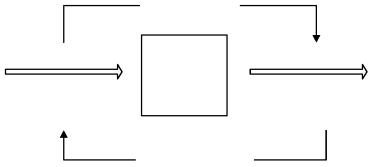


Figure 1 : Search directions for process I/O.

Process variables are stored by classifying them as dependent, controlled or operating. Inputs are operating variables and outputs are controlled and dependent ones. Single variable search is arranged in such a way that if any dependent variable is searched by making a query, the desired value is retrieved with the corresponding scenario stored in the scenario database. This provides a reverse search to find the values of input variables for the selected output variable. The values of controlled variables are automatically retrieved. Dependent variables may be called as query variables in order to distinguish them from other automatically retrieved variables. In multivariable search the user is free to choose any combination of the variables. Both direct and indirect search is possible by choosing corresponding variables.

The data mining of the S_DBMS is concerned with finding trends and patterns in a large amount of data. Records for CFD results in the scenario database hold the data as 3D arrays. A region of any data is classified as either a single cell, discrete cells (collection of single cells) or successive cells (a region in a domain). Figure 2 shows classification of data locations as regions on a 3D geometry having integer coordinates. Any 1D profile within a 3D domain can be searched using any combination of data locations. For example, a typical 1D temperature profile in a combustion chamber, starting from inlet port, going through a maximum and decreasing continuously through the exit port, can be searched by using the combination of discrete cells and a region representing the exit port. The uniformity, minimum, maximum and the average values of any variable at any sub-region defined by successive cells are reported by using a corresponding query command.

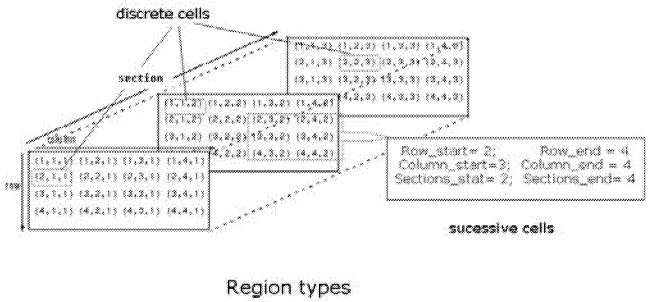


Figure 2 : Selection of Data Regions for Data Extraction from a 3D Geometry Having (4x4x3) Grid Size.

A scenario represents a case defined by a set of operating conditions. Therefore, a scenario database should store as many sets of operating conditions as possible in order to capture the whole operating range. All the values of the input and output variables are stored in records of each scenario. These values can be either CFD simulation results or measurements from process sensors.

The operating conditions of all the scenarios are compared with those in a pre-defined base case scenario. Each scenario is labelled and its case name is added to a case list. All scenarios are visited to retrieve the values of query variables according to the searching criteria. If a specific scenario is chosen from the case list, only a single scenario is visited for searching purposes. Each scenario is defined as an object which independent of the values, unlike entities in a relational database.

The S_DBMS has its own Object Query Language, OQL. It is an easy-to-use SQL-like query language with special features dealing with objects and values. Table 1 gives the syntax of the most commonly used keywords of OQL for the S_DBMS.

Queries need to be prepared by using special keywords of (a) Data Definition Language (DDL) for storing and updating data and (b) Data Manipulation Language (DML) for retrieving data. The dot notation is used to access

components of complex values. The S_DBMS can either process or prepare scenario records by using DDL. It is process independent. No variables or their corresponding items are defined in the S-DBMS. It knows which variables are to be handled from the user supplied DDL keywords. User has full access to change the number of variables, their names or units.

Table 1: The Syntax of Some Keywords of OQL

DDL, Record Preperation
APPEND <record_category> FROM <scenario_source> WHERE<record_size><scenario_no> scenario_case_type> FOR <record_variabletype> <record_variable_numbers> <variable no> <variable name> <variable unit> ;
DDL, Record Processing
CREATE database <database_name> INSERT scenarios <scenario_number> ADD <variable type> <number of variables> <variable no> <variable name> <variable unit> ;
DML, Search
SELECT <Region type> FROM <variable type> WHERE <condition> DEFINE <cell_name> <location>;

The Structure of the S_DBMS

The structure of the scenario database is given in Fig. 3. It consists of (i) a scenario database for storing scenarios, (ii) a statistical database for holding statistical parameters for each scenario and two auxiliary functions, (iii) a data extractor for extracting data from the scenario database and (iii) a statistical analyzer for the application of statistical techniques on extracted data.

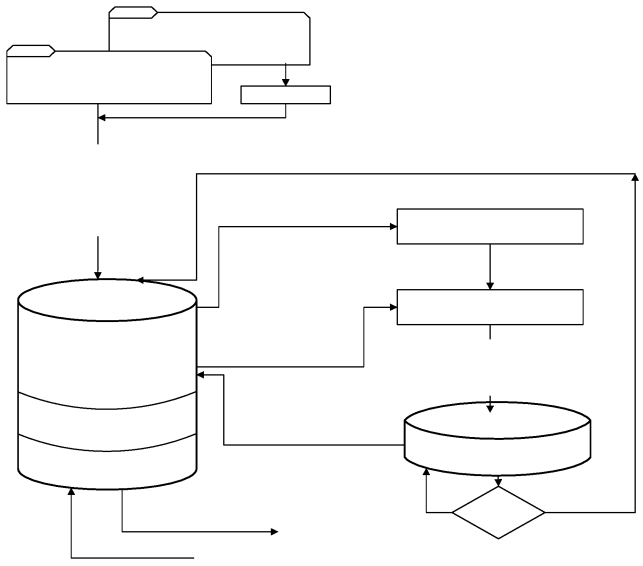


Figure 3: The Structure of the S_DBMS

The CFD results and/or measurements are defined by using the keywords of query DDL. Desired values of variables under consideration, which are treated as query variables, and their data region are specified by using the keywords of query DML. The Data extractor, which is an array processor, extracts required data from the scenario database records and sends it to the statistical analyzer for comparison of the extracted values with the desired values. Some statistical values, standard deviation (StDev), standard error (StErr), mean, minimum, maximum and a matching coefficient, related to that comparison are stored in the statistical database. The matching coefficient is defined in terms of StDev and StErr as shown below

$$\text{Matching_coeff} = \frac{(\text{StDev} - \text{StErr})}{\text{StDev}} \qquad (1)$$

The matching coefficient changes between zero and one. After the matching coefficients of all the scenarios are determined, the scenario having the matching coefficient closest to 1.0 is reported as the fittest scenario.

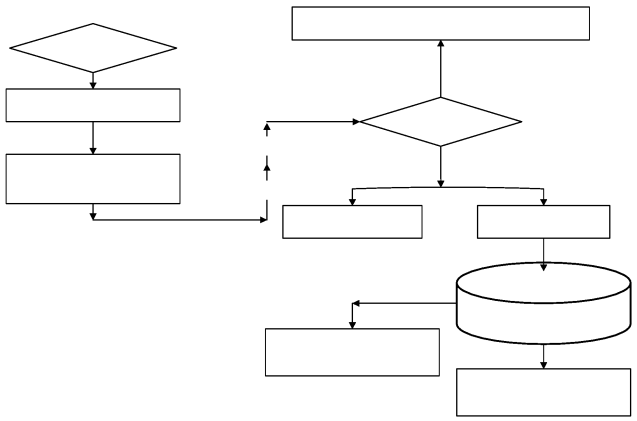


Figure 4 : The Structure of Diagnostics Search Strategy

Diagnostics Search Strategy of S_DBMS

As shown in Figure 4, the diagnostic search strategy of the S_DBMS is based on the identification of faults according to the classification of values of process variables as normal, warning and abnormal. The user can select a scenario representing current operating conditions from the case list of the S_DBMS and make a query by giving the values of process variables that might cause a fault or might trigger multiple faults. The S_DBMS finds a matching coefficient after searching single or multivariable query. If the set of variables is classified as normal or warning the fault sub-database is not called. If the values of variables are abnormally low or high, a possible fault is searched in the fault database. The fault database stores the pre-defined faults with their set of hypothesis that explains the abnormality. The S_DBMS isolates the fault with a matching coefficient. If all fault hypotheses are satisfied the detected fault has a matching coefficient of 1.0. It is less than 1.0 for partially matched hypotheses. Any abnormality having a matching coefficient of very close to zero (typically equal to or less than 0.01) is labeled as undetected novel fault.

APPLICATION OF THE S_DBMS FOR INDUSTRIAL COMBUSTION SYSTEMS

The S_DBMS is first validated on a simple test case, which represents a 3D box shaped combustion system with a coarse grid (2x2x2). It runs on a PC having Intel Pentium CPU 2.8 GHz and 512 Mb RAM. Table 2 shows the query DDL of the scenario database, temp_vs_gasComp, which has three scenarios with three variables, temperature, CO and O_2 gas compositions. The scenario database has nine records. Temperature profile of the process is given in an array structure in one of records of the scenario database. Table 3 gives an answer to a single variable query for finding the scenario where the exit port temperature is equal to 1590 ^oC. It takes less than 1 second of CPU time to get the answer. It finds the scenario, "increase_in_air_temperature_+_100", with a matching coefficient of 0.998. The exit port has a uniformity of 36%.

Table 2. A Simple Scenario Database for CFD Applications

Query DDL
CREATE database temp_vs_gasComp INSERT scenarios 3 ADD dependent_variable(s) 3 1 temperature K 2 co_composition % 3 o2_composition % ADD operating_variable(s) 4 1 fuel_flow_rate kg/s 2 air_flow_rate kg/s 3 fuel_type - 4 air_temperature K ;
A record
Process_name Incinerator Case_Name Increase_in_Air_temperature_+_200 Number_of_operating_variables 4 fuel_flow_rate 16 air_flow_rate 14 fuel_type 1 air_temperature 500 Parametric_change air_temperature Grid_numbers 2 2 2 Temperature 1655. 1595. 1620. 1723. 1423. 1523. 1850. 1855. 890.

Table 3: A Single Variable Query for CFD Applications and Output of the S_DBMS

Query: SELECT Region_successive_cells FROM dependent_variables WHERE dependent_variables.Temperature = 1590. DEFINE Region_name Exit_port Region_boundaries 1 2 1 2 1 1
Output of the S_DBMS: Domain data mining report Variable name: temperature Case Name : increase_in_air_temperature_+_100 Minimum value: 1470 Maximum value: 1805 Mean value : 1585.75 Region: Successive cells : Region Name : Exit_port Minimum value: 1543 Maximum value: 1628 Uniformity(%stdv): 36.15 Mean value : 1587.75 Reference value: 1590 Matching coefficient : 0.998

The S_DBMS is then successfully applied to semi-industrial test furnace of IFRF (Boerstoel et al., 1998). The furnace resembles one compartment of a regenerative glass melting furnace containing one burner. The furnace is modeled using 24x25x37 mesh size for 27 scenarios for different values of burner angle, air temperature and air/fuel ratio. Figure 5 shows the temperature profile for under port burner firing with 20° burner angle, air temperature of 1100 ^oC with 10 % excess air. Table 4 gives an answer to a single variable query for finding a temperature profile at 4 different locations. It takes 1 second of CPU time to get the answer. The S_DBMS is independent from any CFD software. It does not read the grid structure of CFD results. For any kind of query, the user should be able to get the integer coordinates of the locations of interest from post-processor of her/his CFD code.

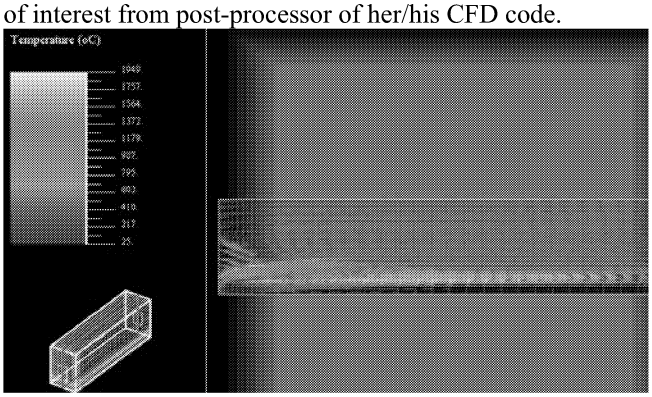


Figure 5: A Temperature Profile of IFRF Furnace.

Table 4: A Query for a Single Variable Profile for CFD Applications and Output of the S_DBMS

Query: SELECT Region.discrete_cells Number_of_discrete_cells 4 FROM variable.dependent WHERE variables.dependent.Temperature = 25. DEFINE cell.name Location_A Cell.location 1 8 1 WHERE variables.dependent.Temperature = 300. DEFINE cell.name Location_B Cell.location 2 8 1 WHERE variables.dependent.Temperature = 1900. DEFINE cell.name Location_C Cell.location 3 9 1 WHERE variables.dependent.Temperature = 1100. DEFINE cell.name Location_D Cell.location 5 13 1;
Output of the S_DBMS: Operating Variables: air_flow_rate 0.188 kg/s burner_angle 20 degree air_temperature 1100 K Domain data mining report Minimum value: 24.85 Maximum value: 1936.93 Mean value : 1363.7 Region: Discrete cells: Region Name : temperature profile Extracted discrete cell(s) Location_A 24.8 Location_B 323.8 Location_C 1801.0 Location_D 1103.0 Reference discrete cell(s) Location_A 25 Location_B 300 Location_C 1900 Location_D 1100 Matching coefficient : 0.872

The S_DBMS is used as the main part of the ECOTHERM supervision module for an incinerator process (ECOTHERM, 2005). The scenario database, which stores steady state data, is filled in using average values of plant data from sensors. The S-DBMS has 66 different scenarios for 30 variables of the process. Table 5 gives an answer to a multivariable query for finding variable V_%H20 = 14.0. It takes 2 seconds of CPU time to get the answer. It is an example for reverse solution. Values of operating variables are reported from a given output value(V_%H20 = 14.) with a matching coefficient of 0.986. Values of controlled variables are also reported to show the relationship between the desired quantities.

TI_1, AI_1 and FI_1 are called as reference variables. It is a simple test for checking fault, named waste_chute_blocked, which is not satisfied due to normal condition of TI_1. The scenario database stores only successful operations happened in the past. It will not retrieve any scenario having any faults. It can only report conditions of variables according to pre-defined normal, warning or abnormal data range. The condition of reference variables gives hint before using fault database for searching a fault.

CONCLUSIONS

In this study, an object-oriented database management system for scenario processing, namely, S_DBMS is developed for supervision, in off-line consulting mode, of industrial combustion processes. It allows searching single or multivariable data and/or the trend of variables among stored scenarios, and reports found value(s) with a matching coefficient. The S_DBMS is applied to a semi-industrial experimental furnace and a real incinerator process. Its CPU time is less than a few seconds on an ordinary PC and it is fast enough to be used for on-line decision support at supervisory level. It is also shown that the S-DBMS can be used for fault detection when the scenario base is connected to a fault database.

Table 5: A Single Variable Query for Historical Data and Output of the S_DBMS

```
Query: SELECT Region.single_cell FROM variable.transient_dependent
WHERE variables.transient_dependent.V_%_H20 = 14.0
DEFINE cell.name Location_A Cell.location 1 1 1 ;
```

```
Output of the S_DBMS:
Case Name : prim_air_is_decreased_-_672nmc/h
Operating Variables:
   FI_1_1_10  4411  nmc/h
  Ram speed  6.1  s/h
  Grate speed _1  10.3  s/h
  Grate speed _2  10.5  s/h
  Grate speed _3  8.3  s/h
  Grate speed 4  11.4  s/h
  Prim_Air  20837  nmc/h
Extracted discrete cell(s),   Location_A     14.1
Reference discrete cell(s),   Location_A     14
 Matching coefficient : 0.986
Controlled variables :
   Variable name: AI_1,  Mean value  : 8.48
   Variable name: FI_1,  Mean value  : 13.62
```

The fault database is filled with 10 faults (Zsigraiová, 2005). A sample fault record is given in below

FAULT waste_chute_blocked
No_of_symptoms 4

1	Waste_level_in_hopper	Pre_Combustion	high
2	AI_1	Post_Combustion	high
3	TI_1	Furnace	low
4	FI_1	Boiler	low

Table 6 gives an answer to a multivariable query for finding process variables TI_1=1000., AI_1= 13., FI_1=7.5 from a chosen scenario, Prim_Air_is_decreased_-_1020nmc/h, as described in Fig.4 to define condition of variables. It takes less than 1 second of CPU time to get the answer.

Table 6: A Multivariable Query for Historical Dataand Output of the S_DBMS

```
Query: SELECT Region.single_cell FROM variable.transient_dependent
WHERE variables.transient_dependent.TI_1= 1000.
DEFINE cell.name Location_A Cell.location 1 1 1
JOIN number_of_variables 2
WHERE variables.transient_dependent.AI_1= 13.
DEFINE cell.name Location_B Cell.location 1 1 1
WHERE variables.transient_dependent.FI_1 = 7.5
DEFINE cell.name Location_C Cell.location 1 1 1
CHOOSE current_scenario 10 Prim_Air_is_decreased_-_2103nmc/h;
```

```
Output of the S_DBMS:
   Operating Variables:
     FI_1_1_10  5083  nmc/h
     Ram speed  13.1  s/h
     Grate speed _1  18.1  s/h
     Grate speed _2  21.4  s/h
     Grate speed _3  16.2  s/h
     Grate speed _4  10.4  s/h
     Prim_Air  19406  nmc/h
   Extracted multivariable
     TI_1  1028.6,  AI_1  7.9,  FI_1  12.6
   Reference multivariable
     TI_1  1000.0,  AI_1  13.0,  FI_1  7.5
   Matching coefficient : 0.95
   Conditions of Reference multivariable
     TI_1  normal
     AI_1  anormal_high
     FI_1  abnormal_low
```

REFERENCES

ACORDE, 1999, "Development of advanced control methodologies using reliable multi-detection sensors for boilers", Project no: BE95_2141, Brite/Euram Programme.

Boerstoel, GP, AM Lankhorst, 1998, "Modelling Combustion Chambers of Glass Melting Furnaces: a validation study", The 18th International Congress on Glass, July 5 to 10, San Fransico.

Carvalho, M.G., N. Kayakol and M. Nogueira, 2003, "Computational intelligence technologies in the design of hierarchical control architecture of glass melting furnaces." *Glass Technology* No.44, 25-29.

Douglas K. Barry, 1996, *The Object Database Handbook: How to Select, Implement, and Use Object-Oriented Databases*, John Wiley & Sons, Inc, New York, 1-28.

ECOTHERM, 2005, Development of evolutionary control technology for sustainable thermal processes, Project no: NNE5-2001-00141, EESD Programme.

Liao, S. H, 2005 "Expert system methodologies and applications—a decade review from 1995 to 2004", *Expert Systems with Applications* No.28, 93–103.

Venkat V, Raghunathan R, Surya N. K, and Kewen Y, 2003, "A review of process fault detection and diagnosis Part III: Process history based methods", Computers and Chemical Engineering No.27, 327_346.

Zsigraiová Zdena, Dr., 2005, Technical University of Košice, Faculty of Metallurgy, Slovakia, Private Communication.

Acknowledgement This research is financially supported by FCT (Fundação para a Ciência e Tecnologia, PRAXIS XXI / BPD / 22023 / 99) in Portugal.

COMPUTER SIMULATION MODEL OF THE MARINE DIESEL ENGINE

Josko Dvornik
Luko Milic
University of Split
Faculty of Maritime Studies
Zrinsko-Frankopanska 38, 21000 Split, Croatia
e-mail: josko@pfst.hr

Enco Tireli
University of Rijeka
Faculty of Maritime Studies
Studentska 2, 51000 Rijeka, Croatia
e-mail: tireli@pfri.hr

KEYWORDS
System Dynamics, Computer simulation, Automatic regulation, Diesel engine, PID-regulators, Heuristics optimization.

ABSTRACT

The System Dynamics Computer Simulation Modeling Methodology is one of the most suitable and effective ways of dynamics modeling of complex non-linear, natural, technical and organization systems. Studying the dynamics behavior of Marine Electric Power Systems, as one of the more complex dynamic non-linear technical systems, requires the application of only the most effective modeling methods.
The aim of this paper is to show the efficiency of the System Dynamics Computer Simulation Modeling of the dynamics behavior of Marine Diesel engine, as one of the most complex and non-linear marine technical systems. The Marine Diesel engine will be presented as a qualitative and quantitative System Dynamics computer models.

SIMULATION MODEL OF THE MARINE DIESEL MOTOR

The mathematical model of the Marine Diesel Motor with turbo compressor:

The mathematical model or level equations of the diesel engine with turbo-compression are:

$$\frac{d^2\varphi}{dt^2} = \frac{1}{T_H^2}\left(-T_{DH}\frac{d\varphi}{dt} - K_{DH}\varphi + T_S\frac{d\chi}{dt} + K_S\chi - T_U\frac{d\alpha_D}{dt} - K_U\alpha_D\right) \quad (1)$$

Where:
φ = relative change of angular velocity,
χ = relative shift of the high-pressure fuel pump cogged shaft,
α_D = relative consumer load change,
T_H = time constant proportional to the moment of inertia,
T_{DH} = time constant opposite-proportional to the moment of inertia of the engine as the object of regulation,
K_{DH} = self regulating and amplification factor,
T_S = motor time constant of inertia,
K_S = motor amplification factor,
T_U = generator time constant of inertia and
K_U = load amplification factors.

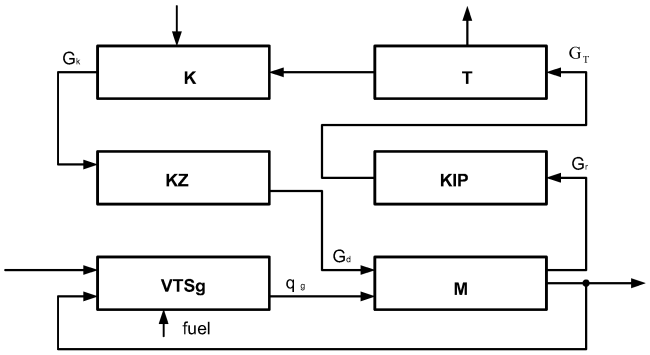


Figure 1. Marine Diesel engine with turbo compressor

SYSTEM DYNAMICS SIMULATING MODELLING OF MARINE DIESEL ENGINE

In the accordance with the System Dynamics Structural-qualitative (Figure 2.) and Mathematical-quantitative (Equations 1.) models of the Marine Diesel Motor with turbo compression, it is possible to work out the next System Dynamics Mental-Verbal-qualitative Simulation Model of the Diesel Motor:
"If the constants: TH=T_H = *time constant proportional to moment of inertia,* TDH= T_{DH}= *time constant opposite-proportional to moment of inertia of the engine as the object of regulation,*
KDH= *self-regulating and amplification factor,* TU=T_U= *generator time constant of inertia* and KU=K_U= *load amplification factor*, then there will be a growth of the D2FIDT2= *acceleration, or the second derivation of the*

relative changing of angular velocity FI= φ will drop". This means that those cause-consequence links have a "negative" (-) dynamics character.

"If the constants: TS=T_S= *motor time constant of inertia* and KS=K_S= *motor amplification factor* grow, then the D2FIDT2= *acceleration or the second derivation of the relative changing of angular velocity FI= φ* will be grow also". This means that those cause-consequence links have a "positive" (+) dynamics character.

"If the variables DFIDT= *speed or the first derivation of the relative changing of angular velocity φ,* FI= *relative changing of angular velocity φ,* ALFAD= *relative consumer load change* α_D and DALFADT= *speed or the first derivation of the relative load change* α_D grow then the D2FIDT2= *acceleration or second derivation of the relative changing of angular velocity φ* will drop". This means that those cause-consequence links have a "negative" (-) dynamics character.

"If the variables DKAPADT= *speed or the first derivation of the* relative shift of the high-pressure fuel pump cogged shaft *χ,* and KAPA= relative shift of the high-pressure fuel pump cogged shaft grow, then D2FIDT2= *acceleration or the second derivation of the relative change of angular velocity φ* will be grow". This means that those cause-consequence links have a "positive" (+) dynamics character. It is important to analyse the dynamics character of the next "exogenous" variables: CFI and MEL.

"If CFI= *nominal (goal's) relative change of angular velocity φ* grows, then the variable DISK1= *discrepancy between CFI and FI= relative change of angular velocity φ* will also grow". This means that these cause-consequence flows have a "positive"(+) dynamics character.

According to the described mental-verbal model it is possible to determine the System Dynamic structural model.

In observed system there are few feed back loops (FBL):

FBL1.(-): D2FIDT2(+)=>DFIDT(+)=>FI(-)=>D2FIDT2

FBL2.(-): FI(-)=>DISK1(+)=>KAPA1(+)=>KAPA(+)=> D2FIDT2(+)=>DFIDT(+)=>FI;

FBL3.(-): FI(-)=>DISK1(+)=>KAPA1(+)=>KAPA(+)=> DKAPADT(+)=>D2FIDT2(+)=>DFIDT(+)=>FI;

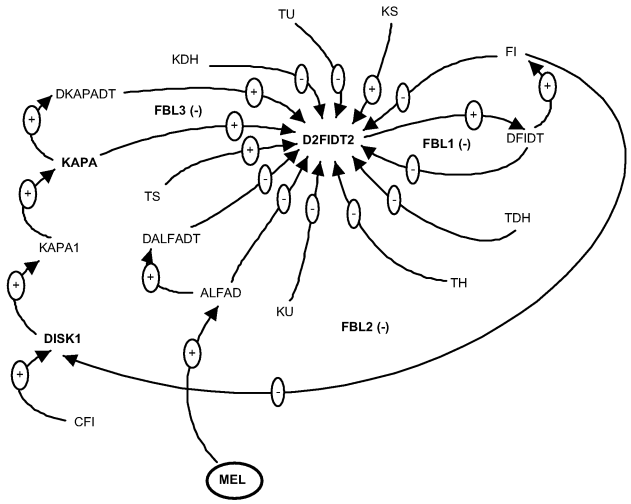


Figure 2. System Dynamics Structural Model of the Marine Diesel Motor

Figure 3. show flow diagrams of model.

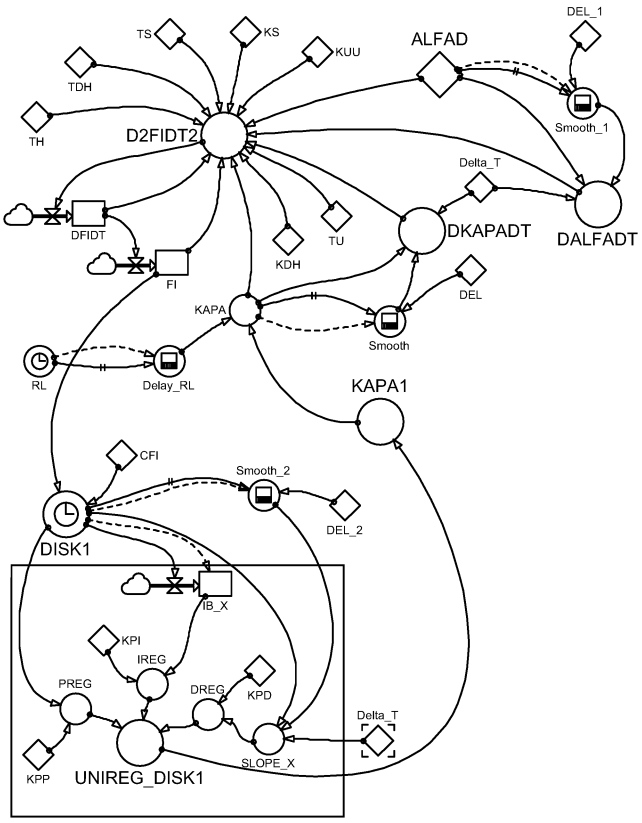


Figure 3. System Dynamics flow diagram of the Marine Diesel Engine in Powersim symbolics

Quantitative System Dynamics computer model of the Marine Diesel Engine

```
* ******************************************
* SIMULATION MODEL OF THE  MARINE DIESEL ENGINE WITH THE TURBO-COMPRESSOR:
*
R D2FIDT2.KL=-DFIDT.K*(TDH/(TH*TH))-FI.K*
          (KDH/(TH*TH))+DKAPADT.K*(TS/(TH*TH))
          +KAPA.K*(KS/(TH*TH))-DALFADT.K*(TU/(TH
          *TH))-ALFAD.K*(KUU/(TH*TH))
*
*  D2FIDT2=RELATIVE ACCELERATION OF ANGLE
          SPEED OF THE DIESEL ENGINE (S-2)
C TDH=2.75
C TH=1
C KDH=1
C KS=1
C TS=.5
C TU=1
C KUU=.8
L DFIDT.K=DFIDT.J+DT*D2FIDT2.JK
N DFIDT=0
L FI.K=FI.J+DT*DFIDT.J
N FI=0
A KAPA.K=KAPA1.K
A DKAPADT.K=SLOPE(KAPA.K,DT)
A ALFAD.K=MEL.K
A DALFADT.K=SLOPE(ALFAD.K,DT)
*SAVE D2FIDT2,DFIDT,FI,KAPA,DKAPADT,ALFAD
*
* PID REGULATOR :
*
A DISK1.K=CLIP(CFI-FI.K,0,TIME.K,1)
C CFI=1
A KAPA1.K=UNIREG(DISK1.K,KPP,KPI,KPD)+1E-30
C KPP=2
C KPI=1
C KPD=0
SAVE DISK1,KAPA1
SPEC DT=1,SAVPER=1,LENGTH=1
```

CONCLUSION

This Computer Simulation Model of the Marine Diesel engine is very suitable education simulator software, especially for marine students and marine system engineers because it provides them with the means to conduct numerous simulations for various productive scenarios.
The authors suggest the use of this modelling approach in designing, for the engineering educational processes as well, which will allow an active and creative participation of students.

REFERENCES

Forrester, Jay W. 1973/1971. "Principles of Systems", MIT Press, Cambridge Massachusetts, USA,

Munitic, A. 1989. "Computer Simulation with Help of System Dynamics", BIS, Croatia,

Munitic, A., 1989, "Application Possibilities of System Dynamics Modelling", *Proceeding of SCS Western Multiconference*, San Diego, California, USA,

Richardson, George P. and Pugh III Aleksander L. 1981. "Introduction to System Dymanics Modelling with Dynamo", MIT Press, Cambridge, Massachusetts, USA,

He:eB4ha, P. A., 1975, "Abtomat4⊥a∗4β cP∗ob∴ix ⊥hepŋet4Peck4x Pctahobok", Moskva, Rusia,

Munitic, A., Milic M. and Milikovic M. 1997. "System Dynamics Computer Simulation Model of the Marine Diesel-Drive Generation Set", *Proceeding of World Congress on Scientific Computation, Modelling and Applied Matematichs*, Berlin,

BIOGRAPHY



Josko Dvornik was born 1978. at Split Croatia, were he finished elementary and high Maritime school. He graduated in 2000. year on theme "*Application on computer simulation dynamics of behavior of ship propulsion system: windlass – asynchronous engine*" at Faculty of Maritime studies, University of Split. Today, he is Master of technical science. He has published over 40 scientific papers on system dynamics simulation modelling. He is a member of the SCS, *The Society for Computer Simulation International,* IASTED, *The International Association of Science and Technology for Development* and SDS, *System Dynamics Society.*

SIMULATION IN PROCESS INDUSTRY

Numerical Modelling and Analysis of Flow and Heat Transfer Processes in a Coke Oven

K. Hellgardt
K. J. Vyas
N. S. Hanspal
V. Nassehi
Advanced Separation Technologies Group
Chemical Engineering Department
Loughborough University, Loughborough, Leicestershire, LE11 3TU, UK

KEYWORDS
Coke oven, Modelling, Simulation, Heat Transfer, Momentum Transfer, Euler time-stepping, Implicit-Theta time-stepping

ABSTRACT
In this paper, we present the study of heat transfer processes and fluid dynamics in a coke oven using the finite element method. The heat transfer processes are represented by a non-linear transient energy balance equation whilst the flow of gas in the coke oven is described by Darcy's law and the continuity equation. A slightly perturbed continuity equation is employed in order to satisfy the Ladyzhenskaya-Babuska-Brezzi stability criterion. C^0 continuous biquadratic 9-noded elements have been used in conjunction with mixed formulation. Streamline upwinding technique has been applied to the heat balance equation in order to stabilise the solution. Gas viscosity, thermal conductivity of coal and refractory bricks, specific heat of coal and permeability of the coke oven bed are all dependent on temperature. Appropriate models for these temperature dependent parameters have been incorporated within the main solution algorithm. It is shown that an Euler time stepping scheme for the Darcy equation along with an implicit-theta time stepping scheme for energy balance equation offers a robust and accurate technique for the solution of governing equations of the described problem. Simulated results are then compared with published experimental data for model validation.

INTRODUCTION
Blast furnaces in steel plants use coke as one of the raw materials. It is defined (Lowry, 1963; Wilson and Wells, 1950) as a fused, cellular, porous structure that remains when bituminous coal and other carbonaceous material is heated in the absence of or in the presence of little air after having removed free moisture and the major portion of the volatile matter. The main industrial application of coke in a blast furnace is to reduce iron oxide to iron and to generate heat for melting both slag and iron. Though, in recent times new techniques in metallurgical industries have been developed and have reduced coke consumption, the demand for coke still stands at 350 million metric tons per year (Gorham Technologies).
The sharp increase in energy costs makes it imperative to have systematic studies of coke making process with a view to reduce cost. Considering the significance of coke in the metallurgical industry it is essential to understand the coke making process thoroughly. Merrick, 1984, Miura *et al.*, 1991 and Tang, 1984 have studied the heat transfer and gas flow processes using finite volume technique which provides the basic ideas about mathematical modelling of coke making process. The aim of this research is to simulate the temperature distribution and gas velocity fields in a coke oven by employing finite element technique in order to predict the thermal and flow behaviour accurately.

MATHEMATICAL MODEL
Thermal Model
In order to model the temperature distribution inside a coke oven, a 2-D transient energy balance equation is considered.
The model is based on number of assumptions listed below:

1. The rate of evolution of volatile matter has been neglected.
2. The local effect of fissuring and of any non-uniformity of bulk density caused by hold ups and segregation are ignored.
3. The effects of boiling, condensation and convection of moisture have also been neglected.

Energy balance equation
Heat transfer within a differential volume of coke oven bed is expressed in the form of a 2-D transient, non-linear and non-isothermal equation,

$$\rho_c c_c \frac{\partial T}{\partial t} + \rho_g c_g \left(v_x \frac{\partial T}{\partial x} + v_y \frac{\partial T}{\partial y} \right) - \frac{\partial}{\partial x}\left(k_x \frac{\partial T}{\partial x} \right) - \frac{\partial}{\partial y}\left(k_y \frac{\partial T}{\partial y} \right) - \dot{S} = 0 \quad (1)$$

where c_c is specific heat capacity of coal, ρ_c is density of coal, k_x and k_y are thermal conductivity of coal in x and y direction respectively, ρ_g is density of gas, c_g is specific heat of gas, v_x and v_y are the velocity of fluid in x and y directions respectively, T is temperature of coal and $\dot{S}$ is the heat source term due to generation of gas.

Flow Model
The coke oven gas is assumed to be an incompressible gas passing through the porous coal matrix. The coke

oven bed consists of a porous region in which flow is described by the Darcy equation and the continuity equation. For a transient system these equations can be represented by

Darcy equation

$$\rho_g \frac{\partial v_x}{\partial t} + \frac{\eta}{K_x} v_x + \frac{\partial p}{\partial x} = 0 \quad (2)$$

$$\rho_g \frac{\partial v_y}{\partial t} + \frac{\eta}{K_y} v_y + \frac{\partial p}{\partial y} = 0 \quad (3)$$

Continuity equation

$$\frac{1}{\rho_g c_s^{\,2}} \frac{\partial p}{\partial t} + \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \dot{m} = 0 \quad (4)$$

where, ρ_g and η are the constant density and viscosity of gas respectively, c_s is the velocity of sound, K_x and K_y are the permeability of the medium in x and y direction respectively, p is the pressure and v_x and v_y are velocity of coke oven gas in x and y direction respectively. The rate of gas generation within the coke oven bed is represented by $\dot{m}$, and is a function of temperature (Wilson and Wells, 1950).

$$\dot{m} = \rho_g \left(6\times10^{-6} T + 0.0013\right) \quad (5)$$

Auxiliary equations

1) Thermal conductivity of coal

The model proposed by Merrick D (1984), is used for calculation of thermal conductivity of coal during the carbonisation process and is represented by,

$$k = \left(\frac{\rho_c}{4511}\right)^{3.5} (T)^{0.5} \quad (6)$$

where, k is the thermal conductivity of coal, ρ_c is true density of coal and T is the temperature of coal

2) Specific heat of Coal

Merrick, 1984 proposed a model for the change in specific heat of coal with temperature during the coke making process,

$$c_p = \left(\frac{R}{a}\right) g(380/T) + 2g(1800/T) \quad (7)$$

where, $g(x)$ is a function of temperature, expressed as

$$g(x) = \left(\frac{\exp(x)}{\left(\frac{\exp(x)-1}{x}\right)^2}\right) \quad (8)$$

In equation (7), c_p is the specific heat capacity of coal, R is the gas constant, T is the absolute temperature and a is the mean atomic weight of coal defined by $\frac{1}{a} = \frac{C}{12} + \frac{H}{1} + \frac{O}{16} + \frac{N}{14} + \frac{S}{32}$, where C, H, O, N, S are the mass fractions of the respective elements.

True density of coal

A direct relation between true density and temperature is not available and hence experimental values as given by Merrick, 1984 have been considered for the simulations. Values for the true density of coal are tabulated in Table 1.

Table 1 - True Density of Coal

Temperature Range, T [K]	True Density, ρ (kg/m^3)
293 ≤ T ≤ 573	1250
573 < T < 673	1330
673 < T < 773	$0.55T + 1110$
773 < T < 1500	$1.20T + 785$

Permeability of coal

The permeability of coal is a function of the temperature of coke. An empirical relation is not available but Miura *et al.,* 1991 conducted experiments on a laboratory scale coke oven and proposed experimental values of permeability of coal as a function of temperature. These are listed in Table 2. The range of permeability variation is from 10^{-10} m^2 to 10^{-12} m^2.

Table 2 - Permeability Values

Temperature Range, T[K]	Permeability Values, K [m^2]
293 < T < 373	10^{-10}
373 < T ≤ 723	$1.92\text{x}10^{-12}\text{T}+1.39\text{x}10^{-9}$
723 < T ≤ 823	$7.60\text{x}10^{-13}\text{T}-5.45\text{x}10^{-10}$
823 < T	$8.0\text{x}10^{-11}$

Viscosity of gas

The equation proposed by Bird *et al.,* 2001 has been considered to calculate the gas viscosity,

$$\eta = 2.6693\times10^{-5} \frac{\sqrt{MT}}{\sigma^2 \Omega_\eta} \quad (9)$$

where η is gas viscosity in gm cm^{-1} sec^{-1}, T is temperature of gas in K, σ is a characteristic diameter of gas atom in Å, M is molecular weight of gas and Ω_η is a constant which is a function of dimensionless temperature, KT/ε, where ε is characteristic energy of collision. The value of Ω_η reduces slowly as temperature increases and a constant value of 0.80 has been used.

Thermal conductivity of silica brick

The coke oven is lined with refractory bricks composed of silica for insulation purposes. The composition of silica brick is made up of 94% silica and 3.13% silicon

dioxide. The thermal conductivity of this insulation layer gradually increases as temperature increases, F. Gantry, 1982. The conductivity can be modelled using the relationship

$$k_B = 0.0007(T) - 0.1966 \tag{10}$$

where k_B, is a thermal conductivity of silica brick and T is the temperature of silica bricks.

The finite element solution scheme

The numerical scheme used in this work is based on the U-V-P method in which both velocity and pressure in the governing equations are the primitive variables and are discretized as unknowns. A domain Ω defined by a smooth continuous boundary Γ is discretized into a mesh of finite elements so that

$$\sum_{e}^{m} \Omega_e = \Omega \text{ and } \sum_{e}^{m} \Gamma_e = \Gamma \tag{11}$$

where m is the total number of elements in the mesh. In the first step the velocity and pressure are determined which are then used to solve the nodal unknown temperatures.

Nine-noded biquadratic C^0 continuous elements with equal order interpolation functions for velocity, pressure and temperature are used in the present mixed formulation. In this scheme, the unknown variables can be represented as

$$v_x \approx \tilde{v}_x = \sum_{i=1}^{k} N_i(x,y) v_{xi}$$

$$v_y \approx \tilde{v}_y = \sum_{i=1}^{k} N_i(x,y) v_{yi} \qquad p \approx \tilde{p} = \sum_{i=1}^{k} N_i(x,y) p_i$$

$$T \approx \tilde{T} = \sum_{i=1}^{k} N_i(x,y) T_i \tag{12}$$

where N_i is the shape function associated with nodes i = 1,…, k (k is the number of nodes per elements)

Streamline upwinding Petrov-Galerkin scheme (Nassehi, 2002) has been employed to develop a stable finite element algorithm for the solution of the energy balance equation. The weight function used in the studies can be represented as

$$W_I = N_I + \alpha N_{I,i} \frac{\overline{v}_i}{|v|^2} \tag{13}$$

where N_I and $N_{I,i}$ are the shape function and its derivative respectively, v is the velocity vector and α is an upwinding parameter.

Implicit-Theta and Euler time stepping schemes have been employed for the temporal discretization of energy equation and flow equations respectively.

Working equations

Energy balance working equations

The finite element working equations for the thermal model can be represented as

$$[M]\{\dot{T}\} + [K]\{T\} = \{F\} \tag{14}$$

where,

$$[M_{ij}] = \iint_{\Omega_e} \rho_c c_c w_j N_i d\Omega_e \tag{15}$$

$$[K_{ij}] = \iint_{\Omega_e} w_j \left[\rho_g c_g v_x \frac{\partial N_i}{\partial x} + \rho_g c_g v_y \frac{\partial N_i}{\partial y} \right] d\Omega_e + \iint_{\Omega_e} \left[k_x \frac{\partial w_j}{\partial x} \frac{\partial N_i}{\partial x} + k_y \frac{\partial w_j}{\partial y} \frac{\partial N_i}{\partial y} \right] d\Omega_e \tag{16}$$

and

$$\{F\} = \int_{\Gamma_e} k_x w_j \frac{\partial T}{\partial x} n_x d\Gamma + \int_{\Gamma_e} k_y w_j \frac{\partial T}{\partial y} n_y d\Gamma_e + \iint_{\Omega_e} w_j \dot{S} d\Omega_e \tag{17}$$

Flow model final working equations

The final working equations for the flow model have been described elsewhere (Ruziwa *et al.,* 2004).

Solution Algorithm

The solution algorithm used in the present work, can be summarised in the following steps:

1. The values of velocity, pressure and temperature are initialised to zero.
2. The physical and material properties, geometrical data and boundary conditions are read from the input data file.
3. The time loop starts with specified time increment.
4. The rate of gas generation is evaluated, which is dependent on the temperature of gas.
5. If the pre-programmed maximum simulation time is not reached then the calculation of pressure and velocity starts.
6. The Darcy equation and the continuity equation are solved to obtain nodal velocities and pressures.
7. The energy balance equation is now solved to obtain unknown nodal temperatures.
8. Convergence is checked using a calculated ratio of the Euclidean norm (Lapidus and Pinder, 1982) between successive iterations to the norm of the solution via the following relationship

$$\sqrt{\frac{\sum \left| X_i^{r+1} - X_i \right|}{\sum_{i=1}^{N} \left| X_i^{r+1} \right|^2}} \leq \varepsilon \tag{18}$$

where, i is the number of the iteration cycle, N is the total number of degree of freedom and ε is a pre-selected convergence tolerance value.

9. If solution does not converge then go to step 3 and also update the rheological properties, thermal properties and permeability.
10. If the current time becomes greater than the maximum pre-programmed time then calculations are stopped, else steps 3 to 8 are repeated.

RESULTS AND DISCUSSION

Physical and Numerical Data

The physical properties of coal are transient and vary continuously as time progresses. However, at the start of simulations the values considered are as follows: initial reference temperature 293.0 K, reference pressure 101.3 kPa, heat capacity of coal 1005 J/kg K, thermal conductivity of coal 1.49 W/m K and bulk density of coal 700 kg/m^3. Hydrogen gas is assumed to be a Newtonian fluid, with density of 1.312 kg/m^3 and viscosity of 8.88×10^{-8} kg/m.s. The coke oven gas comprises of more than 50% of hydrogen gas and hence the properties of coke oven gas have been considered to be the same as that of hydrogen gas. The thermal conductivity of the silica insulation layer is 0.3138 W/m^{2}K.

Computational domain

The coke oven that has been considered for the present studies is of slot type (Wilson and Wells, 1950). Figure 1 shows the computational domain used in the analysis (including dimensions). The width of the silica insulation layer is 0.024m. The total number of nodes in the finite element mesh is 8181 whilst the number of elements is 2000.

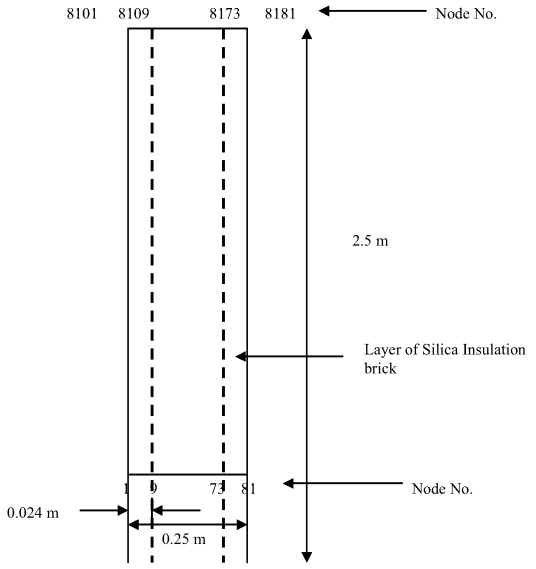


***Figure 1**: Schematic diagram of Coke Oven*

Boundary conditions

A temperature of 1273 K is imposed on the sidewalls and at the bottom. At the top the temperature is imposed as 1073 K. At the bottom end of the coke end the velocity of gas in *y* direction is considered as 0.0m/s whilst on the two sides of coke oven walls the velocity of gas in *x* direction has been taken as 0.0m/s (non-slip). The pressure datum at the top of the coke oven has been imposed as 0.0kPa.

Temperature, Velocity and Pressure profiles

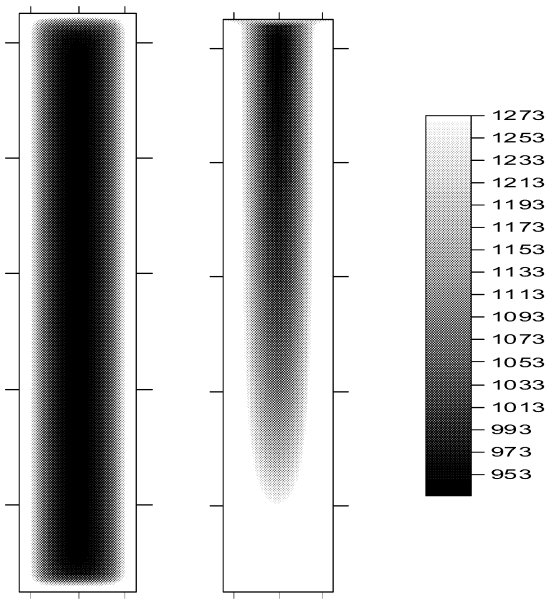


***Figure 2:** a) Temperature distribution in coke oven after 1800s, b) temperature distribution after 11760s*

Temperature profiles: Figures 2a and 2b present the temperature distribution inside the coke oven at different time intervals i.e. 1800s and 11760s, respectively. There are two main modes of heat transfer within the coke oven bed, which have been considered in the present study. The first one is conduction that arises from the direct contact of solid coal and coke particles whereas the second one is convection and is a result of coke oven gas generation. It is seen from the figures that as the time progresses the temperature inside the coke oven increases gradually. It is also observed from the temperature profiles that the portion of coal directly in contact with the walls gets rapidly heated. This is mainly because a constant temperature value of 1273 K has been imposed on the coke oven walls. The temperature distribution becomes almost homogeneous after 15120s. As a result the rate of heat transfer becomes significantly small and consequently no further temperature changes are observed. Combined effects of conduction and convection reduce the coking time drastically. Heat transfer in a coke oven in absence of any gas generation has already been studied by Vyas *at el.* (2004) on a small-scale coke oven bed. It was observed that the coking time is higher in absence of any gas generation. For the sake of model validation we compare our simulated data with experimental ones by Wilson *et al.,* 1950 who present the temperature distribution in a real coke oven (Figure 8) at different distances from the oven wall. The simulated temperature distribution at different distances from the coke oven walls is given in Figure 9. Six different locations i.e. location 1 (centre of coke oven), location

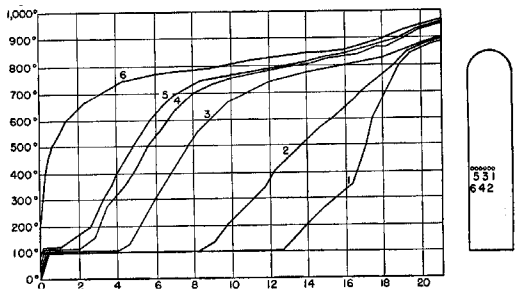


Figure 3: *Temperature distribution in a real coke oven according to Wilson et al., 1950*

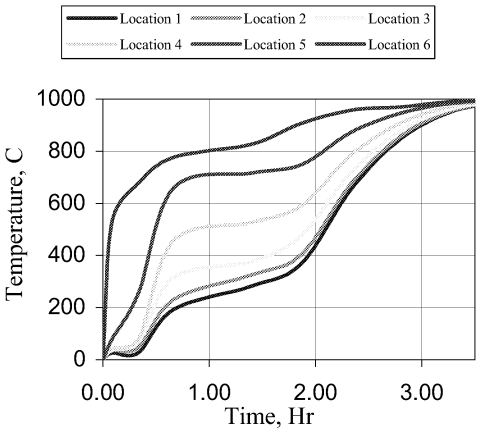


Figure 4: *Simulated temperature distribution in a coke oven*

2 (1/3 from centre), location 3 (1/2 from centre), location 4 (2/3 from the centre), location 5 (5/6 from centre) and location 6 (adjacent to the coke oven wall) have been chosen for comparison purposes. Figure 8 and Figure 9 when compared indicate that the experimental and simulated temperature distributions inside the coke oven follow the same trend. Temperature near the wall increases very rapidly whilst in the centre portion of coke oven the temperature initially remains low and only starts to increase after certain time interval. Further, it is noticed that the order of magnitude of coking time is the same in both cases. In our thermal model, heat losses to the surrounding atmosphere from the gaseous discharge have been neglected. The heat of reaction for the generation of coke oven gas has also not been taken into consideration. Moreover, effects of moisture have also been neglected. All these factors do account fo the small difference in the values obtained for the simulated and the experimental coking times.

Velocity profiles: As the temperature increases with time the velocity at the top of the oven increases. This is due to the increasing rate of gas generation with time. The flow behaviour is of plug type, which is in agreement with the Darcy model. The velocities at the side of the oven walls are smaller compared to the centre line of the oven. The reason for this is that at the oven wall the temperature is very high compared to the centre. This leads to a lowering of permeability value and increase in the viscosity of gas. Both these effects reduce the tendency of gas to escape and hence a lower velocity on the sidewalls. The reason for low velocities in the bottom portion of the coke oven is due to the imposition of boundary conditions i.e. $v_y=0$. For the constant gas source at the bottom of the coke oven, the effect of temperature dependent viscosity on velocity profiles has already been studied by Vyas *et a.,* (2004. They observe that high viscosities result in low velocities near the oven walls whilst in the centre the lower viscosity of the gas due to a lower temperature results in high gas velocities.

Pressure profiles: The pressure drop in the coke oven increases with time from initially 7kPa to 85.0kPa. It was observed that as the time progresses the permeability of the coke first increases with increasing temperature and then starts to decrease. The pressure drops obtained at different time intervals correspond to the permeability model that has been incorporated in the main solution algorithm. In all the cases the pressure gradually decreases from inlet to the exit of the coke oven, which also indicates the direction of flow of the gas from bottom to top of the coke oven chamber.

CONCLUSIONS

The total carbonisation time of coal depends on parameters such as flame temperature, coal preheating temperature, type of coal blend and its properties such as moisture content and bulk density. The gas velocity reduces in the high temperature regions and increases in the low temperature regions due to its change in viscosity. The pressure drop increases gradually from bottom to top as the time progresses. The gas generation causes convective heat transfer and reduces the coking time significantly. The comparison of the simulated and experimental temperature profiles shows a rather good prediction of the behaviour of a coke oven by our model. Incorporation of effects such as rate of evolution of volatile matters, heat of reaction, effect of boiling, condensation and convection of moisture would model the coke oven process even more realistically.

NOMENCLATURE

c Specific heat capacity (J/kg K)
e Element number
k Thermal conductivity (W/m K)
$\dot{m}$ Rate of gas generation (kg/s)
N_i Interpolation function (element shape function)
p Pressure (kPa)
t Time (s)
T Temperature (K)
v Velocity of fluid (m/s)
W_i Weighted function

Greek Symbols

α Upwinding parameter
Γ Boundary of the solution domain
Ω Solution domain

Ω_e An element domain within the solution domain
η Viscosity (kg/m s)
θ Time stepping parameter used in the implicit theta method
ρ Density (kg/m^3)
ε Tolerance value

Subscripts

c Coal
g Gas
x x component
y y component
B Insulation brick

REFERENCES

1. Lowry H H, *Chemistry of coal utilisation*, John Wiley and Sons, London, 1963
2. Wilson P J and Wells H J, *Coal, coke, and coal chemicals*, McGraw-Hill, London, 1950
3. *World coke and coke supply overview*, Gorham Technologies, http://www.goradv.com
4. Merrick D. *Mathematical modelling of the coking process*, Ironmaking Conference Proceedings, Vol 43, 287-302, 1984
5. Miura K, Inoue K, Tankatani K and Nishioka K, *Analysis of steam flow in coke oven chamber by test coke oven and a two-dimensional mathematical model*, ISIJ International, Vol 31, No. 5, pp. 458-467, 1991
6. Tang D L, Leininger G, Williams T, *Modelling and control of the coke making process*, PhD Theses, Purdue University, 1984
7. Nassehi V, *Practical aspects of finite element modelling of polymer processing*, John Willey & Sons Ltd, England, 2002
8. Bird R B, Stewart W E, Lightfoot E N, *Transport phenomena*, 2nd edition, John Wiley & Sons, New York, 2001
9. Ruziwa W R, Hanspal N S, Waghode A.N., Nassehi V. and Wakeman R J., *Computer aided design of pleated cartridge filters for high viscosity applications,*, Transactions of Filtration Society (in print), 2004
10. Lapidius L. and Pinder G.F., *Numerical solution of partial differential equations in science and engineering*, John Willey, New York, 1982
11. Vyas K J, Hanspal N S, Ruziwa W R, Nassehi V & Hellgardt K, *Finite element modelling of heat transfer processes and fluid dynamics in a model coke oven,* 4th European Thermal Science Conference, Birmingham, UK, 28th March-3rd April 2004
12. Gentry F M, *The technology of low temperature carbonization*, http://www.rexresearch.com/coal/ , 1982

INFLUENCE OF SPARE PARTS SUPPLY ON THE MODELING OF MAINTENANCE ORGANIZATION

Petar Stanojević,
Petroleum Industry of Serbia
Otona Zupancica 48/47
11000 Belgrade, Serbia and Montenegro
E-mail: petar.stanojevic@nis.yu

Vasilije Mišković
Military Academy
Neznanog Junaka 12
11000 Belgrade, Serbia and Montenegro
E-mail: vasilije1@ptt.yu

Vladimir Bukvić
Military Academy
Neznanog Junaka 12
11000 Belgrade, Serbia and Montenegro
E-mail: zlatka@beotel.yu

Mirjana Misita
Faculty of Mechanical Engineering
University of Belgrade
Kraljice Marije 16
11000 Belgrade, Serbia and Montenegro
E-mail: mmisita@mas.bg.ac.yu

KEYWORDS
Maintenance organization, simulation model, spare parts supply.

INTRODUCTION

Paper describes complex simulation model of spare parts supply in maintenance sector one manufacturing company.

The problem of spare parts in technical equipment maintenance is more or less but constantly in the centre of attention since the influence of the spare parts supply on technical equipment performance is big. At the same time, spare parts stocks make an expense anyway. On one hand, there is a tendency to increase the spare parts stocks to rise the level of the technical equipment performance and, on the other hand, there is a tendency to reduce the spare parts stocks to cut down expenses. The problem becomes more complicated in multilevel maintenance systems where maintenance work is distributed within levels (eshalons of support) and in space. The distribution includes the distribution of maintenace work betwen levels and different areas and it necessarily includes the distribution of spare parts within levels and areas, but the spare parts distribution does not have to follow the distribution of maintenance work.

The problem of the spare part distribution within levels or according to their range and quantity is most often solved by conducting technological maintenance work on particular levels or by judging the intensity of the demand for a particular spare part taking into consideration the expense criterion. With the developement and application of different methods and models for inventory control, it is strived towards the optimisation of the inventory amount mostly due to the expenses.

MODEL DESCRIPTION

In order to quantify the influence of the spare parts supply on technical equipment performance there has been developed a model for the spare parts supply system and maintenance system. The four-level maintenance system have been taken into consideration (Figure 1).

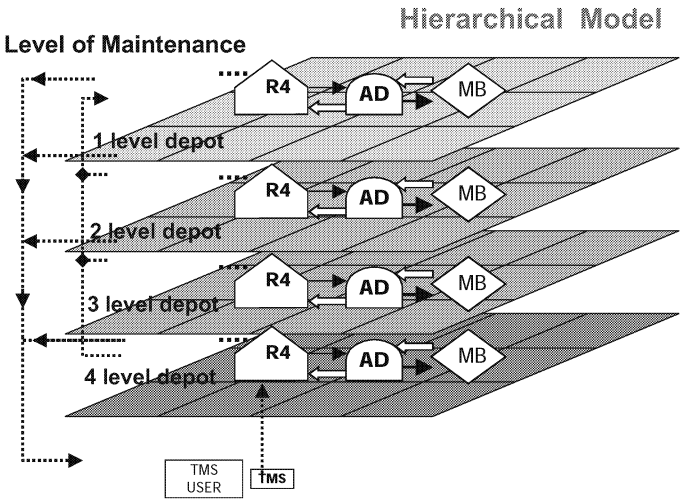


Figure 1. Simplyfied model structure

More detailed model structure (with management bodies, information flow, etc.) is given on Figure 2.

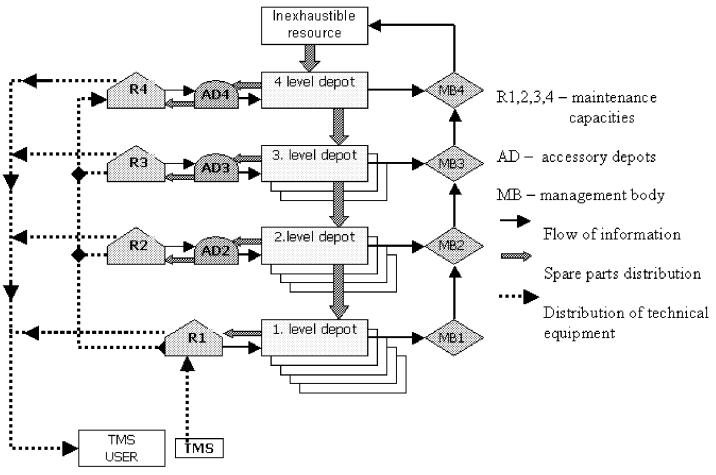


Figure 2. Model of the four-level spare parts supply system and the four-level maintenance system

EXPERIMENT

Figure 3 shows simulation model of four-level organization of maintenance services.

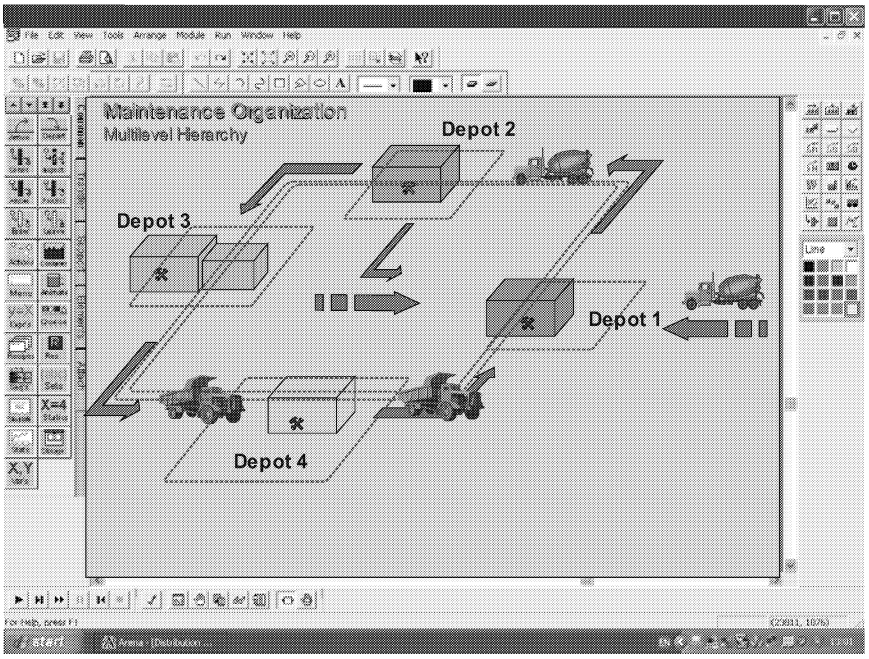


Figure 3. Simulation Model

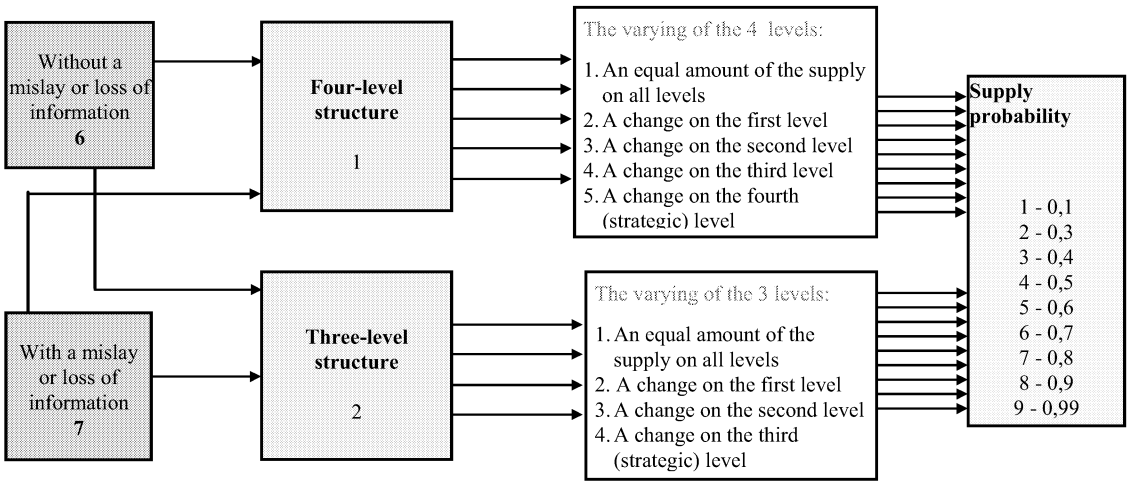


Figure 4. Plan of the experiment

Experiments performed on the model are organised in the way that a simulation was conducted for all possible combinations of versions supply varying and the supply availabilitys 0,1; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9 and 0,99.

The experiment is organised in the following way: Realisation of the experiment represents one simulation for the whole period of the functioning of the system during the observation time needed for one combination of versions. For the purpose of one combination of varsions a number of required experiments was carried out to obtain statistically a great number of the results of the experiments and the processing of the obtained data gave mean values of the number of operable technical equipment. For instance, here is one of the versions:
Informational system: Without a mislay or loss of information (number 6)
Organisational structure: The four-level structure (number 1)
The second supply level (number 3)

Supply of the second level: the supply availability/probability 0,4 ***(number 3)***

Experiment code 613-3

In order to obtain a statistically meaningful number of operable technical equipment, this combination requires a repetition of a sufficient number of experiments.

In order to obtain the dependence of technical equipment performance on the supply of the second level, the experiment 613 is repeated for the supply availbility values 0,1 (613-1); 0,3 (613-2); 0,4 (613-3); 0,5 (613-4); 0,6 (613-5); 0,7 (613-6); 0,8 (613-7); 0,9 (613-8) i 0,99 (613-9). Other values are constant.

The cross-section of the time, in which the reliability of technical equipment is observed, is the moment when the performance of a simulation is interrupted. In the process model it is possible to choose practically any cross-section of the time (one or more time cross-sections) to be observed.

ANALYSIS OF THE RESULTS

Due to their volume, it is not possible to illustrate all the results of experiments even in more elaborate papers. Results consist of the observation of technical equipment performance during intensive exploitation and for a period of time after considerable reduction in exploitation. They are observed according to types of equipment, groups of equipment, maintenance units, levels, for two diffrent informational systems and for twenty diffrent organisational structures, for various loads of maintenance systems, in different cross-sections of the time depending on the purpose of the analysis. Here, the results will be presented for he purpose of illustration and for the model in this paper will be taken only a couple of basic conclusions.

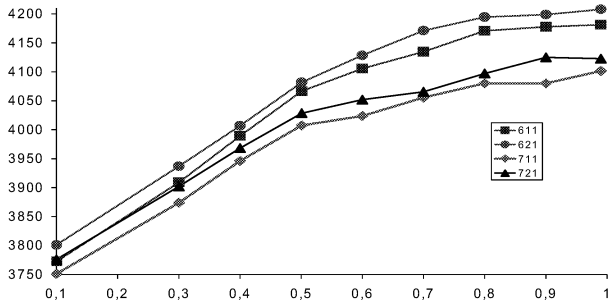


Figure 5. Influence of the spare parts supply of all the levels of the supply system on technical equipment performance

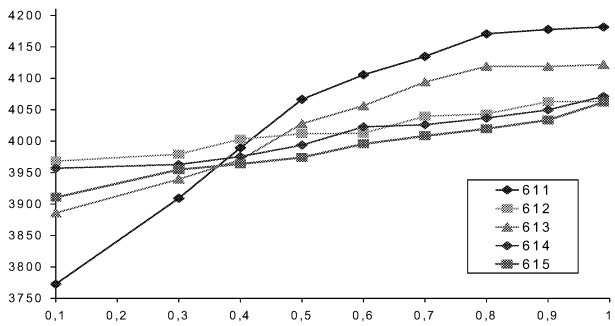


Figure 6. Influence of an individual spare parts supply level technical equipment performance in the four-level structure with a good informational system

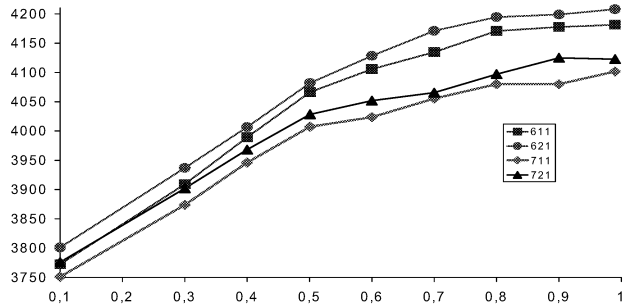


Figure 7. Influence of the spare parts supply of the second level of the supply system on technical equipment performance

The Figure 5 shows that a version where all levels have the same spare parts supply availability it should range from 0,75 to 0,85, depending on the level of technical equipment performance (number of available technical assets). This means that the supply probability of the whole system ranges from 0,9 to 0,95. It goes without saying that only those parts and structures of the equipment are taken into account whose mean time of the first failure is shorter than the time of technical equpiment exploitation.

Figure 6 represent an influence of an individual spare parts supply level on technocal equipment performance in four-level structure with a good informational system. The basic conclusion is that the second level has the biggest influence. This level consists of a relatively wide range of maintenance activities which last for a short period of time and are carried out at a small distance from or on the very spot where the failure occured. This shows in advance that the influence of this level mainly depends on disposable resoources and spare parts availability. Depending on the supply availability/probability of the level, the expected number of operable technical machines and devices is varied about 200 or 9%. As it also depends on the organisational structure, the expected difference is about 2,5% to the advantage of the three-level (echelons) structure.

Figure 7 represents the influence of the second supply level. It is obvious that the influence of the second level in all versions has the same form as when the overall influence is observed. This means that the three-level structure with a good informational system is the best choice. The three-level structure in both versions proves to be better than the four-level structure, whereas the informational system has a bigger influence than the organisational structure.

This paper shows only some basic conclusions with illustrations for particular concrete indicators. Due to the amount of results, the purpose of this paper is not to describe the entire analysis but merely to illustrate how such an analysis can be carried out.

CONCLUSION

The analysis of the influence and determining dependences are very complex processes for they include the analysis and determining the dependence of technical equipment performance on the spare parts supply under the influence of a variety of factors and on the spare parts supply of the whole supply system. It also includes the analysis of the share of the spare parts supply of the particular supply levels in the overall influence, the observation of a direct and indirect influence on technical equipment on the whole but also on technical equipment according to their type. Different types of technical equipment have various methods and different organisation of maintenance. Thus, it is important to determine the method how these dependences and influences should be quantified.

This dependence provides the use the regularity in operative management of the spare parts supply system and in the design and redesign of the spare parts supply system as well as the maintenance system. For instance, it is obvious from the above that when determining the amount of supply, the second level must have rather high spare parts supply (inventory level). As for the operative management, it is important to maintain the spare parts supply of this level as it was projected. The reason is simple - this level has the biggest influence. As for the maintenance work, the second level consists of a relatively wide range of activities of a short duration which are carried out at a small distance from the place or on the very spot where the failure occured. The influence of this level mainly depends on disposable resources and spare parts availability. All of this enables quick restoration of technical equipment into an operational condition and use. The first level very quickly restores technical equipment into an operational condition, however, the scope of its maintenance work is limited, whereas the third and fourth level have a wide scope of maintenance work but they last for a longer period of time and logistics time is long. Thus, a direct influence of these levels is considerably smaller. However, it is necessary to point out one more time that this example reffers to their direct influence and to the period of intensive exploitation. It is also important to emphasise that this procedure is applicable on other periods as well. It can be expected that the results may differ to some extent but the analysis and drawing (tendencies) conclusions are done in the same manner.

Hence, a general conclusion can be made that the method represented in this paper provides a possibility to quantify the influences and dependences, conduct required analysis and draw necessary conclusions.

REFERENCES

Evans, G.N., Towill, D.R., Naim, M.M., 1995. ‘’Business Process Re-engineering the Supply Chain’’, Production Planing & Control, vol. 6., No. 3., p.227-237.

Burlat, P. , Campagne, J.P., Neubert, G., 1998. ‘’Modeling Organizational Structure; a New Challange for Simulation’’, EUROSIM '98, Simulation Congress, Helsinki.

Program package for maintenance system simulation and spare parts supply, 2000, *KTOB, VTA VJ*, Belgrade.

PETAR STANOJEVIĆ was born in Belgrade, Serbia and Montenegro and went on the University of Belgrade, where he studied mechanical engineering and obtained his engennering degree in 1989 and Doctoral degree in 1997. He was working as a Professor on the Military Academy and Mechanical engeneering faculty University of Belgrade. In 2003/2004 he was working as a Deputy Executive Director and Cheef of Cabinet Ministry of Defence. From 2004 he is Corporate Development Director – Petroleum Industry of Serbia ever since.

Using Simulation to Analyse the Procurement of Addictives for Lubricants in the Oil Refinery of Porto, Portugal

J. Manuel Feliz-Teixeira
António E. S. Carvalho Brito

September 2004

GEIN – Section of Management and Industrial Engineering
Faculty of Engineering of the University of Porto, PORTUGAL
Emails: feliz@fe.up.pt; acbrito@fe.up.pt

KEYWORDS: Simulation, Addictives, Lubricants, Transport Echelons, Purchase Policy, Costs.

ABSTRACT

Data in this article results from a study of simulation about the procurement of addictives for lubricants in the Oil Refinery of Porto (Portugal). Here we describe the real case under study as well as the method we have chosen to board it by means of simulation, which finally leaded us to some interesting results regarding the choices on purchasing the addictives from the suppliers. Costs including transportation, holding and stockouts have been considered during the simulation as well as in the final conclusions, which let us recommend a more frequent reordering of materials at the lubricants plant. This we expect to result in the saving of around 0.3M€/year in global costs.

1. Introduction

This Oil Refinery (Fig. 1) is located in Leça da Palmeira, an important suburb of Porto, and is operating since 1970. Even if it already had been used with a processing capacity of 7.5 MTon of crude per year in 1975, since 1982 its capacity has been maintained in the 4.4 MTon/year, mainly due to the installation of a second refinery in the south of the country (GalpEnergia, 2003).

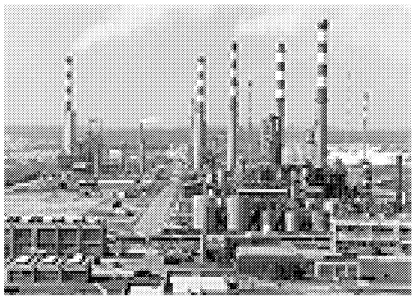


Fig. 1 View of the Refinery of Porto, Portugal

However, since its beginning the refinery included a Lubricants-Base Plant, which at the present has a producing capacity of 150 KTon/year, and where diverse kind of lubricants are produced to the industry and to the general market. The present study concerns with the procurement related with this plant, which is only a part of the refinery.

The study has been conducted with the kind participation of engineer Mr. José Carlos Fernandes, from *GalpEnergia*, the company responsible for the refinery, who at the time was in charge of the *Department of Logistics and Stock Management of Lubricants*. The main objective was to study in more detail the process of purchasing near 200 different additives for lubricant production, ordered from 10 different European suppliers, and with different transportation costs based on the quantity ordered. Notice that final lubricants normally include around 10% of addictives. The Supply Chain in study was then a procurement network whose final node was this plant, as figure 2 illustrates, with the suppliers spreading from England to Spain.

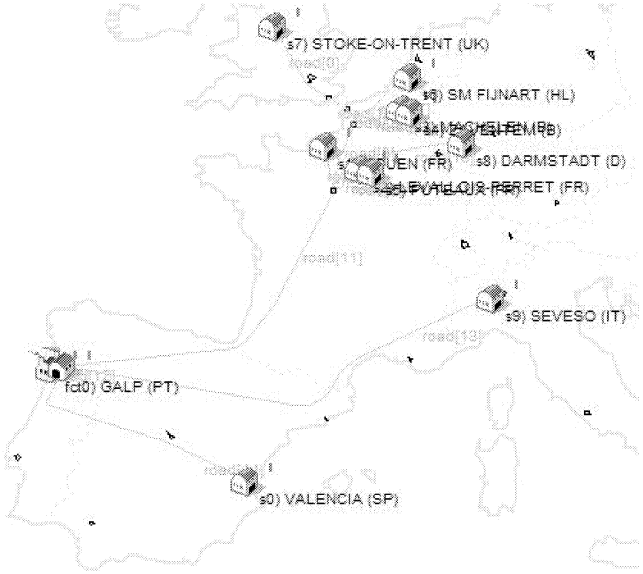


Fig. 2 The Supply Chain in study, with 10 European suppliers

The price due to the transport of one unit of product from each supplier to the plant was dependent on the supplier and on the quantity ordered, so, the primary intention was to seek for the quantity to order that would minimize theses costs at the plant's level. However, later we have opted to consider instead the costs of transportation, products, product holding and product stockouts, since total costs were considered a better financial measure to support practical decisions.

The case was also useful to help on the validation of a Supply Chain Simulator described in a previous article (Feliz-Teixeira & Brito, 2003), as well as to let us conclude about the methodology

being used and about the sort of problems to face while modelling a real system of this kind.

2. The process of building the model

The first important aspect to have in mind when preparing to simulate a real system is to dominate the process of building the model, beginning by the simulation tool. The tool used here was developed in C++ during the Ph.D. studies of Feliz-Teixeira and it was specifically oriented to Supply Chain Simulation, being therefore prepared to represent in a very easy way the elements of this kind of systems. With this tool, to build the model could be reduced to the following 5 procedures:

(1) The representation of the facilities in the "theatre of operations": the facilities have been located in the simulation window without any special geographic precision, as that aspect was irrelevant to the present study. Anyhow, each facility was placed carefully enough to maintain a relative fidelity to its real location, with the help of a map previously drawn in the simulator window (see figure 2). To help on this, *GalpEnergia* has made available a document with the post addresses of the suppliers.

(2) The definition of the transport paths connecting the facilities: these paths have also been traced without great precision in the simulator window, as in this particular case the transport costs were already defined at each supplier by a list of prices per unit transported. In those transport costs the factor distance was already included. Anyhow, the transport paths traced on the simulator try to follow the most probable lines of delivery used by the suppliers, and serve to give an idea of the delivering times as well as to introduce extra variability on the supplier's lead-times. *GalpEnergia* has made available a document with the expected lead-time for each supplier.

(3) The definition and characterization of the products available to the system: this important step let us insert into the simulator, using a dialog like the one of figure 3, the first relevant information about the 195 different addictives with which the plant usually works. Notice that, as it was considered the *product unit* (SKU) for all the products the standard barrel of 200 Kg, the maximum number of SKUs per truck (of 24000Kg) was assumed to be 120.

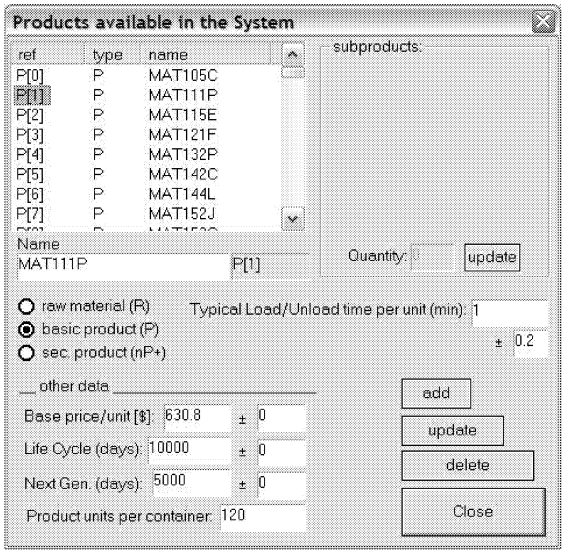


Fig. 3 Dialog Box to insert the products into the simulator

In each product one also included its reference-name and its base price (at the supplier's level), and this list was several times carefully inspected to ensure no mistakes were inserted in the simulation from the base. This data was primarily provided by *GalpEnergia* in the form of an EXCEL document, and directly retrieved from the plant's data base.

(4) The configuration of each facility, including the transport policy and the material's reorder model: once the system was geographically "represented" on the simulator and the available products inserted, the general configuration of each facility had to be established, including the definition of its transport cost policy and the material's reorder model.

In the first case, the suppliers have been configured to use *cost echelons of transport*, which were provided by *GalpEnergia* in the form of costs per Kg based on different ranges of weight, which then have been transformed to express the costs referred to the percentage of occupation on a truck of 24000Kg, considering each SKU a barrel of 200Kg. This data was assigned to the correspondent facility by means of a 5 echelon attribute filled in by the dialog shown in figure 4. The final transport cost of a certain order could therefore be automatically computed as long as its percentage of occupation in the truck was known. More than 100% occupation was made to incur in 10% extra costs, as many real suppliers establish.

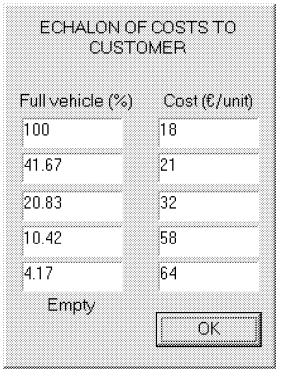


Fig. 4 Dialog Box to define the transport cost echelons

On the other hand, the definition of the material's reorder model and other information related with this aspect was inserted at each facility by means of a dialog like the one shown in figure 5. Each product was chosen to use the reorder model (r, Q), and then it was defined the amount of *safety-stock* and the *quantity to order*, as well as the *setup-cost*, the *Time for Full Satisfaction (TSF)*[1] and, finally, the *supplier*. This was done manually to the 195 products in the list, even if only 108 could be linked to a valid supplier and finally simulated. The suppliers were considered being supplied by perfect instantaneous delivering suppliers.

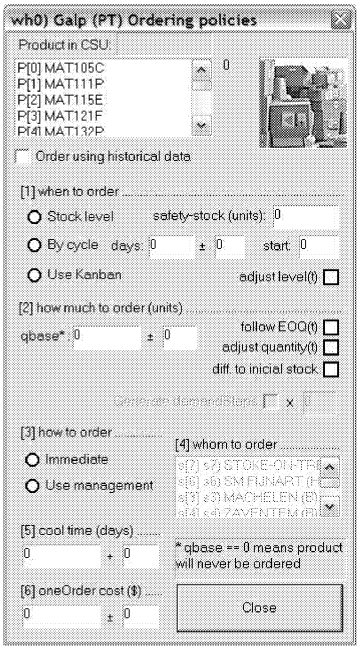


Fig. 5 Reordering Policy configuration

All the facilities were made to operate with an 11.5%/year *bank lending rate*, and null business *margin*, meaning that the purchase cost was only affected by the basic product cost and transport cost.

[1] This simulator uses this measure to retrieve a quantitative idea of clients "satisfaction", applying it not only to last customers but also to any facility in the supply chain. Once established the *Time for Full Satisfaction (TFS)*, any order arriving from the supplier within this delay will be considered 100% satisfaction, otherwise the satisfaction decreases proportionally to the extra time spent…

(5) The characterization of the demand at the plant: the demand information was obtained from documents of the real plant consume during the year of 2003. This data was a monthly historical record. Then, based on such data, two main approaches have been chosen to simulate the case:

1) *Inserting directly the historical demand into the simulator.*
2) *Using a* Normal *statistical approach of this historical data as the demand to insert in the simulator.*

Obviously, in the first case the demand had to be considered monthly, as it is in reality, but, in the second case, by means of the statistical approach, we also could "hint" the consume as if it would be sampled two in two weeks or even weekly. This let us compare results using different frequencies to sample the real demand. However, before inserting this information into the simulator it was needed to transform the 195 lines of original historical data, which were in Kg, into a monthly consume of barrels of 200Kg. Of course, this implied a rearrangement on de monthly demand pattern, as the old demand had to be integrated in time till reaching the amount of a new barrel. This work has been done with the help of a little C program developed to compute such demand transformations for all the 195 independent lines of historical data. Finally, as a result from this approach, we could create the following 4 basic scenarios to be simulated: *hist(30)*, which was driven by the original monthly demand, *stat(30)*, driven by the *Normal* statistic version of the monthly demand, *stat(15)*, considering the statistical demand sampled fortnightly, and *stat(7)* representing the statistical demand sampled weekly. These basic scenarios could later be compared by simulation, each one tested under the 5 different *transport cost echelons* previously considered.

3. Model validation and corrective factors

The simulation process has been planned to help us on reaching two main objectives: first, the confirmation of the validity of the results obtained with the simulator, at least for the present case; and, second, the studying of different approaches to purchase the products, considering the 5 different *transport cost echelons* with which the suppliers operated. This section relates with the first process, that is, the model validation.

The process of validation was restricted to the fact we only had access to documents about the monthly *plant's consume*, *average cost of each product* and *actual inventory level* at the plant's warehouse. From these values we could calculate the *average plant consume* as well as the *accumulated plant consume*, the *accumulated costs of products* and the *average inventory level at the plant's warehouse*, regarding the activity on the year of 2003. These were the 4 realistic parameters with which we later could compare the simulation output in order to validate the model. The following "real" data have been retrieved from documentation based on the 108 products that definitely could be simulated (the others had no assigned a valid supplier inside the network in study):

- Average[2] inventory in 2003 (barrels) = 2484.8
- Average plant consume in 2003 (barrels) = 477.7
- Accumulated plant consume in 2003 (barrels) = 5731.8
- Accumulated costs of products in 2003 (K€) = 2835.3

With the intention to compare this data (*Rj*) with the simulator output data (*Sj*), each of the 4 basic scenarios previously referred was made to run for one year of simulated operations, considering null the transport costs. In each case the results were obtained in the form represented in the next two figures (Figs. 6 and 7), where from data could then be obtained to fill table 1. Notice that the *accumulated cost of products* was considered the *purchasing costs* plus the *stockout costs* (as the unit stockout cost was simply made equivalent to the product cost in this model).

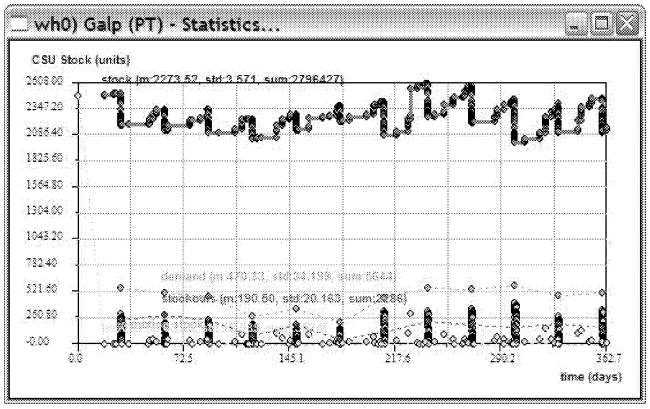


Fig. 6 First view on the inventory level and demand achieved by simulating the scenario *hist(30)* for one year of operations

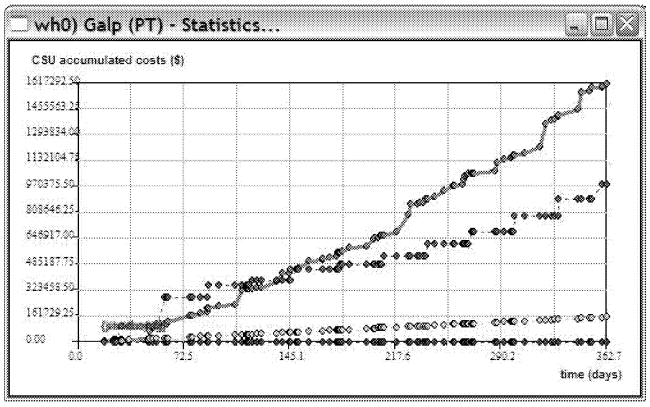


Fig. 7 Accumulated costs obtained by simulating the scenario *hist(30)*, where purchase, stockout and holding costs are represented

	hist(30)	*stat(30)*	*stat(15)*	*stat(7)*
avrg inventory level (barrel)	2273.5	2069.9	2067.1	2096.1
avrg consume (barrel)	470.3	495.2	184.8	70.1
acc consume (barrel)	5644	6437	5358	4607
acc costs of products (K€)	2771.1	2984.9	2491.6	2089.8

Table 1 First results achieved by simulation, for comparing with real data

From the data of this table (*Sj*) and the real data (*Rj*) another table was built (table 2) in which the percent deviations in relation to the real values could be estimated, applying the formula:

$$dev\% = 100\ (Rj - Sj)/Rj \qquad (1)$$

[2] Not really averaged, but instead estimated as the stock level in June 2004

	hist(30)	stat(30)	stat(15)	stat(7)
avrg[3] inventory level (%)	-8.5	-16.7	-16.8	-15.6
avrg consume (%)	-1.5	3.7	-61.3*	-85.3*
acc consume (%)	**-1.5**	**12.3**	**-6.5**	**-19.6**
acc costs of products (%)	**-2.3**	**5.3**	**-12.1**	**-26.3**

Table 2 Estimated deviations between simulated data and real data

Based on these results we could already make some idea on what to expect from future deviations in each scenario. Actually, from this we could conclude to expect a deviation from the real values of less than 3% in the *hist(30)* scenario (what can already provide a validation for the model), as well as expect that the *stat(30)* scenario will inflate the results, while the *stat(15)* and *stat(7)* will deflate them. Consequently, a corrective factor must be applied to each of these scenarios in order that in the end their results can be compared in terms of the same scale of measures. Thinking on *Kj* as the factor by which each parameter must be multiplied to approach its "real" value, we simply have decided to calculate it from the expression:

$$Kj = \text{Average}(Rj/Sj) \qquad (2)$$

Where the *Average()* function was applied only to the accumulated values, since they are considered to be more reliable for validation than simple values, as they act like the *summation of control* used in spreadsheets. By means of such calculations, the following corrective factors have been found:

hist(30)	stat(30)	stat(15)	stat(7)
K_1	K_2	K_3	K_4
1.02	**0.92**	**1.10**	**1.30**

Table 3 Corrective factors for the different scenarios

Which later have been applied to the preliminary results.

4. The simulation's strategy

As the main objective of this study was to find the best purchasing approach based on the *transport echelons* established by suppliers, each scenario have been simulated according to seven different product occupation levels in a truck of 24000Kg. Each level was indirectly establishing the *minimal order quantity* to purchase to the supplier in each particular *order*, as each product quantity on that *order* was increased with steps of 1 unit (SKU) till the entire *order* reached the level's "mark". Notice that the same order can contain several products. Thus, these occupation levels were in fact forcing the *minimal truck occupation* to a certain predefined value. Anyhow, in the same simulation the predefined level was maintained the same for all the suppliers in the network, and not different levels to different suppliers were considered. The following truck occupation levels were tested in the study:

MinOccupLevel = 0%, 4%, 8%, 16%, 32%, 64% and *100%.*

Once the 0% level was chosen, the scenario was made to run during 365 days of simulated operations, along five replications, before data was collected. Then, the new level was chosen and the same process was repeated till the last level (full truck) was simulated. This was done for each of the scenarios in study. Results like those shown in figure 8 were obtained with the simulation of each scenario.

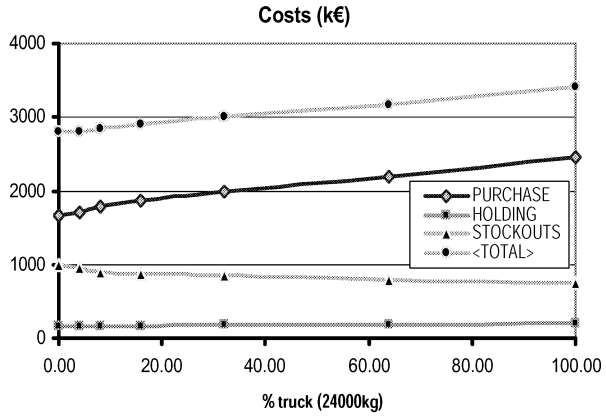


Fig. 8 Variation of costs with different truck minimal occupation levels, obtained by simulating the scenario *hist(30)*

From this graph, the *best occupation level* was chosen as the one which minimized the total costs, which in this case corresponds to 0%, meaning the best option is to purchase exactly what the inventory model requests without worrying about any transportation effects.

Along with these results, also the *average inventory* level was calculated and plotted graphically, as the example of figure 9 shows for the case of scenario *hist(30)*.

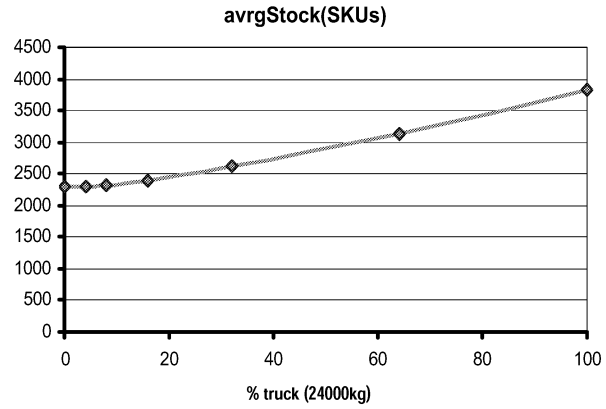


Fig. 9 Variation of average inventory level at the plant's warehouse with different truck minimal occupation levels, obtained with scenario *hist(30)*

And, finally, also the average values and variations observed during simulation on the plant's warehouse performance were computed, in terms of some standard logistic metrics (fig. 10), in this case mainly focused on the *variable costs*, *turnover*[4], *service*

[3] Only used as a qualitative measure

[*] Notice that increasing the demand sampling rate implies a decreasing on the average demand. These values have no special interest for the model validation.

[4] The *turnover* was computed here like in a single product inventory, thus there will be a quite difference from the value one expects from practice

level and *stockout ratio*. This way one could record the behaviour of each measure along the entire spectre of purchasing policies. For example, from this panel one can expect to observe, along these seven different policies, an average value for *service level* of 66.2% with a variation extending from 50.8% to 76.8%.

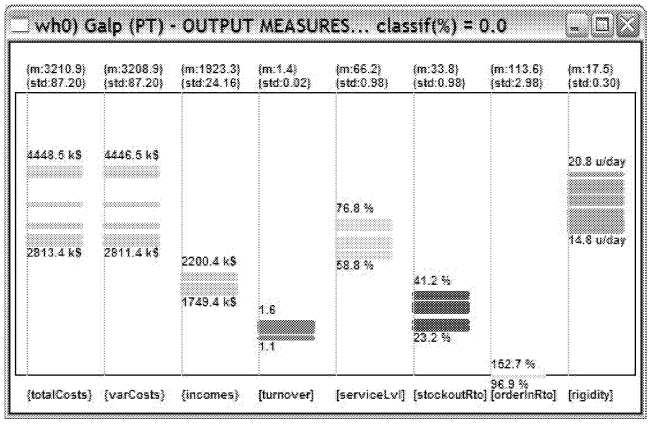


Fig. 10 Plant's warehouse performance obtained with scenario *hist(30)* along the entire spectre of purchasing policies (total runs = 30)

The *best results* obtained with each scenario were then chosen, and this data organized per scenario and joined together in a new table. Finally, the respective corrective factors (*Kj*) were applied to this table and the *final results* obtained, which will be presented in the next section.

Before passing to the next section, however, it is important to make noticed that also the following conditions were considered in the models:

- *BusinessMargin* = 0 at all the facilities, meaning the costs of the products were precisely the same as provided by *GalpEnergia*.
- *No Backorders*.
- *PurchasingCosts* = *ProductCosts* + *TransportCosts*.
- *BankLendingRate* = 11.5 %/year.
- Transportation always available at the suppliers, no need to wait for the arrival of a truck.
- *OrderSetupCosts* = 0, because they were in practice almost null in comparison with the products costs.
- *TimeForFullSatisfaction (TFS)* at the plant was considered 3 ± 0.5 days (indirectly rules the *stockout ratio*)
- *TimeForFullSatisfaction (TFS)* at the plant's warehouse was considered 30 ± 3 days.
- Reorder model *(r=ro, q = Q(t))*, with *ro* the actual safety stock and *Q(t)* to follow the maximum between *average demand* and the *accumulated demand between supplies*.

5. Final results and comments

The four basic scenarios were operating with the inventory model *(r=ro, Q(t))*, as previously referred. Anyhow, to let us test this model with other values, after each of these scenarios has been simulated the *average demand* during the 365 days of operations was computed, as well as the *accumulated demand between supplies*. Then, the original scenario was copied into a new scenario on which a change would be made in the inventory model: *ro* was replaced by this *average demand* and *Q* was substituted by the *accumulated demand between supplies*. Then, the scenario was also simulated. For this reason, the final simulated scenarios became the following eight:

1) *hist(30)* – original historical data, model (r=ro, Q(t)).
2) *hist(30)_ss_qq* – historical data, model (r=avrgDem, Q=accDemSuppl)
3) *stat(30)* – statistical data, model (r=ro, Q(t)).
4) *stat(30)_ss_qq* – statistical data, model (r=avrgDem, Q=accDemSuppl)
5) *stat(15)* – statistical data, model (r=ro, Q(t)).
6) *stat(15)_ss_qq* – statistical data, model (r=avrgDem, Q=accDemSuppl)
7) *stat(7)* – statistical data, model (r=ro, Q(t)).
8) *stat(7)_ss_qq* – statistical data, model (r=avrgDem, Q=accDemSuppl)

And the graph presented in figure 11 shows the results before correction obtained precisely with these scenarios, numbered from 1 to 8, considering their *best occupation level*.

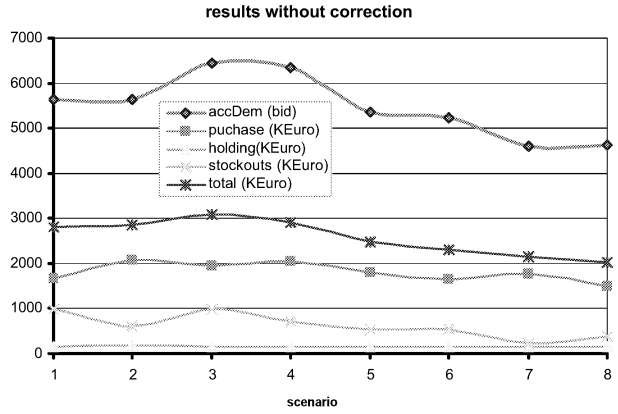


Fig. 11 Results before applying the corrective factors

From this figure the behaviour predicted earlier is now evident. As expected, *stat(30)* scenario inflates the results, while *stat(15)* and *stat(7)* scenarios deflate them. Much of this is a consequence from the slightly differences at the demand generated pattern that accumulate along the 365 days of operations, as also can be noticed from the figure. The upper line represents, in fact, this *accumulated demand*, which ideally should be a horizontal line. It is thus obvious that no conclusions could be retrieved from this data before applying the corrective factors. Doing so, we finally obtain the following costs figure:

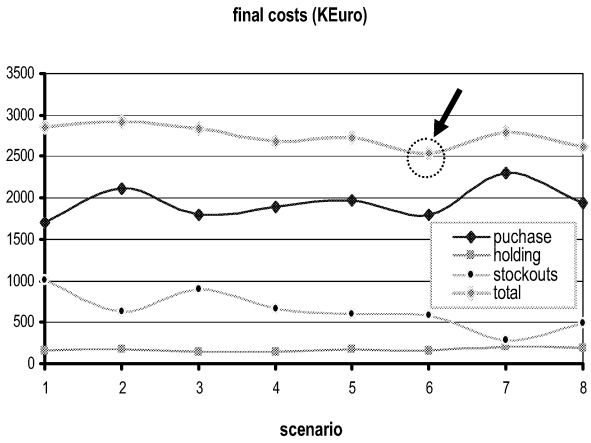


Fig. 12 Final results, concerning accumulated cost during a year

Notice that if this correction would not be applied we could have been induced in mistake by choosing the *stat(7)* scenario as the most interesting option on the preliminary results. In fact, the most interesting is scenario 6, known as *stat(15)_ss_qq*, since it obviously leaded to the minimal total costs accumulated along the

year. Compared with the actual situation, this would mean a save of around 300 K€, measured by the figure. As it was also observed that the *best truck occupation level* for this scenario was 8%, one can finally conclude:

"*The best approach to operate the plant's warehouse would be to ensure a fortnightly inspection of its inventory and, once operating with the inventory model (ro, Qo), to set ro as the average demand (at each product) and Qo the respective expected accumulated demand between supplies, not forgetting to order always at least 8% of a truck. The total expected saves would be around 300 K€ in one year of operation*".

IMPORTANT REMARK: One must not forget that these results have been obtained considering only 108 of the 195 products previously inserted in the simulator, which would represented a more realistic view on the plant's demand.

6. CONCLUSIONS

Any conclusions must count with this last remark. Only those products which could be linked to the list of suppliers considered in these models have been simulated. Consequently, the results presented here will obviously express this fact, meaning that the conditions of their validity do not reflect the overall plant's demand and, probably, they will not fit perfectly with those obtained in practice. Anyhow, they seem consistent and quite interesting, as they confirm important values previously calculated from real data, and seem to point to a higher frequency on inventory revision with the objective of reducing total costs, a trend of nowadays on trying to minimize forecast inaccuracies. Actually, it is in general more difficult to forecast for a long period of time that for a shorter one, at least when the demand is "well behaved" like in the present case.

What concerns the original objective of analyzing the costs variations with the different *echelons of transport*, it was observed that in all the scenarios except the *stat(15)_ss_qq* the best approach was the 0% *minimal truck occupation*, meaning it was to purchase exactly what the inventory model would request without worrying with any transportation effects. For that reason this problem was naturally absorbed by the problem of sampling the demand with different rates, which during the simulation was getting more relevancy than the former issue. This was due to the fact that in most of the cases the original truck occupation was already naturally high, as figure 13 shows for the case of *hist(30)* scenario, with an average value between 30 and 60 %. Lower levels of occupation were thus automatically "ignored" by the simulation, turning this issue less relevant.

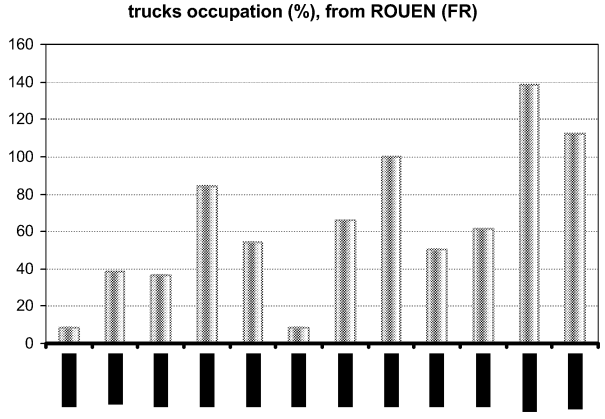


Fig. 13 Original truck occupation in the *hist(30)* scenario

In respect to the methodology, it appeared to be firm and good, and leading to an interesting way not only of validating the model but also of making possible the comparison of different scenarios by means of applying some *corrective factors*. Also the verification of certain aspects of the simulator could be done, because with this study some "bugs" were detected in its software and later repaired, mainly in the procedures for computing the *holding* and the *purchasing* costs.

The most important issue, however, in what concerns a more serious study of this plant, is related with the fact that not all the important products on the real demand could be simulated and that for sure implied a deviation on the results from real data. As shown in figure 14, it was observed that some of the products were extremely expensive and other relatively cheep, but not all of those expensive ones could be simulated in this study. From this figure one can also get an idea of the probable deviation that could have been committed in this approach because of this problem.

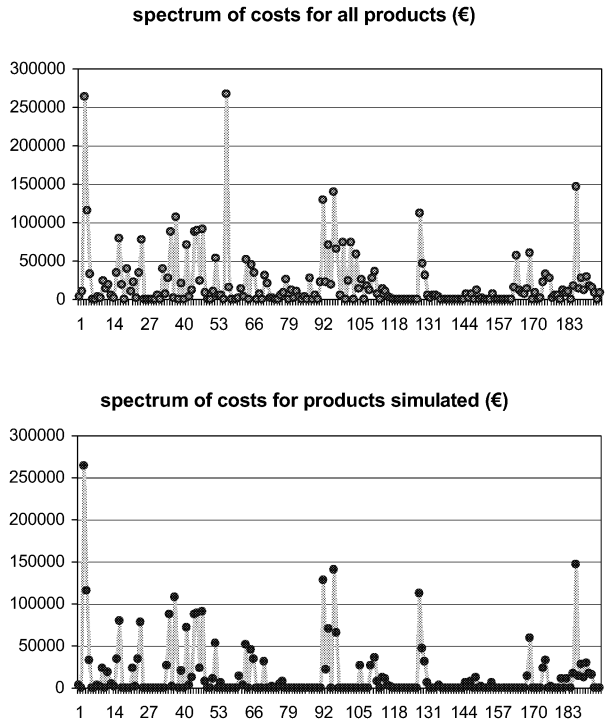


Fig. 14 Comparison of costs involving all the 195 products in the plant and the 108 simulated products

It seems obvious that for future studies first we will have to ensure that the most economical significant products will be represented and turned able to be simulated. In order to improve the simulation of this interesting case, we would in the future need to focus the attention in the following main aspects:

- Ensure the simulation of all A type and B type products.
- Improve the statistical demand generator, maybe.
- Count with more costs: stocking, order setup, etc.

We hope that a more accurate study on this case can be completed in the future, perhaps following the careful analysis of these results that at the present is being made at *GalpEnergia*,

References:

Feliz-Teixeira, J. M., & Brito, A. E. S. C. (2003). *An Approach for Dynamic Supply Chain Modelling.* Paper presented at the 2003 European Simulation and Modelling Conference, Naples, Italy.

GalpEnergia. (2003). *A Galp Energia - História*, 2004, from http://www.galpenergia.com/

Other:

Bause, F., & Beilner, H. (1999). *Intrinsic Problems in Simulation of Logistic Networks.* Paper presented at the 11th European Simulation Symposium, Erlangen, Germany.

Bonney, M. C. (1994). Trends in inventory management. *Journal of Production Economics, 35*(1-3).

Feliz-Teixeira, J. M., & Brito, A. E. S. C. (2001). *Simulação por Computador*: PUBLINDÚSTRIA.

Christopher, M. (1998). *Logistics and Supply Chain Management* (second ed.): Prentice-Hall.

Gonçalves, J. F. (1999). *Gestão de Aprovisionamentos* (revista ed.): PUBLINDÚSTRIA.

Guedes, A. P. (1994). *An Integrated Approach to Logistics Strategy Planning Usig Visual Interactive Modelling and Decision Support.* Unpublished Ph. D., University of Cranfield, U. K.

Heizer, J., & Render, B. (2001). *Operations Management*: Prentice Hall.

Hopp, W. J., & Spearman, M. L. (2001). *Factory Physics*: Irwin McGraw-Hill.

Lambert, D. M., & Pohlen, T. L., "*Supply Chain Metrics*", The International Journal of Logistics Management, 2001, 12(1), pp. 1-19.

MODELLING OF POROUS FLOW IN UNDERGROUND SYSTEMS

Naghmeh Keshavarzi-Roonizi
Rahim Ghorashi
Vahid Nassehi
Department of Chemical Engineering
Loughborough University
Loughborough, LE11 3TU, UK
Email: n.keshavarzi-roonizi@lboro.ac.uk

Diganta Bhusan Das
Department of Engineering Science
University of Oxford
Oxford OX1 3PJ, UK

KEYWORDS
Porous flow, Combined flow, Heterogeneous porous media, Finite volume method, Computer implementation

ABSTRACT

In this paper we discuss the development of a three-dimensional computer model for the quantitative analysis of underground flow. This is a generic model that can be applied to simulate a wide variety of environmental and industrially important problems. Although underground flows are mainly characterised by flow through porous media, conjunctive free/porous regimes are also frequently encountered in sub-surface regions. In this paper we discuss the extension of our model to such combined flow systems and provide an example of the application of the developed software to a case study. We further discuss cases such as flow through unsaturated domains where the advancing flow front needs to be tracked. An effective method for tracking of the flow front is therefore presented. As can be expected the resulting software is a sophisticated program whose effective application requires many hours of training and experience. In order to make the software accessible for the general user, who may only have limited experience in mathematical modelling, we describe the incorporation of the present model into a previously developed computer network system. This system has been structured to include an intelligent front-end for complicated hydro-environmental flow modelling and provides on-line assistance for non-expert users at every stage of an application.

INTRODUCTION

Mathematical analysis of underground flow of viscous fluids has been the subject of extensive research in the past decades. This has resulted in the formulation of a range of equations which describe mass and momentum balance for these regimes. The main criterion for deciding which one of available mathematical models is best suited for a particular situation is the permeability of the porous region in which the flow occurs. In cases where the porous domain consists of densely packed particles the small pores between solid sections allow a fluid to seep through whilst the solid matrix carries all of the stresses in the system. The most appropriate equation of motion under these circumstances is based on the well known Darcy model which relates pressure drop for flow through a porous region to the flow rate (flux) and physical parameters such as domain permeability and fluid viscosity. In contrast, if the permeability of a porous region is high enough to let the porous regime to be, at least, partially dominated by viscous forces the Brinkman model which includes viscous stresses in the momentum balance for the fluid flow is used. It can be said that in the Brinkman model the fluid/solid system is effectively regarded as a pseudo-fluid characterised by an assumed 'effective viscosity'. In addition to the described models, for intermediate situations a combination of these two models have also been used (Bear, 1972; Nield and Bejan, 1992).

By its nature the Darcy model represents creeping (i.e. near zero Reynolds number) flow and cannot take into account inertia effects and no-slip wall conditions. On the other hand the determination of effective viscosity in the Brinkman model is not straightforward and can cause significant uncertainty in the accuracy of the model predictions.

Both models have been used to model combined free/porous flows. Development of effective combined models is usually complicated and specific techniques need to be employed (Das and Nassehi, 2003, Nassehi, 1998). Examples of combined free/porous flow regimes are found in groundwater hydrology, glaciology, conjunctive surface and subsurface flow, mud cake formation during drilling of oil wells, and others (e.g. see, Cieszko and Kubik, 1999 and, Das and Nassehi, 2003). Free flow in large openings under the ground can be laminar or turbulent and a judgment based on conditions of any given problem must be made. However, in this work we have only considered situations where the free regime remains laminar and modelled it using the well known Navier-Stokes equations.

In the present work, a previously developed technique for combining free and porous regimes along arbitrary shape heterogeneous interfaces (Das, 2001; Das, 2002; Das et al., 2002; Das and Nassehi, 2003) has been extended to model industrially important underground flow systems. To maintain the generality of the analysis, the domain of interest is defined to have unbounded/open/flow-through ends.

The propagation of a flow front under the ground, which is a characteristic of unsaturated flow has also been included in the present model. Here a method based on a modified version of the Volume of Fluid (VOF) technique is used (Nassehi, 2002).

Computer implementation of the entire scheme is via a user-friendly network for multi-user applications. The present paper includes a brief description of this system.

MATHEMATICAL MODEL

The continuity equation representing conservation of matter in all types of flow regimes simulated by the present model is based on the following mass balance equation for incompressible fluids

$$\nabla \cdot \mathbf{V} = 0 \qquad (1).$$

Where $\mathbf{V}$ is the velocity vector. However, depending on the nature of the flow, i.e. whether it is free or porous and permeability is lower than a given threshold or not a suitable equation of motion representing the conservation of momentum should be used. We consider the following general form

$$A(\mathbf{V}) + \nabla p = 0 \qquad (2).$$

Where p is the pressure. For $A(\mathbf{V}) = \rho \dfrac{\partial \mathbf{V}}{\partial t} + \mathbf{I} \cdot \mathbf{V}$, where ρ is fluid density and $\mathbf{I}$ is the identity matrix multiplied by fluid viscosity (η) over domain permeability (k), equation (2) represents the Darcy model. For $A(\mathbf{V}) = \rho \dfrac{\partial \mathbf{V}}{\partial t} + \mathbf{I} \cdot \mathbf{V} - \mu_e \nabla^2 \mathbf{V}$, where μ_e is the 'effective' viscosity, equation (2) provides the momentum balance relation corresponding to the Brinkman model. In these models when the porous domain is considered to be heterogeneous viscosity/permeability ratio should be written as a row vector with η / k_{xx} and η / k_{yy} as its components and multiplied by the identity matrix.

Finally for $A(\mathbf{V}) = \rho \dfrac{\partial \mathbf{V}}{\partial t} + \rho \mathbf{V} \nabla \cdot \mathbf{V} - \nabla \eta \nabla \cdot \mathbf{V}$, where η is viscosity of a generalised Newtonian fluid, equation (2) yields the Navier-Stokes equation for laminar free flow in the absence of body forces. For a purely Newtonian fluid, where viscosity is constant μ , the stress term in the operator expression for the free flow can simply be written as $-\mu \nabla^2 \mathbf{V}$. In modelling steady state systems the temporal derivative term in the above definitions vanish.

Combining free flow and Brinkman models does not present any particular mathematical problems and these regimes can simply be simulated conjunctively provided that effective viscosity in the porous region is matched with the fluid viscosity at the interface. However, lack of a second order velocity derivative in the Darcy equation precludes straightforward linking of free and porous equations for this case. A very well established technique for such linking is the use of interface conditions suggested by Beavers and Joseph (1967). Details of the use of this technique for underground flow systems has been reported by Das and Nassehi (2003). Momentum transfer between the porous and free flow domains in this method is manipulated by assuming the existence of a slip layer at the interface separating the regions. The method has also been extended to combined flows in heterogeneous underground regions (Das *et a*l, 2003). We have used this method in the present work.

The advancing flow front is simulated using the VOF (volume of fluid) technique. This method is based on the solution of the front position probability density equation, given as

$$\frac{\partial F}{\partial t} + \mathbf{V}.\nabla F = 0 \qquad (3).$$

Where $0 \le F \le 1$ is called the flow front position function. In this work we have used a modified version of the VOF technique in which the moving front flow is considered as a two-phase regime in which the wet and the dry sections represent the phases (Thompson, 1986). The flow model is solved for the entire domain whilst at each section physical properties relevant to that phase are inserted. A set of fictitious physical values that render the velocity in the dry regions as zero are used and hence the technique is called the pseudo-density method. Simultaneous solution of equation (3) with the flow model generates values of F. A value of F between 0 and 1 (usually $F = 0.5$) is taken as the boundary between phases representing the moving front. Values of physical parameters for each phase in the flow field is related to the position of the front using the following equation

$$y = y_w F + y_d (1 - F) \qquad (4).$$

Where y is a given physical parameter and y_f and y_a are the values of this parameter in the wet and dry regions, respectively, (Nassehi and Ghoreishy, 1997).

The described governing equations are discretised and reduced to algebraic forms using the standard finite volume method (FVM) (Patankar, 1980; Versteeg and Malalasekera, 1995). This method is adopted due to its inherent capability in coping with complex 3-D domain types.

The scheme calculates the flow variables on a 3-D forward-staggered grid where cell faces are ahead of cell nodes, figure1. The computational grid for both free and porous regions involves cubicle cells for storing the flow variables. While the velocity components are evaluated at different

cell faces, pressure terms are predicted at cell nodes such that they act as the driving potential for flow at intermediate cell faces. Detailed descriptions of the mesh and schemes used can be found in Das and Nassehi (2001).

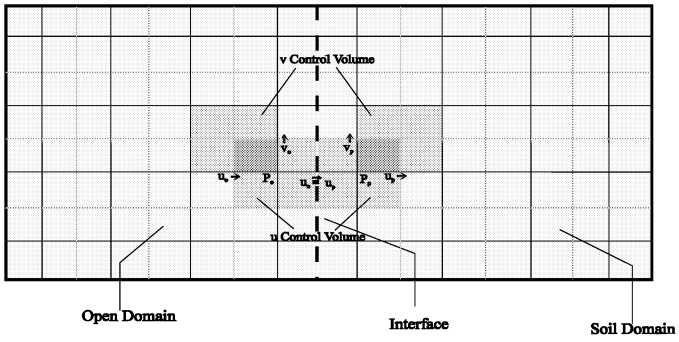


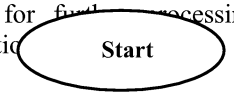
Figure 1- Computational cell used in the finite volume scheme

COMPUTER IMPLEMENTATION

To achieve the goals of flexibility, practicality, speed, multi-user capability and computing economy the use of the described model is based on its implementation via an Information Processing Tool. The detail of the basic system has been previously published (N.Keshavarzi-Roonizi *et al*, 2004). In this work we have extended this system and developed a new architecture for in order to make it suitable for porous and combined free/porous problems. The IT tool consists of three modules.

- Front-end, provides a network node for communication between the users and the software. Quantitative results generated by the model are treated in a back-end module and the final result is provided in clearly tabulated or graphical formats. A log is kept by the system which records the particulars of the application for future comparisons and evaluation of the efficiency of the solution generated for a given problem. The front end will guide the user towards inserting appropriate input data. The system compares given permeability of the porous domain and selects the most suitable mathematical model to carry out computations. The front end facilitates data handling whilst monitoring the suitability and relevance of the data inserted by the users to activate the data generator (simulation) module.

- Data-generator is a systematic library of various modelling software written in FORTRAN and symbolic programming as well as genetic algorithms to solve the model equations. The front-end interface feeds the input data to the modelling sections initiating numerical computations. At the end of the computations the simulation results are passed to the back-end module for final processing required before presentation



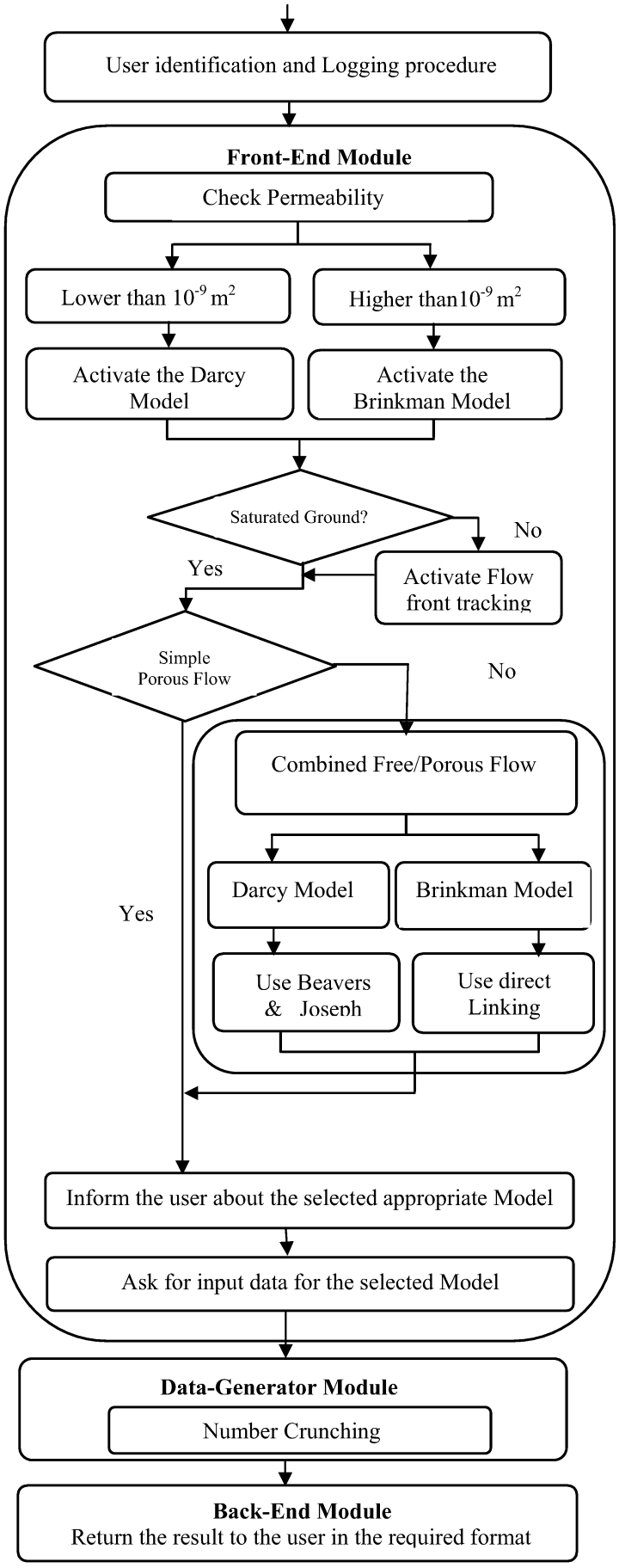


Figure 2- Schematic diagram of the IT system

- Back-end, provides a network node for the import and processing of the data (i.e. results, answers, evaluations) generated by the data generator module. The processed results are logged and returned to a window in the front-end for the utilization by the users. It should be noted that the simulations results obtained via the data-generator mainly consist of large amounts of numerical data and hence the back–end module re-organizes them as pre-formatted tables or graphical out put. The back-end module can also communicate directly with the front-end at the preliminary stages of an application (i.e. problem simulation). For example, it is designed to provide answers regarding the validity of a given input on the basis of its self-consistency or logical status. This feature is mainly utilized to guide the untrained user to obtain required information from the system with ease and efficiency. In figure 2 the general architecture and a basic procedure of the present IT system is shown.

SAMPLE RESULT

To demonstrate the application of the present model we have considered simulation of a combined free/porous flow regime through a domain as shown in figure 3.

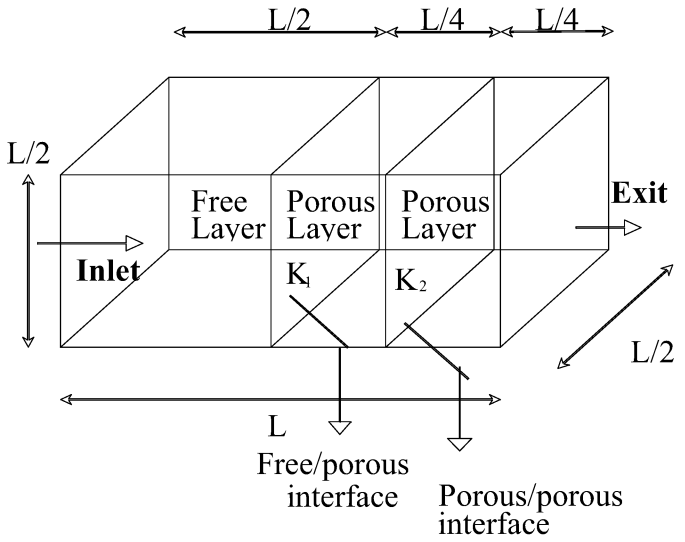


Figure 3- Representative 3-D configurations for free flow and layered porous medium. The direction of the layer is perpendicular to the longitudinal (x-) velocity component

In order to clearly show the simulated velocity field in the following figures only the velocity vectors on the interfaces between free and porous and two porous layers of different permeability are shown. The permeability of both porous regions is taken to be lower than selected threshold and hence porous flow in these simulations is modelled by the Darcy equation. Therefore slip flow at the interface between the free and porous regions are used. However, porous/porous interface is handled normally and only different values of the permeability for these regions have been used. The front-end in the present system guides the user to insert appropriate input at every stage.

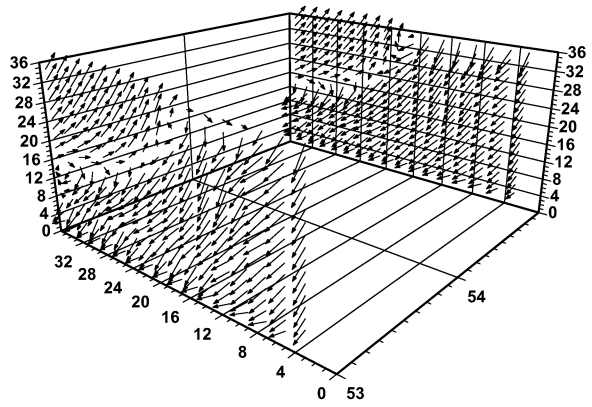


Figure 4- Simulated Velocity Field at Porous/Porous Interface, 19th Time Level
($K_1 = 5\times10^{-9} m^2, K_2 = 15\times10^{-11} m^2$)

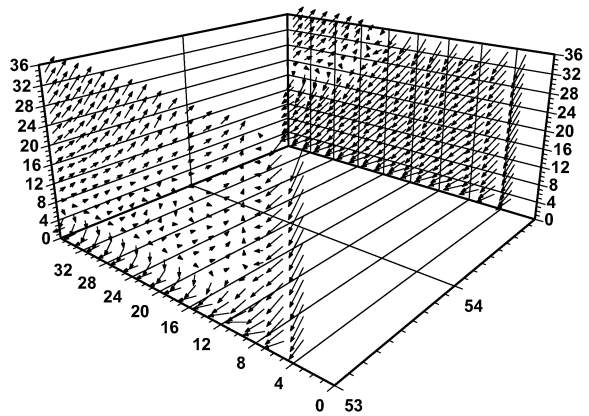


Figure 5- Simulated Velocity Field at Porous/Porous Interface, 19th Time Level,
($K_1 = 15\times10^{-11} m^2, K_2 = 5\times10^{-9} m^2$)

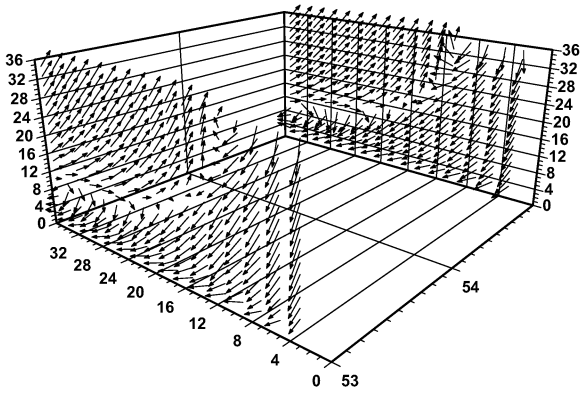


Figure 6- Simulated Velocity Field at Porous/Porous Interface, 37th Time Level,
$(K_1 = 5\times10^{-9}\,m^2, K_2 = 15\times10^{-11}\,m^2)$

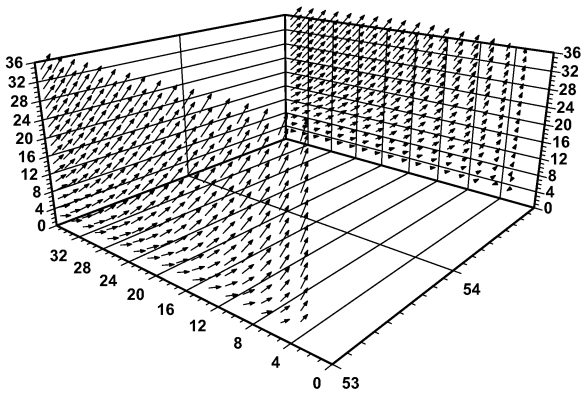


Figure 7- Simulated Velocity Field at Porous/Porous Interface, 37th Time Level,
$(K_1 = 15\times10^{-11}\,m^2, K_2 = 5\times10^{-9}\,m^2)$

As the sample results show the flow pattern at the interface is quite complicated and complex interactions between the regions can be observed. This type of non-uniform flow is expected to occur under the ground as shown by information related to the study of leachates appearing from under ground regions.

ACKNOWLEDGMENT

The authors wish to thank Engineering and Physical Research Council of UK (EPSRC) and Loughborough University for PhD studentships which enabled the development of the model and information processing system described in this work.

REFERENCES

Bear.J. 1972. *Dynamics of Fluids in Porous Media.* American Elsevier Publications, Inc., Newyork, Reprinted (1980): Dover Publications Inc., New York.

Beavers GS, Joseph DD. 1967. "Boundary conditions at naturally permeable wall". *Journal of Fluid Mechanics* No.30(1): 197–207.

Cieszko M, Kubik J. 1999. "Derivation of matching conditions at the contact surface between fluid-saturated porous solid and bulk fluid". *Transport in Porous Media* No.34(1–3): 319–336.

Das. D. B, Nassehi V. 2003. " Influence of multiple interfacial boundaries on contaminants' mobility in subsurface regions". *Water Resources Research No.*39(3): 1072. DOI: 10Ð1029/2002WR001506.

Das D. B., Nassehi V., Wakeman R.J., 2002 "Advances in Environmental Research". 7 (1), 35-58.

Das D. B. 2002. "Hydrodynamic modelling for goundwater flow through permeable reactive barriers". *Hydrological Processe. No.*16: 3393–3418

Das. D.B. Hanspal N.S. and Nassehi V. 2003. "A Numerical Study of the hydrodynamic Cnditions in Coupled free and Heterogeneous Porous Domains" . *Proceedings of Computational Methods in Water Resources Conference*, 13th -17th June Chapel Hill, North Carolina, USA.

Das D.B. 2001. "Computer Modelling of contaminants mobility in subsurface regions", PhD thesis, Loughborough University.

Keshavarzi-Roonizi N., Ghorashi R. and Nassehi V., 2004, "An Information Processing Tool for the Management of Natural Water Systems", *Proceedings of the 2nd International Conference on Industrial Simulation, ISC 2004* , Malaga, June 7-9, pp 40-44.

Nassehi V. 2002. *Practical Aspects of Finite Element Modelling of Polymer Processing*, John Wiley & Sons, Ltd, Chichester, West Sussex.

Nassehi V., and Ghoreishy M. H. R. 1997. " Simulation of Free Surface Flow in Partially Filled Internal Mixers" *International Polymer Processing* No.4 (Dec), 346-353.

Nassehi V. 1998. "Modelling of combined Navier-Stokes and Darcy flows in crossflow membrane filtration" Chemical Engineering Science No.4(Feb), 1253-1265

Nield D. A; Bejan A.; 1992. *Convection in Porous Media.* Springer-Verlag: New York.

Patankar S.V.; 1980. Numerical heat transfer and fluid flow. Hemisphere, Washington DC.

Thompson E. 1986. "Use of pseudo-concentrations to follow creeping viscous flows during transient analysis" International Journal of Numerical Methods in fluids, No.6, 749-671.

Versteeg H.K.; and Malalasekera W.; 1995. *An introduction to Computational fluid dynamics: the finite volume method.* Addition Wesley Longman Ltd, Essex.

ARCHITECTURE FOR SIMULATION & KNOWLEDGE BASED SUPERVISION FOR PROCESS FACILITIES

Hans-Werner Schmidt and Ulrich Saffert
Elpro Prozessindustrie und Energieanlagen GmbH
Marzahner Straße 34
D-13053 Berlin, Germany
E-mail: { hans-werner.schmidt | ulrich.saffert }@elpro.de

Gerhard Schreck and Alexander Sabov
Fraunhofer-Institute for Production Systems and Design Technology
Pascalstrasse 8-9
D-10587 Berlin, Germany
E-mail: {gerhard.schreck| alexander.sabov}@ipk.fraunhofer.de

Matthias Pleßow
GFaI e.V.
Rudower Chaussee 30
D-12489 Berlin, Germany
E-mail: plessow@gfai.de

KEYWORDS
Simulation, SCADA, Supervisory Control, XML, OPC, SOAP

ABSTRACT

In this paper, a system architecture is presented for simulation and knowledge based supervisory control. A process control system is extended with components for knowledge based modeling, simulation and visualization in order to support the operator with high level management functions. A methodology is introduced which uses interface generation mechanisms and modern IT technologies like XML and SOAP to extend the fixed infrastructure of current systems with a platform independent implementation, allowing distributed systems and web-based communication. The operation mode of the central control and management component is explained in detail. The main aspects are the distinction between different use cases and synchronization mechanisms for different cycle times.

INTRODUCTION

The increasing complexity of modern process plants and the demands for energy conservation, product quality, environmental protection, safety and reliability make new approaches to process automation necessary. Beside decision-making and control procedures based on mathematical models, and solvers for multi-criteria optimization tasks, knowledge based systems involving the experience of human operators are a promising approach.

SCADA systems (supervisory control and data acquisition) are usually introduced for high level process control and automation. They play an important integration function in distributed, multilevel control environment, and provide a common view of the system and the corresponding operator panels required for process operation. Typical functionalities include activation of control actions, monitoring of process states, recording of alarms and events, emergency shutdown, etc.

The human operating team is still responsible for high level process management which includes tasks like assessing process situations, selecting operating points, considering different modes of work and use of resources, reacting to changing requirements and demands. Operator decisions are based on his knowledge of process situations, process trends, and process control. Thus, the experience gained in the past plays an important role in reaching high level performance.

The aim of the simulation and knowledge-based methods is to combine objective information from the facility, e.g. plant models and available historical operation profiles, with operation knowledge, which means experience-based evaluation and assessment through an operator team. The mapping of known process situations to approved control strategies and actions leads to case oriented supervisory control (Lisounkin et al. 2004).

In the following chapters the components of a simulation and knowledge based process supervision system, interface generation mechanisms, and integration with conventional SCADA systems are presented.

SYSTEM ARCHITECTURE

Several components must be added to the infrastructure of a process control system in order to support the operator with a knowledge based supervision system. The components can be grouped according to the phase in which they are needed, the development phase or the runtime phase. This section gives a brief introduction of each component

with a description of its purpose. The tools for knowledge acquisition, modeling and visualization are described in more detail in: (Pleßow et al. 2005), while aspects of knowledge modeling and processing are discussed in: (Krüger et al. 2005).

Development phase

The development of the supervision systems starts with modeling the topology of the facility with the **Topology Editor**. The editor contains a library of facility specific process units with predefined process parameters. Units can be dropped on a sheet and connected to form the model that is used later for visualization and simulation purposes.

The **Knowledge Editor** is used for the acquisition of operation knowledge. Knowledge is described in terms of knowledge objects, basically containing a set of rules to apply and actions to take when the rules are violated.

The **Profile Editor** defines different profiles for the desired output of the process facility. Profiles contain the variation of process variables over specific time periods.

The **Schedule Editor** allows the setup and configuration of the runtime environment. Components of the runtime environment are combined with the models and data created with the Toplogy-, Knowledge- and Profile Editor into runtime modules.

Runtime phase

Figure 1 shows the components of the runtime phase. The **Scheduler** is the central control and management station of the system, connecting the existing process control system with the supervision system. Typical tasks are the initiation, monitoring, timing and data exchange of components.

The **Topology Viewer** extends the visualization of the process control system with knowledge and simulation relevant information, e.g. intermediate results of the calculations. It reads in the simulation model created with the Topology Editor and visualizes the dynamic behaviour of the process.

The **Knowledge Viewer** visualizes the possible and applied rules of the supervision system. This allows the operator to retrace the actions and the reasons they were executed.

The **Profile Viewer** gives the operator an overview of the relevant profiles. The currently selected as well as all available operation profiles are drawn as a diagram on the screen.

The **Simulator** forecasts the process values for the supervision components. The simulation model is a sub-model of the topology model. Based on the current process values, the output values for a defined time frame are computed.

This allows the **Inference Machine** to take preventive actions based on the knowledge provided. The current process states are checked constantly for violated rules.

The **Interaction Server** allows the dynamic connection of multiple simulators and inference machines to the system architecture. Simulation and knowledge models can be split into subsystems and still appear as one system to the scheduler by encapsulating the coordination of the subsystems.

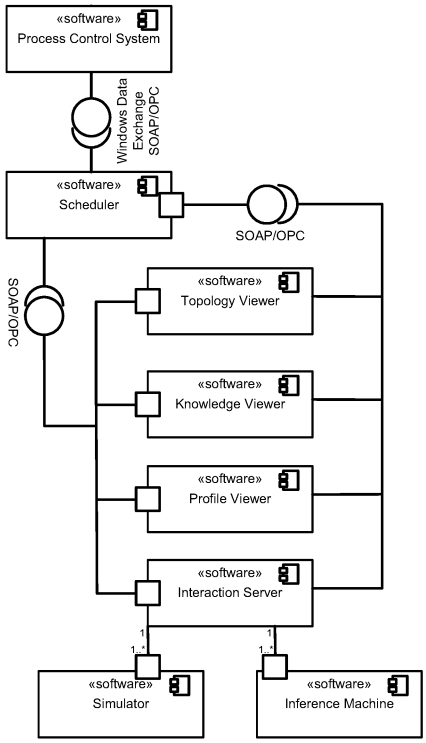


Figure 1: Runtime environment of the knowledge based supervision system

INTERFACE GENERATION

Components of the system architecture presented are developed using different tools and programming languages. Therefore the components must provide interfaces for control and data exchange in a generic way.

Requirements

The conventional approach using plugins and Dynamic Link Libraries in SCADA software does not satisfy the requirements of the system because it would limit the accessibility to a specific operating system and make function calls difficult (calling convention, data types).

Also, the underlying communication infrastructure must allow the distribution of components on different machines for accessibility and performance reasons. There are plenty of architecture/infrastructure tools for

distributed systems available (CORBA, DCOM, JAVA RMI) but most of them require high license costs or limit communication to a specific operating system.

When designing the architecture, great attention must be paid to common interfaces and protocols used for communication in process facilities. OLE for Process Control (OPC: http://www.opcfoundation.org) is a world-wide standard in industrial automation, defining protocols for data access, alarm & events, batch processes, data exchange and historical data access. The history of OPC has its origin in system dependent OLE COM and DCOM technologies, typically C++ implementations. Currently SOAP (http://www.w3.org/TR/soap) and XML based (http://www.w3.org/XML) specifications are being developed which can be used to offer components of process facilities as a Web-Service over the internet.

OPC XML Gateway

The components of this architecture are based on the current state of OPC XML specifications, extended for the start sequence of the system with proprietary initialization functions. In order to connect existing OPC based components like SCADA software to the system architecture, a gateway component was created with the following properties:

- Providing a C++ based server interface for connection with non-XML clients;
- Mapping of C++ function calls to Web Service calls;
- Providing a Web Service Client interface for connection with OPC XML-DA Server.

Figure 2 shows the component diagram for the gateway.

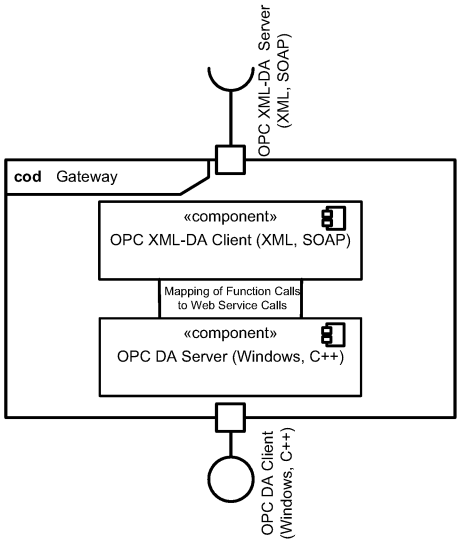


Figure 2: OPC XML Gateway

Interface generation mechanism

Providing a system specific interface involves multiple steps. The OPC interface specification together with the initialization functions are provided by a Web Service Definition Language file (WSDL http://www.w3.org/TR/wsdl), describing:

- The service endpoint
- An abstract set of operations
- An abstract definition of the data

The generation of the system specific interface can be done using code generation tools like gSoap for C++ or Axis for Java. Figure 3 shows an activity diagram for the interface generation. Source code skeletons are generated for the server whose functionalities need to be implemented. Source code stubs are generated for the client which can be used to make Remote Procedure Calls (RPC) to the server.

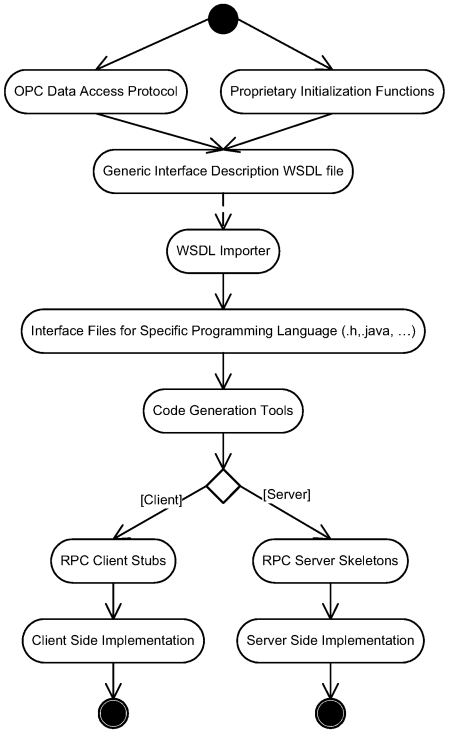


Figure 3: Automatic interface generation technique

TASK SCHEDULING

Due to the fact that the scheduler has a centralized role it is the only client in his domain view. Its major tasks are:

- Sending the appropriate models to the components at startup;
- Cyclic calls of the visualization and computation components of the runtime environment;

- Monitoring each component to assure that the system is running properly;
- Providing all components with the current data.

An assignment must be defined between data source and target for a data exchange between two components. The assignment topology varies depending on different use cases. The scheduler can distinguish between them by using "Functional Units." A Functional Unit manages different modules, whereas each module is an instance in the domain of one component (e.g., a system model or an observer with a specific view).

In most use cases, the output results of computational components are connected with the input values of the visualization components. A cyclical call of the components must happen in a defined order. For this a Sequence Module has been defined, that can be used by Functional Units. Figure 4 shows the developed architecture:

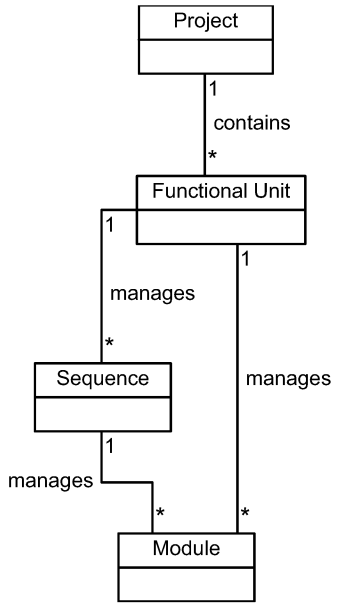


Figure 4: Class diagram of the module management architecture

Cycle times

The demands of typical process data bases require special handling of cycle times for the components. Access to process data is mostly realized by reading groups of process values for performance and consistency reasons. Therefore, the cycle times should be defined with optimized execution and idle times, but in a synchronized manner using clock pulses.

The scheduler defines the base clock pulse and performs error checking and timeout monitoring within one pulse. To assure synchronized execution times, cycle times t_c for components must be multiples k of the base time t_b:

$$t_c = t_b * \prod_{i=1}^{c} k_i, \quad c = 1,2,... \tag{1}$$

$$k_i : \quad < i, k_i > \in K,$$

$$K = \{< i, k >\}, \quad k \in N, \; i = 1,2,...$$

Figure 5 shows an example with two tasks running with different cycle times.

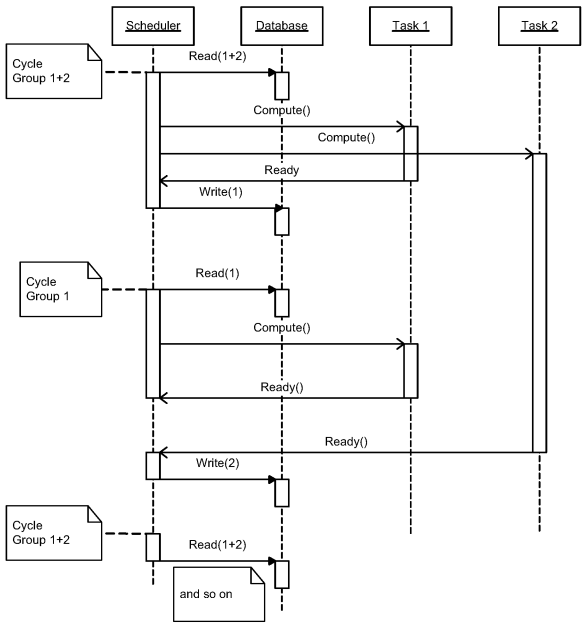


Figure 5: Scheduling sequence diagram

Scheduler

Figure 6 shows a screenshot of the scheduler component. The left side shows the structure as a tree with a Functional Unit as a root node. On the right side, the modules are listed with their current monitoring status. The data for the simulator and inference machine has been combined into a "Knowledge+Model" module for the interaction server.

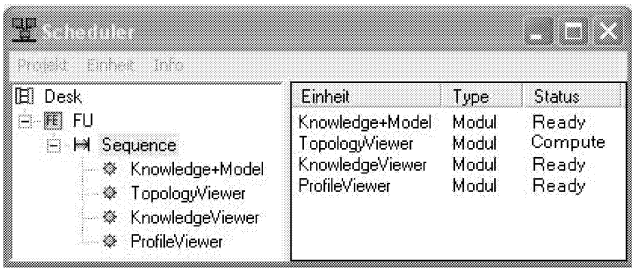


Figure 6: Screenshot of the scheduler component

Schedule Editor

The information about the structure, cycle time and time-out values, the process data base and the assignment of sources and targets are stored in an XML based project file created with the Scheduler Editor. The creation of the file requires the following steps:

- Loading a list with available process values,
- Loading of the modules and associating description files,
- Defining the structure of the modules and
- Connecting module variables via drag & drop.

CONCLUSION

Extending a process control system with simulation and knowledge based supervision requires the development of numerous tools. These tools must be coordinated and integrated into the existing infrastructure.

The interface generation mechanism presented allows rapid development of such tools using various programming languages and platforms.

The current development in XML based specifications for OPC allows the integration of supervision components in process control systems as Web Services.

A central control and management component allows for a configurable architecture depending on different use cases. This enables simulation and visualization components to be developed and tested as standalone tools.

The described approach has been implemented and approved for applications in water supply facilities. Further research will be done in the transfer to other application domains.

ACKNOWLEDGEMENT

The presented architecture was developed in the framework of the R&D project "Akquisition, Management und Integration von Prozesswissen in eine modellbasierte Prozessführung als Repräsentant einer neuen Generation von wissensbasierten Systemen in der Leittechnik," (http://amaryl.ipk.fraunhofer.de) partially funded by the Federal Ministry of Education and Research, Germany. We thank our industrial partner Elpro Prozessindustrie und Energieanlagen GmbH, Berlin, for extensive contacts with companies of the water industry.

REFERENCES

Gayet, P. et al. 2003. “Deployment and Integration of Industrial Controls: the Case of LHC Cryogenics Controls.”. In *Proceedings of ICALEPCS2003.* Gyeongju, Korea, 463-465.

Krüger, J. et al. 2005. “Knowledge Modeling and Processing for Supervision of Process Facilities”. In *Proceedings of the Industrial Simulation Conference 2005 (ISC-2005, Jun.9-11).* Berlin, Germany.

Lisounkin, A. et al. 2004. “Knowledge Acquisition from Historical Data for Case Oriented Supervisory Control.”. In *Emerging Solutions for Future Manufacturing Systems.* Edited by Luis M. Camarinha-Matos, Springer, 499 – 506.

Paiuk, J. et al. 2002. “Advanced Model Based Process Supervision in Hot Steel Milling.” Proceedings of *b´02, IFAC 15th World Congress (*21-26 July, 2002). Barcelona, Spain.

Pleßow, M. et al. 2005. “Tools for Knowledge Acquisition, Modeling and Visualisation applied to Process Supervision”. In *Proceedings of the Industrial Simulation Conference 2005 (ISC-2005, Jun.9-11).* Berlin, Germany.

Robinson, M. et al. 2002. “Building Plug & Play Power Applications Using Abstract Object Modeling.”. In Proceedings of the 35th Hawaii International Conference on System Sciences. Hilton Waikoloa Village, Island of Hawaii.

Youngpil, C. et al. 2002. “Modeling of Information Architecture using XML in the Agent-Based Distributed Control System.”. In *Proceedings of the 6th International Conference on Engineering Design & Automation.* Maui, Hawaii, 178-183.

TOOLS FOR KNOWLEDGE ACQUISITION, MODELING AND VISUALIZATION APPLIED TO PROCESS SUPERVISION

Rainer Fröhling, Michael Pocher and Matthias Pleßow
Society for the Promotion of Applied Computer Science (GFaI)
Rudower Chaussee 30
D-12489 Berlin, Germany
E-mail: { froehling | pocher | plessow }@gfai.de

Alexei Lisounkin
Fraunhofer-Institute for Production Systems and Design Technology
Pascalstrasse 8-9
D-10587 Berlin, Germany
E-mail: alexei.lisounkin@ipk.fraunhofer.de

KEYWORDS
Software architectures, tools, modeling, acquisition of process knowledge, visualization of knowledge processing, network, graph drawing

ABSTRACT

The operators of local care and disposal facilities (water companies, gas supply nets etc.) want to optimize the managing of their plants. This can be achieved by optimization of the handling, based on simulations and processing rules. For this purpose the existence of a suitable mathematical and information-technological model of the plant structure as well as the process knowledge are required.

This paper describes concepts and software-technological solutions for the acquisition and graphical preparation of this topic. The presented tools allow the description of the structure of a technological plant in a graphic-interactive way, for an abstract data model could be developed. Furthermore the plant-intrinsic technological elements and their interactions are illustrated. Finally it is visualized, in which way acquired formalized process knowledge can be prepared and integrated into the data model. This knowledge can be used for the knowledge-based guidance of plants.

INTRODUCTION

The structure and function of care and disposal networks and comparable technological plants (e.g. water companies) become an increasing importance within different forms, particularly by technical designs (CAD systems) or geographical information systems.

Furthermore data base systems, in which operating data and account parameters are stored, are playing a large role. In these cases the network architecture (topology of the net) is contained only implicitly and is not opened for use. There is a growing need for a new concept, which illustrates the structure of a supply network and the knowledge about the technological process adequately. In addition, the relevant equipment components with their technological parameters are required to be described and displayed graphically. Furthermore it is necessary to describe the relations of resources to each other.

In order to model and test plants as well as to acquire and manage processing-knowledge, the described tools supply the necessary environment. With its assistance one shows, how technological and process engineering knowledge can be integrated by dealing with formal methods into a model-based processing.

The product family presented in this article permits both a horizontal integration of the technological data needed with tasks of project engineering as well as vertical process integration by the possibility of letting different tools access over a defined interface to the data of a universal model.

All tools introduced here are based on a framework named CASTool (**C**omputer **A**ided **S**chematics **Tool**box), developed in the GFaI. This framework forms the basis for the development of graphical engineering systems. These are characterized particularly by the use of complex layout mechanisms, comfortable interaction and navigational methods as well as of configuration and expendabilities of the developed systems. Particularly important is the presence of an internal network model for the illustration of the plant topology [Schmidt, H.-W. et al. 2005].

In the following this base framework and tools developed for respective planning and design phases are specified more precisely.

BASIC FRAMEWORK

The development of the product family described below is based on the work on different R&D projects. Especially the implementation of a base framework (CASTool) including a network model (**ELADO** – **E**xtended **La**yout **D**ata M**o**del) allows the development of intelligent graphical editors within the area of graphical engineering systems for net-like schematics.

In general terms, the framework provides mechanisms to store net-like, hierarchically structured application data, to

add application-specific functionality, and to define varying types of object and system behavior. The developed applications can be used either as standalone applications or as distributed applications in a heterogeneous environment. In order to cover areas of usage, which are as large as possible, the system has been developed to run on various system environments by means of platform-independent libraries.

In order to be able to develop miscellaneous engineering solutions, the system architecture is designed in a way that enables adding or modifying of its functionality. For this purpose, the framework can be augmented by application-specific services such as special layout methods or other engineering components. These methods or components are added dynamically to run time via a defined interface. This allows them to be registered, managed (registry engine) and carried out (command factory) using various modules, without having to alter the framework itself. Furthermore it is very important to develop solutions that are user-friendly and ergonomic. For this purpose, various configuration mechanisms have been implemented within the system (configuration engine), which are primarily used for the design of the GUI (graphical user interface) of a concrete application or layout management (in the form of layout plans that formulate the calls of individual analysis, layout and evaluation methods). A lot of layout mechanisms (layout engine) and interaction issues (GUI) are integrated in this system. Moreover, by the integration of script interpreters, simple functionality and system behavior can be defined at system run time through scripts.

To create and modify net-like schemes, a network model was implemented, based on graph models similarly to mathematical graphs. But this model needs to extend. Networks differ from graphs for they have connection points (pins) with a defined position on the shape of the component. Furthermore, components that are corresponding to the graph vertices can also contain networks of their own. This means that hierarchically structured models can easily be represented. This core structure of the network model is called domain model in Fig. 1.

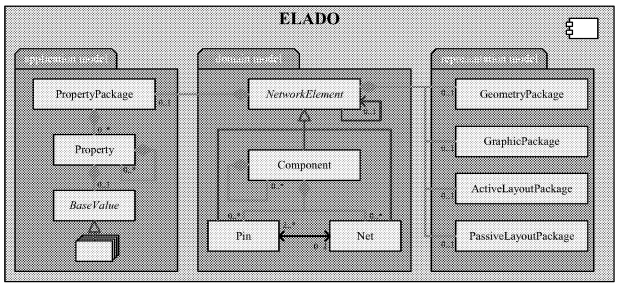


Figure 1: ELADO – The network model

In order to be able to develop different applications, and in particular for their application-specific data to be incorporated into the model (e.g. technology data, business data), the **ELADO** has been augmented with an application model. These application-specific data are modeled in a property tree whereby diverse types of basic data are implemented as derived classes in the leaves of the tree (application model).

The model also includes representation information. This part of the model is used to store information about the graphic depiction of network elements, their shape and positions, and the layout-specific behavior (representation model).

MODELLING OF TOPOLOGY

The development of a supervision system starts with modeling the topology of the facility. For this purpose the **Topology Editor** was developed. It allows not only embodying a graphical representation of the topology. Furthermore an internal model of the facility including all resources with resource parameters is created.

In a first step the resources will be created. For this purpose a **Resource Editor** is integrated into the system. It creates libraries of prototypes of resources, which involve a graphical representation. Furthermore structural information is included, especially for the definition of the connection points between the symbols and the nets. The creation of the values of the technological parameters of the resources is very important. These information is used by other modules like the simulator etc.. The visualization of resources is done by the integration of special behavior of symbols in the same line of dynamic visualization.

The created resources will be used in the **Topology Editor**. Here they are connected to an abstract schematic visualization on a graphic-interactive way. With this step the modeling process ends. The following figure shows the result of the modeling of waterworks.

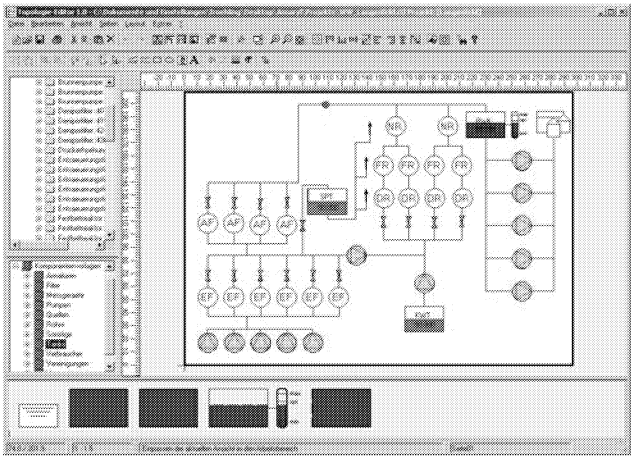


Figure 2: Topology Editor with created topology

Additional the user has the possibility to define particular parts of scheme for a later visualization in the runtime phase.

A special XML-like format was defined for the export of the generated topology (CTL – CASTool Markup Lan-

guage). This format extends traditional formats for the description of graph structures, especially for the definition of pins, graphic visualizations and technological parameters of resources.

A complex layout mechanism, an intelligent interaction system as well as a simple solver for design rules are integrated in the editor. This alleviates the user's work with the system. The application is getting easier and more comfortable.

GENERATION OF CONSUMPTION PROFILES

To enable optimum management for supply and disposal systems (gas, water, electricity etc.) consumption profiles are used for the forecast of consumption of specific resources. Due to this circumstance it is necessary to use experiences which allow forecasting by a combination with time intervals like daytime etc..

With the **Profile Editor** a tool has been developed and implemented which facilitates the user to create new profiles as well as to modify existing profiles. Furthermore the profiles can be organized in a clear and comfortable manner.

Consumption profiles run over a variable time horizon that fixed by the user. They are visualized in the **Profile Editor** as graphical representations and structured hierarchical in a tree structure. The leafs of the tree are representing the particular profiles.

The graphical visualization of daytime profiles is rendered by polylines defining the values of a process parameter over 24 hours. To edit a daytime profile the particular visualization of the value in the form of a horizontal line can be moved by the mouse pointer in vertical direction. The following figure demonstrates the creation of a daytime profile in the **Profile Editor**.

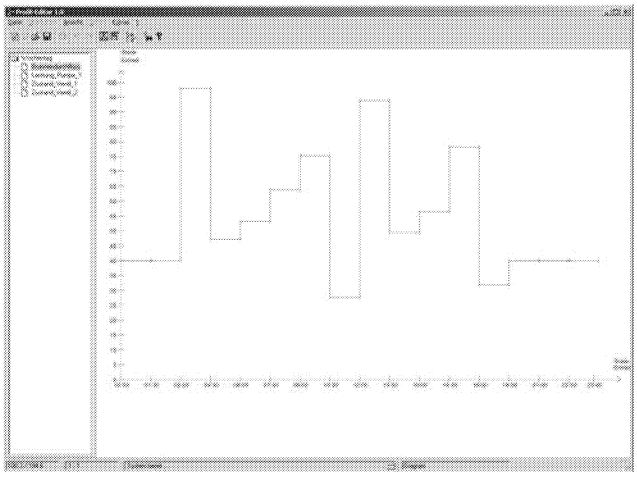


Figure 3: Profile Editor with profile

The **Profile Editor** was designed with the implementation to get an output file in the csv-format known from database applications. This is necessary for the further use of profiles in the **Knowledge Editor**. In this tool the user has the possibility to load existing profiles, which were generated earlier within standard database applications.

ACQUISITION OF KNOWLEDGE

In the engineering phase of a system for process supervision the knowledge must be formalized and acquired. For this experiences of human surveillance operators can be used. To transform this knowledge into rules for process supervision a model was concerted and implemented. This model is similar to the concept of mathematical implications ("if condition is true, then perform action"). But this simple definition of rules is extended for some important features. For example the main condition to be solved can be a Boolean representation of simple conditions. Furthermore rules have a so-called priority, defining the order of their execution. A general view of the model is given in the next figure.

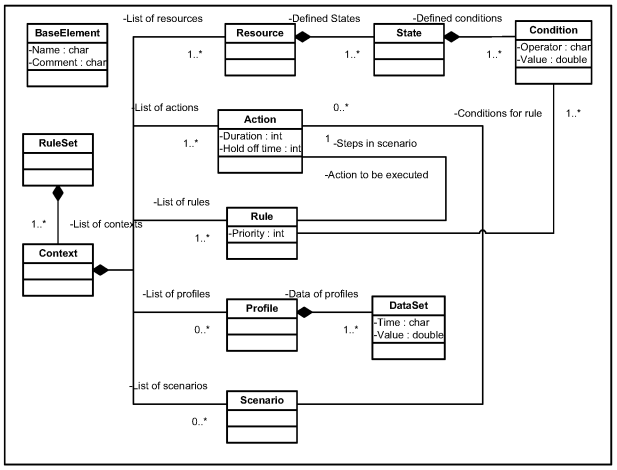


Figure 4: Class diagram of process knowledge

It is important to connect the knowledge, acquired in rules, with a concrete description of topology. For this purpose links to resources and resource parameters are implemented (states). These states are combined with defined values through different operators (MIN, MAX, DELTA etc.), so the user can define conditions for the execution of rules. Furthermore the rule set can be filled with consumption profiles. This enables the system to work close to a defined operating point.

Figure 5 demonstrates an example of a defined rule set in the **Knowledge Editor**.

The rules are organized in so called rule contexts. These contexts are a set of rules, which apply to specific times with different modes of work. For example the user can define different contexts for weekdays and Sundays.

Within the system the creation and visualization of rules in contexts a tree visualization is used. Hereby all necessary data input is converted in form-like views. To provide an optimized general view to the user, different

graphical kinds of visualization are implemented in the editor. This is illustrated in Figure 6.

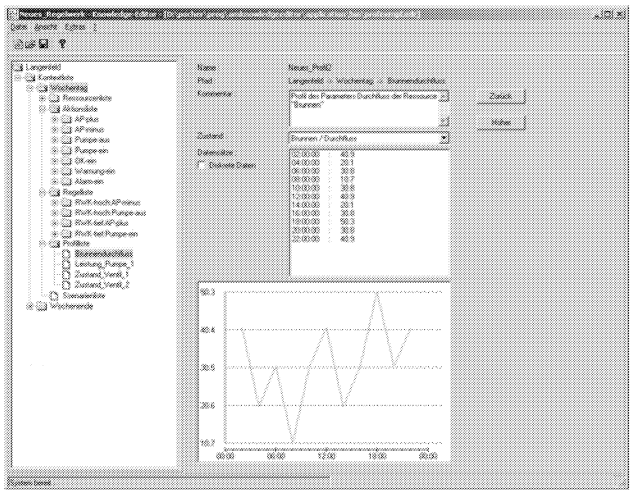


Figure 5: Knowledge Editor

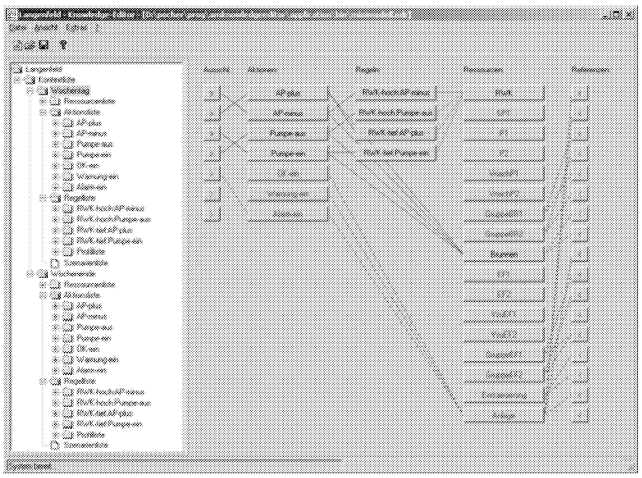


Figure 6: Graphical representation of a rule context

In the left part of the GUI a tree visualization of the rule set is shown. In the right part the content of a rule context in a graphic manner is visualized. Hereby are shown the actions to be executed and the resources, involved in particular rules. By the use of different colors and line styles information like not used resources or not executable actions can be differentiated at the first glance. The symbols are implemented as buttons. By the activation of this buttons the user can navigate through the rule set.

If the rule set is created completely, the user can export it in a XML description. This is necessary for the input of the Interaction Server and the visualization of profiles in the runtime phase. To support the user while creating correct rules, specific design rules are implemented. If these rules are not solved, the user will get a message from the system.

VISUALIZATION OF PROCESS BEHAVIOUR

With the **Topology Editor** simplified visualizations of a topology for the runtime phase can be generated. These different views can be loaded into the **Topology Viewer** during the runtime phase of the system. In this way the developer can generate views with a visualization of those resources, which process parameters should be observed at runtime. Due to this task symbols with dynamical behavior are used. This is the advantage against visualization possibilities of traditional process supervision systems.

To resolve specific industrial norms for the connection with the **Scheduler** an OPC interface based on the SOAP protocol was implemented. Through this interface information about the state of specific resources is transferred and visualized. The following figure shows a visualization of the topology of a facility build during the development phase and defined for representation at runtime phase.

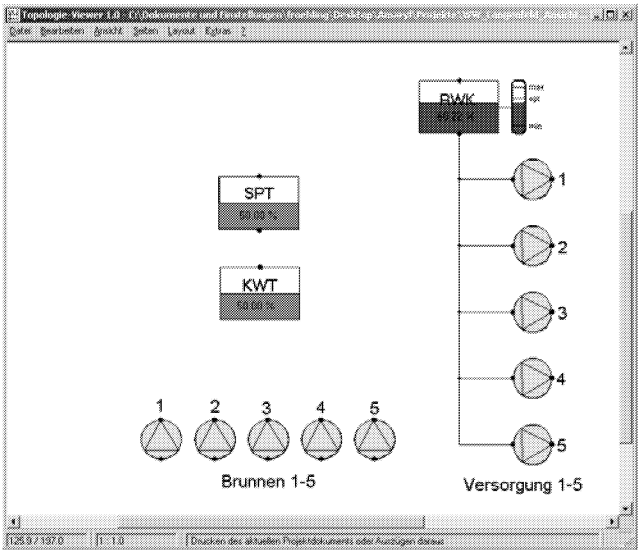


Figure 7: Resources view in runtime phase

VISUALIZATION OF PROCESS SIMULATION

In the runtime phase the computed results of the **Interaction Server** are to be visualized in the **Knowledge Viewer**. This is conducted by so called decision trees. These trees characterize the content of the computed decisions. The **Knowledge Viewer** also has an OPC interface for the connection with a real process (or the **Scheduler**).

For the internal description of decisions a model was developed and implemented, which represents a part of the content of the registered rule set. The basic structure of this model is shown in Fig. 8.

The tree structure of decisions is realized by a list of possible decisions, following the previous decision. Furthermore every decision has an internal time stamp representing the computed execution time of the resolved rule.

From the model of the set of rules each decision knows the rule, whose processing led to the decision. The decision knows also resources with its states, concerned by the rule.

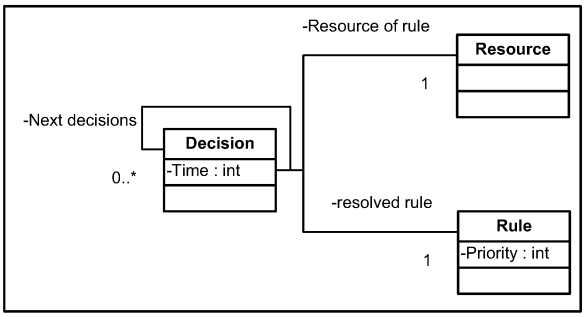


Figure 8: Class diagram of decisions

With this information the **Knowledge Viewer** generates different kinds of visualization of Interaction Server's computation. This is demonstrated in Fig. 9.

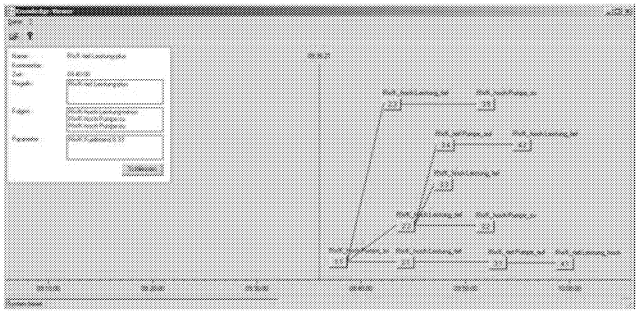


Figure 9: Graphical view of a decision tree

In the right part of the GUI the decision tree is visualized. The symbols of particular decisions are shown as buttons. By pressing these buttons the user is able to visualize the content of the specific decision in connection with the applied rule and information about the referred resource in text form at the left side.

VISUALIZATION OF PROFILES

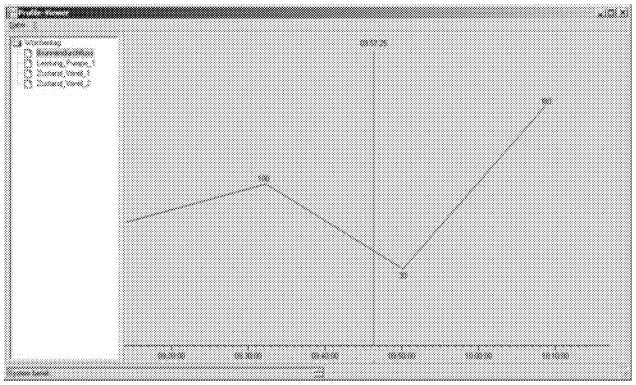


Figure 10: Profile Viewer with visualized profile

To show the progress of execution of operation profiles a **Profile Viewer** was realized.

It is connected to the **Interaction Server** through an OPC interface that receives information about the rule context to be executed at a specific time. This information is a part of the process knowledge registered in the **Knowledge Editor**.

The **Profile Viewer** gives the possibility to visualize all profiles registered in a rule context. They are prearranged in a tree view. The selected entry of this tree (the selected specific profile) is visualized as a time dependent diagram. This kind of graphical description is demonstrated in Fig. 10.

CONCLUSION

The presented tools permit the support of different steps of a production stage (horizontal integration) as well as different production stages (vertical integration). It is possible to build a complete internal model of the topology of waterworks and its resources. Hereby information about human operators experience is included. Hence, technological knowledge can be integrated in a model and knowledge based process supervision. Due to the use of an internal network-like model a simple and effective possibility of a modification of the facility is guaranteed. During the work with the system in the runtime phase the knowledge about the process can be completed and integrated into the system. It is proofed as evident that the underlying framework CASTool and the network-model ELADO are suitable for graphical modeling of complex structures. The framework was qualified and is also usable in other areas [Augenstein et al. 2004].

ACKNOWLEDGEMENT

The presented results arose in the context of the R&D project "Akquisition, Management und Integration von Prozesswissen in eine modellbasierte Prozessführung als Repräsentant einer neuen Generation von wissensbasierten Systemen in der Leittechnik (Amaryl)" (see http://amaryl.ipk.fraunhofer.de). The authors are grateful to the Federal Ministry of Education and Research for the financial support of this project.

REFERENCES

Augenstein, E.; Wrobel, G.; Kuperjans, I.; Plessow, M.: "TOP-Energy – Computational Support for Energy System Engineering Processes. " In: Tsahalis, D. T. (ed.): *Proceedings of the 1st International Conference "From Scientific Computing to Computational Engineering"*, University of Patras, Athen, 2004.

Schmidt, H.-W. et al. 2005. "Architecture for simulation & knowledge based supervision for process facilities". In: *Proceedings of the Industrial Simulation Conference 2005 (ISC-2005, Jun.9-11).* Berlin, Germany.

MODELLING AND SIMULATION IN THE TEXTILE INDUSTRY

FIBRE AND POLYMER SIMULATION

DEVELOPMENT OF A MATHEMATICAL MODEL FOR THE DRYING OF TEXTILE FABRICS

Pradeep K. Agrawal and Li Xue
School of Chemical & Biomolecular Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332-0100, USA
Email: Pradeep.Agrawal@ChBE.GaTech.Edu

KEYWORDS

Industrial Drying, Model Design, Model Parameter Estimation, Model Validation, Dryer Optimization

ABSTRACT

In drying of high pile fabrics such as carpet, large excess air appears necessary to provide driving force for removal of moisture. In such demanding operations, a quantitative understanding of the drying rates and the effect of key variables on the drying rates requires a good working model. This work is aimed at developing a mathematical model for the drying of high pile textile fabrics. The model is based on the temperature profiles obtained at different layers of the carpet as it travels through the dryer. These profiles aided in the configuration of a mathematical model that incorporates the essential elements of the drying process. The drying consists of two phases: (i) removal of free moisture that is governed by simultaneous convective heat and mass transfer, and (ii) restricted moisture removal which is controlled by the diffusion of moisture through a porous carpet pile. A laboratory set-up allowed independent estimation of the model parameters. These parameters were diffusion coefficient and the critical moisture content. The mathematical model was applied to the industrial carpet drying process. The results matched the model prediction successfully and allowed optimization of the dryer zones.

INTRODUCTION

Drying is an intermediate step in the dyeing and finishing operations in the textile industry and uses a substantial amount of energy. About one half of the energy used in a typical textile finishing mill is used in the drying processes. Due to the increasing fuel cost, it is imperative that the dryers be operated as efficiently as possible. The use of excess air is a major source of energy losses in a dryer operation. Measurements have shown that as much as 60-80% of the total energy used in a dryer operation is wasted in exhaust. The high energy loss is a direct consequence of not having a knowledge of the drying rate curves in general and roles of various operating parameters in particular. Besides, the complex relationships among these factors exacerbate the problem and result in difficulty of controlling the drying operation in a timely and precise fashion. A typical temperature profile of the fabric consists of three regions. Zone I represents the initial heating of the fabric. Zone II represents the removal of the free moisture from the fabric. During this period, the temperature of the fabric usually remains constant. This period is called the constant drying rate period. Zone III is for removing the restricted or bound moisture from the fabric. The temperature of the fabric increases in this zone. The drying rate decreases as the moisture content of the carpet deceases, defined as the falling drying rate period.

Although several efforts have appeared in the literature to model the drying of porous materials in general and textile fabrics in particular (Brock and Gorton, 1964; Francis and Wepfer, 1993; Francis and Wepfer, 1996; Geankoplis, C.J., 1993; Lee et al., 2002; McCabe et al., 2001), none of these models allow independent determination of the model parameters. This work is aimed at developing a mathematical model that can quantitatively describe the evolution of transient temperature and moisture content profiles in a carpet tile. In order for a mathematical model to have predictive ability, it should be based on the physical features observed in specific applications. Another desirable characteristic would be to have a model where all or most of the model parameters can be determined from independent experiments. An attempt was made to examine the physical characteristics of the carpet tile drying. This included a time-temperature profile in an industrial dryer as well as scanning electron microscopy (SEM) results from the tile cross-section. The SEM results showed that the polyurethane layer would be impermeable to moisture (closed cells) and thus all the moisture is expected to reside within the Nylon carpet face and within the primary backing/latex coating at the backend of the Nylon pile. A mathematical model was developed taking into consideration these features as well as the general drying behavior observed during industrial operation.

MATHEMATICAL MODEL

The constant rate drying period is characterized by having the same evaporation rate throughout the period. It is physically characterized by a thin film or pool of unbound water at the solid's surface. When the porous solid is

exposed to a drying stream such as air, evaporation will occur at a constant rate such that the heat transferred to the surface equals the latent heat of vaporization. Consequently, the temperature remains constant in this region. In this regime, the water within the nylon fibers moves readily to the surface to keep it wet. For the constant drying rate period, it is generally assumed that the resistance to the mass transfer is within the boundary layer between the air and the drying surface. The mass transfer rate is governed by the driving force for water vapor between the nylon-air interface (y) and the air (y_a):

$$-m_f \frac{dw}{dt} = k_c C M_w (y - y_a) \quad (1)$$

The modeling work of heat balance takes the approach of a lumped parameter analysis. The heat transfer resistance to conduction within the carpet tiles is neglected, and all layers were assumed to be at the same temperature. The purpose of this assumption was to simplify the resulting heat transfer equations by leaving the derivatives of temperature only with respect to time.

$$-H_v m_f \frac{dw}{dt} + (m_f Cp_f + m_f w Cp_w)\frac{dT}{dt} = h(T_a - T) \quad (2)$$

The term on the right side represents the convection heat transfer from the air; the first item on the left side represents the latent heat for water vaporization; the second item on the left side represents the sensible heat for heating the carpet and the water contained in the carpet. The mass transfer coefficient can be correlated to convective heat transfer coefficient by Chilton-Colburn analogy, which is based on the thermal boundary layer and the concentration boundary layer analysis, and the heat and mass analogy.

When the moisture content in the carpet reaches the critical moisture content, w_c, the drying process enters the so called falling rate regime. During this period, the rate of moisture movement from the interior of Nylon pile to the pile surface begins to decrease; dry spots begin to appear on the solid surface as the moisture evaporation front recedes into the carpet tile. As the evaporation front continues to retreat into the porous medium, the liquid concentration behind the evaporation front remains almost constant. The driving force for this moisture diffusion lies entirely between this evaporation front and the Nylon pile surface. The falling rate period may be described by diffusion of water vapor through the nylon pile or/and through the interface of latex and primary backing. Generally this is referred to as bound moisture removal.

LABORATORY SET-UP

Figure 1. shows a schematic of the laboratory set-up for a bench scale dryer in order to measure the drying rate curves for carpet tiles and to use these data to determine the model parameters such as transport properties and critical moisture content.

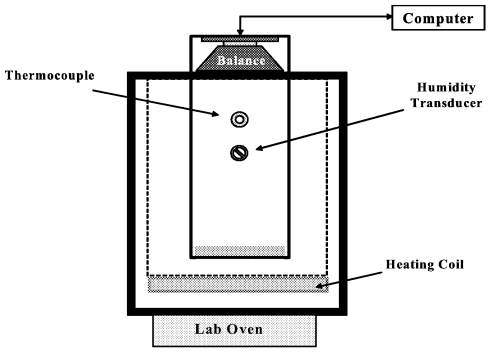


Figure 1: Experimental set-up for drying curve measurement in a bench scale oven

Experimental set-up consisted of three main components: an oven, an electronic balance, and a PC-based data acquisition software. The oven is a Cole Parmer laboratory oven (model number G05053-10), which can be heated to a temperature of 220 ^{o}C (with an accuracy of 1 ^{o}C). A sample holder frame was constructed to hold the carpet tile sample (18×10 cm^2). The electronics balance (Denver Instrument, Model number APX-4001) was placed on top of the oven. The sample holder was suspended into the oven through the two holes drilled in the top surface of the oven, such that the sample holder rested on top of the electronic balance. The electronic balance has a load capacity of 4000 g ± 0.1 g. The software used for data acquisition was Balance Talk XL Version 5.0 through RS232 connections (Labtronics Inc.). The air was supplied by Airgas Company. The gas flow was controlled by AALBORG mass flow controller with capacity up to 500 ml/min. Oven temperature and humidity were continuously monitored with time during the experiments. Oven chamber humidity was measured by a transducer (Ohmic Instruments Co., model AHT-200-01) with working temperature range up to 200 ^{o}C. The oven temperature was measured by a K-type thermocouple with an accuracy up to ±0.1^{o}C inserted close to the sample. The drying experiments were carried out at constant temperatures; three different temperatures were utilized 100 ^{o}C, 120 ^{o}C, 140 ^{o}C.

RESULTS AND DISCUSSION

The carpet samples used in this study were three different tufted carpet tiles produced by the Milliken & Company in LaGrange, Georgia. Three different types of carpet tiles were investigated: Roadrunner (RR), Midnight Sparkle (MS), and Grand Plaza (GP). Among them, RR and MS are loop piles, and GP is cut pile. Weight of the carpet tile, oven temperature and humidity were monitored and recorded during the experiment. When the weight of the carpet tile no longer changed for about three minutes, the

experiment was stopped. The sample could be considered "dry" at this point.

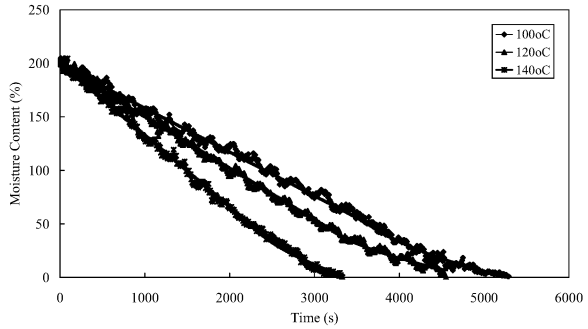


Figure 2: Moisture Content Profiles of the RR carpet tiles at different temperatures

Figure 2 shows the moisture content vs. time data for the Roadrunner carpet files at three different temperatures (100°C, 120 °C, and 140 °C). In each case, the initial moisture content was 200% (by weight of the Nylon pile). The moisture content decreased almost linearly in the beginning. As the tile approached near dry conditions, the rate of drying started to fall which is indicated by the decreasing value of dw/dt. Also, shown in Figure 2 are the mathematical model curves which were useful in reducing the effect of noise in the data. Due to the different characteristics of the two drying periods, the key model parameters are determined separately from the different parts of the experimental data. Convective heat and mass transfer coefficients were calculated from data obtained in the constant drying rate period. Since convective heat and mass transfer coefficients are correlated by the Chilton-Colburn analogy, there is only one independent parameter. On the other hand, the diffusion coefficient was determined from the data in the falling drying rate period. The critical moisture content, w_c, was determined from the mass versus time plot where the transition occurs from a straight line to a curved line. All drying curves in this study displayed these two drying rate periods. It thus provides a good basis for the determination of the transport properties and the critical moisture content analysis. The critical moisture content values for the three different tile products are remarkably in a narrow range (0.50-0.55), as shown in Table 1. The fact that the critical moisture content changes negligibly with temperature should be largely expected. As stated above, the critical moisture content should only be a function of the microstructure of the Nylon fibers, primary backing, and the type and the amount of latex. In the case of three different products used in this study, the latex and primary backing remain the same.

The values of the diffusion coefficients for the three different types of carpet under three different temperatures 100 °C, 120 °C, and 140 °C were consistent with each other. It indicates that, for the three different types of carpet in this study, the diffusion coefficient is only dependent on the temperature, not the specific type of carpet. The values of critical moisture content (w_c), diffusion coefficient (D), and heat transfer coefficient (h) at three different temperatures (100 °C, 120 °C, and 140 °C) have been obtained. Since critical moisture content (w_c) and diffusion coefficient (D) can be considered to be dependent on the microstructure of the Nylon pile as well as that of the region near the primary backing, latex coating, and Nylon tufts, it is safe to assume that these values are also applicable under industrial drying conditions.

Table 1: Critical moisture contents of the three different types of carpet at different temperatures

Products	100 °C	120 °C	140 °C
Roadrunner	50.6%	51.2%	49.8%
Midnight sparkle	51.6%	51.3%	49.3%
Grand plaza	55.5%	53.1%	53.4%

The convective heat transfer coefficient (h) obtained in the laboratory drying condition is likely to be very different from that in the industrial dryer. This parameter depends on the flow conditions as well as the flow geometry. Impinging flow creates very high gas velocities (10 ~ 100 m/s) at the Nylon pile surface, which cannot be duplicated or matched in the laboratory dryer. Therefore, this parameter would have to be adjusted to fit the industrial dryer data. In any case, these values of heat transfer coefficient are 10~15 fold higher than those found in the laboratory studies.

Having all the model parameters for simulation of industrial dryer operation, testing the model validity is the next logical step. Figure 3 shows three different sets of temperature profiles for the Nylon pile in RR carept file. There were ten zones in the industrial dryer, each zone being 3.05 m in length. The temperature of gas phase in the first 6 zones was 165.5 °C, whereas the temperature in the last 4 zones was 148.9 °C. The three sets of data were obtained on three different Roadrunner tiles and on three different days. Thus, the experimental data shown in Figure 3 represent a good indication of product and process variability. From the experimental data in the constant drying rate period (10-90 sec), a heat transfer coefficient valve was fitted and it was about 15 times the valve of h in the laboratory oven. This was as expected. Using this valve of h, together with the diffusion coefficient and the critical moisture w_c, we can obtain model prediction. The agreement between the model prediction and the data is remarkable. The fact that there is a good fit in the constant drying rate region is not as surprising since the h value was obtained by fitting the rate equations. What is particularly noteworthy is the

model prediction after the tile has moved beyond the constant drying rate region, because moisture diffusion controls the drying rate here. The diffusion parameter (D) and the critical moisture content (w_c) were obtained from the laboratory scale drying experiments, and the time scale of these experiments was at least an order of magnitude higher than the industrial drying conditions. The excellent fit (both qualitative as well as quantitative) provides a good validation of the model. Most likely, it also reflects the fact that the model is developed based on a physical description of the drying process. This sets the stage for using the model's abilities to predict and improve the performance of the industrial dryer. A successful outcome here has provided additional support for model validity.

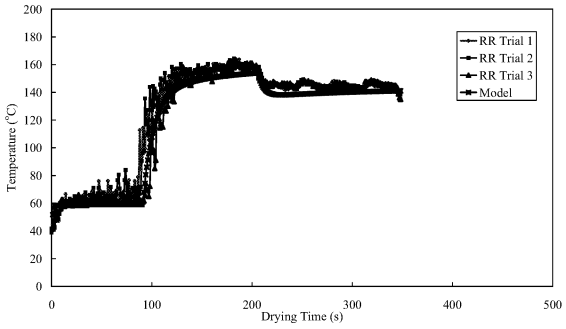


Figure 3: Model fitting of the Nylon pile temperature profiles of RR carpet tile

CONCLUSIONS

The mathematical model is validated in several ways: (1) the model parameters obtained from laboratory experiments offer a picture consistent with the physical aspects as expected, i.e., heat transfer coefficient values vary only slightly with temperature, and diffusion coefficient is a power function of temperature. These also remain relatively unchanged between different tile products, indicating similar microstructure; (2) with only one adjustable parameter (convective heat transfer coefficient h), the model applies to the industrial dryer operation quantitatively.

LIST OF SYMBOLS

C Total gas phase concentration, mole/m^3

C_{p_f} Specific heat of fabric, J/kg – K

C_{p_w} Specific heat of water, J/kg – K

h convective heat transfer coefficient, W/m^2 – K

H_v Latent heat of water evaporation, kJ/kg

k_c convective mass transfer coefficient, m/s

m_f mass of fabric per nit surface area, kg/m^2

M_w molecular weight of water (18)

t independent variable, time

T fabric temperature (^oC)

T_a ambient gas temperature (^oC)

w moisture content/100

y mole fraction of moisture at the evaporation front

y_a mole fraction of moisture in ambient air

REFERENCES

Brook, J.D. and C.W. Gorton. 1964. "Through – Flow Drying of Tufted Textile Materials." *Textile Research Journal*, No. 34, 1-13.

Francis, N.D. and W.J. Wepfer. 1993. "Experimental and Numerical Analysis of the Drying Characteristics of Modular Carpet Tiles." *Textile Research Journal*, No. 63 (1), 1-13.

Francis, N.D. and W.J. Wepfer, 1996. "Jet Impingement Drying of a Moist Porous Solid." *Int. Journal of Heat and Mass Transfer*, No. 39 (9), 1911-1923.

Geankoplis, C.J. 1993, "Transport Processes and Unit Operations", 3rd edition, *Prentice Hall.*

Lee, H.S, W.W. Carr, and H.W. Beckham. 2002. "A Model of Through-Air Drying of Tufted Textile Materials". *International J. of Heat and Mass Transfer*, No. 45 (2), 357-366.

McCabe, W.L., J.C. Smith, and P. Harriott. 2001. "Unit Operations of Chemical Engineering". 6th edition, *McGraw-Hill.*

BIOGRAPHY

Pradeep Agrawal was born in Rampur (U.P.) India. He graduated from the University of Roorkee (now I.I.T., Roorkee) with highest honors with a B.S. in chemical engineering. He received an M.S. (1977) and Ph.D. (1979) in chemical engineering from the University of Delaware. He joined Georgia Institute of Technology as a faculty in the School of Chemical and Biomolecular Engineering and is presently associate professor and holds the Lee and Dorothy Tanner Faculty Fellow position. He is a consultant to the Milliken and Company

A SYNERGISTIC APPROACH BY MRI AND CFD FOR THE STUDY OF MICRO-FLOW IN TEXTILES

Johannes Leisen
Georgia Institute of Technology
School of Polymer, Textile and Fiber Engineering
Atlanta/GA 30332-0295
U.S.A.
E-mail: johannes.leisen@ptfe.gatech.edu

Marco Märtin, Thomas Schipotinin, Elke Rodehüser, Karin Farber, Peter Farber
Fachbereich Maschinenbau und Verfahrenstechnik
Hochschule Niederrhein
47805 Krefeld
Germany
E-mail: peter.farber@hsnr.de

KEYWORDS
Magnetic Resonance Microscopy, MRI, Computational Fluid Dynamics, CFD, Micro-flow

ABSTRACT

A synergistic approach using techniques of Computational Fluid Dynamics (CFD) and Magnetic Resonance Imaging (MRI) is demonstrated using a simple model system, which may be considered a coarse representation of a woven textile structure. A simple flow cell suitable for high resolution flow MRI experiments is described. The design of this cell is such, that it may be easily incorporated as "virtual flow cell" into CFD calculations. A qualitative comparison between results from both techniques shows a good agreement. However, significant deviations between experiment and simulation may occur as a result of small deviations of the flow geometry of the "ideal" geometry. Techniques of MRI provide an ideal tool to detect areas within a flow cell, where these effects are likely to occur.

INTRODUCTION

The transport of fluids through textiles is of great importance for many aspects regarding the design, application and functionally of textiles. Examples range from effective dyeing processes over filtration materials to high performance fabrics with improved wearing comfort. A common requirement for the understanding of all these processes is insight regarding the transport of momentum through fibrous assemblies. Often, this information is difficult to obtain experimentally. Traditionally, one had to rely on "try and error" methods. For instance the conditions of a dyeing process have to be optimized by thoroughly inspecting the dyed goods with respect to an even coloration.
A powerful approach to obtain insight into momentum transport within textiles is based on **C**omputational **F**luid **D**ynamics (CFD) (Ferziger and Pedric 2002, Farber and Dahmen 1998). With this methodology it is in principal possible to predict details regarding the distribution of pressure and velocities of fluid momentum in textiles to any desired accuracy. However, limitations are still imposed by computer hardware. Moreover, CFD calculations are always based on specific assumptions; hence in order to be meaningful they need to be validated by experiments.
This experimental validation may be provided by **M**agnetic **R**esonance **I**maging (MRI) with a high spatial resolution, which is also referred to as **N**uclear **M**agnetic **R**esonance (NMR) Microscopy, (Blümich 2000). This technique is ideally suited for the investigation of fluid distributions in textiles. Generally contrast, i.e. the intensity of the detected MRI signal, is entirely due to fluids contained within a fibrous substrate and not the fibers itself. There are several factors relating contrast to the physico-chemical structure of a fluid trapped within a textile (Leisen and Beckham 2001). One of these factors is the displacement of fluid molecules due to self diffusion or flow. The resolution routinely achievable with dedicated commercial MRI equipment (so called NMR microscopes) is in the order of 40 μm. Higher resolutions beyond 10 μm are possible. However, they require lengthy measuring times and sometimes even special instrumental developments. Therefore, NMR microscopy may not be able to achieve a resolution, which is sufficient to depict individual fibers in textiles. Nevertheless, the technique is able to selectively measure many relevant properties of fluids entrapped in textiles so that it has a great potential for many applications in the field of textile science.

ASSESMENT OF TEXTILE GEOMETRIES

Any CFD calculation for the assessment of flow phenomena in textiles will require the inclusion of the textile geometry into its respective algorithm. Hence, the first step of any CFD calculation will involve the generation of a respective geometry for instance by methods of computer aided design (CAD). These methods often demand a technique able to measure the three dimensional spatial structure of the textile. Traditionally, a destructive approach is needed in order to access interior structures in usually opaque textiles, which are not accessible to optical light. This involves the microscopic characterization of a textile after physical "slicing". This slicing process is experimentally demanding and time consuming. It can be avoided by the use of techniques which are able to directly measure the three-dimensional structure of opaque objects. These techniques are micro-

CT, which is based on X-rays (Shen et al. 2004) and NMR-Microscopy (Blümich 2000). While micro-CT is able to directly measure the three-dimensional distribution of fibers this is not possible with the MRI technique. Due to their limited molecular mobility, polymer fibers generally do not provide a measurable contrast using most MRI techniques. Hence, the fabric will need to be immersed into a suitable contrast giving fluid. This may be any proton containing fluid, for instance water. Often, it may be advantageous to add a surfactant to the fluid in order to ease the fluid ingress into all pores of the textile. Of further advantage may be the addition of small amounts of paramagnetic compounds such as $CuSO_4$ or $MnCl_2$. This leads to changes in the relaxation behavior of nuclear spins such that MRI data may be recorded in a shorter time span.

Figure 1 shows an example of a technical fabric used in paper machinery (courtesy of AstenJohnson, Kanata/Canada and Heimbach, Düren/Germany). The geometry of this fabric may be readily assessed by methods of MRI. Figure 1.B shows a three-dimensional data set in a CAD format (STL), which has been generated from a 3D MRI data set consisting of isotropic voxels. It is obvious that some imperfections exist within the CAD data. For instance the surface of the fiber appears rougher and less even than in the original sample. This is largely due to the limited resolution achievable by NMR microscopy. Nevertheless, the existing data set may be of great value when it comes to the generation of more accurate geometries using CAD software. Hence, even if the resolution achievable by MRI may not be sufficient to obtain CAD data suitable for the direct import into CFD software packages the information obtained with this method may be of great help when it comes to the generation of the geometry using CAD software.

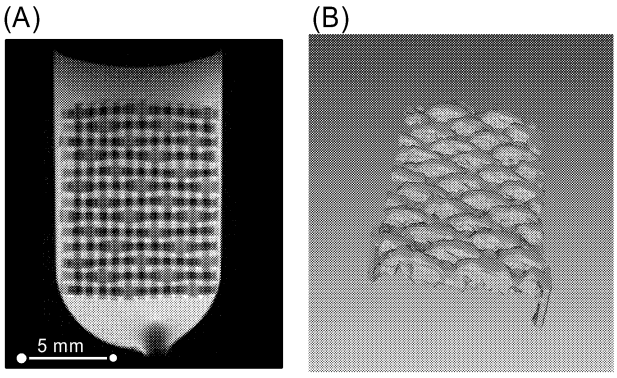


Figure 1: MRI Micrographs Used for the Assessment of a Technical Fabric:

(A) Single-slice two-dimensional image of the fabric immersed in a sample tube containing water doped with $CuSO_4$ (SE-technique, TE = 12 ms, TR = 50 ms). Dark contrast refers to areas void of the water providing a means to record the structure of the fabric, which by itself does not provide any contrast measurable by MRI.

(B) Three dimensional structure of the fabric as obtained after conversion of 3D MRI data (GE-technique, TE = 6 ms, TR=50 ms) in a CAD format (STL).

It should be also noted that flow MRI experiments as described below usually require setup and calibration steps, which are performed with nonmoving fluids, i.e. under conditions where the flow is switched off. Images identical to that of figure 1.A are generally available as a result of these steps. Hence, if momentum transport through the fabric is to be measured by MRI, information regarding the textile geometry will also be available.

FLOW THROUGH A MODEL SYSTEM

Design of a Flow Cell for a Synergistic CFD and NMR Study

Prior to measurements of industrially relevant textile samples it is important to demonstrate the feasibility of both MRI and CFD to obtain comparable results. For this task we have designed a flow cell with a simple geometry (cf. figure 2.). The design of the flow cell was chosen to achieve well defined conditions for both experiment and simulation. This is a very crucial step. If for instance conditions such as the geometry and flow velocities with respect to the area under investigations are not defined, results from both methods may be largely differ and their comparison is not justified.

The geometry and relative dimensions of the “virtual flow cell” used for the CFD simulations is depicted in figure 2.A. It consists largely of a simple tube. The porous structure of a textile obstructing a tubular flow was coarsely approximated by a solid block with five circular channels located at the center of the flow cell. We are well aware that a comparison of this structure with any textile structures is problematic. However, the simple structure serves the purpose of this study well. Data were obtained to establish the comparability between CFD results and flow MRI data obtained on a length scale relevant for flow phenomena in textiles with coarse structures in the mm to sub-mm range. The geometry for the flow cell was generated by the mesh generating software package ICEM CFD 5.0 (Ansys, Otterfing/Germany) and CFD simulations were performed within the Fluent 6.1 software environment (Fluent Deutschland, Darmstadt/Germany). The system model of figure 2.A consisted of a numerical mesh of 307.132 hexahedral cells and the solution domain extended over a length of 23.34 mm. Assuming steady, isothermal, laminar and incompressible flow the three dimensional Cartesian continuity and momentum equations were solved approximately with the segregated solver for water with an ambient temperature of 22 °C. Density and dynamic viscosity were linearly interpolated from values taken from reference (VDI-GVC 2002): Density ρ = 997.746 kg/m3 and dyn. viscosity η = 0.000956992 kg/ms). The fully developed laminar stationary pipe flow of 0.0054 m/s velocity magnitude in the axis was assumed (volume flow rate 8 ml/min) as inlet boundary condition. The Reynolds number based on the pipe diameter and the average velocity magnitude of 0.0027 m/s was 22. All walls were set to no-slip stationary walls of zero velocity magnitude. At the outlet a constant pressure difference of zero Pa to an operating pressure of 1 bar was used. Iteration convergence was checked by monitoring the residuals and the area weighted averages of

all degrees of freedom in the inlet and outlet of the small pipes in the probe. The simulation experiments were found to have converged after 960 iterations with first order upwind and additional 370 iterations with a second order upwind discretization algorithm. The results are considered to be accurate, however further calculations in order to provide verification according to best practice guidelines (Casey ad Wintergerste 2000) are currently underway.

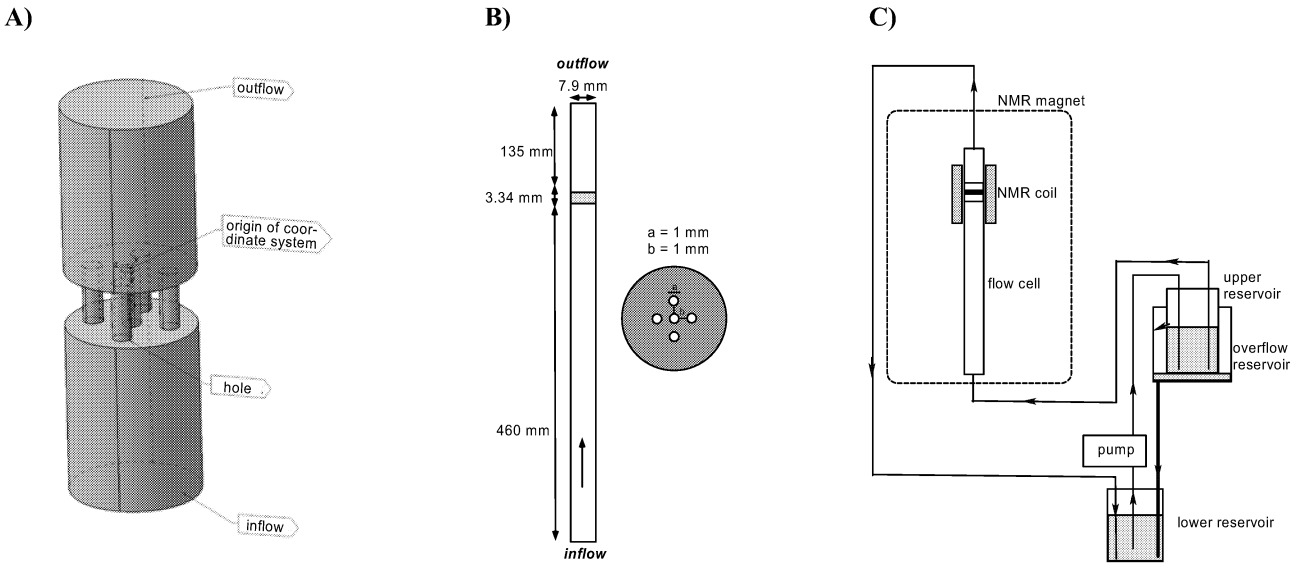


Figure 2: System Design for a Comparative Flow-MRI – CFD Study:
A) "Virtual flow cell" as used for CFD-studies.
B) Flow cell as used for NMR studies constructed out of a glass tube and a Teflon insert acting as obstruction.
C) Flow apparatus to provide a homogeneous nonpulsed flow through the flow cell in B) while performing MRI experiments.

The conditions within the above mentioned "virtual flow cell" were chosen to exactly match the geometry and dimensions of a flow cell constructed for MRI experiments (cf. figure 2B). In order to achieve a very regular non-pulsating flow, the flow was generated gravimetrically by siphoning water through the flow cell (cf. Figure 2.C.). The flow apparatus, which was used for this task, is based on a design described by Yao et al. (Yao et al. 1995). The overall flow through the cell was 8 ml/min. This was measured gravimetrically by determining the weight of water leaving the flow cell during a defined time interval. In order to eliminate any uncertainties regarding velocity profiles at the in and outlet of the flow cell, the MRI flow cell was designed such that it consisted of a rather long glass tube. The in- and outlets are sufficiently far away from the area under investigation, which is the five channel obstruction realized by a disk fabricated out of Teflon. Consequently, a profile as expected for the laminar flow of a Newtonian fluid in a tube can be expected for most areas within the cell. Due to the large number of required unit cells it is not feasible to simulate the whole flow cell used for NMR experiments. Only a region of interest with dimensions depicted in Figure 1.A. was used for the CFD calculations. However, this virtual flow cell was chosen long enough that it was appeared safe to assume conditions of laminar pipe flow at positions corresponding to the in- and outlet of the virtual flow cell.

MRI-flow experiments were performed using a Bruker DSX400 NMR spectrometer with NMR Microscopy accessories (Bruker-Biospin, Rheinstetten /Germany). A commercial probe (Bruker) was used in conjunction with a 10mm rf-coil insert. The flow cell was inserted into this coil insert, such that the area of interest was contained within the field of view accessible by this coil. The NMR microscope was controlled through the Paravision Software environment (Bruker). This software was also used for the analysis of all experimental data. Two major classes of techniques of flow MRI have been described, which are tagging techniques and techniques based on phase encoding (Blümich 2000). The phase encoding technique provides detailed images regarding velocity vectors at all positions within the measured image (Bieljic et al 2001). Images based on the conceptually simpler tagging techniques provide a direct visualization of transport phenomena occurring at a selected location within the flow cell. This technique is therefore well suited for the qualitative investigation and visualization of flow phenomena. Hence, within the scope of this paper, only data obtained by the tagging technique will be presented. Tagging experiments consist of a series of tagging pulses, which lead to the destruction of magnetization components usually located in a cross or grid shaped pattern. The MR image is recorded at t_{tag}, a defined delay time after the application of the tagging pulses. This results in an MR-image where tagged areas will appear with a largely reduced contrast. Any momentum transport during the tagging time t_{tag}, will result in a shift of position of this area. Hence, observing position changes of the tag as a function of

t_{tag} provides an intuitive insight into flow occurring at a selected position. This insight may even be enhanced by assembling individual images recorded for a series of equidistant t_{tag}-values into a movie. It should be noted that MRI tagging refers to changes of magnetization states at specific locations within the sample. Hence, any type of marker is not required and MRI-tagging may be considered not to interfere at all with the momentum transport.

Results

Figure 3 provides results obtained from CFD simulations of the flow cell of figure 2.A. From this figure it can be seen that flow up to about 3 mm before the obstruction can be described as a laminar flow within a tube. Increased velocities are observed within the range of the inflow into the channels. These increased velocities occur as a direct result of large velocities encountered within the channels. It appears that within the channels the laminar flow profile as described by the Hagen Poiseuille Law is formed within less than 0.5 mm of the channel inlet. It is interesting to note that the outflow of the channels leads to larger areas of deviation from the tubular laminar flow pattern than observed for the inflow area. Especially negative flow velocities i.e. flow against the main flow directions are found in the outer rim of the sample as seen in Figure 3B and C.

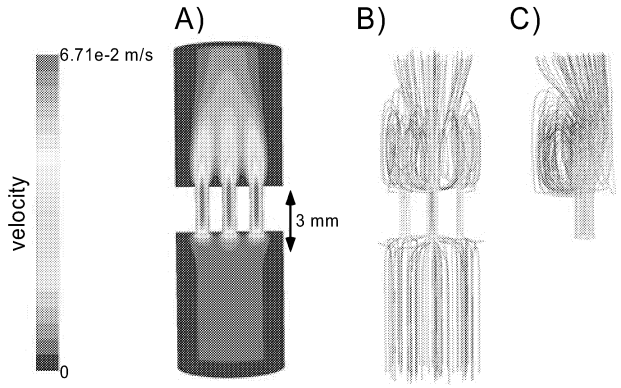


Figure 3: Results from CFD Simulating Flow in the Virtual Flow Cell as Depicted in Figure 2A:
A) 2D map of velocity magnitudes within a center slice through the flow cell symmetrically centered to contain three channels.
B) 3D representation of the path lines within the virtual flow cell.
C) Section of B) depicting the occurrence of stationary laminar separations.

The results for the area directly behind the phantom may nurture the suspicion that this area is highly sensitive to minor modifications in the overall flow geometry. Any minor variation in the geometry of the phantom is going to lead to an uneven momentum transport through the channels, which then may cause significant changes in the overall distribution of velocities with respect to the finding of our simulation.

Figure 4 provides a series of MR-images obtained with the above mentioned tagging techniques. The momentum transport may be observed by comparing the change of the position of the tag within each column. Visible along with the cross shaped tag are up to two tags with minor intensities, which appear to represent the progress of the momentum transport at longer tagging times. These "secondary" and "tertiary" tags are caused by incomplete relaxation of magnetization during a rather short repetition time TR=500 ms of the experiment. They may be eliminated through the appropriate choice of measuring conditions, however they also provide additional information, namely the change the tag position after the time t_{tag}+TR and t_{tag}+2*TR. Note also that characteristics of the flow are not only apparent from the position change of the tag, but also from the overall contrast observed throughout the image. This is especially visible in the pattern behind the obstruction, which indicates a beam like expulsion of the fluid. The origin of contrast forming these patterns is due to two effects: (i) the dephasing of magnetization during the Echo Time TE due to flow is not refocused in a simple Spin Echo sequence and leads to a reduction of the observed signal intensity (Blümich 2000). And (ii) any inflow of fluids into the observed image slice (thickness = 1 mm) during the repetition delay TR is leading to an increase in signal intensity. Effect (ii) may be eliminated by choosing a sufficiently large value of TR. However, we did not choose to do so, since the contrast in the displayed images provides additional qualitative insight into the occurring flow phenomena.

The analysis of figure 4 confirms the findings with respect to the inflow area into the obstruction. The tag placed at 4.2 mm before the obstructions provides evidence of laminar flow. The tag 1.5 mm before the obstructions exhibits still some characteristics of the laminar flow. In addition, it is apparent that some inflow into the channels occurs with increased velocities. Note that the overall image contrast until shortly before the obstruction is largely homogeneous. This is a further strong indication of a highly laminar flow. The bright area at the very bottom of each image is not attributed to flow effects but to hardware effect, namely an inhomogeneous excitation profile of the rf coil. The outflow area after the obstruction shows a contrast pattern of individual beams. The higher contrast intensities (represented by brighter grey scales) indicate an inflow of fluid into the image plane, which represents a slice with a thickness of 1 mm. This inflow may be explained by the circulatory flow patterns found by CFD (cf. figure 3.C and 3.D). In fact, preliminary data obtained with the phase encoding techniques have also provided evidence of these flow patterns (data not shown). Momentum transport in the area between individual beams is occurring on a rather slow time frame as can be seen from figure 4.C. The area of the tagging line between beams does not change position during 200ms. On the other hand, areas of the tag within the beams are already deformed during t_{tag}=5ms. They have disappeared after 200ms. This can be explained by the fact that a majority of the tagged

magnetization has left the selected slices due to momentum transport. Note that for all images within figure 4 the beams are not symmetrical. This is a consequence of the deviation of the flow cell from an ideal geometry possibly due to the slightly tilted position of the obstruction. Figure 4.D provides evidence for the findings from CFD with respect to the relatively long distance in the outflow areas until tubular laminar flow is reached. While the simulations predict that this flow pattern is reestablished at around 7 mm after the obstruction, the NMR data prove that the laminar momentum transport profile is largely reestablished at 8.1 mm above the obstruction. However, it should be noted that the results from CFD depend critically on the position of the outflow boundaries. Further calculations, where the outflow boundary is located further downstream will need to be conducted in order to obtain unambiguous results. The nonsymmetrical shape of the deformed flow tag in figure 4.D provides also a clear indication that some deviations from the tubular flow do exist. This effect still must be caused by the deviation of the obstruction from the ideal sample geometry.

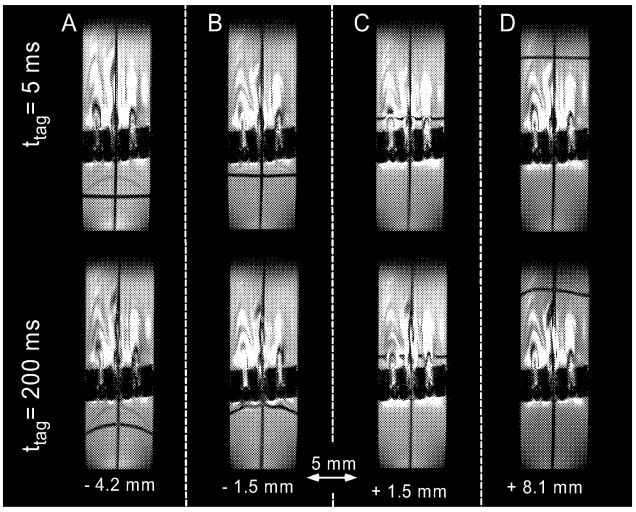


Figure 4: Tagged MR-Images as Used to Visualize Flow Phenomena in the Flow cell of Figure 2.B:
Each column depicts as series of two tagging images for t_{tag}= 5 ms (upper row) and t_{tag}= 200 ms (lower row), respectively. Columns A –D correspond to different positions above and below the obstruction to the laminar flow through the tube as labeled. Images were recorded using a Spin Echo sequence (TE = 12 ms, TR = 500 ms, slice thickness: 1 mm) preceded by a tagging sequence consisting of two sinc-shaped soft pulses with a duration of 3 ms, which were successively applied with gradients in the horizontal and vertical direction.

CONCLUSION

The above qualitative comparison between flow MRI experiments and CFD simulation provides a valuable insight into the flow process in a model system. In this study a good qualitative agreement was found between a “real” experimental flow cell and a corresponding “virtual” flow cell. A meaningful comparison between both methods is possible since the experimental flow cell and the virtual flow cell used for CFD calculations were carefully coordinated. While CFD has the potential to make very accurate predictions regarding flow patterns, flow MRI assumes the role of validating these simulations as well as providing insight to the extent, where deviations between simulations and the reality may occur for instance as a consequence of non-ideal sample geometries.
Results with respect to the inflow into a flow obstructing object are very consistent for both methods, however it is also apparent that small deviation of the experimental setup from an ideal geometry may have some effects on the flow pattern observed at the outflow of the obstruction. This finding may have a direct technical relevance for instance with respect to the dyeing or laundering process of a woven fabric structure. The front side of this fabric hit with a laminar flow is expected to experience a regular momentum transport, while the backside is more subject to irregular flow patterns.

REFERENCES

Blümich, B. 2000. *NMR Imaging of Materials.* Clarendon Press, Oxford.

Bijeljic, B; M.D. Mantle, A.J. Sederman, L.F. Gladden and T.D. Papthanasiou. 2001."Slow flow across macroscopically rectangular fiber lattices and an open region: Visualization by magnetic resonance imaging." *Physics of Fluids 12*, No.12, 3652-3663.

Casey, M. and T. Wintergerste (editors). 2000. *Quality and Trust in Industrial CFD - Best Practice Guidelines, Version 1.0., ERCOFTAC Special Interest Group.*

Farber, P. and N. Dahmen. 1998. "About Fluid Flow and Mass Transfer in Textile Finishing Machines." *ASME Fluids Engineering Division Summer Meeting,* Washington, DC, June 21 - 25.

Ferziger, J. H. and M. Peric. 2002. *Computational Methods for Fluid Dynamics.* Springer, Berlin.

Leisen, J. and H.W. Beckham. 2001. "Quantitative Magnetic Resonance Imaging of Fluid Distribution and Movement in Textiles." *Textile Research Journal* 71, No.12, 1033-1045.

Shen, H.; S. Nutt, and D. Hull. 2004. "Direct observation and measurement of fiber architecture in short fiber-polymer composite foam through micro-CT imaging." *Composites Science And Technology* 64,No.13-14, 2113-2120.

VDI-Gesellschaft Verfahrenstechnik und Chemie-ingenieurwesen (GVC). 2002. *VDI-Wärmeatlas - Berechnungsblätter für den Wärmeübergang*, Springer, Berlin.

Yao, S.; M. Costello, A.G. Fane, and J.M. Pope. 1995. "Non-invasive observation of flow profiles and polarisation layers in hollow fibre membrane filtration modules using NMR micro-imaging." *Journal of Membrane Science* 99,No.207-216.

ACKNOWLEDGEMENT

Support from the National Textile Center (U.S.A.) and the Ministry of Science of the State of Nordrhein-Westfalen (Germany) is greatly acknowledged.

USING NEURO-FUZZY FOR PREDICTION RING SPUN YARN STRENGTH FROM COTTON FIBERS PROPERTIES

A.Shams-Nateri
Assistant Professor
Textile Engineering Department Guilan University, Rasht, Iran
E-mail: a_shams@guilan.ac.ir

KEYWORDS
Cotton, Fiber, Yarn, Relationship, Neuro-fuzzy

ABSTRACT

This paper describes advance techniques that can be used to Predict ring spun yarn strength from HVI/FMT measured properties of cotton fiber such as span length, bundle strength, fineness, uniformity ratio and maturity of cotton fibers. A neuro-fuzzy network and multiple-linear regression techniques and Sitra's expressions have been used to predict yarn strength. The results conclusively prove the superiority of neuro-fuzzy method over multiple-linear regression techniques and Sitra's expressions.

INTRODUCTION

One of the most important production processes in the textile industry is the spinning process. Started with cotton fiber. Yarn has been produced in rotor or ring spinning machine. The quality of the resulting yarns is very important in determining their application possibilities. The two most important characteristics of yarn are its tenacity and elongation. Prediction formulae to related fiber properties to yarn quality have been an interesting field of work for several research workers for more than three decades (Sette et al. 1997; Ramey et al. 1977).

Sitra has developed expressions for prediction yarn strength from fiber properties in the year 1989(Chellmani et al. 1995). In the prediction expressions fiber properties measured with conventional instrument. In next step, they establish prediction expressions for strength of spun yarns from cotton fiber properties using HVI test system. The following expression has been derived for yarn Lea CSP under good working condition for carded count:

$$Lea \cdot CSP = 280 \times \sqrt{FQI} + 700 - 13C \qquad (1)$$

and for combed count:

$$Lea \cdot CSP = (280 \times \sqrt{FQI} + 700 - 13C) \times (1 + \frac{W}{100}) \qquad (2)$$

Where FQI is fiber quality index and arrived from the following formula:

$$FQI = \frac{L \times S \times m}{F} \qquad (3)$$

Where
L: 50% span length (mm)
S: Bundle strength in g/tex
F: Fineness (micronnaire value)
m: Maturity
C: yarn count
W: % waste extracted during combing

If the maturity coefficient values are not readily available, as a rough approximation the yarn CSP values may be arrived from the following formula:

$$Lea \cdot CSP = 250 \times \sqrt{\frac{LS}{F}} + 590 - 13C \qquad (4)$$

For yarn strength, RKm value defined as:

$$RKm = \frac{Single\ \ Yarn\ \ strength}{Tex\ \ Count\ \ of\ \ Yarm} \qquad (5)$$

RKm value can be predict from CSP value by following formula:

$$RKm = \frac{Lea.CSP}{150} \qquad (6)$$

Chellamani et al reported a similar relationship between fiber properties and yarn strength (Chellmani et al. 1990). The relationship between the yarn strength and fiber properties was studied and a Fiber Quality Index was derived to overcome inconsistencies within the basic fiber properties. It was then used to predict the strength of the yarn and significant correlations were obtained.

This work reports a new method for prediction ring spun yarn strength from cotton fiber properties.

ANFIS

The fuzzy inference system (FIS) is a popular computing framework based on the concepts of fuzzy set theory, fuzzy IF/THEN rule and fuzzy reasoning to transform an input space into an output space. The basic structure of fuzzy inference system consists of three conceptual components:

1. A rule base
2. Database or dictionary which defines the membership functions used in the fuzzy rules
3. Reasoning mechanism, which performs the inference procedure upon the rule and a given condition to derive a reasonable output or conclusion.

A fuzzy system can be created to match any set input/output data. This can be done with an adaptive neuro-fuzzy inference system (ANFIS). ANFIS is about taking a fuzzy inference system and training it with a

backpropagation algorithm, well known in the artificial neural network (ANN) theory, based on some collection of input/output data (Rumelhart et al. 1986; Marjoniemi et al. 1997; Jang 1993; Nariman-Zadeh et al. 2001).

Then the back-propagation algorithm can be used to adjust the weights and biases of networks in order to minimize the sum-squared error of the network. In summary, the input variables *x* and y define the input space that is partitioned into fuzzy subspaces with a given number of rules. The input MFs with unit weights is modified during training (Marjoniemi et al. 1997; Jang 1993; Nariman-Zadeh et al. 2001).

EXPERIMENTAL

We select 22 different quality cotton fiber samples. The cotton fiber properties such as spun length, fiber bundle strength, maturity; micronnaire was measured by HVI/FMT test system. Ring spun yarn has been spun in Rieter spinning system and their quality measured. The properties of cotton fiber and quality of spun yarn are shown respectively in Tables 1 and 2.

Table 1. Properties of cotton fibers

No.	Cotton fiber properties			
	50% span length (mm)	Micronnair	Maturity ratio	strength (g/tex)
1	14	2.69	0.46	20.21
2	13.54	4.6	0.86	23.38
3	13.83	4.04	0.82	18.47
4	13.76	3.79	0.87	17.99
5	11.57	3.79	0.88	20.88
6	13.03	3.47	0.75	21.11
7	13.48	4.53	0.86	22.86
8	13.38	3.39	0.66	16.65
9	13.17	3.34	0.66	16.28
10	14.77	3.69	0.81	19.77
11	13.78	3.55	0.82	18.86
12	12.6	3.21	0.72	19.58
13	14.37	5.19	0.9	18.54
14	13.38	3.61	0.72	18.58
15	13.3	3.04	0.54	19.99
16	13.97	3.58	0.66	19.26
17	13.41	3.06	0.6	19.42
18	13.26	3.33	0.63	19.36
19	13.72	3.48	0.69	19.64
20	14.76	3.56	0.76	20.2
21	13.82	3.84	0.64	19.69
22	14.72	4.31	0.86	20.4

Table 2. Properties of ring spun yarn

Sample No.	Yarn properties	
	Count (Ne)	RKm
1	41	14
2	39	16.95
3	40.2	16.56
4	40.2	17.6
5	39.5	14.57
6	39.8	15.37
7	40.6	16.24
8	40	14.47
9	40.7	17
10	40	16.3
11	39.9	15.4
12	39	15.8
13	39.2	15.7
14	40	15.07
15	29.8	16.04
16	24	13.9
17	29.5	14.78
18	29.3	15.96
19	38.8	15.47
20	40.4	14.36
21	40.6	14.43
22	28.9	16.7

Results and Discussions

At first, the relationship between fiber properties and strength of spun yarn was study by using sitra expression (Equation 1). The predicted value of yarn RKm for various samples is given in Table 3.

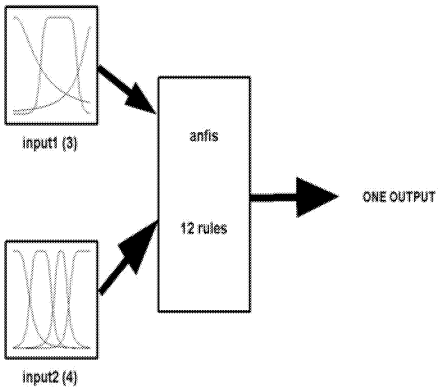
Two multiple linear regressions used for correlation yarn strength to cotton fiber properties. In first method, Lea CSP of yarn was calculated from 50% span length (L), Bundle strength in (S), Fineness (F), Maturity(M) and yarn count (C) by Equation 14:

$$RKm = 9.73 + 0.229 \times L - 0.08 \times F + 3.7 \times M - 0.027 \times S + 0.0249 \times C \quad (14)$$

In second method, RKm value of yarn was calculated from fiber quality index (FQI) and yarn count (C) by equation 15:

$$RKm = 12 + 0.0527 \times C + 0.0308 \times FQI \quad (15)$$

In neuro-fuzzy method, ANFIS system applied to prediction yarn's strength from cotton fibers properties. ANFIS system was designed with two input, 81 IF-THEN rule and one output(Figure 1). Two input nods referred to fiber quality index (FQI) and yarn count. one output nod referred to spun yarn strength. First input had three Gaussian shape membership function and second input had four Gaussian shape membership function. Network was trained by back-propagation algorithm. After training, the neural network has been tested with all samples. The predicted results are shown in Tables 3 and 4.



Figures 1. ANFIS system structure used for prediction yarn strength from cotton fibers properties

Table 3. Actual and predicted RKm value of ring spun yarns by Sitra method and neuro-fuzzy method

Sample No.	Actual Value	Sitra Method	Neuro-fuzzy Method
1	14	14.76	14.00
2	16.95	16.31	16.95
3	16.56	15.29	16.56
4	17.6	15.92	17.60
5	14.57	15.89	14.57
6	15.37	16.28	15.41
7	16.24	16.09	16.24
8	14.47	14.16	14.47
9	17	13.96	17.00
10	16.3	16.81	16.30
11	15.4	16.34	15.36
12	15.8	15.84	15.78
13	15.7	14.62	15.70
14	15.07	15.01	15.07
15	16.04	15.58	16.05
16	13.9	16.40	13.90
17	14.78	16.12	14.78
18	15.96	15.80	15.95
19	15.47	15.61	15.48
20	14.36	16.72	14.00
21	14.43	14.39	16.95
22	16.7	17.28	16.56

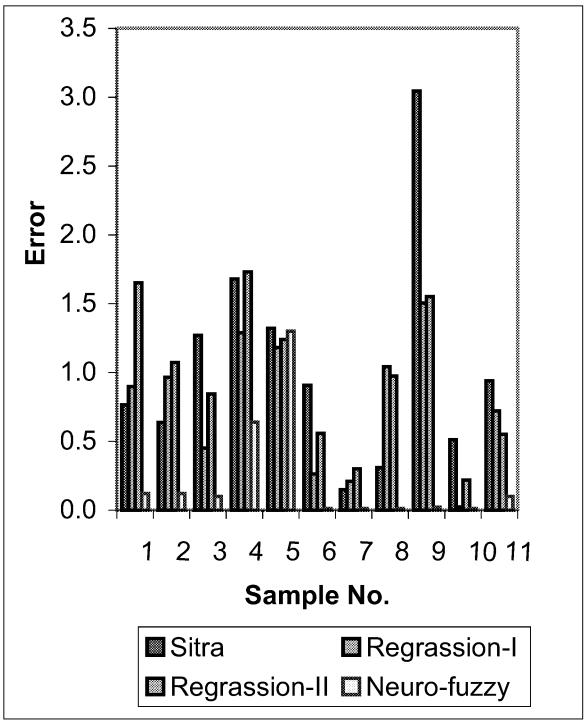
Table 4. Actual and predicted RKm value of ring spun yarns by multiple linear regressions

Sample No.	Actual Value	Regression Method I	Regression Method II
1	14	14.90	15.65
2	16.95	15.98	15.88
3	16.56	16.11	15.72
4	17.6	16.31	15.87
5	14.57	15.75	15.81
6	15.37	15.63	15.93
7	16.24	16.03	15.94
8	14.47	15.51	15.44
9	17	15.49	15.45
10	16.3	16.28	16.08
11	15.4	16.12	15.95
12	15.8	15.47	15.76
13	15.7	16.41	15.49
14	15.07	15.66	15.64
15	16.04	14.73	15.03
16	13.9	15.16	14.79
17	14.78	14.99	15.13
18	15.96	15.04	15.04
19	15.47	15.58	15.69
20	14.36	16.10	16.09
21	14.43	15.43	15.54
22	16.7	16.11	15.37

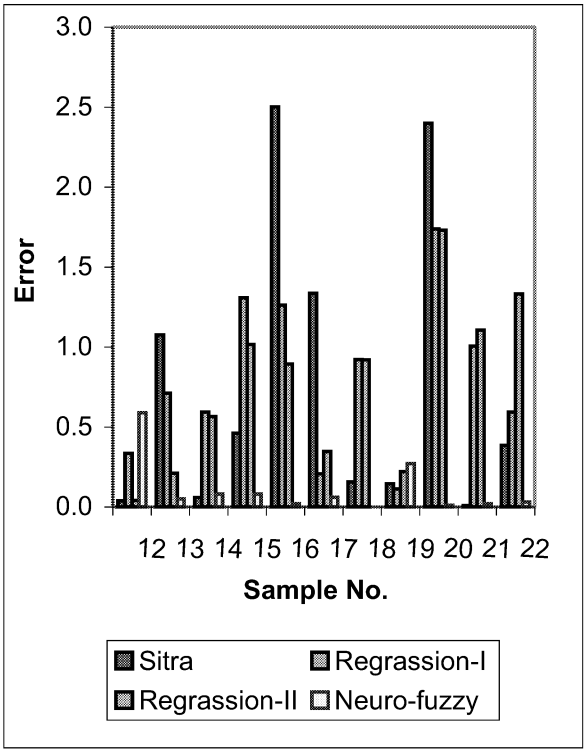
Samples number 1 to 19 has been used for calculation formulae and neural network training and testing. Samples number 20 to 22 only has been used for testing each method. The prediction error of all methods is shown in figures 2 and 3. Detected results also were evaluated by sum of error as different between actual and predicted values. The summation of error in Sitra method was 20.09, in first regression method was 17.33, in second regression method was 19.07, and in neuro-fuzzy method was 3.47.

CONCLUSIONS

In this work, a new advanced technique has been used to predict ring spun yarn strength from HVI/FMT measured properties of cotton fiber. The results show better agreement between the actual and predicted values in neuro-fuzzy method than regression and Sitra method.



Figures 2. Compare absolute error in different methods



Figures 3. Compare absolute error in different methods

ACKNOWLEDGEMENT

The author would like to express their grateful thanks to Babakan Textile Corporation and Mr. N. Shigh-ol-Eslami for their help in processing the samples.

REFERENCES

Sette, S.; L. Boullart ; L. Van Langenhove and P. Kiekens 1997. " Optimizing the Fiber-to-Yarn Production Process With a Combined Neural Network/Generic Algorithm Approach". Textile Research Journal, 67(2), 84-92.

Ramey H.H. ; R. Lawson R. and S.Worley 1977. " Relationship of Cotton Fiber Properties to Yarn Tenacity". Textile Research Journal, 47(10), 685-691.

Chellmani K.P. ; K. Gnanasekar ; M.S. Ravindran and T.V. Ratnam 1995. " Fiber Yarn Relationships Using HVI/FMT Measured Fiber Properties". The South India Textile Research Association (SITRA), Vol.41, No. 1.

Chellamani P. ; Indra Doraiswamy; T.V. Ratnam 1990. " Fiber Quality and Yarn Strength Relationships". Indian Journal of Fiber & Textile Research, Vol. 15, 1-5.

Rumelhart G. E. ; G.E. Hilton and R.G. Williams 1986. " Learning Internal Representation by Error Propagation, in: Parallel Distributed Processing". Vol.1, Chp. 8, Cambridge, MA, MIT. Press .

Marjoniemi M. and E. Mantysalo 1997. " Neuro-Fuzzy Modeling of Spectroscopic Data. Part A: Modeling of Dye Solutions". J.S.D.C., Vol.113, 13-17.

Marjoniemi M. and E. Mantysalo 1997. " Neuro-Fuzzy Modeling of Spectroscopic Data. Part B: Dye Concentration Prediction", J.S.D.C., Vol.113, 64-67.

Jang J.S.R. 1993. " ANFIS: Adaptive-network-based fuzzy Inference System". IEEE Trans. On Sys., Man. Cyb., 23.

Nariman-Zadeh N. and a. Darvizeh 2001. " Design of Fuzzy System for the Modeling of Explosive Cutting Process of Plates Using Singular Value Decomposition". WSES 2001 Conf. On fuzzy sets and fuzzy systems (FSFS, 01) , spain(Feb).

A NEW TECHNIQUE TO MEASURE RAYLEIGH'S VISCOSITY PARAMETERS OF TEXTILE

Hatem Charfi and André Gagalowicz
INRIA Rocquencourt
Domaine de Voluceau Rocquencourt
B.P. 105 - 78153 Le Chesnay Cedex - France
E-mail: Hatem.Charfi@inria.fr
E-mail: Andre.Gagalowicz@inria.fr

KEYWORDS
Damping, Rayleigh model, parameter identifiaction, textile materials

ABSTRACT

In this paper, we present an experimental work using MOCAP system allowing to compute damping parameters and to measure their contribution for the simulation of cloth in free fall movement.
Energy damping is an important phenomenon to consider for the 3D simulation of warp and weft materials, since it has a great influence on the movement realism.
This phenomenon can be generated either by friction between moving cloth and air, or by friction between the warp and the weft of the fabric.
The contribution of this paper is to determine viscous parameters of cloth using precise trajectory data of a real cloth.

INTRODUCTION

Cloth simulation is becoming an important research domain in computer graphics. In fact, it is involved in a wide range of applications such as movie films including synthetic actors, video games and online garment trade. A great deal of work on simulating the motion of cloth, and generally of fabric, has already been done (Terzopoulos and al. 1987)(Eischen and Bigliani 2000)(Volino and al.1995) and several cloth simulators have been developed (Baraff and Witkin 1998)(Breen and al. 1994)(Choi and Ko 2002). The motion of fabric is determined by its resistance to bending, stretching, shearing and by aerodynamical effects such as friction and collisions (Bridson and al. 2002). It is also determined by how energy is damped.
Realism of a simulation is usually used as a criterion to evaluate the accuracy of simulation and energy damping plays an important role in this search of realism (Bhat and al. 2003).
This paper presents a new experimental way to determine damping parameters in cloth simulation.

FABRIC MODEL

We model fabric (limited to warp and weft textile materials) using a mass-spring system (Provot 1995). Thus, cloth is approximated with a quadrilateral mesh of particles. Various types of springs are used to connect masses to each others. There are three types of springs :

- Tensile springs : connect two neighbor masses in the warp (or weft) direction.
- Shearing springs : connect two neighbor masses in the diagonal direction to the warp (or to the weft).
- Bending springs : connect four masses forming two adjacent facets.

The tensile and shearing springs forces are in the plane of the fabric, whereas the bending springs forces are orthogonal to the plane of the fabric.
These springs have to be fed with correct parameters to meet the realism that we look for simulation.
The Kawabata Evaluation System (Kawabata 1980) is widely used as the industry standard for fabrics measurements (limited to warp and weft textile materials): it operates typical deformations on cloth samples and evaluates the resulting forces, which provide us with the parameters to put on the springs.

Damping Model

The damping model used is the Rayleigh damping model. Its mathematical formula is :

$$[C] = \alpha[M] + \beta[K] \quad (1)$$

where $[C]$ is the damping matrix (n x n), $[M]$ is the mass diagonal matrix (n x n), $[K]$ is the stiffness matrix (n x n), α and β are the damping constants, and n is the total number of masses used to model the fabric.
However, our mechanical model uses 3 different types of springs. So, the stiffness matrix $[K]$ is decomposed as the

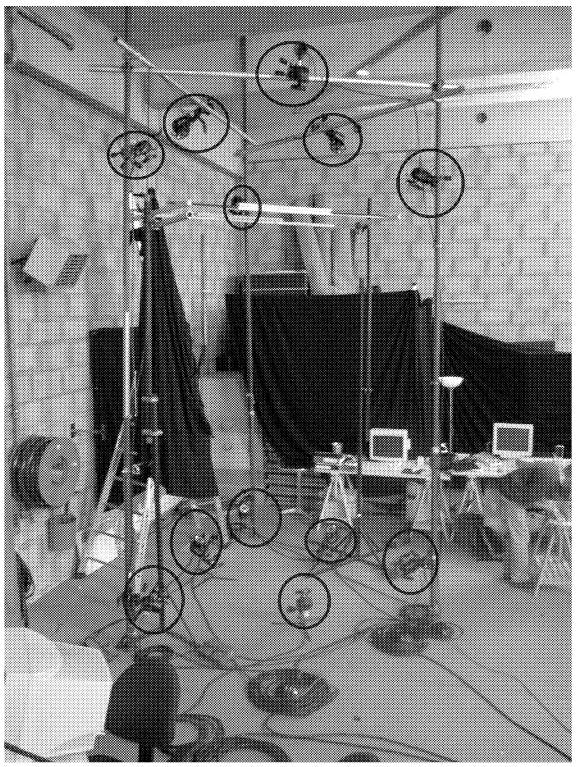


Figure 1: *12 cameras of the Motion Capture System (MOCAP)*

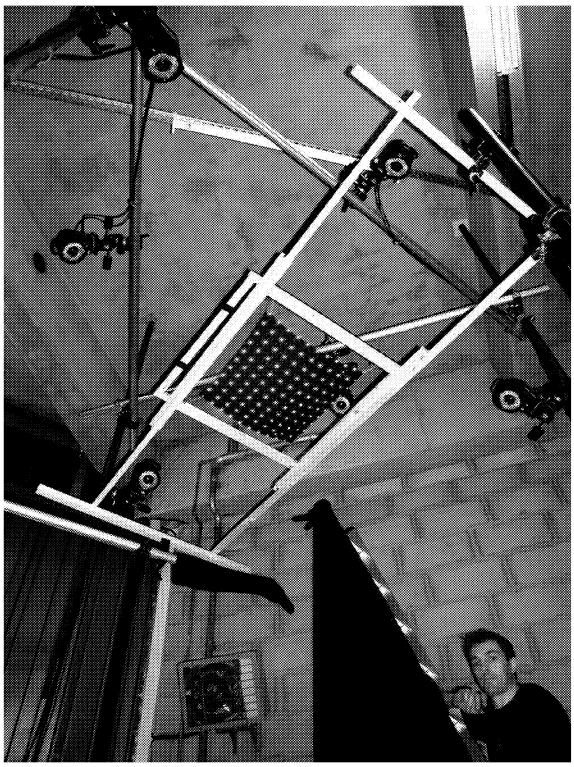


Figure 2: *A sample of fabric with reflective markers*

sum of 3 stiffness matrices.

$$[K] = [K_b] + [K_{sh}] + [K_t]$$

Equation (1) becomes :

$$[C] = \alpha[M] + \beta_b[K_b] + \beta_{sh}[K_{sh}] + \beta_t[K_t]$$

The linearity of Rayleigh's model makes it possible to derive the equation above. The computation of the 3 constants β_b, β_{sh} and β_t allows the computation of the damping force for each spring type.
The determination of the constant α allows the computation of damping produced by the movement of masses in the air, i.e. the viscous damping between the fabric and the air.
The total damping force is :

$$F_{damp} = [C]V$$

where V is the velocity vector of all masses.

EXPERIMENTAL SETUP

The experiment performed in collaboration with ATTITUDE-STUDIO consists in dropping a piece of fabric in free fall and measuring its trajectory using a motion capture system (MOCAP) (see figure (1)).
The viscous parameters are then obtained by the adjustment of the simulated trajectory of this fabric computed by our simulator, to the real trajectory. A sample of 50cm by 50cm of a fabric (woven in warp/weft) is cut. Reflective round markers are stuck on both sides of the fabric in order to form a regular grid.
Then, the fabric is thrown in a free fall and the MOCAP system starts recording the successive positions of the markers (see figure (2)).

DAMPING PARAMETERS IDENTIFICATION

Given the data collected by the MOCAP, it is possible to compute the speed of each mass i and its acceleration (by finite differences). The fundamental principle of dynamics (F.P.D) is then written for each mass i:

$$\forall i, m_i A_i = F_i$$

where F_i is the sum of external forces applied on mass i.

Identification Of The Parameter Of Viscous Damping With The Air

In order to make this identification, all springs are omitted. In fact, the viscous damping force between fabric and air is applied only on masses (springs have a null weight).
Writing the F.P.D for each mass i, we obtain the following equation :

$$m_i A_i = m_i g + F_{damp}$$

where g is the gravity and F_{damp} the viscous damping force of the air. F_{damp} can be written using the Rayleigh model :

$$F_{damp} = \alpha_i m_i V_i$$

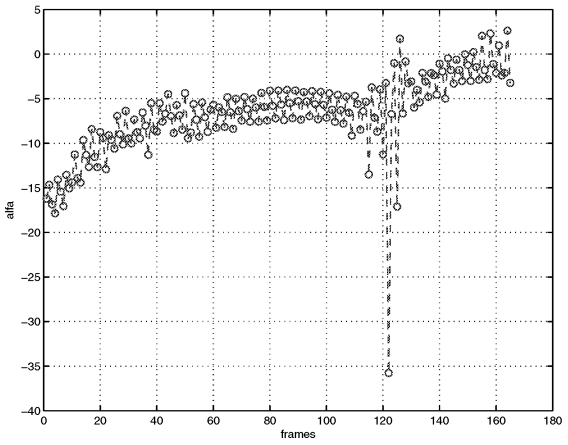


Figure 3: *Viscous damping parameter of the air per frame*

Let

$$F_{error} = m_i A_i - m_i g$$

We have to compute the α that minimizes

$$\Phi(\alpha) = \| F_{error} - F_{damp} \|^2$$

$$\begin{aligned} \Phi(\alpha) &= \| F_{error} - \alpha_i m_i V_i \|^2 \\ \frac{d\Phi}{d\alpha} &= -2 * ((F_{error} - \alpha_i m_i V_i).m_i V_i) \\ &= 0 \\ \alpha_i &= \frac{(F_{error}.V_i)}{m_i \| V_i \|^2} \end{aligned}$$

So, for each mass and for each frame, we obtain an α_i.
As the textile material is homogeneous, all α_i are equal and do not depend on the speed.
So, we compute α_f for each frame as the mean of the α_i of this frame and then, we compute α as the mean of the α_f in the *viscous* part of the movement.

In fact, let's analyse the example shown in figure (3). The part of the movement between the frames 0 and 40 corresponds to the beginning of the free fall movement. The speed of the fabric is still very low and the movement of the fabric is still polluted by the launch (very noisy data).
Beyond frame 100, the fabric has a chaotic turbulent movement and the interaction type between the air and the fabric can no longer be modelled using the Rayleigh model.
So, we compute α as the mean of the α_f in the *viscous* part of the movement, ie between the frames 40 and 100.
Indeed, in this part of the movement, the fabric has already acquired a minimum speed that enables us to measure a force of viscous friction with the air (since this force is proportional to the speed of the masses).
In addition, we observe that the movement of the fabric on the video is slowed down in this part without having turbulent or chaotic movements, and we know that α_f does not depend on the frame, so we have to restrict the computation to the horizontal part of figure (3).

Identification Of The Parameter Of Viscous Damping Of Bending Springs

After computing the viscous friction parameter α between the fabric and the air, we include the bending springs in the simulation.
Hence, the model evolves and allows to take into account forces between two adjacent facets.
A bending spring connects 2 adjacent facets (4 masses) and models the reaction of fabric to bending. Bending forces are very weak compared to tension forces or shearing forces, so errors induced by bending springs are much smaller as well. That is why we have added these springs first to the model (before adding shearing and tensile springs).
We write the F.P.D for each mass i:

$$m_i A_i = m_i g + \alpha m_i V_i + F_i^b + F_{damp}(i)$$

where $F_{damp}(i)$ is the viscous damping force of the bending springs.
F_{damp} has this expression :

$$F_{damp}(i) = \beta_i^b (K_b V)(i) \tag{2}$$

where $K_b = \frac{dF^b}{dP}$, F^b is the vector of forces produced on masses by the bending springs and P is the position vector of all masses.
Let $F^b(P_i)$ be the vector of forces produced on mass i by the bending springs.

$$F^b = \sum_i F^b(P_i)$$

$$F^b(P_i) = \sum_r F_r^b(P_i)$$

where $F_r^b(P_i)$ is the vector of forces produced on mass i by the bending springs connected to mass i.

$$F_r^b(P_i) = \mathcal{M}_r^{Kaw} \frac{d\theta}{dP_i}$$

where $\mathcal{M}_r^{Kaw}$ is the torque intensity produced by the spring, given by *Kawabata* and θ is the angle between the two facets of the bending spring r.
So, K_b is a 3n by 3n matrix whose (i, j) bloc (3 by 3) is:

$$\frac{dF^b(P_i)}{dP_j} = \sum_r \frac{dF_r^b(P_i)}{dP_j}$$

where

$$\frac{dF_r^b(P_i)}{dP_j} = \mathcal{M}_r^{Kaw} \frac{\partial^2 \theta}{\partial P_i \partial P_j} + \frac{\partial \mathcal{M}_r^{Kaw}}{\partial \theta} \left(\frac{d\theta}{dP_i}\right) \left(\frac{d\theta}{dP_j}\right)^T$$

The equation (2) shows that the computation of the damping force of a bending spring on a mass i uses proporties of other masses (its neighbors).
Thus, in order to compute the ith element of vector $K_b V$ using the matrix K_b and the vector V, we have to start computing the product $K_b V$ (a partial bloc computation is much more complicated to do in this case).

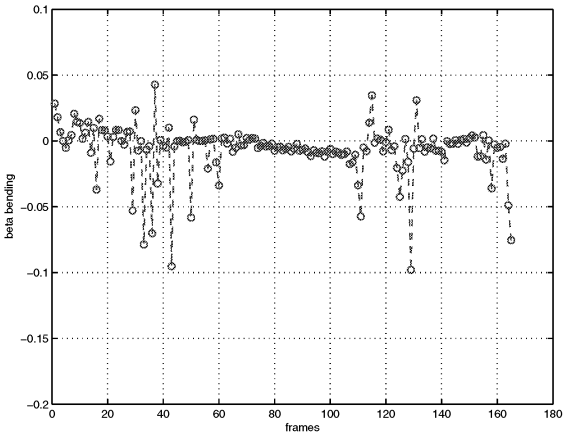


Figure 4: *Viscous damping parameter of bending springs*

The approach will be slightly modified comparatively to the one used for computing α (viscous damping parameter of the air), since we will directly search for a β^b for each frame.
Let

$$F_{error} = MA - Mg - \alpha MV - \beta^b K_b V - F^b$$

where M is the matrix of masses, A the acceleration vector and F^b is the bending force vector.
We search for β^b that minimizes

$$\Phi(\beta) = \| F_{error} - F_{damp} \|^2$$

$$\begin{aligned} \Phi(\beta) &= \| F_{error} - \beta^b K_b V \|^2 \\ \frac{d\Phi}{d\beta} &= -2 * ((F_{error} - \beta^b K_b V).K_b V) \\ &= 0 \\ \beta^b &= \frac{(F_{error}.K_b V)}{\| K_b V \|^2} \end{aligned}$$

Figure (4) shows the β^b value found for each frame.
Between frames 0 and 40, the fabric has a free fall movement. However, its speed is not high enough to observe any phenomenon related to viscosity. In fact, figure (4) shows that values of β^b are oscillating around 0 in this area.
Beyond frame 100, the fabric has a turbulent and chaotic movement and the Rayleigh model is not able to model precisely the interactions in this area.
Consequently, β^b is computed as before, as the mean of β^b in the *viscous* part of the movement, ie between frames 40 and 100. (see figure (4))

Identification Of The Parameter Of Viscous Damping Of Shearing Springs

we use the already determined parameters α and β^b to compute the air and bending springs damping forces. We include the shearing springs in the fabric model.
We write the F.P.D for each mass i:

$$m_i A_i = X + F_i^{sh} + F_{damp}(i)$$

where

$$X = m_i g + \alpha m_i V_i + F_i^b + \beta^b (K_b V)(i)$$

and $F_{damp}(i)$ is the viscous damping force of the shearing springs.
F_{damp} has this expression :

$$F_{damp}(i) = \beta_i^{sh} (K_{sh} V)(i)$$

where $K_{sh} = \frac{dF^{sh}}{dP}$, F^{sh} is the vector of forces produced on masses by the shearing springs and P is the masses position vector of all masses.
Let $F^{sh}(P_i)$ be the vector of forces produced on mass i by the shearing springs.

$$F^{sh} = \sum_i F^{sh}(P_i)$$

and

$$F^{sh}(P_i) = \sum_r F_r^{sh}(P_i)$$

where $F_r^{sh}(P_i)$ is the vector of forces produced on mass i by the shearing springs connected to mass i.

$$F_r^{sh}(P_i) = F_r^{Kaw} \frac{ds}{dP_i}$$

where F_r^{Kaw} is the spring force intensity given by *Kawabata* and s is the stretch of the spring r.

$$s = l - l_0$$

where l is the spring length, and l_0 its free length.
So, K_{sh} is a 3n by 3n matrix whose (i, j) bloc (3 by 3) is:

$$\frac{dF^{sh}(P_i)}{dP_j} = \sum_r \frac{dF_r^{sh}(P_i)}{dP_j}$$

where

$$\frac{dF_r^{sh}(P_i)}{dP_j} = F_r^{Kaw} \frac{\partial^2 s}{\partial P_i \partial P_j} + \frac{\partial F_r^{Kaw}}{\partial s} \left(\frac{ds}{dP_i}\right) \left(\frac{ds}{dP_j}\right)^T$$

We will use the same approach as that one used for computing β^b. Let's start by computing a β^{sh} for each frame.
Let

$$F_{error} = F_{error}^b - \beta^{sh} K_{sh} V - F^{sh}$$

where

$$F_{error}^b = MA - Mg - \alpha MV - \beta^b K_b V - F^b$$

We search for the β^{sh} that minimizes

$$\Phi(\beta) = \| F_{error} - F_{damp} \|^2$$

$$\begin{aligned} \Phi(\beta) &= \| F_{error} - \beta^{sh} K_{sh} V \|^2 \\ \frac{d\Phi}{d\beta} &= -2 * ((F_{error} - \beta^{sh} K_{sh} V).K_{sh} V) \\ &= 0 \\ \beta^{sh} &= \frac{(F_{error}.K_{sh} V)}{\| K_{sh} V \|^2} \end{aligned}$$

Figure (5) shows the β^{sh} value found for each frame. We compute β^{sh} as the mean of β^{sh} in the *viscous* part of the movement, ie between frames 40 and 100.

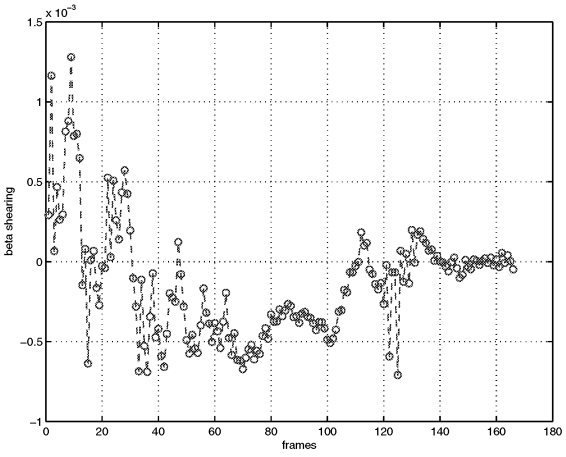


Figure 5: *Viscous damping parameter of shearing springs*

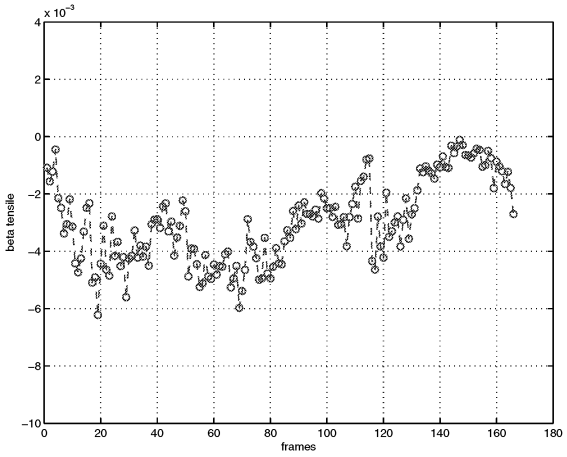


Figure 6: *Viscous damping parameter of tensile springs*

Identification Of The Parameter Of Viscous Damping Of Tensile Springs

We use the already determined parameters α, β^b and β^{sh} to compute the air, bending springs and shearing springs damping forces. We include the tensile springs in the new textile model.
We write the F.P.D for each mass i:

$$m_i A_i = X + Y + F_i^t + F_{damp}(i)$$

where

$$X = m_i g + \alpha m_i V_i + F_i^b + \beta^b (K_b V)(i)$$

$$Y = F_i^{sh} + \beta^{sh}(K_{sh}V)(i)$$

and $F_{damp}(i)$ is the viscous damping force of the tensile springs.
F_{damp} has this expression :

$$F_{damp}(i) = \beta_i^t (K_t V)(i)$$

where $K_t = \frac{dF^t}{dP}$, F^t is the vector of forces produced on masses by the tensile springs and P is the masses position vector of all masses.
Let $F^t(P_i)$ be the vector of forces produced on mass i by the tensile springs.

$$F^t = \sum_i F^t(P_i)$$

and

$$F^t(P_i) = \sum_r F_r^t(P_i)$$

where $F_r^t(P_i)$ is the vector of forces produced on mass i by the tensile springs connected to mass i.

$$F_r^t(P_i) = F_r^{Kaw} \frac{ds}{dP_i}$$

where F_r^{Kaw} is the spring force intensity given by *Kawabata* and s is the stretch of the spring r.

$$s = l - l_0$$

where l is the spring length, and l_0 its free length.
So, K_t is a 3n by 3n matrix whose (i, j) bloc (3 by 3) is:

$$\frac{dF^t(P_i)}{dP_j} = \sum_r \frac{dF_r^t(P_i)}{dP_j}$$

where

$$\frac{dF_r^t(P_i)}{dP_j} = F_r^{Kaw} \frac{\partial^2 s}{\partial P_i \partial P_j} + \frac{\partial F_r^{Kaw}}{\partial s} \left(\frac{ds}{dP_i}\right) \left(\frac{ds}{dP_j}\right)^T$$

We use the same approach as that one used for computing β^{sh}. Let's start by computing a β^t for each frame.
Let

$$F_{error} = F_{error}^{sh} - \beta^t K_t V - F^t$$

where

$$F_{error}^{sh} = MA - Mg - \alpha MV - \beta^b K_b V - F^b - \beta^{sh} K_{sh} V - F^{sh}$$

We search for the β^t that minimizes

$$\Phi(\beta) = \| F_{error} - F_{damp} \|^2$$

$$\begin{aligned} \Phi(\beta) &= \| F_{error} - \beta^t K_t V \|^2 \\ \frac{d\Phi}{d\beta} &= -2 * ((F_{error} - \beta^t K_t V).K_t V) \\ &= 0 \\ \beta^t &= \frac{(F_{error}.K_t V)}{\| K_t V \|^2} \end{aligned}$$

Figure (6) shows the β^t value found for each frame. As for the shearing part, we compute β^t as the mean of β^t in the *viscous* part of the movement, ie between frames 40 and 100.

RESULTS

Using the optimized parameters found, we can evaluate the improvement of the simulation.
We can first compute the error force without taking into acount damping $F_{error}^{without-damp}$ at each step of the identification, and compare it with $F_{error}^{with-damp}$.

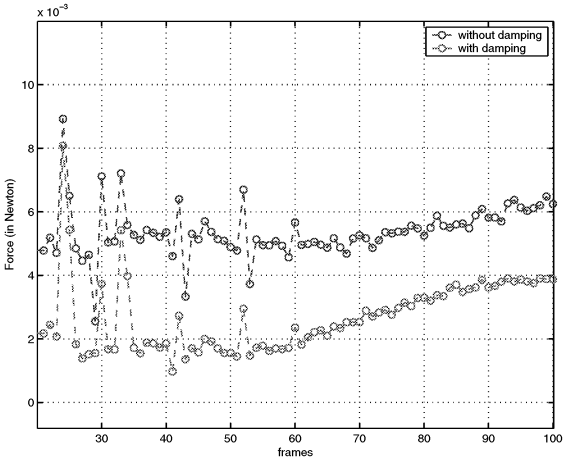


Figure 7: *Comparaison between* $\|F_{error}^{without-damp}\|$ *and* $\|F_{error}^{with-damp}(airdamping)\|$

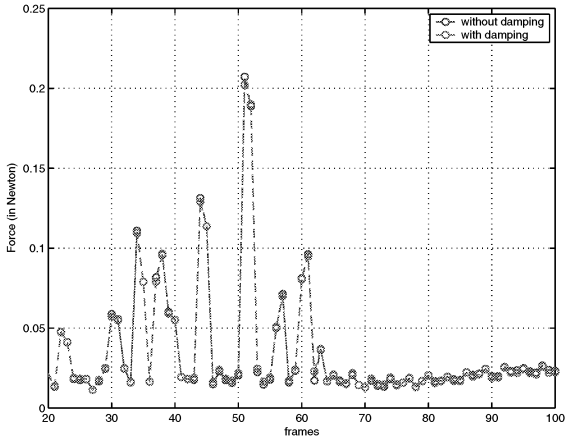


Figure 8: *Comparaison between* $\|F_{error}^{without-damp}\|$ *and* $\|F_{error}^{with-damp}\|$ *(bending)*

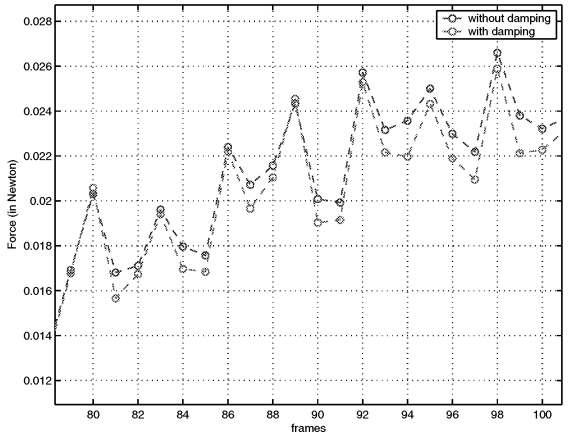


Figure 9: *Zooming : Comparaison between* $\|F_{error}^{without-damp}\|$ *and* $\|F_{error}^{with-damp}\|$ *(bending)*

Air Damping

At the first step of the identification (Damping parameter of the air), we can compute the error force including damping

$$F_{error}^{with-damp} = m_i A_i - m_i g - \alpha m_i V_i$$

and compare it with the error force without damping

$$F_{error}^{without-damp} = m_i A_i - m_i g$$

(see figure (7)).
We can notice that the norm of $F_{error}^{with-damp}$ is smaller than the norm of $F_{error}^{without-damp}$, which validates our work. According to figure (7), we notice that damping due to viscous friction with the air allows to decrease the error of about 50% on average for this example.

Bending Spring Damping

We use the already determined parameter α. We compute the error force taking into account the bending spring damping and compare it with the error force without damping. (see figure (8))
Figure (8) shows that the difference between the two error forces is very small. If we zoom on the area located between frames 80 and 100 (figure 9), we notice that the bending spring damping really brings an improvement (decrease of the error) though it is very small.
According to this analysis, we can neglect the bending springs viscous damping while simulating the fabric movement in order to decrease the simulation time.

Shearing Spring Damping

We use the already determined parameters α and β^b. We can compute the error force taking into account the shearing spring damping and compare it with the error force without damping. Figure (10) shows that the difference between the two error forces is very small between frames 40 and 50. In this part of the movement, the shearing damping does not give any improvement to the simulation.
Between frame 50 and frame 100, the $F_{error}^{with-damp}$ norm is always smaller than $F_{error}^{without-damp}$ norm. On average, shearing spring viscous damping decreases the error by 3%.
In this case, we cannot conclude whether the use of shearing spring viscous damping improves the realism of cloth simulation or not.

Tensile Spring Damping

We use the already determined parameters α, β^b and β^{sh}. We can compute the error force taking into account the tensile spring damping and compare it with the error force without damping
Figure (11) shows that tensile spring viscous damping decreases the error almost for each frame between frames 40 and 100.
The error decrease is most important between frames 50 and 90. On average, tensile spring viscous damping decreases the error by 9%.

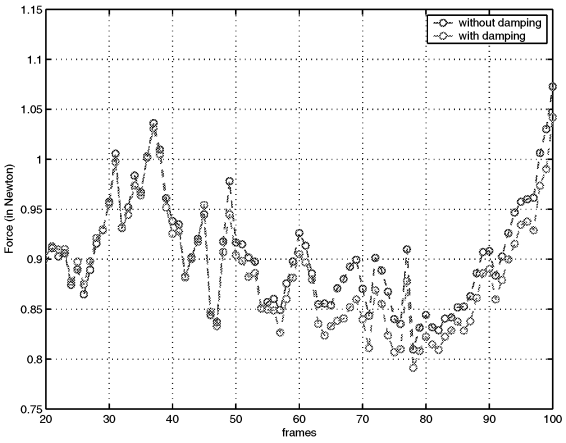


Figure 10: *Comparaison between* $\|F_{error}^{without-damp}\|$ *and* $\|F_{error}^{with-damp}\|$ *(shearing)*

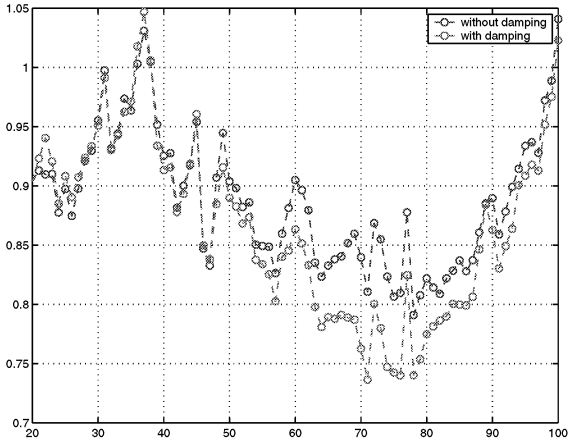


Figure 11: *Comparaison between* $F_{error}^{without-damp}$ *and* $F_{error}^{with-damp}(tensile)$

Other Results Summary

The same experiments have been done with other types of fabrics.
The results are summed up in tables 1 and 2.

Table 1: Damping parameters

	α		β^{bend}		β^{shear}		$\beta^{tensile}$	
	mean	std	mean	std	mean	std	mean	std
fabric 11	-7.0	1.9	-6.8e-3	2.0e-2	-3.1e-4	2.9e-4	-3.9e-6	4.2e-6
fabric 12	-5.9	2.4	-5.2e-3	1.5e-2	-4.0e-4	4.0e-4	-2.7e-6	3.0e-6
fabric 13	-7.2	2.2	-3.1e-4	2.0e-3	-1.8e-4	4.6e-4	-4.0e-6	5.3e-6
fabric 21	-7.2	2.3	-2.0e-4	5.7e-4	-4.3e-4	6.7e-4	-4.6e-7	5.7e-7
fabric 31	-7.4	1.1	-8.4e-4	3.1e-3	-1.1e-3	6.1e-4	-2.2e-8	6.0e-8

Simulation

We have also used the damping parameters found for our study fabric, and we have simulated a free fall movement using a cloth simulator. Then we have compared the position of the simulated piece of fabric with the real one.
We observe that in the viscous part of the movement, the simulated cloth follows faithfully the trajectory of the real cloth. see figure (12)

Table 2: Error decrease using optimized parameters

	α error decrease	β^{bend} error decrease	β^{shear} error decrease	$\beta^{tensile}$ error decrease
fabric 11	50%	0.3%	3%	9%
fabric 12	45%	0.3%	1.1%	6%
fabric 13	51%	0.1%	1.4%	3.3%
fabric 21	48%	0%	2.3%	0.5%
fabric 31	72%	0.2%	36%	0.8%

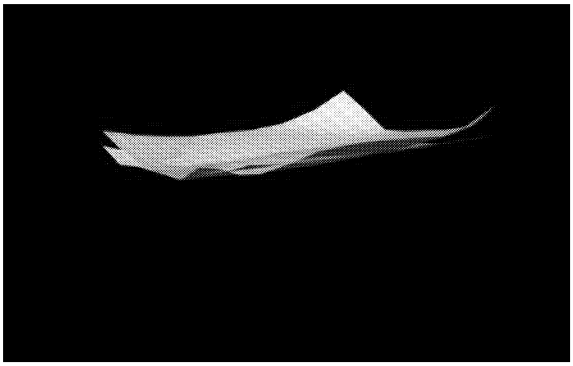


Figure 12: *Comparaison between the real fabric (yellow) and the simulated one (green) falling down*

DISCUSSION

This paper describes experiments allowing the measurement of damping parameters for cloth simulation, in the case of warp and weft materials.
We captured the behavior of pieces of fabric in a free fall movement using a MOCAP system.
Then an optimization framework was used in order to compute damping parameters of the fabric.
The validation of these measurements was performed by comparing real and simulated fabric free falls.

1. **Air damping:** results obtained for the damping parameter α of the fabric with the air show that we can use an average value equal to -7. When using an optimal α, we can decrease the error (numerical error made by ignoring air damping) by 50%.
   So, air damping has an important influence on the realism of cloth simulation.

2. **Bending damping:** results obtained for bending damping parameter β^{bend} show that bending damping does not decrease the error. So, there is no difference if we add bending damping or not in the simulation. Thus, we will ignore bending damping in future cloth simulation which allows some gain in computing time.

3. **Shear and Tensile dampings:** These parameters model an inner phenomenon in the fabric, so they depend on the mechanical properties of the fabric. Fabric 11, 12 and 13 in tables 1 and 2 are three experiments with the

same fabric. We notice, that the computed (β^{shear}) and ($\beta^{tensile}$) have almost the same value.
Shear and tensile damping decrease the error force, so they add some realism to coth simulation. However, the amount of error decrease depends on the experiment.

REFERENCES

Baraff D.; and A.P. Witkin. 1998. Large steps in cloth simulation. In *Proceedings of SIGGRAPH 98, Computer Graphics Proceedings*.

Bhat K.; C. Twigg; J. Hodgins; P. Khosla; Z. Popović; and S. Seitz. 2003. Estimating cloth simulation parameters from video. In *Proceedings of ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA 2003)*.

Breen D.; D. House; and M. Wozny. 1994. Predicting the drape of woven cloth using interacting particles. In *Proceedings of SIGGRAPH 94, Computer Graphics Proceedings*.

Bridson R.; R. Fedkiw; and J. Anderson. 2002. Robust treatment of collisions, contact and friction for cloth animation. *ACM Transactions on Graphics (ACM SIGGRAPH 2002)*.

Choi K.J.; and H.S. Ko. 2002. Stable but responsive cloth. *ACM Transactions on Graphics (ACM SIGGRAPH 2002)*.

Eischen J.; and R. Bigliani. 2000. Continuum versus particle representations. In *Cloth Modeling and Animation*. A.K. Peters.

Kawabata S. 1980, *The standardization and analysis of hand evaluation*. The Textile Machinery Society of Japan.

Provot X. 1997. *Animation réaliste de vêtements*. PhD thesis, Université de Paris 5.

Terzopoulos D.; J. Platt; A. Barr; and K. Fleischer. 1987. Elastically deformable models. In *Computer Graphics (Proceedings of ACM SIGGRAPH 87)*.

Volino P.; M. Courchesne; and N. Magnenat-Thalmann. 1995. Versatile and efficient techniques for simulating cloth and other deformable objects. In *Proceedings of ACM SIGGRAPH 95*.

GA for Optimisation of Fibre Steering in Composite Laminates

Xavier Legrand*, Don Kelly+, Alan Crosky+

*Laboratoire de Génie et Matériaux Textiles (GEMTEX)
Ecole Nationale Supérieure des Arts et Industries Textiles (ENSAIT)
9 rue de l'Ermitage BP 30329, 59056 Roubaix, France

+ University of New South Wales (UNSW)
Department of Mechanic
School of Materials and Mechanical Engineering
NSW, Australia

xavier.legrand@ensait.fr, d.kelly@unsw.edu.au , a.crosky@unsw.edu.au

KEYWORDS

Genetic Algorithms, Optimization, Bolted joints, Finit Elements.

ABSTRACT

A genetic algorithm is applied to define the optimal orientation of fibres in a uni-directional laminate in which the fibres are allowed to vary continuously across the domain. The domain is divided into two-dimensional finite elements and anisotropic properties corresponding to a carbon fibre laminate with all layers aligned in the zero element axis direction are applied to the laminate. The orientation of the material axis on each element is then prescribed as an independent variable for the Genetic Algorithm.

1- Introduction

Fibre steering techniques have been applied to fibre reinforced polymer laminate structures to increase the bearing strength of bolted composite joints [1], to increase the stiffness of a C-section beam of carbon fiber composite [2], and to improve the strength of a panel with an internal cutout subjected to a shear load [3]. In these applications the fibres have been aligned either with the principal stress trajectories or with load paths aligned with the total stress vectors defined in [4]. A finite element analysis is executed with a quasi-isotropic laminate and the principal stress or total stress vectors are plotted. A fourth order Runge Kutta algorithm is then used to define a contour tangent to the stress vectors equivalent to a particle trace parallel to velocity vectors in fluid flow. These contours define the trajectories for the fibres and are then returned to the FEA for refined analysis or digitised for a fibre laying robot [5].

In this paper a Genetic Algorithm has been programmed to define the optimal orientation of fibres. The domain is divided into two-dimensional finite elements and anisotropic properties corresponding to a uni-directional carbon fibre laminate are defined for the plate. The orientation of the material axes for each element are then set as variables for the genetic algorithm. The fitness criterion is then set in terms of strain energy and the population is allowed to evolve to define the orientation of the fibres. In this application the fibre orientation is assumed constant on each element and the vector showing the orientation on the element is defined at the centroid of the element. The orientation of the material axis on each element is then prescribed as an independent variable for the Genetic Algorithm. The fibre trajectories can then be defined using the Runge Kutta algorithm.

The GA is optimisation technique based on a simulation of Darwin's theory of life evolution. From a randomised population, the GA evolves (selection, cross-over, mutation) based on several fit goals in order to get an acceptable population. The advantage of GAs is that they give the designer a family of near optimal designs with a small variation in their performance index instead of a single design. In this paper, first the problem is fully discussed and the implementation using a finite element model for analysis explained. Then, what are GAs is shortly remembered, and all issues of use of GA made in this work are discussed.

The application considered in this paper is a bolted joint or lug. Bolted joints are a preferred method of connection in engineering construction because of the ease of disassembly and repair. But hole in a laminated composite leads to stress concentration, and then laminate could fail by bearing of the bolt on the hole surface. In high strength applications bolted lugs are used in helicopter rotors and for actuators applying control loads to composite control surfaces on aircraft. Fibre steering has been used to increase the stiffness and strength of the joints in these structures but the trajectories of the fibres are either placed along principal stress trajectories or are defined by tests. It is intended that the GA approach being developed in this paper can be used to check the trajectories that have been used and perhaps propose some new configurations that will improve stiffness and strength.

2. Bolted joints and finite element analysis

The classical way to improve the bearing load that can be carried by the plate is to increase the number of plies locally and so to increase local thickness. This work proposes to optimise fibre steered according one main objective, which could be to increase global stiffness of the joint. The definition of the trajectory of the fibre has then to be defined remote from the joint as it contributes to the marshalling of the loads from the panel to transfer to the joint. The trajectory at the hole is important to improve the failure loads but may be of secondary importance for the stiffness of the global structure. To reinforce this concept we will consider two types of joints. The first will rely only on the carbon laminate to resist the load, the second will include a metal bush to improve the wear characteristics of the joint and improve the bearing strength at the hole surface.

The first model adopted for this study is illustrated in Figure 1. It is a half plate with one hole. One edge is fixed. Loads are applied on the hole in plane to simulate the action of a bolt.

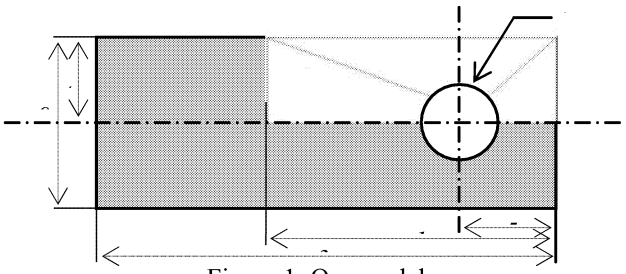
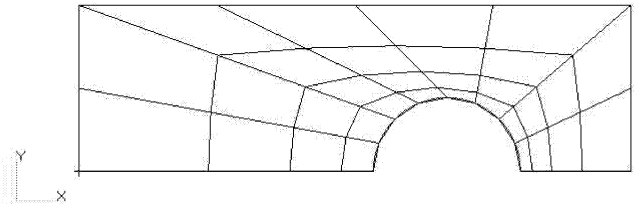
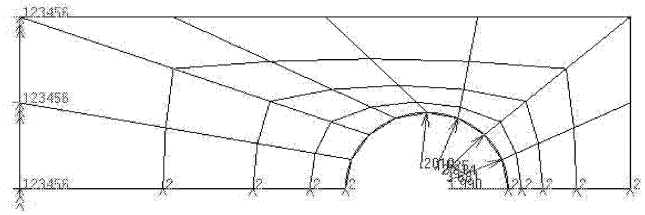


Figure 1: Our model

Using Patran, all the modelling is defined with the meshing (Figure 2-a) and the boundaries conditions, which are loads and degrees of freedom of edges (Figure2-b).



(a)



(b)

Figure 2: Mesh and Boundaries Conditions

In order a be more efficient with what happening around the hole, a mesh seed is defined with a increasing element edge length away from the hole. It leads a mesh of 32 finit elements. It is a coarse mesh with large elements but it is sufficient for the preliminary work attempted here. Also, a finer mesh with 128 elements has also been used in order to insure that algorithm is not sensitive to mesh size as shown in Figure 4-a.

The loads are computes to simulate a sinusoidal stress on half of our half hole. Beside, fixing degree of freedom 2 (no displacement on Y) simulated plate symetry. The left edge is fixed.

With such a model of our problem, it need to defined a specific material for each finite element. This specific material has the properties of a unidirectionnal fibre composite (UD) with a variable orientation. So, every finite element get mechanical property in function of the orientation of fibre in as illustrated in Figure 3.

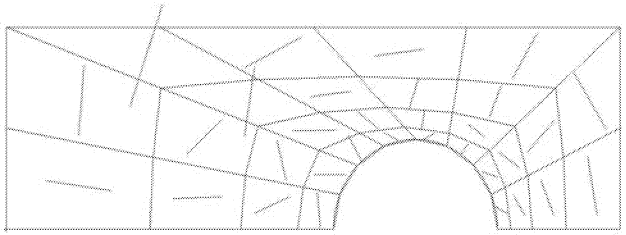


Figure 3: One set of Finite Element fibre orientation

A difficulty with this modelling strategy is the definition of the load distribution on the surface of the hole. A second model was therefore defined in which the surface has normal displacement fixed and and non-zero displacements or lodas prescribed on the remote left hand edge of the model, as shown in Figure 4-b.

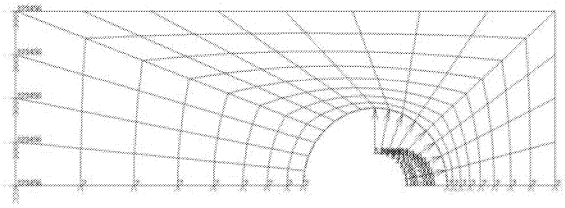


Figure 4: Second refined model boundary conditions.

The loads are set in model 1 in order to model the effect of a bolt or pin into the hole actioning on X direction. So, it is lead to a sinusoidal stress drawn on figure 7 or to the equivalent loads, which are used. In this configuration work is done by the applied loads acting over the displacements calculated by the finite element analysis. If the genetic algorithm favours members of the population that have lower strain energy evolution should lead to configurations that maximise the stiffness of the joint.

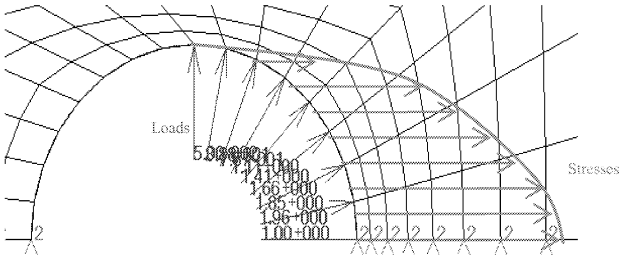


Figure 7: load and equivalent stresses on the hole for model 1

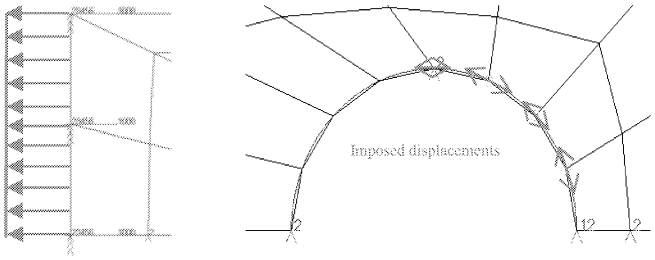


Figure 8: details of left edge and hole edge imposed displacements

For the model in Figure 8 only displacements are prescribed. In this configuration higher stiffness of the laminate will lead to higher reactions at the left hand edge of the model that do work on the prescribed displacements. So, the maximisation of strain energy with fixed displacement means will lead to a configuration with maximum stiffness.

3. The Genetic Algorithm

GA is an optimisation method appeared in early 70s. It is based on simulation of the Darwin's theory of species evolution. Contrarily of others classical optimisation method, GA could not stack on local optimum. Indeed, GAs combine both exploration of research space and exploitation of visited points.

GA operates on a given population of individuals. One individual is a possible solution of the given problem into the environment. To operate, GA uses 4 elements: encoding structure of individual, operators to affect individuals (mutation, cross-over), a fitness criterion to determine the goodness of individual (fit) and selection function (selection). So, from one population, GA builds a new one globally with a better fit to criterion.

Constraints in GA are handled in three ways: data structuring control, repair operator and penalty functions. Data structuring control refers to defining the design variables so that the optimiser always produces feasible designs. Here, for example, we model only half of the laminate, which guarantees the symmetry. An other control guarantees the balance conditions.

Since only the fittest individuals survive and reproduce, the genes of weaker individuals would disappear gradually. So, it leads that if the environment (fit law) does not change during the process, then, finally population will converge to a state where every individual has the fittest genes.

Encoding:
GA operates on encoded representations of the potential solution, equivalent to those chromosomes of individuals in nature.
A mesh of FE models the plate. Each FE has is own orientation of fibre. In this work, the all FE with theirs own orientation is called orientation map. The orientation map is encoded as a file filled with value of each FE orientation. For example, one individual could have an orientation map also call chromosomal stuff as illustrated if Table 1.

Chromosome	...	15	16	17	18	19	20	21	22	23	24	25	...
Individual #	...	1	34	175	44	163	104	137	64	170	175	178	...

Table 1: Example of one individual

So, am individual is a group of chromosomes, which represent the plate orientation map, and each chromosome is the orientation of its corresponding FE.

Fit:
A fit function needs to be defined in order to evaluate every individual. Then, a fit value is set to each individual reflecting how good it is.
Because the final orientation map is strongly depending of fit criteria, the choice of such a criteria is very important. One other aim of this job is to determine the optimising set of orientation map according to the problem. So, it is to find the good fit either the best-fit criteria.
To start, it had been decide to combine stress and strain to define the fit criteria. So, the global Strain Energy was the first fit criteria. But after several tries, le summation of the Strain Energy Square of each element seemed to be better because it focused more on element with high strain energy. Set the fit criteria as the maximum of the strain energy of each element increased more the focusing on the elements with high strain energy. So, the GA had to minimize the defined fit criteria.
Then in order to improve the load on the hole edge, this former had been fixed and le previous fixed edge had been submitted to loads.

Selection:
In nature, only the best individuals survive. Selection function models this fact. An individual able to survive to the next generation is an individual, which has a good fitness.
So, the GA has to select several individuals able to survive to the next generation. The selection method used is the tournament. GA takes two individual and keeps the one of the two, which has got the best fitness.
This is done to build only half of the next generation. The other part is done by le function cross-over.

Cross-over:
Like in nature, to generate the next population, GA makes individuals to reproduce. It needs to define a cross-over

function. The cross-over is an operator used to create one child from two parents. In this GA, the cross-over used is a "mask" cross-over statistically (p% probability) choosing more the chromosomes of the better of the two parents. It is illustrated in Table 2. This kind of cross-over is also called uniform cross-over with p% probability.

Parent 1	...	21	156	34	175	44	163	104	137	164	170	175	178	...
Parent 2	...	19	44	30	135	110	60	91	112	161	2	6	10	...
Mask (p%)	...	0	1	1	0	1	1	1	1	0	0	1	1	...
Child	...	19	156	34	135	44	163	104	137	161	2	175	178	...

Table 2: mask cross-over operator

Mutation:
In order to avoid to quick convergence of the population to a local optimum, mutation function causes sporadic and random alteration on the chromosomes. This permits a regeneration of lost genetic materials.

Convergence:
The convergence for an optimisation algorithm permits to determinate when the running could be stop, it is to say, when the best optimum (optimum optimorum) is reach. Using GA, there is not any good criterion to say that algorithm reach or not the optimal solution. This is one main problem of GA. The only thing that it could be say is that, after computation of several populations, the last one is composed by several very good individuals according to the fit criterion.

4. Results and discussion

a. simulation result

The results prensented in this chapter are for an all composite laminate with load applied to the bearing surface as illustrated on figure 9.

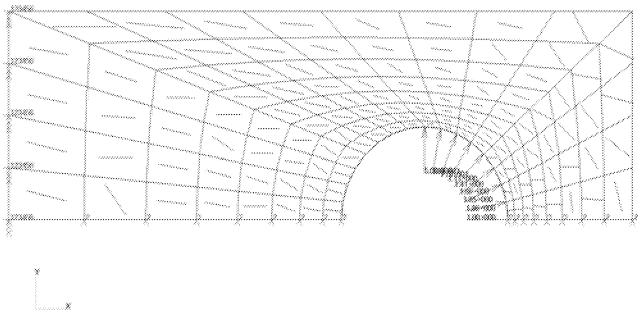


Figure 9: the studied problem

The Genetic Algorithm has been launched to design the trajectories of fibres for a continuous steered pattern in order to improve high strength and high stiffness configurations for carbon fibre structures. The algorithm used the PCL programming functions in MSC.PATRAN to modify the direction of the material axes on elements in a finite element mesh and executed MSC.PATRAN to determine the stress and strain to define the strain energy in the structure.
Even with relatively coarse meshes the applications ran for several days on a high performance 2.8GHtz personal computer.
The raw result is presented on figure 10.

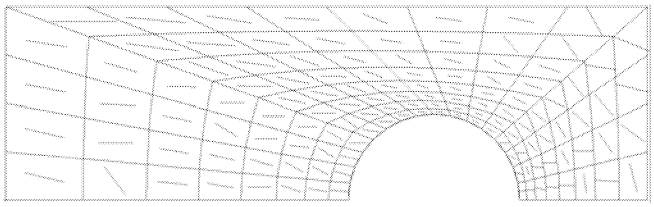


Figure 10: orientation set by GA

b. interpretation of simulation result

It is not easy from simulation result give by Patran to determinate where and how the tow should pass. The first read of a simulation result shown in figure 10 could give us the fibre steered pattern of figure 11. But, taking on account that tow is about 5min wide and that the plate is not covered everywhere by fibre steered, it leads to the figure 12.



Figure 11: First fibre steered pattern

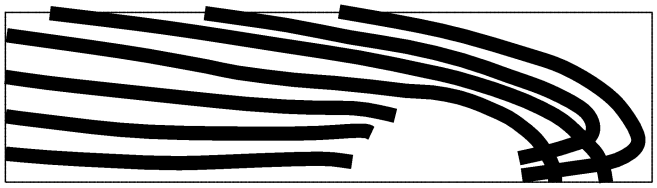


Figure12: Second fibre steered pattern

5. Conclusion

A Genetic Algorithm has been proposed to design the trajectories of to define high strength and high stiffness configurations for carbon fibre structures. It défines the optimal orientation of fibres in a uni-directional laminate in which the fibres are allowed to vary continuously across the domain.
The algorithm used the PCL programming functions in MSC.PATRAN to modify the direction of the material axes on elements in a finite element mesh and executed MSC.PATRAN to determine the stress and strain to define the strain energy in the structure or any function of strain energy used to compute the fit criterion. An automated analysis procedure is required because the Genetic Algorithm required large numbers of analyses to evolve the

population according to the fitness criterion and, even with relatively coarse meshes, the applications ran for several days.

A feature of the algorithm is the requirement to interpret the results of the execution of the programs. It is realised that this interpretation could be enhanced by colour coding the vectors that indicate the direction of the fibres in Figure 10. This is a development that will be attempted in the future. The application however successful and an interesting conclusion emerged for an optimal configuration for a highly loaded joint that includes a metal bush. It is hoped that it will be possible to confirm this conclusion be preparation and testing of a set of specimens.

Acknowledgement
The authors wish to acknowledge the support for this research from the Cooperative Research Centre for Advanced Composite Structures Ltd.

References

1. R. Li, D.W. Kelly and A.G. Crosky, 'Strength improvement by fibre steering around a pin loaded hole', in *Composite Structures*, vol 57(1-4), pp. 377 - 383 (2002).
2. M.W. Tosh and D.W. Kelly. “Fibre steering for a composite C-beam.” *Composite Structures* 53 pp 133-141. (2001).
3. R. Li, D.W. Kelly, S. Arima, R.A. Willgoss & A.G. Crosky. 'Fiber steering around a cut-out in a shear loaded panel', in *Proceedings, 47th International SAMPE Symposium*, eds B M Rasmussen, L A Pilato, H S Kliger, Long Beach, CA, USA, Covina, CA, USA, pp. 1853 – 1861 (2002)
4. D.W. Kelly, P. Hsu and M. Asudullah. “Load Paths and Load Flow in Finite Element Analysis.” *Engineering Computations* V. 18, No. 1/2 pp 304-313 (2001)
5. D. Kelly, R. Willgoss, R. Li, and A. Crosky. “A manufacturing procedure for composite laminate production using precision fiber placement”. *Proceedings of SAMPE2001* “2001: A Materials and Processes Odyssey”. eds SAMPE, ISBN 0-938994-90-5 Vol 2 pp2495-2506 (2001).
6. M.C.Y. Niu, “Composite Airframe Structures, Practical Design Information and Data”. ConMilit Press Ltd (1992)

ELECTRICAL PROPERTIES OF PANI-DYNEEMA® CONDUCTIVE FIBRES – FUZZY LOGIC MODELLING

Bohwon Kim*, Vladan Koncar* and Claude Dufour+
*Laboratoire de Génie et Matériaux Textiles (GEMTEX),
Ecole Nationale Supérieure des Arts et Industries Textiles (ENSAIT),
9 rue de l'Ermitage BP 30329, 59056 Roubaix, France
+Institut d'Électronique de Microélectronique et de Nanotechnologie (IEMN),
Cité Scientifique, Avenue Poincaré, BP 69, 59652 Villeneuve d'Ascq, cedex, France
kim.bohwon@ensait.fr , vladan.koncar@ensait.fr, claude.dufour@univ-lille1.fr

ABSTRACT

Fuzzy logic tool has been used in order to model electrical properties of conductive polymer fibres developed in our laboratory. Conductive fibres were obtained using experimental process based on chemical coating. Polyaniline was treated using dodecylbenzene sulfonic acid (DBSA) to improve the solubility and the dispersion of polyaniline in xylene. Then, chemical PANI coating on Dyneema® yarns were performed by absorption of fibre through PANI solution using sol-gel process. The electrical conductivity and morphological properties of conductive fibres were investigated. These fibres are expected to be used to realize connections in smart clothing or conductive fabrics for electromagnetic shielding applications.

KEYWORDS

Conductive fibre, Fuzzy logic modelling, Polyaniline, DBSA, Sol-gel process.

INTRODUCTION

Generally, fuzzy logic, neurocomputing, and genetic algorithms may be viewed as the principal constituents of what might be called soft computing. Unlike the traditional, hard computing, soft computing is aimed at an accommodation with the pervasive imprecision of the real world. The guiding principle of soft computing is: exploit the tolerance for imprecision, uncertainty, and partial truth to achieve tractability, robustness, and low solution cost. In the same time, phenomena that influence the electrical conductivity in the doped conductive polymers, used in order to create the conductive textile fibres, are complex and highly non linear. Different parameters as the morphological properties, the PANI concentration of solution, fibre length and others play important roles in the electrical conductivity. Therefore, soft computing tools and particularly the fuzzy logic approach seems to be well adapted to model electrical properties of conductive polymer fibres.

Conductive polymers exhibit interesting electrical properties and wide variation in colour due to their conjugated double bonds chain structures, which derive from both their conducting or neutral (non-conducting) forms. However, they are inherently insoluble and infusible due to their strong intermolecular interactions. High quality conducting blends with conventional polymers by melt mixing [1-5] or by solution casting [6-9] are still in a development stage.

Among the conjugated conducting polymers, polyaniline (PANI) has attracted much interest worldwide because of its excellent environmental, thermal and chemical stability. Polyaniline exists in a variety of oxidation states ranges from leucoemeraldine (completely reduced) to emeraldine (partially oxidized) to pernignaniline (completely oxidized). Most broadly studied forms of polyaniline are the emeraldine base form and the emeraldine salt form. With simple doping process by charge transfer or by the acid/base (protonation) reaction, polyaniline converts the semiconducting emeraldine base form (10^{-5} S/cm) to the metallic emeraldine salt form (10^{2} S/cm) [10, 11].

The soluble conductive form of PANI may be obtained by protonation with functionalised protonic acid, denoted as $H^+(M^-R)$ [12]. H^+M^- is protonic acid group and may be sulfonic acid, carboxylic acid, phosphonic acid, etc. The proton of the protonic acid reacts with imine groups of polyaniline and the M^-R group serves as the counterion. R is an organic group that can be compatible with nonpolar or weakly polar organic solvents e.g N-methyl-2-pyrrolidione (NMP), Dimethylsulfoxide (DMSO), xylene and m-cresol. Efficient doping methods of PANI with dodecylbenzene sulfonic acid (DBSA) or camphor sulfonic acid (CSA) were already reported [13]. The PANI chains tend to form extended chain conformation in presence of DBSA; electron transport between polymer chains should be enhanced. Moreover, long flexible alkyl chains of DBSA may enhance the crystallization rate of polyaniline. In the polyaniline/solvent system, the film-like agglomerates of polyaniline take place on the solution/solid interface when the solvent is removed. This spontaneous molecular assembly has been used to coat conductive polymers on the material surface, and has been successfully applied to fibres and textiles.

In this paper a doped polyaniline has been used in order to prepare conductive fibres with textile properties. The main objective of our research work was to obtain an improved electrical conductivity without modifying the mechanical (textile) properties of fibres. These textile properties are very important because the conductive fibres should be transformed in textile structure by weaving, knitting or other manufacturing process. The method developed to synthesize conductive fibres is detailed below:

Sol-gel process
An approach for the synthesis of conductive fibres has been developed based on a 'sol-gel' process. In general, the sol-gel process involves the transition of a system from a liquid 'sol' (mostly colloidal) into a solid 'gel' phase [14]. PANI coating was carried out during the impregnation of PET yarns in the PANI solutions (Figure 1). Then, the electrical resistance and the morphology of the conductive fibres obtained were investigated.

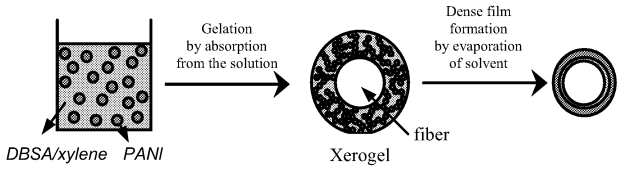


Figure 1: Overview of the sol-gel process used in this study [14]

EXPERIMENTAL

Materials

Emeraldine salt form polyaniline (PANI-ES) was supplied by Sigma-Aldrich Chemical Company. Dodecylbenzene sulfonic acid (DBSA, Fluka) was used as a dopant and surfactant to obtain an homogeneous conducting solution of PANI in xylene. Ultrahigh molecular weight polyethylene (UHMW-PE, Dyneema®, 220 dtex) was supplied by DSM Company in Nederland.

Coating process

The method developed to synthesize conductive fibres called "sol-gel process". The sol-gel process has been performed using an apparatus that can control the temperature and the taken-up speed. This process may contain several bathes filled with PANI-DBSA/xylene solution. The temperature of these baths varies from 30 to 100°C. PANI-DBSA mixtures with 25 % PANI-ES were dissolved in xylene to prepare 4, 5, 7 and 9 wt % PANI solutions at 90°C for 3 hours. Then, they were treated in the ultrasonic bath at 80°C for 2 hours. PANI solutions containing the dispersed PANI particles which size goes down to 3 µm and partially dissolved particles have been obtained. Dyneema® yarns were withdrawn through the baths in the presence of a dry airflow along the coating surfaces to enhance solvent evaporation. The taken-up speed of yarn is adapted to the drying dynamics and for our experiments; it was set to 6 cm/min. Then, the electrical resistance and the morphology of the conductive fibres obtained were investigated.

Electrical and morphological characterization

The electrical resistance of conductive yarns was measured using the "two-probe test technique" with an Agilent 34401A multimeter at 20-25°C, 60-65 HR%. The measurements were repeated 20 times for different yarn lengths from 1 cm to 20 cm and mean values were computed. Detailed images of the PANI/Dyneema® yarn surfaces were obtained using FIB (Focus Ion Beam lithography, FEI Company, Strata TM DB 235) and SEM (Scanning Electro Microscopy, JEOL 100 CX model ASID4-D).

RESULTS

All the electrical resistances of PANI/Dyneema® conductive yarns obtained by coating with PANI solutions are presented in the Table 1. The measurements have been repeated 30 times for different yarn lengths from 1 cm to 20 cm and average values have been computed. To measure their effects on electrical resistance (R) of PANI/Dyneema®, we have introduced two controlled variables;

- Fibre length (L),
- Concentration of PANI solution (C).

The effects of fibre length on electrical resistance with different concentrations are shown in the Figure 2. Moreover, the Figure 3 resumes the electrical properties of conductive fibres in function of two input parameters, fibre length and PANI concentration. It is observed that the resistance generally increases with fibre length at all concentrations. For the same fibre length, higher resistance is obtained with lower concentration. Resistance increases linearly when the solution is highly concentrated (between 7% and 9%) but for low PANI concentrations strong non linearity may be observed. For the PANI/Dyneema® 4% conductive fibre the electrical resistance does not follow the general principle defined by the expression $R = \rho l/S$, where ρ is the fibre specific resistivity, l the length and S the surface of the cross section.
To understand this phenomenon, several parameters should be taken into account:

- Coating quality (morphology of a conductive layer),

- Concentration of used PANI solution,
- Spreading of PANI molecules onto the fibre surface.

The phenomenon of the conductivity is influenced by PANI molecular chains. Interactions among these PANI chains are important to the movement of electrons. We observed that the coating of PANI on the fibre surface was not visibly homogeneous in the case of PANI/Dyneema® 4%. Hence, the thickness of the PANI layer on the fibre surface is less uniform for lower concentrations. There are some areas where the thickness is very week disabling contacts among PANI chains and making conductivity lower that is supposed to explain the non-linear resistance behaviour in function of the length. On the other side, as we can see in the Figure 4, the PANI film forms a compact conductive network on the Dyneema® surface when the fibre is coated with 9% of PANI solution. These continuous PANI chains interactions may increase the linearity and enhance electrical conductivity of fibres.
One of the important parameter of organic coating is the concentration of used solution. The PANI molecules spreading onto the fibre surface is very important. At 4% of PANI solution, there are minor amounts of molecules of PANI which attach onto the fibre surface, comparing with that of 7% and 9%. In general, a uniform distribution of PANI onto the fibre surface should be obtained with high concentrations of solution (over 7%). The standard deviations for different solution concentrations and fibre lengths decrease with the increase of the solution concentration (not shown in the Table 1). Thus, it is obvious that the homogeneity of the conductive layer is better for higher solution concentrations.

Table 1: Electrical resistance (ohm) of PANI/Dyneema® conductive fibres

	PANI solution (%)			
Length (cm)	4%	5%	7%	9%
1	316,28	107,29	133,21	23,60
3	691,66	263,69	261,76	64,97
5	2283,07	456,15	391,34	126,02
10	4759,03	1255,51	761,16	241,68
20	9572,19	2640,79	1803,60	462,19

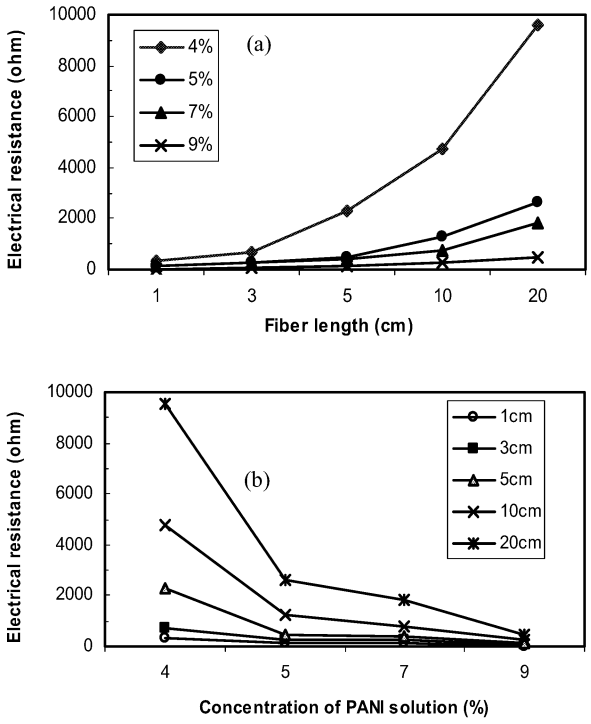


Figure 2: Electrical resistance of PANI/Dyneema®

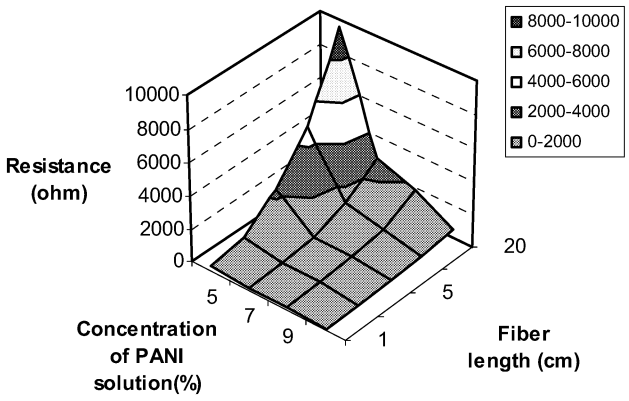


Figure 3: Electrical properties of PANI-Dyneema®

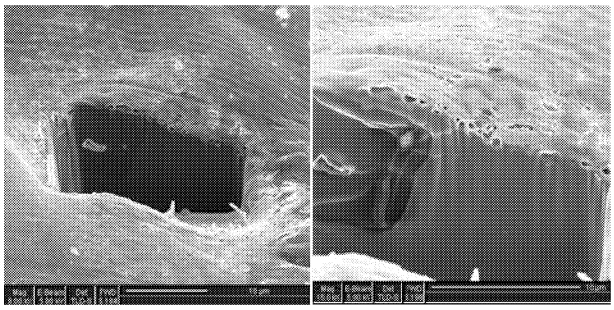


Figure 4: Cross section of the PANI film on surface of Dyneema® fibres (Electron scanning microscopy)

FUZZY LOGIC MODELLING OF CONDUCTIVE FIBRES ELECTRICAL PROPERTIES

This section deals with the modelling of electrical and mechanical properties of conductive fibres presented in previous section.

Electrical model Dyneema® PANI fibres

Fuzzy sets (inputs) have been defined using the Table 1 as follows:

- PANI concentration: low 4 (4%), midlow 5 (5%), midhigh 7 (7%) and high 9 (9%)
- Fibre length: 1, 3, 5, 10 and 20 (cm).

A membership function (MF) is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. The input space is sometimes referred to as the universe of discourse, a fancy name for a simple concept. Membership functions are built on the Gaussian distribution curve: a simple Gaussian curve and a two-sided composite of two different Gaussian curves. The two functions are gaussmf and gauss2mf (Fuzzy Toolbox, Matlab®). These functions (Figures 5 and 6) are optimal for electrical properties modelling and permit the setting of the rule surface corresponding to the experimental Figure 3.

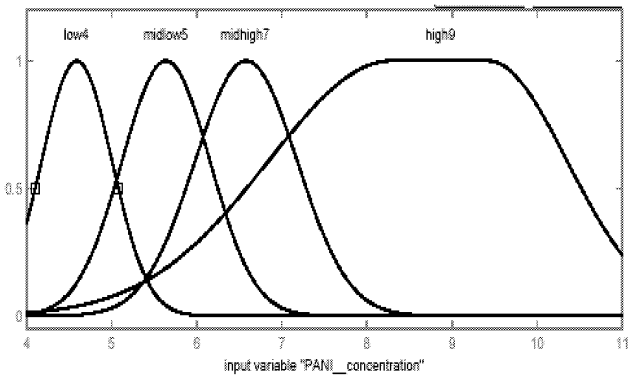


Figure 5: PANI concentration membership function

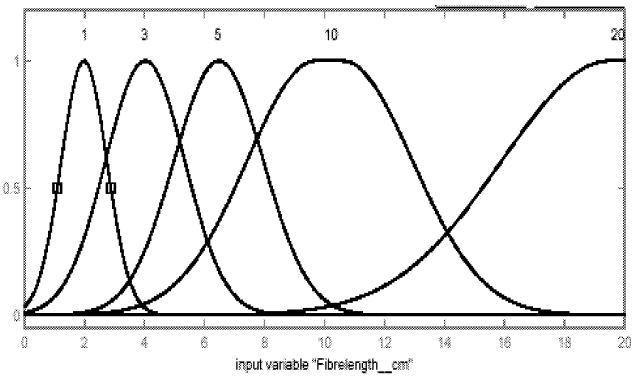


Figure 6: Fibre length membership function

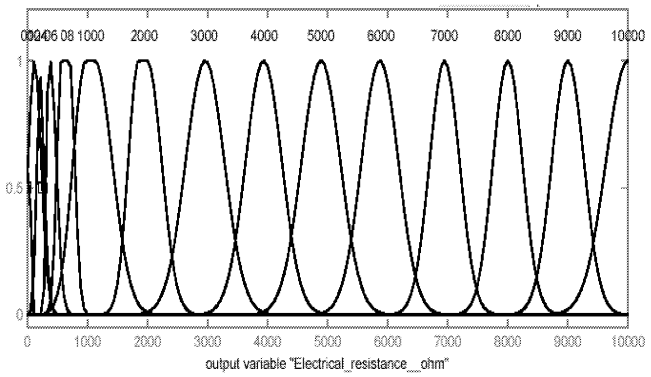


Figure 7: Output – electrical resistance

Fuzzy sets and fuzzy operators are the subjects and verbs of fuzzy logic. These if-then rule statements are used to formulate the conditional statements that comprise fuzzy logic. A single fuzzy if-then rule assumes the form if x is A then y is B where A and B are linguistic values defined by fuzzy sets on the ranges (universes of discourse) X and Y, respectively. The if-part of the rule "x is A" is called the antecedent or premise, while the then-part of the rule "y is B" is called the consequent or conclusion.
In our case the rules are defined using the data from the Table 1 in order to approximate as good as possible the experimental data. For instance:

Rule : If PANI concentration is low4 and fibre length 5 then Electrical resistance is 2283,07502 ohm.

Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping then provides a basis from which decisions can be made, or patterns discerned. The process of fuzzy inference involves all of the pieces that are described in the previous sections: membership functions, fuzzy logic operators, and if-then rules. There are two types of fuzzy inference systems that can be implemented in the Fuzzy Logic Toolbox: Mamdani-type and Sugeno-type. We have used the Mamdani-type one.

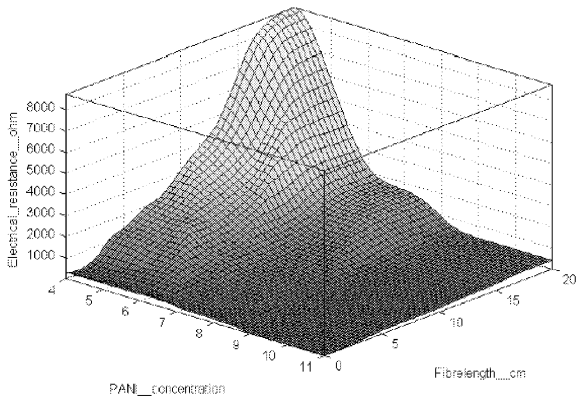


Figure 8: Fuzzy rules surface

Finally the surface obtained from the two inputs (PANI concentration and Fibre length) generating the Electrical resistance output is given in the Figure 8.

CONCLUSION

This paper describes fuzzy logic modelling of conductive fibres prepared by sol-gel coating processes. The advantage of this model is that it can easily take into account all nonlinearities and complex phenomena related to electrical properties of conductive polymers. The obtained model enables fast computation of conductive fibre starting from the concentration of the PANI solution used in order to coat textile fibres and from the length of conductive yarn. The model developed in this paper is the first approach that can be improved by introducing the adaptive and learning functions from new data obtained experimentally.

REFERENCES

[1] M. Zilberman, A. Siegmann and M. Narkis, *J. Macromolecule Sci.- Physic,* **B37** (3), 301-318 (1998).
[2] O. T. Ikkala, J. Laakso, K. Vakiparta and E. Virtanen, *Synth. Met.*, **69**, 97-100 (1995).
[3] I. L. Hosier, A. S. Vaughan, D. Patel, S. J. Sutton and S.G. Swingler, *IEEE Transactions on dielectrics and electrical insulation*, **8**, 698-704 (2001).
[4] J. P. Yang, P. Rannou, J. Planès, A. Pron and M. Nechtschein, *Synth. Met.*, **93**, 169-173 (1998).
[5] M. Zilberman, G. I. Titelman, A. Siegmann, Y. Haba, M. Narkis and D. Alperstein, *J. Appl. Polym. Sci.*, **66**, 243-253 (1997).
[6] G. I. Titelman, M. Zilberman, A. Siegmann, Y. Haba and M. Narkis, *J. Appl. Polym. Sci.*, **66**, 2199-2208 (1997).
[7] C. L. Gettinger, A. J. Heeger, D. J. Pine and Y. Cao, *Synth. Met.*, **74**, 81-88 (1995).
[8] Alan. J. Heeger, *Synth. Met.*, **55-57**, 3471-3482 (1993).
[9] C. Y. Yang, Y. Cao, P. Smith and A. J. Heeger, *Synth. Met.*, **53**, 293-301 (1993).
[10] Y. Cao, P. Smith and A. J. Heeger, *Synth. Met.*, **55-57**, 3514-3519 (1993).
[11] Reghu M., Y.Cao, D. Moses and A. J. Heeger, *Synth. Met.*, **55-57**, 5020-5025 (1993).
[12] Y. Cao, Jinjin Qiu and Paul Smith, *Synth. Met.*, **69**, 187-190 (1995).
[13] R. V. Gregory, *Handbook of conductive polymer*, second edition, Marcel Dekker. Inc., New York, 1998 Chap. 18.
[14] C. Jeffrey Brinker and George W. Sherer, *Sol-gel Science*: *The physics and chemistry of sol-gel processing*, Academic Press, Inc., New York, 1990, pp 787 837.
[15] L.Vangheluwe, Study of the Time Dependent Mechanical Properties of yarns for Weavning, Doctoral thesis, University Ghent, (1992).
[16] L.Vangheluwe, *Relaxation and inververse relaxation of yarns after dynamic loading*, Textile Research Journal; **63(9)**, 552-556 (1993).
[17] W. Zurek, S. Aksan, *A rheological model of viscose rayon*, Journal Appl. Polym. Scien., **19**, 3129-3137 (1975).
[18] A. Manich, *Viscoelastic behavior of polypropylene fibres*, Textile Research Journal, **69(5)**, 325-330 (1999).
[19] X. Legrand, *Contribution à la modélisation dynamique et à l'identification de tissus techniques*, Thèse de Doctorat, Université des Sciences et Technologies de LILLE, (2000).
[20] P. Bruniaux, J.M. Castelain, *Traction model - A methodological approach*, ISC 2003 Valencia (Spain), 523-530 (2003).
[21] H. Sabhi, Etude expérimentale et modélisation mécanique et numérique du comportement des tissus de fibres de verre lors de leur préformage, Thèse de Doctorat, Université de Franche-Comté, (1993).

BIOGRAPHY

Dr. Bohwon Kim, Laboratoire de Génie et Matériaux Textiles (GEMTEX), Ecole Nationale Supérieure des Arts et Industries Textiles (ENSAIT), 9 rue de l'Ermitage BP 30329, 59056 Roubaix, France,
E-mail :kim.bohwon@ensait.fr ;
Phone : +33 3 20 25 75 70, Fax : +33 3 20 27 89 94
www.vfic.net,

Professor, Dr. Vladan Koncar, Laboratoire de Génie et Matériaux Textiles (GEMTEX), Ecole Nationale Supérieure des Arts et Industries Textiles (ENSAIT), 9 rue de l'Ermitage BP 30329, 59056 Roubaix, France,
E-mail :kim.bohwon@ensait.fr ;
Phone : +33 3 20 25 89 59, Fax : +33 3 20 27 25 97
www.vfic.net, http://www.mtmc.ensait.fr

Professor, Dr. Claude Dufour, Institut d'Électronique de Microélectronique et de Nanotechnologie (IEMN) /DHS, Cité Scientifique, avenue Poincaré BP19, 59652 Villeneuve d'Ascq Cedex, France,
E-mail :claude.dufour@univ-lille1.fr;
Phone: +33 3 20 19 79 08, Fax:+33 3 20 19 78 78

MICROSCOPIC MODELLING OF POLYMERS AND THEIR INTERACTIONS

Peter Ahlström
Kim Bolton
Mikael Skrifvars
School of Engineering
University College of Borås
SE-501 90 Borås
Sweden
Email:peter.ahlstrom@hb.se

ABSTRACT

The modeling and prediction of properties of polymers, the main constituent of textile materials, is discussed. Two examples of detailed microscopic modeling of polymers and their properties are given. One is the modeling of structure and properties of poly(propylene oxide) (PPO). The other one is an ongoing study of polyethylene (PE) as an insulation material. Furthermore, we discuss an approach of modeling the different length scales present in polymers and their applicability to polymer materials.

INTRODUCTION

In order to understand the properties of textile materials, a profound insight in the properties of their constituents is a crucial step. Since the central constituents of most textile materials are polymers our interest is focussed on their properties and their interactions with smaller molecules which could be additives, dyes, water etc.

The structure of a polymer molecule can be described at different levels, the primary structure being the order of monomers of different types, the secondary structure being their local ordering into structural elements like helices, loops etc., whereas the tertiary structure is the ordering of these elements into an overall structure. This hierarchal description is especially relevant for biological polymers consisting of several different types of monomers in a given order like proteins, which consist of some twenty different types of amino acids with a genetically given order. For many industrially relevant polymers, however, only one monomer type is present. This fact both simplifies and complicates the modeling of those polymers. The presence of only one monomer type makes it easier to model the interactions between different groups. On the other hand it leaves a much broader range of plausible structures whereas the structure of, e.g., a protein normally is in principle determined by the primary structure. (An example where this, however, is not the case but several structures are possible is the case of BSE, “mad cow disease”, where the presence of a misfolded protein molecule causes other proteins to misfold and aggregate)

When we wish to model a polymer molecule from an atomistic description we have to take into account that the number of possible conformations of such a long molecule is so huge that it is impossible to find *the* native structure. Normally exist several conformations with energies close to the minimum energy. Also the time scales for important processes are much longer than can be modeled with present or foreseeable computers using an atomistic model. Instead, one has to approach the problem by concerted modeling at several length and time scales. This may involve a microscopic (atomistic) description and possibly even a quantum mechanical description as the most detailed level.

The next step then has to be a *mesoscopic* description in which several atoms are lumped together and interact via an effective potential and a simplified dynamics, examples of this are given by Müller-Plathe (2002) and Theodorou (2004). The parameters used for simulations in this mesoscopic description are often deduced from the microscopic description and some results can be compared with the more detailed simulation to tune the mesoscopic model.

In order to obtain macroscopic properties for different materials the results from the micro- and mesoscopic modeling (like Young's modulus, viscosity etc.) can be used as input in a finite element modeling of the material.

Some properties do, however, not depend on, e.g., the global structure of the polymer but more on the local structure and dynamics. Examples of this may be the penetration of small molecules through polymers, local relaxation properties such as measured by spectroscopic techniques. An example that will be discussed here is the neutron diffraction results for relaxation and local dynamics of poly(propylene oxide), PPO, also known as poly(propylene glycol). Another example is the penetration of water in polyethylene (PE) high voltage insulators.

Starting from these examples we will describe some recent and ongoing polymer modeling research at the University college of Borås. We will discuss the work in conjunction with the modeling of textile materials. We will also discuss the bridging of the length and time scale gap from the microscopic (atomistic) scale to the macroscopic scale of everyday life using a hierarchy of multiple scales for the modeling.

Most polymers used for technical applications are amorphous but many (like PE and most fibres) contain crystalline regions which are very important for the functioning of the polymers. As we will se these regions pose a problem to molecular modelling methods but less so to mesoscopic and macroscopic modelling methods.

MODELLING METHODS

Molecular Dynamics Simulations

In this work we will mainly use the molecular dynamics (MD) method when performing the atomistic modelling. In this method the atoms are displaced according to the classical equations of motion (Newton's second law)

$$\boldsymbol{a}_i = \frac{\boldsymbol{F}_i}{m_i} \qquad (1)$$

where $\boldsymbol{a}_i$ is the acceleration of atom i (with mass m_i), $\boldsymbol{F}_i$ is the force on that atom due to all other atoms in the system. The equations of motion were the integrated using the Leap-frog algorithm that is another formulation of the Verlet algorithm (cf. Frenkel and Smit, 2001).

Monte Carlo simulations

In Monte Carlo (MC) simulations as used in this work the atoms are moved randomly from one configuration to create a new configuration and the energy difference between the old and new states determines whether the new configuration is to be accepted. This method has been extended in several ways (see, e.g., Frenkel and Smit, 2001). For the study of interactions between water and polymers in this work we use the Gibbs ensemble (Panagiatopolous, 1987). In this method two boxes are simulated which are in equilibrium with each other. This is ascertained by moves involving the swapping of molecules and the exchange of volume between the two boxes.

Force fields

The forces in the simulation of PPO were obtained from quantum chemical calculations to which a potential function was fitted (Smith et al., 1998). This approach yields a good force field at a moderate expense in simulation time. However, there are disadvantages in that the higher order effects like polarisabilities are hard to reproduce. These can be included in a classical simulation at a price of more than a doubling of the computational resources needed or by even more elaborate quantum chemical calculations.

Another alternative that was used in the simulation of water clusters is to use forces obtained by fitting a potential function such that resulting properties from a simulation reproduces macroscopic data. This method was used to construct the water model SPC-E (Berendsen 1987), used in the simulations of water made in order to study the interactions between polyethylene and water. This method has the advantage over the quantum chemical approach that also higher order effects, like polarisations etc., are taken into account by this fitting procedure - but in an effective way.

Initial structure

As mentioned above, one of the main obstacles in atomistic polymer modelling is the very large configurational space available to a polymer molecule. This large space makes it virtually impossible to directly equilibrate a simulated sample, i.e. to let the sample reach a set of states that is representative for the overall polymer configuration. That makes the choice of initial structure crucial to most of the results. There exist several methods to generate representative structures for amorphous polymers. One of the most popular and efficient is end-bridging Monte Carlo (Mavrantzas, et al., 1999). In this method parts of the polymer molecules are swapped between different chains in order to facilitate the approach to equilibrium.

Many properties, like penetration properties for small molecules, local dynamics etc. are, however, largely determined by the local structure of the molecule and are not very sensitive to the overall fold of of the polymer as long as it has liquid-like properties. Since that is the case for PPO close to room temperature, we have used another approach to obtain representative structures for PPO as described by Carlsson et al. (1998). The structure of the polymer backbone was initiated by means of a random walk. This structure was then adjusted to fit neutron diffraction data by means of a reverse Monte Carlo simulation before the actual equilibration was started. This was performed during a simulation that lasted for several nanoseconds, first with a simplified force field and then with the more detailed quantum chemical force field used during the analyzed part of the simulation.

RESULTS AND DISCUSSION

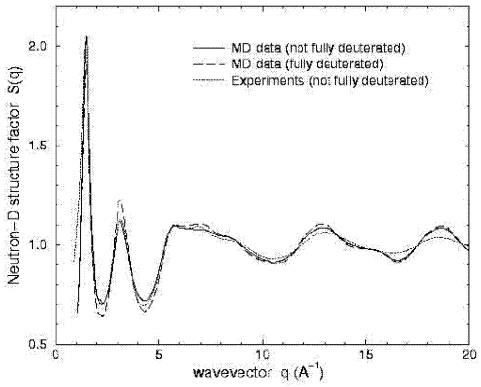


Fig. 1 Static Structure Factor for Poly(propylene oxide) as Obtained from Neutron Diffraction and MD Simulation. For MD Simulations There are two Curves, one Corresponding to a Fully Deuturated Sample and one Corresponding to a Sample with Approximately the Experimental Degree of Deuteration. (From Ahlström et al, 2000)

Poly(propylene oxide)

The molecular dynamics simulations of poly(propylene oxide) were used to calculate several static and dynamic properties. To verify the correctness of the simulations, the static structure factor was calculated and compared to neutron diffraction experiments. This comparison showed a discrepancy that later could be explained by incomplete deuteration of the experimental sample.

For the penetration properties of small molecules into polymers, the occurrence of free volumes is crucial and a recent study using the CAVA method (Račko et al., 2005) based on these simulations reveals the occurrence of

percolated empty space through the sample (Račko and Bartos, 2005). These results are in agreement with experimental studies using positron annihilation techniques. However, from the experimental studies alone it was not possible to deduce the shape of the cavities, a crucial property for the percolation studies. We note that these studies together can be used as a first step in studying the permeability of the sample to ions which is of great importance for the use of the polymer in, e.g., lithium accumulators as used in cellular phones.

The dynamic properties of the polymer were also studied by both MD simulations and quasi-elastic neutron scattering (Tengroth et al., 2005). By using simulations it was possible to interpret the scattering results and to separate the local dynamics from the backbone dynamics of the polymer.

Water and polyethylene

One important application of simulations is the study of the interaction between polymers and smaller molecules which could be additives, like softening agents, stabilizers of different kinds, solvents and other molecules like water. In order to model the interaction between these substances and the polymers, atomistic detail is needed. On the other hand, this interaction mainly depends on *local* properties of the polymer and not on the global structure of the polymer. Hence one could regard an amorphous polymer as a liquid for these studies. However, the large amount of available structures has to be taken into account. Thus one will need to investigate several different snapshots from a simulation which uses refined sampling techniques.

An example of those studies is our ongoing study of the interactions between polyethylene in high voltage insulators and water. In that case the behaviour of the small penetrant is important and a study of the equilibrium between water liquid and vapour using the Gibbs ensemble method mentioned above has been performed as an initial step. Here the clustering properties of water in vapour were studied (Johansson et al., 2005a and 2005b). These results can be used to find the distribution of water clusters of different sizes and configurations. This is a property crucial to the calculation of the second (and higher) virial coefficients of water but also highly relevant to water penetration into polymers since the water vapour properties are similar to the behaviour of water in a hydrophobic environment, like polyethylene. The next step in the PE studies is the replacement of the empty vapour box with a box containing polyethylene.

As an initial study a series of PPO configurations from the simulation above will be investigated and the outcome compared to the CAVA analysis (Račko and Bartos, 2005). Then samples of the more complicated polyethylene will be introduced and results compared. The final aim of this project is to understand mechanisms for the degradation of polymer insulators for high voltage applications. Hence electric field effects will be crucial and we will introduce a strong electric field to study the effects of that on the penetration of water into polyethylene.

This will be done with series of different configurations of PE in order to sample several possible configurations similar to the works by Fukuda (2000). He used, however, a test particle insertion method whereas we intend to use a physical water model and Gibbs ensemble simulations to study the equilibrium. The amorphous PE configurations can be obtained from e.g., connectivity-altering MC (Mavrantzas et al.,1999), possibly combined with MD simulations. The simulations will also be used to obtain the structural changes of PE due to an electric field and the presence of water. Starting from that knowledge new (polymer bound) voltage stabilizers may be designed.

Since PE at ambient temperatures is a semi-crystalline polymer, another question that has to be resolved is how to handle the effect of crystalline regions. This may require the simulation of a large sample in order to have a sample of all types of regions (amorphous and crystalline).

OUTLOOK AND TEXTILE APPLICATIONS

All textile materials are based on polymers, either natural like cellulose in cotton and proteins in silk or synthetic like polyamides (nylon) or polyesters. They form fibres containing both amorphous and crystalline regions. This mixture will make it rather difficult to model textile fibres. A multiscale approach involving scaling at all levels, microscopic (atomistic), mesoscopic and macroscopic modelling will be needed so that the different regions can be treated correctly and kept in realistic proportion to each other.

Microscopic modelling

Already with "traditional" microscopic modelling several of the interesting properties of textile materials can be calculated if only the structures of the involved molecules are known. So can elasticity properties of the amorphous polymer be obtained from either direct calculation (by calculating the tension when the polymer molecules are elongated) or from the fluctuations in one simulation with no (or constant) tension applied. However, to obtain a representative sample of the structure as needed and then often a mesoscopic or even macroscopic approach (see below) is needed to obtain such a structure.

Mesoscopic modelling

The atomistic modelling has, as mentioned above limitations in that it can not handle processes taking more than a fraction of a microsecond, since that would be too demanding of computational resources. One way to go is to use coarser models where several atoms are described by one "super-atom" which then interacts with an effective potential with other "super-atoms". Also the dynamics is changed in this type of "mesoscopic" simulations to account for the missing degrees of freedom, like disspative particle dynamics.

On the other hand, the detailed modelling is necessary to catch the details and it often is necessary to be able to go backwards and resolve real atoms from the "super-atoms" (cf., e.g., Müller-Plathe (2002)) in order to be able to calculate the desired properties. This process works well when the structural units that are transformed to beads in the mesoscopic simulation are internally rather inflexible, like the hexose rings in cellulose, but less well for highly flexible polymers, like PE.

Macroscopic modelling

In order to obtain the relevant textile properties the micro- and mesoscopic modelling need to be combined with more traditional modelling like finite element methods from which the properties of a whole system (like a thread or a piece of fabric) can be obtained. The output from this modelling is then compared to experimental results and parameters may be adjusted.

Concerted Multilevel Modelling

The above three steps (microscopic, mesoscopic and macroscopic modelling) all yield results that can be compared to experimental results. This was clearly shown above in the microscopic simulations of PPO where diffraction results were faithfully reproduced by the simulations that at the same time could give a microscopic explanation to the different features in the diffraction results. Using the insights obtained in this way the modelling is calibrated and possible inaccuracies in the modelling (and experiments) can be remedied.

In a similar way the predictions from mesoscopic modelling based on either microscopic modelling or deduced from macroscopic models can be compared to experimental results related to mesoscopic properties (like viscosity etc.) and the mesoscopic modelling is then used to produce the parameters needed By doing so each level can be firmly founded in the more detailed levels and in the less detailed levels as well as compared to experimental findings. The result will then give a reliable picture of the textile material and its properties. At a later stage when the methods are more refined experimental verification at each step will become unnecessary and materials with desired properties may be designed completely in the computer.

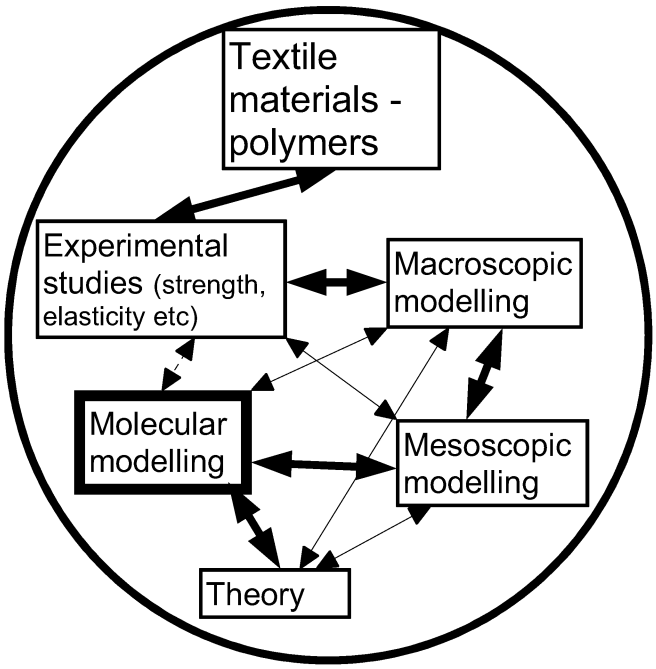


Fig 2.A Schematic Drawing of the Pathways in Concerted Multilevel Modelling. Thick Arrows Correspond to Main Path of Materials Development Whereas Dotted Arrows Correspond to Verification Paths.

CONCLUSIONS

The atomistic studies discussed in this work are all examples of studies where the detailed structure is of large importance for the investigated properties. On the other hand, these properties were not highly dependent on the "large"-scale structure of the polymer and hence a rather crude model of the global structure of the polymers could yield a good insight in the properties under study. So we could make a molecular interpretation of neutron diffraction results of PPO which would have been impossible without a detailed microscopic model.

Also the preliminary results on water are encouraging and we think that the microscopic modelling gives a deep insight in the behaviour of polymers and small molecules. However, the modelling of slow processes including the equilibration of a polymer structure or the handling of mixed crystalline/amorphous polymers like polyethylene and many fibrous polymers is still is beyond reach for the microscopic methods and hence we need to use a mixture of different methods involving both microscopic and mesoscopic methods.

Since textile materials are rather complex we believe that their correct modelling requires a multilevel treatment involving the full scale from microscopic modelling of the constituent molecules to a macroscopic modelling of the full materials. These modelling levels have to cooperate in a self-consistent way. Then the properties of many fibrous materials will be predictable and new materials may be invented.

REFERENCES

Ahlström, P, Wahnström, G., Carlsson, P., Schantz, S., Brodin, A., Maurer, F. and Torell, L. (1998) Low frequency vibrations in monomers, dimers and polymers of propylene glycol *Phil.Mag.B*, 77, 699-707.

Ahlström, P.,. Borodin, O., Wahnström, G., Wensink,E., Smith, G.D. (2000) Molecular Dynamics studies of poly(propylene oxide) using a new quantum chemical force field*J.Chem.Phys*., 112 10669-10679.

Berendsen, H.J.C., Grigera, J.R., Straatsma, T.P. (1987). The missing term in electrical pair potentials, *J. Phys. Chem.* 91, 6269.

Carlsson, P., Swensson, J., Börjesson, L, Torell, L.M., McGreevy, R.L., Howells, W.S (1998), Structural properties of poly (propylene oxide) from diffraction experiments and reverse Monte Carlo simulation, *J. Chem. Phys.*, 109, 8719.

Frenkel, D., Smit, B., (2001), Understanding Molecular Simulation: From Algorithms to Applications, Academic Press.

Fukuda, M (2000) Solubilities of small molecules in polyethylene evaluated by a test-particle-insertion method, *J.Chem.Phys.* 112, 478-486.

Johansson, E., Bolton, K., Ahlström, P. (2005a) Simulation of water vapour clusters in direct equilibrium with liquid water, *Comp. Phys. Commun.,* in press.

Johansson, E., Bolton, K., Ahlström, P. (2005b) Simulations of Vapor Water Clusters at vapor-liquid equilibrium, *J. Chem. Phys.,* in press.

Mavrantzas, V.G., Boone, T., Zervopoulou, F., Theodorou, D.N. (1999) End-bridging Monte Carlo: An ultra-fast algorithm for the equilibration of the condensed phases of long polymers, *Macromolecules*, 32, 5072-5096.

Müller-Plathe, F (2002) Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back, *CHEMPHYSCHEM* 3: 754-769

Panagiotopoulos, A.Z. (1987) Direct Determination of Phase Coexistence Properties of Fluids by Monte Carlo Simulation in a New Ensemble. Molecular Physics, *Molecular Physics* 61, 813

Queyroy, S, Neyertz S, Brown D, Müller-Plathe, F (2004) Preparing relaxed systems of amorphous polymers by multiscale simulation: Application to cellulose, *Macromolecules* 37 7338-7350

Račko, D., Bartos, J., personal communication (2005)

Račko, D., Chelli, R., Cardini, G., Bartoš, J.,Califano, D. (2005) Insights into positron annihilation lifetime spectroscopy by molecular dynamics simulations *Eur.Phys.J. D* **32**, 289

Smith, G.D., Borodin, O., and Bedrov, D. (1998) Quantum Chemistry Based Force Field for Simulations of Poly (propylene oxide) and Its Oligomers *J. Phys. Chem* A102, 10318 .

Tengroth, C., Engberg, D., Carlsson, P., Ahlström, P., Wahnström, G., Howells, W.S., Börjesson, L. (2005) The segmental and rotational dynamics of PPO, above the glass-transition, investigated by Neutron Scattering and Molecular Dynamics simulations, *Soft Materials,* in press.

Theodorou, D.N. (2004) Understanding and predicting structure-property relations in polymeric materials through molecular simulations, *Mol. Phys.* 102, 147

AUTHOR BIOGRAPHIES

PETER AHLSTRÖM was born in Stockholm, Sweden. After studies at the University of Lund in Sweden and the ETH in Zürich he obtained his masters of engineering degree in 1983 and his PhD in physical chemistry in 1988. 1989-1992 he was a post-doc at the University of Groningen in the Netherlands and 1993-1999 he was employed as assistant professor in chemical physics at Chalmers University of Technology i Göteborg, Sweden. Since 1999 he is employed by the University College of Borås where he was appointed associate professor for chemical physics in 2003.

KIM BOLTON was born in Durban, South Africa and obtained his BSc(Hons) and MSc in Chemistry from the University of Durban, Natal in 1991. He obtained his PhD in Physical Chemistry at Göteborg University, Sweden in 1996, and currently has a position as associate professor at the University College of Borås, Sweden.

MICHAEL SKRIFVARS was born in Helsinki, Finland, and obtained a PhD degree in polymer science at University of Helsinki in 2000. He worked as a researcher and project manager at the company Neste Chemicals from 1986 to 1999, after which he joined the R&D Institute SICOMP in Sweden. Since 2003 he is professor in polymer technology at the University College of Borås in Sweden.

TEXTILE WEAVING SIMULATION

COMPUTER MODELING OF ELECTROSTATIC WEB FORMATION

Yiyun CAI
College of Textiles
North Carolina State University
Raleigh, NC 27695
USA

Yong K. KIM
College of Engineering
University of Massachusetts Dartmouth
N. Dartmouth, MA 02747
USA

Abdel-fattah A. SEYAM
College of Textiles
North Carolina State University
Raleigh, NC 27695
USA

KEYWORDS
Modeling, Numerical Methods, Nonwoven, Electrostatic, Fiberweb Formation, Fiber Orientation

ABSTRACT

Current web formation technologies used by the nonwoven industry do not provide definite control over fiber orientation and it is difficult to generate lightweight fiberwebs with these processes. This research deals with development of new method of fiberweb formation using electrostatic forces where fiber orientation and fiber orientation distribution are positively controlled. A computer simulation for electrostatic web forming process provides better insights to this process and serves as powerful tool for engineering development. Results obtained from the model with variables such as geometric configurations, electrostatic field strength, airflow, and fed fibers' initial conditions will be used to develop a prototype of electrostatic web forming system.

1. INTRODUCTION

The lack of controlling fiber orientation and the fiber orientation distribution (FOD) of fiberwebs leads to nonwoven fabrics with significant difference between the strengths in cross and machine directions. Currently, the nonwoven industry uses four main technologies to form fiberwebs from staple fibers: carding, carding and cross-lapping, air lay, and wet lay. In such fiberweb formation methods, different trials with different processing parameters are conducted and the fiber orientation distribution is measured for each trial either directly or through the analysis of fabric properties. This lengthy procedure is often repeated until some acceptable FOD and fabric properties are achieved within the limitations and inherent variability of the fiberweb forming system. In addition, traditional methods lack the ability to generate lightweight fiberwebs. The limitation in producing lightweight fiberwebs is due to the non-uniformity in basis weight produced by these traditional techniques.

Our research addresses the development of a new technology for producing fiberwebs with positively controlled fiber orientation and FOD by the use of electrostatic forces acting on charged fibers and improving the industry's ability of generate low weight structured webs.

Earlier efforts in fiber transfer, orientation and separation in electrostatic spinning were reviewed and analyzed by Dogu [Dogu 1975, 1976]. Robert Jr. *et al* reported improved electrostatic spinning system equipped with a new twisting and fiber feeding device [Robert 1984]. In these electro-spinning systems, they employed corona charging methods. These systems experienced difficulties in controlling fiber orientation and speed. This may be due to the fact that corona charging is not reliable and charges imparted on the fibers are varying. Yao invented an apparatus of electrostatic opening and short fiber separation. This device aligns cotton fibers in a carded web using electrostatic field forces acted upon by a pair of metal belts, which are connected to a high DC tension generator [Yao 1994].

Nagi-Zade and Grosberg studied fiber orientation in electrostatic fields, and reported that fibers with only field-induced polarization do not have enough force available to straighten such fibers and the orientation of these "uncharged" fibers would take approximately one second [Nagi-Zade and Grosberg 1973]. In a recent work, Hou studied short fiber motion in electrostatic fields by high-speed digital photography. It was found that short fibers with a conductive finish in an electrostatic field are oriented perpendicular to the equi-potential line and move along the field line [Hou 2001]. It is essential to treat fibers with a conductive surface finish. Generally, it is desirable to impart a surface resistivity of 10^5– $10^7\Omega$ for direct charging.

Required time to charge cotton fibers in an electrostatic field are strongly correlated to the field strength. For example, a cotton fiber resting on a positive pole acquires a charge amounting to 2-4 x 10^{-11} C in 23 msec and moves toward the negative pole at an average speed of 7.1 m/sec [Nagi-Zade and Grosberg 1973]. This amount of charge is insufficient for fiber orientation and movement required for electrostatic web forming process. Moreover, charge transferred from the electrode to polyester and other synthetic fibers are far lower than the amount acquired by cotton fibers. In this research conductive surface finish and direct charging are employed for positively controlling the fiber motion that allow for the control of FOD and fiberweb basis weight and its uniformity.

It is evident that proper arrangement of electrodes, direct charging and a conductive surface finish are needed to

provide the positive control of fiber orientation in a web minimum or no 'fiber hook' formation. However, the precise control of fiber orientation and motion with electrostatic forces to form webs with predetermined FOD has significant challenges. The problems stem from non-uniformity in surface finish applied, fiber geometrical parameters, and deviation of the electrostatic field lines from the theoretically predicted one and other variations.

To achieve the goal of our research, it is essential to conduct computer modeling of fiber delivery, fiber movement in electrostatic field, and the fiber orientation in the web forming zone. Results obtained from the model using process variables (electrostatic field strength, air flow, fiber geometrical parameters, fiber feed conditions, etc.) will be used to develop a prototype of electrostatic web forming system.

Mathematical models were constructed by integrating the governing equations with a model that describes the fiber configurations. The model considered different electrostatic field strengths and airflow field. The geometrical configurations of a modeled space (the space in which fiber movement and web formation take place) were deduced. The modeled space configuration is the determining factor in designing and engineering the real equipment. A numerical method was developed for solving the model. For result analysis, visualization programs were established to illustrate and animate the simulated fiber movements.

Fiber delivery and orientation with different initial and boundary conditions were analyzed using the computer modeling results. These results proved that the electrostatic field could effectively provide the force needed to positively control the fiber orientation.

2. COMPUTER MODELING

2.1 The Mathematical Model

Figure 1 represents one element, *ds*, of a moving fiber. The basic governing equation of this element can be written as

$$\rho_l \frac{\partial \vec{v}}{\partial t} = -\frac{\partial \vec{F}}{\partial s} + \vec{f}_{surf} + \vec{f}_{body} \qquad (1)$$

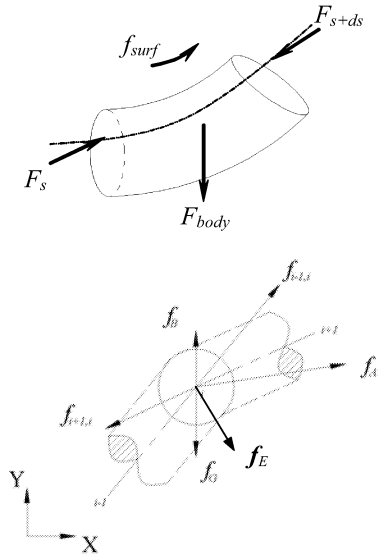
where ρ_l: fiber linear density; a: element acceleration of speed; F_s: concentrated force; f_{surf}: surface force; F_{body}: body force; v: element velocity, f_{body}: body force of unit length. In computer simulations, the concentrated forces F_s and F_{s+ds} on this element are considered as interactions from elements that connected to it; surface force f_{surf} is air drag; body force F_{body} is a combination of the fiber's gravity, electrostatic force, and buoyancy.

The selection of an applicable description of the fiber configurations is the key that links the above discussed governing equations to computer simulations, in which fibers are studied in discrete forms. In this research, the configurations and characteristics of a fiber are described as spheres connected by inflexible zero-diameters, zero-mass rods, which are depicted in Figure 1(b). The air drag and electrostatic forces only act on the spheres. The fiber was considered to be inextensible. Therefore, the following dynamic equation can be used to describe the movement of one single sphere of the fiber in the electrostatic field

$$m\vec{a}_i = \vec{f}_{i-1,i} + \vec{f}_{i+1,i} + \vec{f}_a + \vec{f}_b + \vec{f}_g + \vec{f}_e \qquad (2)$$

where m: mass of a single sphere; a_i: velocity of the sphere; f_a: air drag; f_b: buoyancy; f_g: gravity; f_g: electrostatic force; $f_{i-1,i}$: action from sphere i-1 to i; $f_{i,i+1}$: action from sphere i to i+1.

Combined with other constraints and initial and boundary conditions, the above equation then can be integrated with the basic governing equations, and generate the discrete form equations which are used in the computer model.



(b)

Figure 1 (a): Forces acting on a fiber element;
(b): Forces acting on one sphere

2.2 Numerical Solution

Numerical algorithms were developed in solving the derived equation sets. These equations form the basis to predict the movement of the fiber at a particular instance of time and its trajectory. The method used can be applied to two–dimensional models and three–dimensional models.

In order to solve the equations, one objective function should be selected for the model. It was defined as:

$$f_{OBJ} = \sum_{i=1}^{K} (L_i^{(n+1)} - 1)^2 < C \qquad (3)$$

This function was selected because during the whole movement, the extension of a fiber is neglected and so the length of every single segment (L_i in Equation 3) should remain as a constant value, as C is a predetermined convergence criterion. The solutions are robust during computations. Computation codes were written in C++.

2.3 Data Visualization and Analysis

Visualization programs are developed to display the data from solutions. The program displays every single position of the fiber and its trajectory in the modeled space at any instance of time. They are developed using the C++ with OpenGL. They are also capable of generating animations of the fiber movements.

3. RESULTS AND DISCUSSIONS

Figure 2 depicts the geometrical configuration of the model, H1 is the height of the electrodes, H2 is the height at which the fiber is discharged from the feeding elements, which is formed of rollers and aspirator.

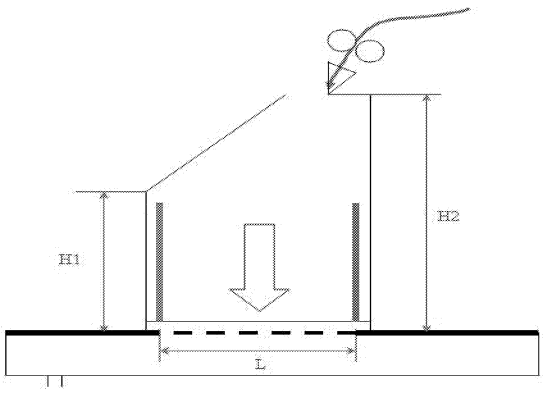


Figure 2. The geometrical configuration of the model

Fibers in the model are simulated as cotton fiber. The parameters are listed in the following Table 1. The simulated fiber is considered to be negatively charged with a surface density of 3.0 X 10^{-8} C/cm^2 contact with high negative potential at aspirator surface. The fibers enter the modeled space at an initial speed, which is actually determined by the aspirator's configuration and parameters, and then fall into the influences of the electrostatic field.

Table 1: Fiber Properties

Fiber Density:	1.55×10^{3} kg/m^3
Fiber Diameter:	1.5×10^{-5} m
Fiber Length:	2.0×10^{-2} m

Different factors, such as fiber's initial speed, fiber's incident angle, and electrostatic field were studied through computer modeling. Both two-dimensional and three-dimensional results were obtained from the computation. This paper reports three-dimensional results. Results can be used to carry out studies on the optimization of system's geometrical and operational parameters. Resultant data were analyzed, and displayed by the visualization program, which are shown in the following snapshots.

3.1 Optimization of the Electrodes' Height

At the beginning the proposed height of the positive electrode (H1 in Figure 2) is 0.5 m, modeling results shows that at the field strength of 20 KV/m, 0.15 m is enough for the fiber to be aligned to the direction of the field lines. Figure 3 shows the fiber movement inside the original field. Fiber enters the field at speed of 25 m/s and incident angle of 20 degree. It is obvious that a 0.5 m height is sufficient to provide orienting zone volume under this condition.

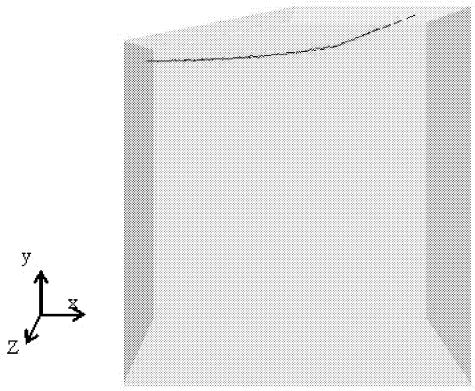


Figure 3. Fiber movement in the original design

3.2 Fiber's Initial Conditions

3.2.1 Incident Angle

With different incident angles, the trajectories of the fibers are very different, Figure 4(a) and 4(b) show the movements with incident angles of 20 degree and 35 degree respectively. Both fibers have an initial speed of 25 m/s. and field strength is 20 KV/m. The modeling results show that the end point of the trajectory changes significantly when the incident angle increases from 20 to 35 degree. The end point of the fiber entered with 35 degree angle is 4.7 cm lower than the fiber entered with 20 degree angle.

Figure 4(c) shows a different situation with incident angle of 90 degree. In this case the fiber was drop into the alignment zone perpendicularly; results show that the fiber hits the bottom of this system before it gets aligned to the field lines. In this case the H1 and or field strength has to be increased to provide space and time for the fiber to get oriented as desired. This issue is dealt with later in the paper.

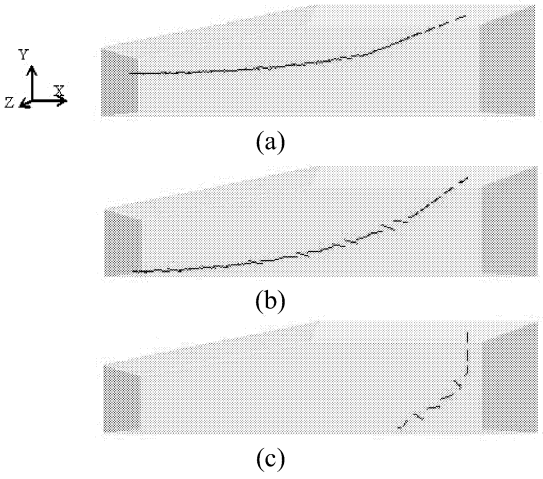


Figure 4. Fiber movements according to different incident angles

3.2.2 Initial Fiber Entry Position

Computer modeling also was used to study the fiber's initial entry position, which could have important impacts on the configuration of the electrostatic field. When the field strength is strong enough, which is 20 KV/m. In this model, changes in the horizontal direction within a certain range did not cause significant change of fiber alignment at the end point. Figure 5 shows the movement of a fiber that started 5 cm closer to the positive electrode than the fiber shown in Figure 4(a). At the end point, both fibers are aligned to the field lines (the angle is close to 0 degree).

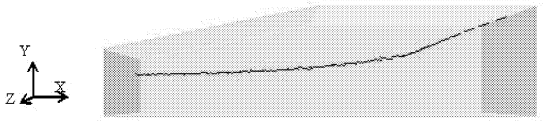


Figure 5. Fiber's movement at a different fiber discharge point

3.2.3 Initial Fiber Speed

With the same incident angle, the fiber with a higher initial speed requires a longer travel distance along the direction perpendicular to the field line, as shown in Figure 6, which is the modeling result of a fiber with a initial speed of 45 m/s with a incident angle of 20 degree, compared with the movement shown in Figure 5, which is of a fiber with a initial speed of 25 m/s, the end point is 5.2 cm lower. The time length of this movement is 0.023 sec, the frame rate of these snapshots is 0.001 sec/frame.

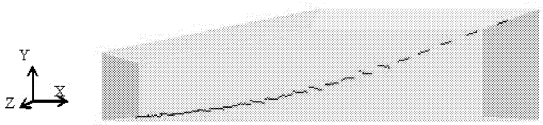


Figure 6. Fiber movement at initial speed of 45 m/s

3.3 The Influence of Airflow

The equipment used in this research has vacuum suction floor, which causes an airflow, which has a direction as shown in Figure 2 by the arrow, the airflow speed is about 1 m/s. Computer modeling predicted that the end point of the fiber will be 1.8 cm lower than the end point of a fiber moving in the electrostatic field with zero airflow speed, provided other conditions are the same: initial speed 25 m/s, incident angle 20 degree, field strength 20 KV/m, and same fiber discharge position.

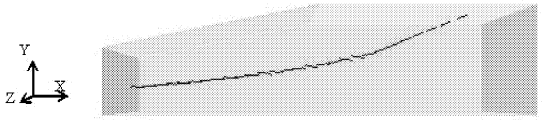


Figure 7. Fiber movement under the influence of electrostatic field and airflow

3.4 Electrostatic Fields

Electrostatic field strength is determined by the power voltage, configurations of electrodes, and the geometric parameters. Two-dimensional results, which are not shown reported here, proved that stronger field strength will speed up the alignment of the fiber, but with same geometric configuration, this is going to require a power supply with higher voltage. Computer modeling results show that for above configurations, 20 KV/m is quite enough to align fibers because the angles at the end points are close to zero to the field lines. This is revealed in Figure 8(a), which shows the Fiber movement under field strength of 50 KV/m. Fiber alignment at the end point is almost the same compared with field strength of 20 KV/m.

In another case, as shown in Figure 8(b), when the incident angle is 90 degree, field strength of 150 KV/m helps the fiber movement to avoid the situation which is depicted in Figure 4(c), in which the field strength is 20 KV/m. With high field strength of 150 KV/m, the fiber can get aligned before it hits the system's bottom.

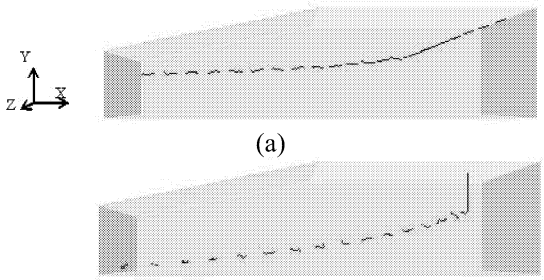


Figure 8. Fibers movement under field strength of 50 KV/m and 150 KV/m

4. CONCLUSIONS

Computer modeling shows that the fiber delivery and orientation in electrostatic fields are determined by several factors, such as geometric configurations, electrostatic field, airflow, and fiber's initial conditions. Fibers leaving the feeding system with relatively smaller incident angle (20 degree in this research) need shorter alignment zone height to be aligned, but changes in the horizontal direction within a certain range did not cause significant change of fiber alignment at the end point. With the same incident angle and field strength, high fiber initial speed will need an increase in the alignment zone height; the vacuum suction will also require an increase in the alignment zone height. Stronger electrostatic field strength improves the alignment of fibers with large incident angles. These results can be used to optimize the equipment of the process.

The findings from computer modeling can serve as a powerful tool for not only understanding how to align fibers to form fiberwebs using electrostatic field, but also engineering the fabric with predetermined geometry and hence predetermined properties.

REFERENCES

Dogu,.I. (1975) "Fundamentals of Electrostatic Spinning Part I." *Textile Research Journal*, **35**,521-532

Dogu,.I. (1976) "Fundamentals of Electrostatic Spinning Part II." *Textile Research Journal*, **36,** 676-691

Hou, Y. (2001) MS Thesis Fall 2001, University of Massachusetts at Dartmouth

Nagi-Zade, A. T., and Grosberg, P. (1973) "The Manipulation of Fibers by Means of a Charged Field." *Journal of the Textile Institute*, **64**, 431-436

Robert, Jr., K.Q., A. Baril, R.B. Reif, P.E. McCrady, (1984) "New Electrostatic Spinning Concept for Cotton." *Journal of Engineering for Industry*, **106**, 247-252

Yao, S. C., (1994) "Electrostatic Opening and Short Fiber Separation Apparatus for Carding Machines", US Patent 5,327,617

CORRESPONDING AUTHOR ADDRESS:

Dr. Yong K. Kim
Chancellor Professor
Textile Sciences Department
University of Massachusetts at Dartmouth
Dartmouth, MA 02747 USA
Email: ykim@umassd.edu

BIOGRAPHY

Yiyun Cai, a post-doc in Textile and Apparel Technology and Management at NC State University. He received a Ph.D. in fiber and polymer science from NC State University in 2003; he also received a M.S. degree in aerospace engineering in 1999 from Beijing University of Aeronautics and Astronautics (China). Yiyun's research interests include computer modeling of textile processes, electrostatic properties in textile systems, and innovative in textile manufacturing.

Yong K. Kim, a Chancellor Professor of Textile Sciences at UMass Dartmouth, joined the faculty in 1981 when he earned a Ph.D. in fiber and polymer science from NC State. Yong also earned an M.S. in textile engineering from Seoul National University (Korea) in 1974. His research interests include textile process design and manufacturing systems, electro fiber coating, mechanics of fibrous structures, and micro- and nano-composite materials.
ykim@Umassd.edu
508-999-8452
www.tesumassd.org

Abdelfattah Seyam, a Professor in Textile and Apparel Technology and Management at NC State since 1991, earned a M.S. Degree in Textile Eng. in 1978 from Alexandria Univ. (Egypt) and a Ph.D. in Fiber and Polymer Science from NC State in 1985. He then held research positions at Burlington Inc. and Valdese Textiles. His research interests include mechanics of woven fabrics and technologies for apparel automation and web forming. Currently, he is exploring new research areas in direct garment manufacturing using meltblown technology, micromachines in textiles, and integrating electronic components in smart textile structures.

Neural Network Modeling of Fabric Pilling Evaluation

Binjie XIN
Jinlian HU
Institute of Textile & Clothing
The Hong Kong Polytechnic University,
Hung Hom, Kowloon
Hong Kong
E-mail: tcxinbj@polyu.edu.hk

KEYWORDS
Neural network, Image processing, Textile, Pill evaluation.

ABSTRACT

In quality control of modern textile industry, it is necessary to develop some objective methods to evaluate cloth appearances, such as pilling, wrinkling and seam puckering. This paper presents a new method to evaluate the fabric pilling based on artificial intelligence techniques (image analysis and neural network). In this work, we first digitalize the fabric appearance using digital camera, after that, we extract the essential features of pills on fabric surface using image analysis techniques, which could characterize the fabric appearances objectively. At last, we apply neural network models to establish the correlationship between these essential features and final rating grade of fabric pilling. The experimental results in this paper show that good correlation can be achieved between objective and subjective evaluation data. It proves that artificial intelligent system is able to simulate the functions of human eyes and brain. The aim of this paper is to develop an artificial expert system for the quality control of fabric appearances, so as to liberate human being from the tedious, repeatable, time-consuming subjective judging works. It will help us to build up an objective evaluation and quick response management system for modern textile industry.

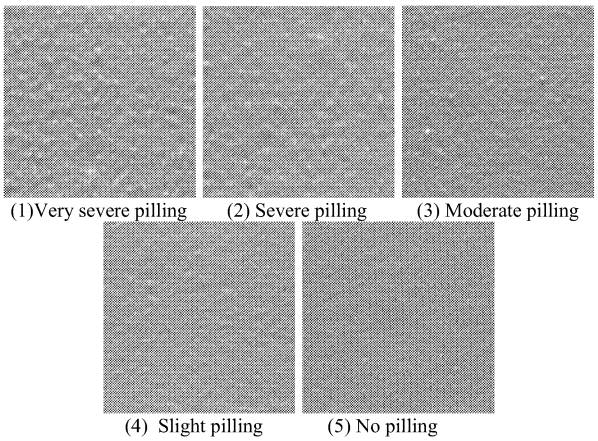
INTRODUCTION

Pilling on fabrics is a well-known phenomenon and this unpleasant appearance can seriously compromise the cloth's acceptability. The generation of pills begins with a migration of fibers to the external part of yarns, so that fuzz emerges in the fabric surface. Due to friction, fuzz gets entangled forming pills, which remain suspended from fabric by long fibers. Pilling is an effect mainly caused by wear and tear that spoils considerably the original appearance of a fabric, and many researchers are investigating how to improve the pilling resistance ability of fabric, and also make studies on how to evaluate its pilling grade objectively. Some good attempts on pilling evaluation using image analysis techniques have been reported in referrences(Cooke et al.1985, Konda 1988, Krebs 1989, Ramgulam 1993, Xu 1997, Hector et al.1998, His et al. 1998, and Wagner 1999); however most of these researches focused on the algorithms of image analysis how to identify and characterize fuzz or pills, only a few works done on mathematical judging models. Our previous research(Xin 2002)already successfully developed a set of image analysis algorithms to identify pills and define some essential features (number, area, contrast and density) to characterize pills. In this paper, we further introduce an artificial intelligent method to simulate the mechanism of subjective evaluation based on two-layer feed-forward backpropagation neural network models (Richard,1987).

METHODOLOGY

The development of artificial intelligent system based on neural network includes four main steps: data collection, neural network modeling, system training, and validation.

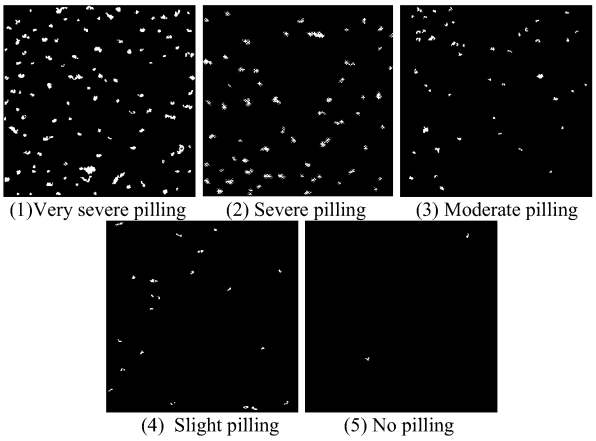
Data collection



Figures 1: Standard Pilling Images

In this project, we select a plain knitted wool fabric, which is weighted 150g/m2 with 16 wales/inch and 22 courses/inch, folded yarns dyed in gray color. Each fabric sample was cut into approximately 105mm (43/16 in.) square and subjected to laboratory abrasion on a Random Tumble Pilling Tester for certain time according to ASTM Test Method D3512. In order to get different pilling grade for evaluation, we have 90 fabric samples, which are subject to different abrasion time: from 0 to 60minutes. After digitalization of fabric appearance image using digital camera, a set of image processing algorithms(image enhancement, histogram fitting, template matching and image segmentation) are developed to identify pills from fabric surface, Figure 1and 2 show original standard pilling image and its related binary

image after image segmentation separately, and then pill properties including number, total area, average area, contrast and density are measured by using image analysis techniques(Xin 2002). These five features have been reported to be related with the subjective rating grade, so in this paper we select them as the input independent parameters of neural network model. The subjective evaluation results are from 10 experienced judges, in order to improve the reliability of their subjective evaluation, the average judging grade is chose as the final subjective rating grade.



Figures 2: Image Segmentation of Pilling Standards

Neural network modeling

We design three kinds of neural network with different neurons functions as the following specifications (programming in Matlab), and investigate that which neurons function could simulate the human judging behaviour with the best performance; all of them are 5 inputs and 1 output feed-forward backpropagation network as shown in Figure 2.

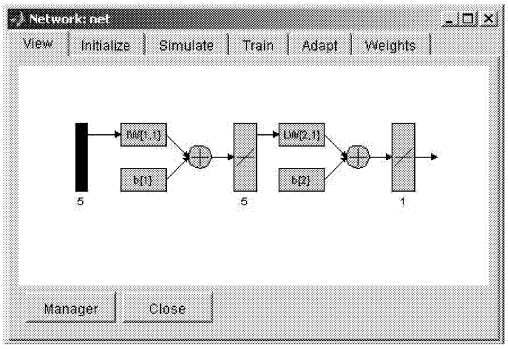


Figure 2: Design of Neural Network

To be more explicit, consider a two-layer feedforward NN model consisting of a stream of input vector X of 5 neurons; i.e. $\{x_1, x_2, x_3, x_4, x_5\}$, to the input layer, and an output layer with a single neuron. The network has the general form:

$$Y = f(X,\Theta) = f\left\{b_0 + f_1\left(\sum_{i=0}^{n} w_i x_i + b_i\right)\right\} \quad (1)$$

where Y represents the generated output response variable; Θ denotes the overall parameters space; w_i denotes the connecting weight of the ith input neuron; b_0 and b_i are the bias nodes; f_1 and f are the activation functions of first layer and second layer respectively.

Model 1: Two layers, the first layer have five PURELIN neurons; the second layer has one PURELIN neuron.

$$f(x) = x; \qquad f_1(x) = x; \quad (2)$$

Model 2: Two layers, the first layer have five LOGSIG neurons; the second layer has one PURELIN neuron.

$$f(x) = x; \qquad f_1(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}; \quad (3)$$

Model 3: Two layers, the first layer have five TANSIG neurons; the second layer has one PURELIN neuron.

$$f(x) = x; \qquad f_1(x) = \frac{e^x}{1 + e^x}; \quad (4)$$

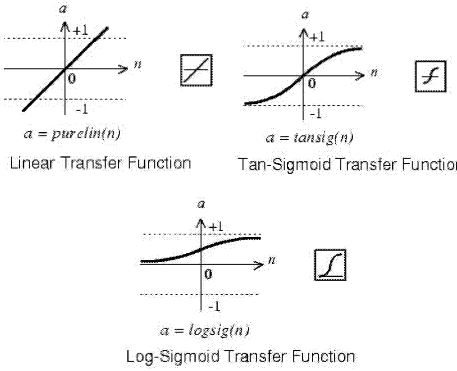


Figure 3: The Definition of Transfer Functions

System training

Out of the 90 data set available, 45 of them is selected as training data to the network and the rest is used as validation data for examining the performance of the trained model. We first input the training data to the network and perform gradient descent using the Levenberg-Marquardt Algorithm. Parameters are adjusted iteratively until global error function converges to some specified minimum so that good generalization is achieved. Some important considerations are the number of training iterations, the learning rate and the momentum coefficient.

Suppose the estimated grade y is governed by the underlying function $f(x;\theta^*)$, where x is the set of input variables and θ^* represents the true values of the parameter vector θ from the parameter space Θ for the function which models

the process. With n observations, the simulation process is represent by:

$$y_i = f(x_i;\theta^*) + \varepsilon_i, i = 1 \sim n, \theta^* \in \Theta \quad (5)$$

The least-square estimate of θ^* is the $\hat{\theta}$ which is obtained by minimizing the error function:

$$S(\theta) = \sum_{i=1}^{n} [y_i - f(x_i;\theta)]^2 \quad (6)$$

The Levenberg-Marquardt algorithm uses this approximation to the Hessian matrix in the following Newton-like update:

$$\theta_{k+1} = \theta_k - [J^T J + \mu I]^{-1} J^T e \quad (7)$$

Where, J is the Jacobian matrix that contains first derivatives of the network errors with respect to the weights and biases, and e is a vector of network errors. When the scalar μ is zero, this is just Newton's method, using the approximate Hessian matrix. When μ is large, this becomes gradient descent with a small step size. Newton's method is faster and more accurate near an error minimum, so the aim is to shift towards Newton's method as quickly as possible. Thus, μ is decreased after each successful step (reduction in performance function) and is increased only when a tentative step would increase the performance function. In this way, the performance function will always be reduced at each iteration of the algorithm.

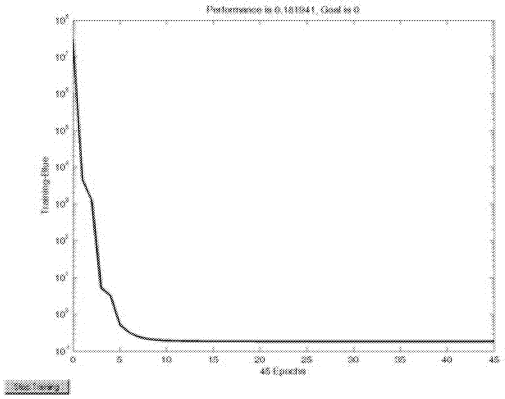


Figure 4: The Training Performance of Model 1 (PURELIN + PURELIN)

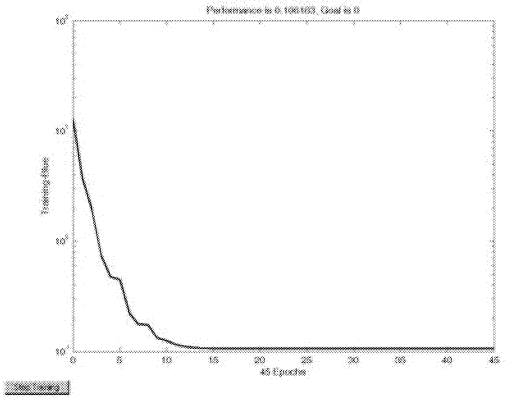


Figure 5: The Training Performance of Model 2 (LOGSIG + PURELIN)

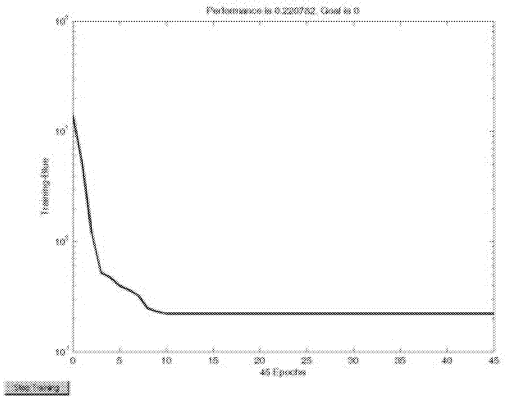


Figure 6: The Training Performance of Model 3 (TANSIG + PURELIN)

We trained three kinds of neural network models separately, and found that model 2 has the minimum training error, its final performance is only 0.106163, while the performance of other two models are 0. 181941 and 0.220782, it proves that model 2 with LOGSIG neurons function has the best training performance. So we choose model 2 as our neural network model.

Validation

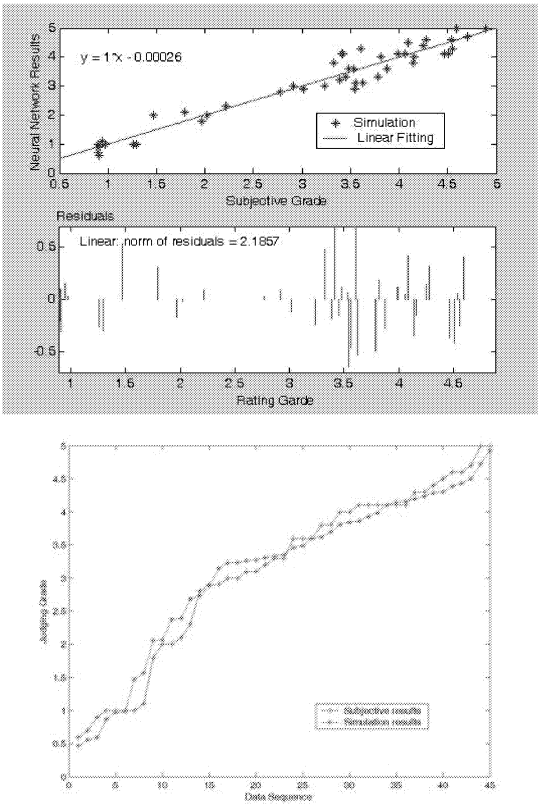


Figure 7: Validation of Neural Network Model

After we trained the neural network models, we chose the left 45 data set for validation of this system as shown in

Figure 7. The experimental results show that the artificial intelligence system developed in this project could achieve very well consistent with the human evaluation results.

CONCLUSION

An artificial intelligence experts system based on neural network method is developed for objective evaluation of fabric pilling evaluation. We tried three kinds of two layer feed-forward backpropagation neural network models based on three different transferring functions. The final performance of three different neural network models show that Model 2 based on LOGSIG transfer function achieved the best simulation results with the minimum estimation error. The testing results for validation show the artificial intelligent system developed in this project has very good correlationship with human being operation. It proves that it is workable to replace the traditional subjective evaluation method based on human eyes and brain by using modern artificial intelligence.

REFERENCES

Cooke, W.D., Pilling Attrition and Fatigue. Text. Res. J., 55, 409-414 (1985).

Hector, C. A., Millan, M. S., Torres Y., and Navarro, R., Automatic Method Based on Image Analysis for Pilling Evaluation in Fabrics, Optical Engineering, 37(11) 2937-2947 (1998).

His, C.H., Bresee, R.R., and Annis, P.A., Characterizing Fabric Pilling by Using Image-analysis Techniques Part I: Pill Detection and Description, J.Textile Inst. 89(1), Part I, 80~95, (1998).

Konda, A., J.Textile March. Soc. Japan, Vol 36, No 3, 96-101 (1988).

Krebs, C.J. Ecological Methodology, Harper & Row, New York, NY, USA, (1989).

Ramgulam, R.B., The Objective Assessment of Fabric Pilling, Part I: Methodology, Textile Inst., 84(2), 221-223 (1993).

Richard P. Lipmann, an Introduction to Computing with Neural Nets, IEEE ASSP Magazine, April, 1987.

Wagner J.R. & Associates, Development of an Image Analysis Algorithm For Assessing Pilling, 1999 Nonwovens Conference, TAPPI Proceedings, 305-311 (1999).

Xin B., Hu J., Yan H., Objective evaluation of fabric pilling using image analysis techniques, Text. Res. J., Dec 2002, 72, 12.

Xu. B., Instrumental Evaluations of Fabric Pilling, J.Text.Inst. (88), 488-500 (1997).

BIOGRAPHY

BINJIE XIN received the BA degree in textile engineering from the University of Qingdao, the MS degree in textile material from Donghua University. He is currently an research associate at the Hong Kong Polytechnic University. His research is focused on the application of image analysis and artificial intelligence in textile field.

JINLIAN HU obtained her Bachelor's and Master's degrees from China and her PhD from UMIST, UK. She is currently an associate professor at the Hong Kong Polytechnic University. Her research is focused on smart material, textile structure and mechanics, cloth simulation, computer vision and artificial intelligence.

TEXTILE INDUSTRY INVENTORY AND SALES MANAGEMENT SIMULATION

DEVELOPMENT OF AN INVENTORY MANAGEMENT SPREADSHEET SIMULATOR FOR A LARGE APPAREL COMPANY

Jeffrey A. Joines, Kristin A. Thoney, and Ben R. Martin
North Carolina State University
College of Textiles, Campus Box 8301
Raleigh, NC 27695 USA
E-mail:JeffJoines@ncsu.edu

ABSTRACT

Despite carrying large amounts of inventory, the customer service levels of a large apparel company remained unacceptable. A simulation study using actual demand, finished goods inventory, and forecasts was conducted to evaluate the average inventory and service levels of different proposed inventory policies. The best policy from the simulation study set inventory targets at the SKU level while taking into account forecast inaccuracies. A pilot implementation of this inventory policy was successful. The company adopted the policy which ultimately resulted in large savings.

1 INTRODUCTION

The U.S. textile/apparel industry is facing many challenges in its effort to compete in the global marketplace. These challenges include demand variation, long lead times, and raw material supply fluctuations. Many companies with long lead times are forced into make-to-stock situations in order to respond to their customer demand quickly and maintain market share. Too much inventory causes higher costs and may lead to product obsolescence/spoilage, while too little inventory may cause stockouts and low customer service levels.

In an effort to retain market share, a large apparel company is attempting to lower finished goods inventory while maintaining or increasing customer service levels. The company was unable to support acceptable customer service levels, even with large inventories. As a result, the company would try to raise inventory levels across the board until customer service reached an acceptable level. In most cases, this approach resulted in increased costs without really addressing the service problems. In fact, the products having high service levels continued to have high service levels, but with higher inventory levels. In contrast, products with low service levels had an increase in service, but many still were not at acceptable levels.

This paper discusses the development of alternate inventory models at the SKU level, analysis of the models through simple spreadsheet simulations, and the implementation of a new inventory policy. Section 2 provides a complete description of the problem. In Section 3, the development and analysis of the inventory models are discussed. Section 4 discusses the pilot implementation of the inventory policy selected, and Section 5 presents the conclusions.

2 DESCRIPTION OF THE PROBLEM

This section describes specific details of the problem addressed in developing the inventory model for the company.

2.1 Customer Ordering Pattern

Customers can place orders for items at any time for any given quantity. This presents a significant problem for finished goods planning. If the forecast is not reasonably aligned with customer demand, then the company will have trouble fulfilling the orders unless there is sufficient inventory on hand.

2.2 Forecasting

In this company, forecasts are made at the category level because aggregating all the SKUs into one item tends to result in greater forecast accuracy (i.e., one is better able to predict that a customer might order 1000 of category A than predict an order of 12 for SKU 1). The company then uses a SKU distribution mix (e.g. 10% of the demand is always for SKU 1) to explode the forecast at the category level down to the SKU level for production planning purposes. Currently, forecasts are made on a monthly basis in an effort to lower forecast error but planning is done weekly.

2.3 Forecast Bias

An important issue to accommodate in the finished goods inventory planning for this apparel company involves their significant forecast bias. Forecast bias is defined as the sum of the forecast errors over a given period of time. Forecast error is defined as the difference between the forecasted demand and the actual demand for a given time period. If the forecast is unbiased, then the forecast errors will be randomly distributed yielding a bias near zero. In the case of the apparel company, due to financial and other influences, the forecast may be biased upward or downward for long periods of time. Significant biases over long period of times will result in too much inventory in the case of a positive bias or poor service due to a lack of inventory in the case of a negative bias.

2.4 Long Lead-Times

The lead-time of the products of the apparel company averages six weeks but can take as long as twenty weeks. In an environment where the lead-time is constant, very little inventory is required to maintain customer service levels. However, in the situation of highly variable lead times, large inventories must be carried in order to support customer service levels when the lead-time is longer than expected.

2.5 *ABC* Inventory Targets

The company's existing inventory policy establishes inventory targets using a classic *ABC* strategy where sixty-five percent of the volume is classified as *A*, twenty percent of the volume is classified as *B*, and fifteen percent of the volume is

classified as *C*. All *A* items receive the same target level, all *B* items receive the same target, and all *C* items receive the same target. Generally, the *C* items will be assigned a higher inventory target than the *A* items due to the expected volatility of demand in a low volume item (i.e., *A* items are highly demanded and therefore produced on a regular basis).

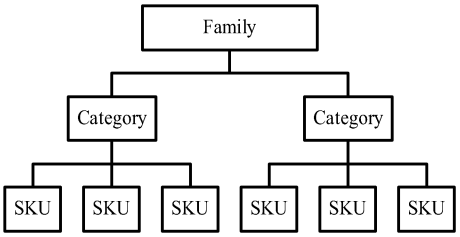


Figure 1: Hierarchical Grouping of Apparel Items

The *ABC* classification is established at the category level or the style/color designation, just as the forecasts are based. A category can contain many stock keeping units (SKU) and is part of a family/style as shown in Figure 1. Therefore, all SKUs within a category will receive the same inventory target. This presents a problem because individual SKUs have significantly different shipping patterns and thus volatilities. In fact, conducting *ABC* analysis at the SKU level revealed that within an *A* category, there will be *A*, *B*, and *C* SKUs. This can be seen in Figure 2 which shows a very popular core style and color that is classified as an *A* category. As you can see, most of the highly demanded SKUs have very high service levels, while the fringe items (*C* SKUs) do not. This causes the overall service level to be low for the *A* category.

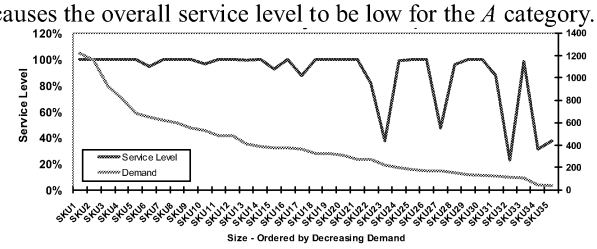


Figure 2: Implications of Setting Target Levels at Category Level (*A* Category Example)

In order to increase the level, the inventory for the entire category is increased. Therefore, the category *ABC* inventory target will assign too much inventory to some SKUs and too little to other SKUs. High inventories and low service levels are a result of this policy. The primary way to raise service levels using this policy is to increase the category inventory target to a level where *C* SKUs within the category have a high service level. This results in needlessly raising inventory on the *A* and *B* SKUs that are already being serviced well.

2.6 SKU Proliferation

At this apparel company, there are more than 25,000 SKUs for which finished goods inventory must be planned. For an inventory system to be used in practice on this many SKUs, the system needs to be simple enough for most employees to understand, yet complex enough to handle the myriad of possibilities that may arise from the large number of SKUs that are stratified in *ABC* classifications.

2.7 Target Weeks of Supply

The company used a target weeks of supply (TWS) as their current category inventory policy. For example, if the style/color was an *A* category, the company might set the TWS at eight weeks, which meant the company would produce based on carrying eight weeks of forecasted demand. Because of the SKU proliferation mentioned earlier, the TWS could more easily be maintained at the category level by the various planners and would be impossible at the SKU level. Therefore, a more automated way of determining TWS was needed if inventory levels were to be determined at the SKU level.

2.8 Other Issues to Consider

There are several other issues that confound finished goods inventory planning at this company. Some of the more important items are given in the following list:

- Little correlation between the forecast and actual demand;
- Manufacturing capacity limitations;
- SKU distribution mix error;
- Targeted service level is 95% for all SKUs regardless of volume.

3 INVENTORY MODEL DEVELOPMENT

3.1 SKU Level Justification

From the problem description given in previous sections, it is clear that the *ABC* Style/Color level inventory targets were not working and that they should be replaced by *ABC* Style/Color/Size targets. To further justify the move from a category level to a SKU level inventory policy to upper level management, our team did some analysis on a large percentage of their Style/Colors. The team looked at both sourced (i.e., items which are sourced completely from the far east) and internal categories (i.e., items which are made in plants owned by the company in the US, Caribbean, and Central America). Figure 3 shows the results of the analysis by reclassifying the SKUs into *ABC* based on SKU level analysis. The *ABC* analysis was done at the SKU level using the same breakdowns (i.e., 65% of demands are *A*s, 25% are *B*s, and the last 10% are *C*s) across all three types of categories. Figure 3 demonstrates, that an *A* category contains a significant number of *A* SKUs but also has several *C* SKUs as well. Even the *C* category contains a few *A* SKUs.

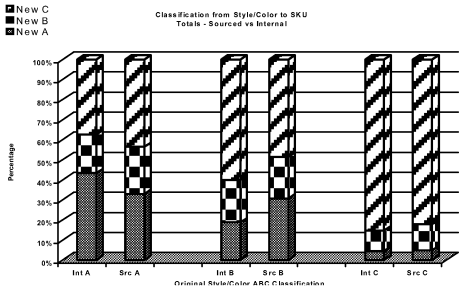


Figure 3: SKU ABC vs. Category ABC Analysis

3.2 Inventory Models

A series of SKU level inventory models were developed over a six month period using input from the company. Five dif-

ferent inventory models (A-E) were selected as candidate SKU models to be used in further analysis. Model E represents the way the company is currently setting the TWS and will be used as a benchmark. All of the models incorporate two components: safety stock calculation and demand estimate. Models A and B came from a study performed by a consultant group. They were quite complicated and never implemented by the company. They cannot be shown for proprietary reasons. Models C and D were developed by our team based on some standard textbook inventory models with some slight modifications. Based on our dealings with the company, the models had to be simple enough to understand to be truly embraced. Equations 1 and 2 represent the safety stock and demand estimation calculation for Model D.

$$\text{Safety Stock} \quad z \quad {}^{13\text{wks}}_{D} \quad \sqrt{L} \quad MAD_{13\text{wks}} \qquad (1)$$

$$\text{Demand Estimate}_t \quad {}^{t\ L}_{i\ t\ 1} F(i) \qquad (2)$$

The z represents the standard normal value for the desired service level (e.g., 95%), $_D$ is the standard deviation of the actual demand over the past 13 weeks, L is the lead time that depends entirely on the product, $F(i)$ is the forecasted demand at period i, and MAD represents the mean absolute deviation of the forecast error. If the bias has been negative during the past 13 weeks, the model will add the MAD to the safety stock. Otherwise, the model will subtract the MAD if it has been positive. Model D is similar to the standard textbook model, the primary difference being the addition of the MAD term. The demand estimate at time t represents the sum of the forecasted/estimated demand over the next L periods.

3.3 Spreadsheet Simulator

The company had very little simulation experience. This, coupled with the fact that the company believed their situation was very unique, led them to assume a simulation based on statistical distributions for demand, lead time variability, and production would not be beneficial. In order to determine which of the models was better for this apparel company, a simple spreadsheet simulation analysis was performed using their data to drive the simulation. The demand variability was generated based on actual demand and forecasts over the past two years. Table 1 shows the input to the simulation model.

Table 1: Input data used for the simulation

Starting inventory
Historical demand for the previous two years
Forecasts for the same period
Lead times for the various products and production methods
Supply chain fluctuations that can occur
Desired service level for each SKU

By using historical demand, the simple spreadsheet simulator determines what would have happened had the company followed the various inventory models under the assumptions made. The simulation assumes a perfect supply (i.e., what you order/produce will arrive exactly L weeks later. Also, it assumes the demand is perfectly met for the first L weeks. However, the statistics are not collected until after this warm up period is finished. If demand cannot be met for a particular week, a partial order is shipped and the backorders become cancels. The company's policy is that cancels are carried until they are filled (i.e., they become additional demand the next week). Figure 4 shows the algorithm used in the spreadsheet simulator. The simulation was programmed using Microsoft's VBA inside of MS Excel. This provides users with a visual front end that gives them the ability to easily specify style, color, lead-time, etc., download data, and then run the model.

1) Download the input data (i.e., actual demands, forecasts, inventory levels) from the database based on the user specified style/color
2) Specify lead time of product, target service level, etc.
3) For each SKU and model, the following is automatically performed
   a) Fill the demand for the first L weeks
   b) Initialize statistics so that they do not include the warm up period
   c) For each week t
      i) Calculate DemandEstimate_t
      ii) Calculate SafetyStock_t
      iii) $\text{TargetInventory}_t = \text{DemandEstimate}_t + \text{SafetyStock}_t$
      iv) $\text{Production}_t = \max(0, \text{TargetInventory}_t - \text{EndingInventory}_t - \text{WIP}_t)$
      v) $\text{Receipts}_t = \text{Production}_{t-L}$
      vi) $\text{WIP}_t = \text{WIP}_{t-1} + \text{Production}_t - \text{Receipts}_t$
      vii) $\text{EndingInventory}_t = \text{EndingInventory}_{t-1} + \text{Receipts}_t - \text{Demand}_t$
      viii) $\text{Cancels}_t = \text{abs}(\min(0, \text{EndingInventory}_t))$
      ix) If $\text{EndingInventory}_{t-1} > 0$
      $\text{CustomerShip}_t = \min(\text{Demand}_t, \text{EndingInventory}_{t-1} + \text{Receipts}_t)$
      Else
      $\text{CustomerShip}_t = \min(\text{Receipts}_t, \text{Demand}_t, + \text{Cancels}_{t-1})$
      x) Calculate service level
4) Simulation automatically generates the following outputs
   a) Individual worksheet for each SKU and model showing all of the generated data, summary of the results in terms of the two metrics (service level and average inventory level)
   b) Summary table of all SKUs in terms of category for each model
   c) Summary graphs for average inventory and service level at the category level (see Figures 4 and 5)

Figure 4: Spreadsheet Simulation Algorithm

The safety stock and demand estimate for the individual models is calculated in Steps *3ci* and *3cii*. Again, note the DemandEstimate_t is the estimate of the total demand for the next L periods (i.e., demand over the lead-time). Based on the target inventory, ending inventory, and work in process (WIP), the production orders for the week can be calculated. Then, the WIP is updated based on the production and receipts for the week, which is the production order for the t-Lth week since a perfect supply is assumed. The ending in-

ventory (Step *3c-vii*) may go negative which represents a backorder and a cancellation. The amount shipped to the customer is based on the current inventory level, demand, receipts, and cancellations. The system automatically generates summary statistics and graphs as seen in Figures 5 and 6 for all five models for each of the SKUs.

3.4 Results

One hundred different style/color categories were selected to perform the analysis. The styles ranged from core to seasonal products, small number of SKUs (9) to a large number (35), department to mass stores, as well as different product lines in order to capture all the characteristics of very diverse lines. From Figures 5 and 6, one can see the service levels compared with the average inventory.

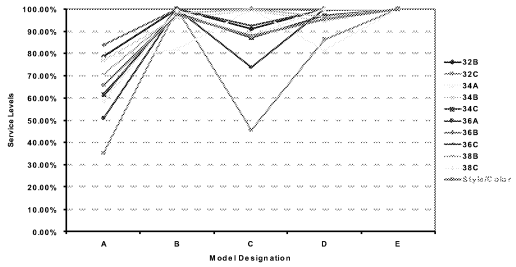


Figure 5: Service Level for Representative Category

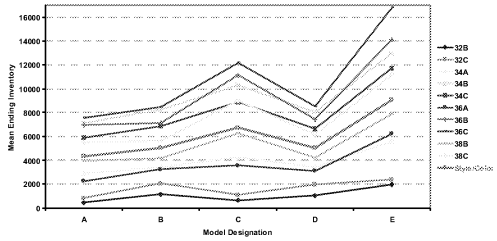


Figure 6: Mean Inventory for Representative Category

Upper level management was convinced by the simulation results that setting the inventory targets at the SKU level would reduce the average inventory and maintain or improve service levels in many cases. They agreed that using a statistical model to determine appropriate safety stock to meet required service levels was better than the current policy. Model D (see Equations 1 and 2) was shown to statistically perform the best across all the evaluated categories at the SKU level and was developed with the company's input during Fall 2002. Due to confidentiality, the actual statistical results cannot be shown. The model performs the best due to the inherent forecast bias that exists. As the forecast becomes better, the MAD term will drop out.

4 PILOT TEST

From the simulation results, a methodology has been developed that provides a direct relationship between inventory and service levels as well as establishes customized targets for each SKU. A pilot test was designed and implemented to evaluate the capabilities of the model under real conditions. Part of the pilot was to see if the current systems could be used to drive a statistical model. Much of our time was spent cleaning up the data before actually using it. The pilot was designed to determine the effect of the new targets and the effects of replacing the company's forecast with a moving average forecast, an idea that management wanted to experimented with. In addition, the pilot was untaken to help determine factors that resulted in desirable and undesirable behavior in both inventory and service levels and make recommendations for wide scale deployment of SKU level inventory targets.

Figure 7 shows the process that was used. The forecast, 13 weeks of historical demand, and target service level were input into the statistical model manually. The number of units required was converted into a weeks of supply (WOS) and then uploaded into the i2 planning engine which then planned production. As a result, the plants sent finished goods to the distribution center.

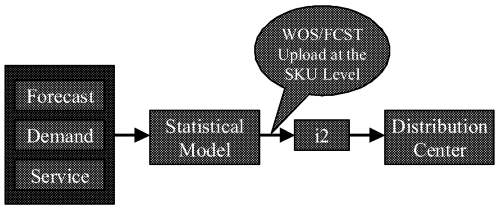


Figure 7: Process Used

The following four test groups, each with 6 style/colors, were used: NF/NT – New Forecast/New Targets, OF/NT – Old Forecast/New Targets, NF/OT – New Forecast/Old Targets, and OF/OT – Old Forecast/Old Targets. Figures 8-10 show results of the new model on three categories. Figure 8 shows the inventory levels dropping by a large percentage while maintaining service level. Figure 9 shows a similar trend for many weeks. However, after several weeks of decreasing inventory, the planner became nervous and scheduled more production, ignoring the model (intervention). Figure 10 shows inventory levels dropping but service was quite poor during periods 49-52. However, plenty of bulk inventory existed in the DC.

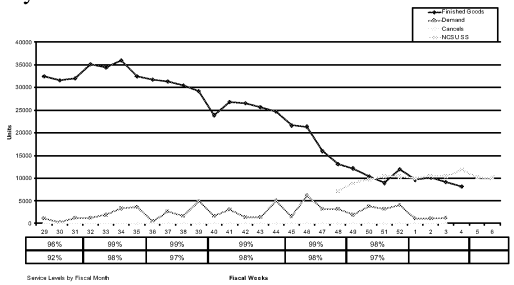


Figure 8: Pilot Working Well

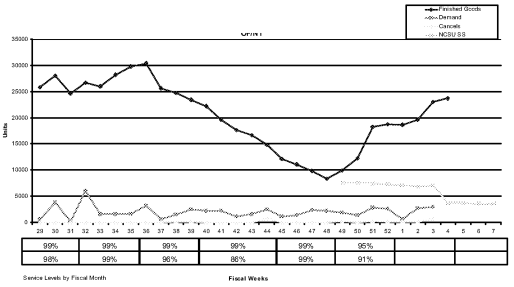


Figure 9: Pilot Working Well Until New Inventory Policy Abandoned

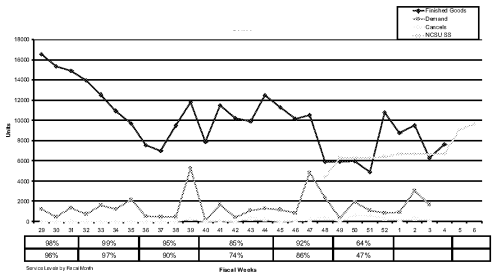


Figure 10: Poor Service Levels but Plenty of Bulk Inventory Existed

Figure 11 shows a summary of the results from the pilot test. The test was designed such that each method would have similar categories in similar plants for comparison. At this point, it was difficult to really distinguish between the various models and the production problems that occurred.

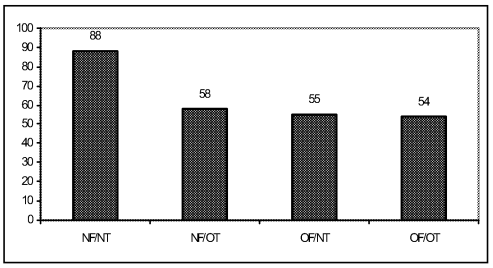


Figure 11: SKUs at 95% Service for Pilot Test

Several problems were discovered as depicted by the large boxes in Figure 12 which needed to be addressed before a system-wide implementation. In virtually all cases, the model called for sufficient finished goods inventory to meet demand. In some cases, the individual planners had too much control and did not believe in the model and intervened. The minimum order quantities (MOQ) on the sew plants were to low to allow small lots and had to be fixed. In many cases, there was enough finished goods in bulk, but it took the DC longer than expected to package it, etc. Figure 13 shows the finished goods inventory from two different database systems. One database system is used by the forecasting department, while the other one is used by planning for production purposes. These discrepancies had to be reconciled before actually implementing the inventory policies.

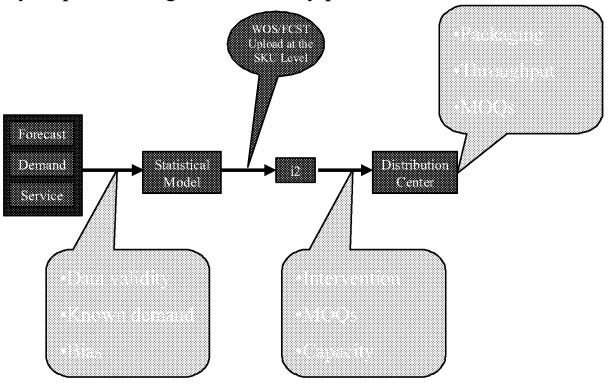


Figure 12: Problems Discovered

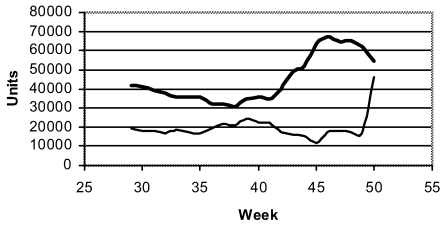


Figure 13: Finished Goods Inventory Discrepancy

5 CONCLUSIONS

An inventory model was developed taking into account the specific characteristics of the apparel company's business. The fact that the model was relatively simple allowed our team to be able to explain it to managers, regardless if they had background in inventory theory. By using the spreadsheet simulator we developed in conjunction with the company's data, our team was able to convince management of the benefits of implementing our inventory model.

The model has been operational at the apparel company since December 2003 and now controls approximately 25,000 SKUs. Figure 14 shows the inventory levels for a category the year the new policy was implemented. The inventories have been reduced by 50% and the service level has been increased to 93%. The new inventory system has resulted in significant savings for the apparel company.



Figure 14: Inventory Levels of a Category the Year the New Policy was Implemented

The current model assumes a constant lead-time when in fact there is a quite a bit of variability. However, at the time, the company did not track the actual lead-times (i.e., cut kits left the USA to be sewn and just showed up). Equations 3 and 4 shows a simple modification that can take into account the lead-time variability. In the actual system, an extra week was added to the lead-time estimate which is roughly implementing the model in Equation 4.

A new Monte Carlo simulation analysis will be run during the summer to evaluate the effects of the lead-time variability on inventory and service levels.

$$\text{Safety Stock} \quad z \sqrt{{}^{2}_{D}L \quad {}^{2}_{L}D} \quad MAD_{13\text{wks}} \qquad (3)$$

$$\text{Safety Stock} \quad z \quad {}^{13\text{wks}}_{D} \quad \sqrt{L \quad z \quad {}_{L}} \quad MAD_{13\text{wks}} \qquad (4)$$

REFERENCES

Seila, A. 2004. "Spreadsheet Simulation." Procedings of the 2004 Winter Simulation Conference, ed. Ingalls, Rossetti, Smith and Peters, 41-48.

Zipkin, P. Foundations of Inventory Mangement.

IMPLEMENTATION OF THE NEWSBOY METHOD FOR THE SALES FORECASTING OF THE APPAREL INDUSTRY

Thomassey Sébastien and Douillet Pierre
GEMTEX-ENSAIT
9 rue de l'ermitage
59100 Roubaix, France
E-mail : sebastien.thomassey@ensait.fr

KEYWORDS
Sales forecasting, Newsboy problem, decision support system

ABSTRAT

The supply chain management is a crucial success factor for companies. Due to the specific constraints in the textile-apparel industry, the sourcing and replenishment procedures require the implementation of efficient decision support system. Amongst these systems, sales forecasting models are essential for the optimization of the supply chain management. However, in order to take the right decisions and to decrease the financial risks, it is also recommended to take into account the notions of uncertainty and costs in the forecasting procedure. The proposed system is based on a forecasting system adapted to the constraints of apparel industry, which has been developed in previous work. The performed forecasts are then optimized according to the uncertainty and the purchase and selling prices of each item. This optimization is founded on the Newsboy method. Our system is tested on a real data base relative to a important ready to wear retailer.

INTRODUCTION

Due to the competitive environment and the globalization, retailers, especially in the apparel industry, require the use of efficient decision support systems to manage their supply chain. However, the specificities and constraints of the textile-apparel industry strongly complicate the implementation of such systems. In deed, retailers have to deal with:

- A large quantity of items,
- Items with short lifespan (6 to 8 weeks),
- The substitution of the most items each collection,
- Long replenishment lead time,
- A consumer's demand which is very fluctuated and strongly influenced by numerous exogenous factors.

For these companies, the main issue is the trade off between the costs, the sourcing lead time and the final inventory levels.

Actually, many supply chain management tools, such as Manufacturing Requirement Planning, Distribution Requirement Planning and Enterprise Resource Planning, are enabled the improvement of the production scheduling, the sourcing, and the retailing of apparel items. However, in order to take into account of the previously mentioned constraints, to decrease inventory level and to increase profit margins, these tools have to be implemented with efficient and reliable forecasting methods. In deed, the sales forecasting is a success factor to optimize the textile-apparel supply chain (Mattila, 2002). Forecasting methods based on data mining are particularly suitable to take into account the constraints of the apparel industry (May-Plumlee, 2001)(Thomassey, 2004a).

However, an accurate forecast is not sufficient to optimize the supply procedure, and it is necessary to use efficient sourcing methods, such as Quick Response or Vendor Management Inventory, which are able to smooth out some forecast errors (Zhao, 2001). Lastly, it is also essential to take into account the uncertainty of the forecast in the decision process. In deed, orders based on forecast, should take into account the risks due to the errors of the forecast and the costs.

The aim of this article is to propose a decision support tool for retailers of the apparel industry, which enables the computation of the optimal quantity to order during the selling cycle. Our system is composed of a forecasting model and a method of profit margin optimization. The forecasting model, developed in previous work, is based on clustering and classification techniques, and enable the treatment of new items without historical data. The optimization of order quantities, which is founded on the Newsboy method (Khouja, 1999), enables the maximization of profits. This maximization takes into

account the uncertainty of the forecasting model and the cost and the selling price of items.

FORECASTING MODEL

The implementation of a suitable forecasting system is often difficult especially in the apparel industry. In deed, the short life span and the frequent substitution of the items strongly complicate the forecasting procedure.

Many forecasting models, such as regression models, exponential smoothing, neural networks, or fuzzy inference systems, have been developed in different domains and achieve satisfactory results (Kuo, 1999). However, most of them require historical sales of several years long, and are consequently, unsuitable for our context. In order to obtain usable historical data, which enabled the seasonality and the influence of the explanatory variables treatment, it is necessary to aggregate the data by families (Thomassey, 2004b). In order to obtain forecast by items (or SKU level), a forecasting model, adapted to the constraints of the textile-apparel industry, has been developed in previous work data mining techniques are required (Thomassey, 2004a).
The principle of this model, which performs forecasts for new items one season ahead, is composed of two stages (figure 1):

1. a clustering procedure (k-means method) determines the representative sales profiles (prototypes),

2. a decision tree (C4.5 algorithm) assigns future items to the appropriate prototype of sales, elaborated in step one, according to their descriptive attribute (price, life span, starting date of the sales).

The straightforward k-means algorithm is simplistic but reasonably effective to carry out the training for the decision trees. It enables a clustering of the instances into disjoint classes from numeric attributes. Different final clusters can be found because of the random initialization of the center of classes. Consequently, to increase the chance of finding the global optimum, it is a simple matter to repeat the whole algorithm several times with different starting points.
Decision trees are very popular and powerful tools in data mining (Duch, 2004)(Tsujino, 1995) and several significant works confirm their efficiency (Tsujino, 1995)(Lee, 1996). Indeed, they are both accurate and interpretable. In many applications, the structural description of the knowledge is as important as the ability to perform well on new examples (Fiordaliso, 2000).
The fact that decision trees are efficient in term of performance and easily interpretable is a major argument for their use in our problem.
The decision tree learning is performed by the C4.5 algorithm (Quinlan, 1993). This method is usually chose because of its predictive accuracy (Last, 2004) and the numerous possibilities in term of pruning, treatment of both numerical and nominal data.

The number of clusters strongly influences the quality of the partition and the efficiency of the decision tree. This number can be fixed manually, but this strategy is relatively complex when items are numerous. Consequently, this parameter is automatically optimized with a 10-fold cross validation procedure (Kohavi, 1995).

The forecasts of the profiles of the new items enable then the optimization of the quantity to order via the Newsboy technique.

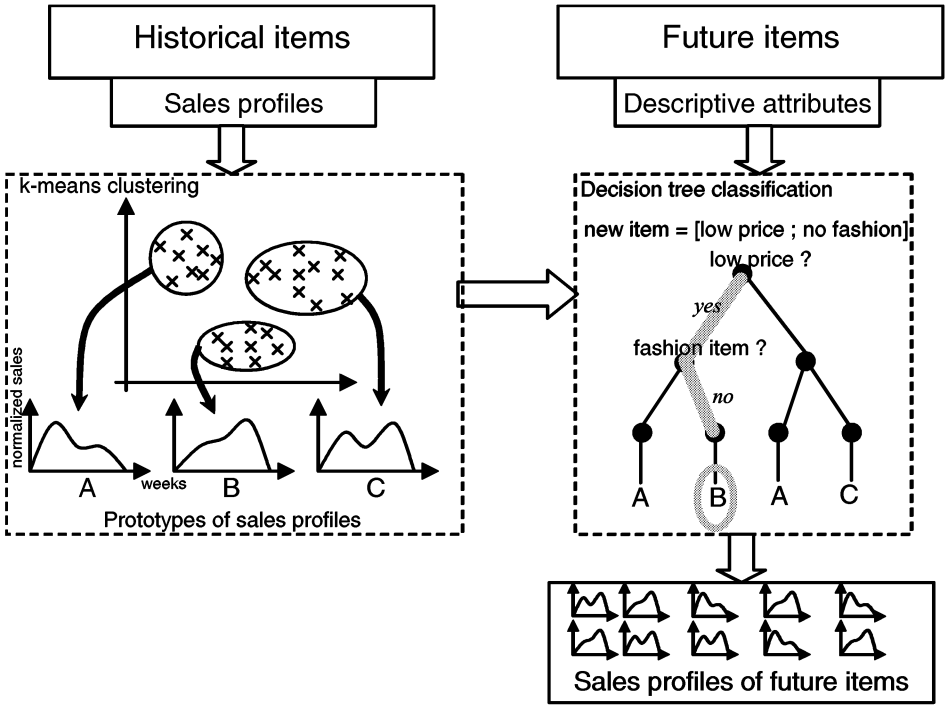


Figure 1. Principle of the forecasting model

OPTIMIZATION OF THE QUANTITY TO ORDER

Previous work (Thomassey, 2004a) has only taken into account the sales prototype of the future item which was selected by the forecasting model. This prototype is the mean profile of historical items which belong to the corresponded cluster. The proposed method takes into account further information such as the standard deviation (characterizing the uncertainty of the forecast) of sales profiles including in the cluster, the selling price and the purchase price of each future item. Thus, the prototypes, which are assigned to future items, are fitted to optimize the quantities to order, and thus to increase the profit margin and decrease the risk. The principle of this method is based on the Newsboy technique.

The well-known Newsboy problem consists in finding the quantity to order which maximizes the profit at a fixed period. For this issue, the expected cost *C(Q)* and the optimal quantity *Q**, function of the order quantity *Q*, are given by the following formula (Lau, 1997a):

$$C(Q) = c_0 \int_0^Q (Q-x) f(x) dx + c_u \int_Q^\infty (x-Q) f(x) dx$$

$$\text{and} \quad F(Q^*) = \frac{c_u}{c_u + c_o}$$

where c_0 and c_u are the item unit underage and overage cost, respectively; f and F are item demand density function and cumulative density function defined by the mean and variation coefficient, respectively.

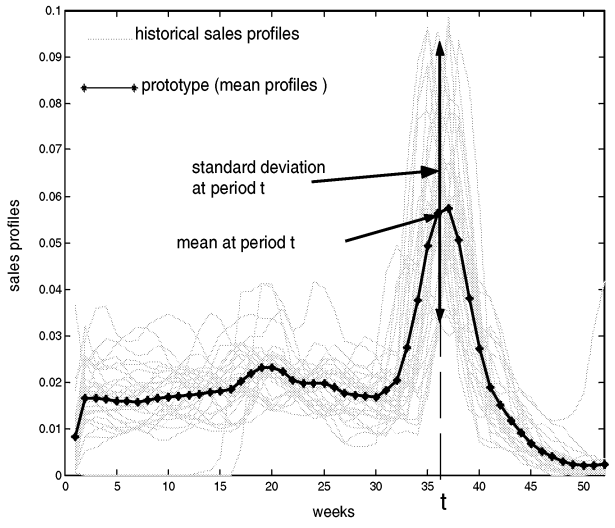


Figure 2. Example of cluster

These equations are defined for a simplified context and many extensions of are proposed in the literature: multi-items (Voros, 2001), multi-levels (Houtum, 1996), multi-contraints (Lau, 1997b),... These extensions enable a better simulation of real-world problems. However, due to the limited available data, we only implement in this article the simple Newsboy method.

The chose density functions correspond to the normal distribution with mean and standard deviation defined from the clusters of historical sales profiles: at a fixed period, the mean is the value of the prototype curve and the standard deviation is represented by the dispersion of profiles of items which belong to the cluster (figure 2). Underage (item bought but not sold) and overage (lost sales) costs are only characterized by the selling price and the purchase cost.

EXPERIMENTATION AND RESULTS

Our system is evaluated on a real database of a ready-to-wear retailer. This database is composed of three years of sales: the two first years (482 items) are used as historical data, and the third year (286 items) enables the evaluation of the results.

The results are then compared with other forecasting models. Due to the specific context of the study: items frequently substituted, short selling cycle,...), it is very difficult to implement classical time series forecasting models.

The optimum number of clusters is 14. Thus, the sales of the 482 historical items are summarized by 14 prototypes (mean profile). Each of the 286 new items is then assigned to one of the 14 prototypes optimized according to the cost data of the considered item.

Figures 3 and 4 show the optimized sales profiles of two items which are assign to the same cluster. The well-known results of the Newsboy problem obviously appear in the comparison of these 2 figures. In figure 3, the ration purchase price / selling price is low (0.25). Since, this item is cheap for the retailer, it is not problematic if too many items are bought then the demand is uncertain. Consequently, the mean profile is increased when the standard deviation is high and is decreased when the standard deviation is low. In figure 4, the ration purchase price / selling price is high (0.75), thus it is suitable to decrease the quantity to order when the demand is uncertain. The rules of optimization are inversed compared with the previous item.

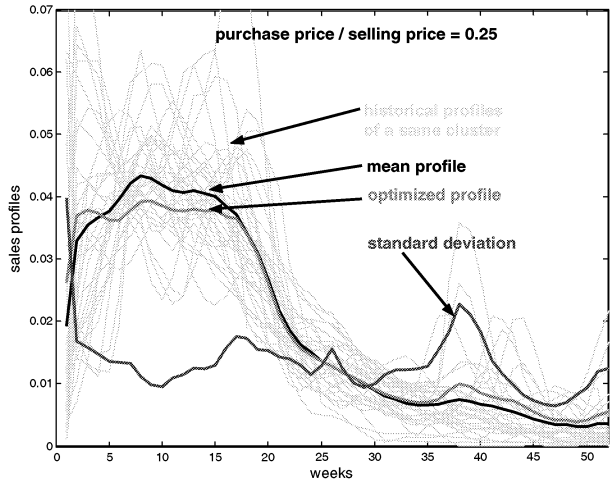


Figure 3. Example of forecast with a ratio "purchase/selling price" = 0.25

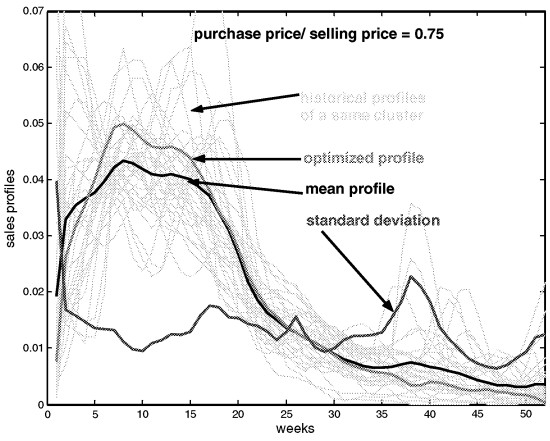


Figure 4. Example of forecast with a ratio "purchase/selling price" = 0.75

CONCLUSION

In this article, we have proposed an original decision support system suitable for the apparel industry. This tool is composed of:

- A forecasting model which is useful to estimate sales profiles of new items for which have no historical sales data,
- An optimization of the quantity to order which takes into account of the uncertainty of the forecasts and the costs of the items.

The forecasting model is based on k-means clustering and decision tree classification. The optimization of the quantity to order relies on the Newsboy method.

Our system is tested on a real data base composed of 486 historical items and 285 future items. Due to the lack of data, especially in term of underage and overage costs and sourcing strategy (frequency and number of replenishment), we only use here a simplified method of the Newsboy. Therefore, we only consider here the normal distribution; it could be also interesting to take into account of other distribution such as the lognormal distribution. However, this system enables us to understand the impacts of the uncertainty and the costs on the profit and the optimum quantity to order.

The implementation of a more complex system with more realistic constraints and data will be the subject of future works.

REFERENCES

Duch W., Setiono R., Zurada, J.M., (2004) "Computational Intelligence Methods for Rule-Based Data Understanding", *Proceddings of the IEEE,* vol. 92, n°5, p. 771-805.

Fiordaliso A., (2000) "Autostructuration of fuzzy systems by rules sensitivity analysis", *International Journal for Fuzzy Sets and Systems*, vol. 118, n° 2, p. 281-296.

Houtum G.J., Inderfurth K., Zijm W.H.M., (1996) "Materials coordination in stochastic multi-echelon systems", *European Journal of Operational Research*, vol. 95, p. 1-23.

Khouja M., (1999) "The single-period (news-vendor) problem: literature review and suggestions for future research", *Omega*, vol. 27, p. 537-553.

Kohavi R., (1995) "A study of cross-validation and bootstrap accuracy estimation and model selection", *Proceeding of 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada, Morgan Kaufmann, San Francisco, p. 1137-1143, 1995.

Kuo R.J., Xue K.C., (1999) "Fuzzy neural networks with application to sales forecasting", *Fuzzy Sets and Systems*, vol. 108, p. 123-143.

Last M., Maimon O., (2004) "A compact and accurate model for classification", IEEE Transactions on Knowledge and Data Engineering, vol. 16, n° 2, p. 203-215.

Lau H.S., Lau A.H.L., (1997b) "Some results on implementing a multi-item multi-constraint single-period inventory model", *International Journal Production Economics*, vol. 48, p. 121-128.

Lau H.S., (1997a) "Simple formulas for the expected costs in the newsboy problem: An educational note", *European Journal of Operational Research*, vol. 100, p. 557-561.

Lee K.C., Oh S.B., (1996) "An intelligent approach to time series identification by a neural network-driven decision tree classifier", *Decision Support Systems*, vol. 17, p. 183-197.

Mattila H, King R., Ojala N., (2002) "Retail performance measures for seasonal fashion", *Journal of Fashion Marketing and Management*, vol. 6, n° 4, p. 340-351.

May-Plumlee T., Little T.J., (2001) "Consumer purchase data as a strategic product development tool", *Journal of Textile, Apparel, and Technology, Management*, vol. 1, n° 3, p. 1-10.

Quinlan J.R., (1993) *C4.5: Programs for Machine Learning*, San Francisco : Morgan Kauffman.

Tsujino K., Nishida S., (1995) "Implementation and refinement of decision trees using neural networks for hybrid knowledge acquisition", *Artificial Intelligence in Engineering*, vol. 9, p. 265-275.

Thomassey S., Happiette M., Castelain J.M., (2002) "Textile items classification for sales forecasting", *4th European Simulation Symposium and Exhibition Simulation in Industry*, Dresde, Germany, October p. 23-26.

Thomassey S., Fiordaliso A., (2004a) "A hybrid forecasting method of new item sales based on clustering and decision trees", *Decision Support System,* article in press.

Thomassey S., Happiette M., Castelain J.M., (2004b) "A short and mid-term automatic forecasting system - Application to textile logistics", *European Journal of Operational Research*, vol. 161, n° 1, p. 275-284.

Voros J., Szidarovszky F., (2001) "Determining the production order of two seasonal products under capacity constraint", *International Journal of Production Economics*, vol. 73, p. 117-122.

Zhao X., Xie J., Lau R.S.M, (2001) "Improving the supply chain performance : use of forecasting models versus early order commitments", *International Journal of Production Research*, vol. 39, n° 17, p. 3923-3939.

BIOGRAPHY

SEBASTIEN THOMASSEY received the M.Sc. degree from the Ecole Nationale Supérieure des Arts et Industries Textiles (ENSAIT) in Roubaix in 1999 and the Ph.D. in automation from the University of Lille 1 (France) in 2002. He is assistant Professor in logistic and automation at the ENSAIT since 2003. His research interests include classification and sales forecasting of textile items.

PIERRE DOUILLET studied mathematics and computer science at ENS Cachan. He received his M.S. and Ph.D. degrees in computer science from the Science University of Paris (Jussieu, Paris VI). He was formely "professeur agrégé" then "professeur de chaire supérieure" in mathematics at College level. He is currently an Associate Professor at the ENSAIT (Textile Engineers Institute, Roubaix). His research interests include formal computing, approximations, network performances and modeling. He authored two books on formal computing and several research papers on modeling.

.

LATE PAPERS

The Design, Fabrication and Image Processing of a 3 DOF Passive Biped Walking Robot

Ahmad Bagheri
Amir Hajiloo
Salar Basiri
Rahman Yousefi Moghaddam

University of Guilan, Department of Mechanical Engineering
Po.Box: 3756 – Rasht, Iran
Email: bagheri@guilan.ac.ir
ahajiloo2000@yahoo.com
salarbasiri@noavar.com
usefi@guilan.ac.ir

Abstract:

Designing and manufacturing mechanisms that are able to move on a slope according to the gravity has been an issue of the interests of researchers recently. In this paper the designing and manufacturing a Passive biped robot with 3 degrees of freedom is presented. This robot is able to move down on slope with a minimum tangent based on the gravity and without any controller or any supplier. After presenting the governing equations and the dynamic of the robot, a computer simulation is done. Finally a comparison between the results from the physical model and the results from the numerical simulation is done to check the validity of the project.

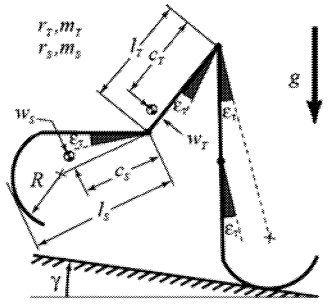
INTRODUCTION

Most terrestrial multi-cellular animals, including humans, can walk. Some people also hope for useful walking robots. Increasing the knowledge of walking has a wide range of applications. Since human motion is controlled by the nervous system and powered by muscles, the role of nerves and muscles is obviously of interest. But one way to understand the role of nerves and muscles is to learn how much can be done without them. Human walking, for example, might be modeled for some purposes as an uncontrolled mechanical process. One idea behind the case is that the role of the nerves and muscles in walking might be more to gently guide than imposingly control. It has been shown by McGeer that stable human-like walking motions can be exhibited by uncontrolled sets of rigid links powered only by gravity [1]. McGeer's work in passive dynamic walking has shown that coordination in human walking does, to some extent, rely on the passive dynamics of the legs. If one's goal is to study active and humanlike robotic walking, it is natural that one would also take advantage of the passive dynamics of the robot's limbs to reduce the energetic and computational cost of coordinating the walker's motion. A simple and open loop powered walking scheme was studied by McGeer (1989, 1990, and 1991).

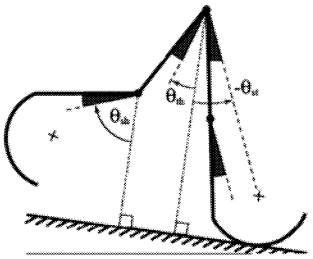
The approach here was originally pioneered by McGeer (1989-1993), who demonstrated that a two-dimensional, four-link mechanism, which somewhat resembles human legs, is capable of stable, human-like gait down a shallow slope with no activation (besides gravity) and no control. McGeer's passive-dynamic theory of bipedal locomotion describes gait as a natural repetitive motion of a dynamical system or, in the language of nonlinear dynamics, a limit cycle.

Passive Walking Mechanism

The two-dimensional kneed walking machine we study here, essentially is a copy of McGeer's design, is shown schematically in Figure (1). It consists of a swing leg (not in contact with the ground) and a stance leg (touching the ground), connected by a frictionless joint at the hip. Extra mass is generally added at the hip serving as a crude model of an upper body. Each leg (assumed identical to the other) is composed of a rigid thigh and shank. The stance knee is locked. For kneed walkers, the swing knee is a frictionless joint with a knee-stop preventing hyperextension between knee strike and heel strike. The knee stop also prevents the stance leg from hyper extending. Straight-legged (knee less) walkers may be viewed as obtained from kneed walkers by permanently locking the knees [6,7].



(a) Model Parameters



(b) Dynamic Variables.

Figure 1 - Description of McGeer's Kneed Walking Model.

The motions of the walker are governed by standard rigid body dynamics. In our way of formulating these, all of the equations described below are based on angular momentum balance about various points. A walking step starts right after heel strike, just when the old stance leg (which is now the swing leg) starts swinging. Equations of motion for the three-link mode are integrated forward in time until knee strike is detected. Assuming no rebound (perfectly plastic knee collision), jump conditions then determine the post-knee strike state. Equations of motion for the two-link mode are now integrated forward in time until heel strike is detected. Assuming no rebound (perfectly plastic heel collision) and no ground impulse at the trailing foot, jump conditions are used to determine the post-heel strike state from the pre-heel strike state.
The state space of this model describes by $\{\theta_{st}, \theta_{th}, \theta_{sh}, \dot{\theta}_{st}, \dot{\theta}_{th}, \dot{\theta}_{sh}\}$, where θ_{st}, θ_{th} and θ_{sh} are stance angle, thigh angle and shank angle, respectively. The simulated walking cycle is shown schematically in Figure (2).



Figure 2 - Simulation of Gait Cycle.

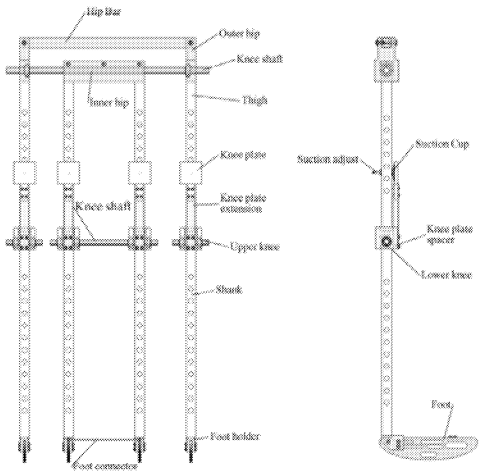
Specifications of the Fabricated Biped

The physical characteristics of the fabricated biped are shown in Figure (3). This robot is designed and manufactured based on the McGeer Biped which was first designed and manufactured in Cornell University in the USA [8].



Figure 3 - Passive Biped Walker of the University of Guilan

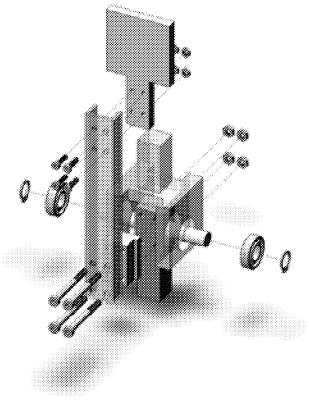
This sort of walkers has 4 legs – 2 couples. The internal couple moves together while the external one does the same so it will prevent the robot from falling asides. Thus it can be assumed as a 2-D Biped walker model with a good approximation**.** Different components of the robot are shown in Figure (4) separately and in detail.



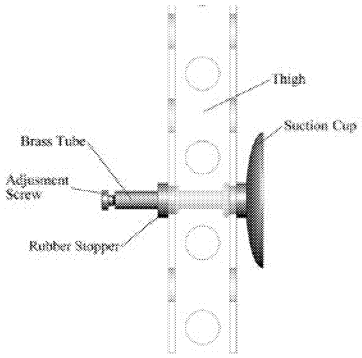
(a) Front View (b) Left View

Figure 4 - Biped Walker Model in Details

The robot's foot is circular and so it will make an easy rolling contact with the slope. As mentioned before the knee mechanism is one of the important parts of the robot. In this prototype a plane is attached to the knee that is connected to the knee joint in one end. The knee plane is in contact with a suction cup which locks the knee in supporting phase and also prevents the knee from over bending. The knee mechanism is shown in Figure (5).



(a) Knee joint



(b) Knee Locking system

Figure 5 - Knee Mechanism in Detail

The robot parts are made of aluminum and some machinery are done on them.

Dynamics Algorithm

The Newton-Euler equation has been used to obtain the equations of motion of this model and will yield the following equations, after some rearrangements:

$$
\begin{aligned}
{}^{i+1}\omega_{i+1} &= {}^{i+1}_{i}R\,{}^{i}\omega_i + \dot{\theta}_{i+1}\hat{z}_{i+1} \\
{}^{i+1}\dot{\omega}_{i+1} &= {}^{i+1}_{i}R\,{}^{i}\dot{\omega}_i + {}^{i+1}_{i}R\,{}^{i}\omega_i \times \dot{\theta}_{i+1}\hat{z}_{i+1} + \ddot{\theta}_{i+1}\hat{z}_{i+1} \\
{}^{i+1}v_{i+1} &= {}^{i+1}_{i}R\left({}^{i}\omega_i \times {}^{i}p_{i+1} + {}^{i}v_i\right) \\
{}^{i+1}v^{c}_{i+1} &= {}^{i+1}\omega_{i+1} \times {}^{i+1}p^{c}_{i+1} + {}^{i+1}v_{i+1} \\
{}^{i+1}\dot{v}_{i+1} &= {}^{i+1}_{i}R\left({}^{i}\dot{\omega}_i \times {}^{i}p_{i+1} + {}^{i}\omega_i \times ({}^{i}\omega_i \times {}^{i}p_{i+1}) + {}^{i}\dot{v}_i\right) \\
{}^{i+1}\dot{v}^{c}_{i+1} &= {}^{i+1}\dot{\omega}_{i+1} \times {}^{i+1}p^{c}_{i+1} + {}^{i+1}\omega_{i+1} \times ({}^{i+1}\omega_{i+1} \times {}^{i+1}p^{c}_{i+1}) + {}^{i+1}\dot{v}_{i+1}
\end{aligned}
\qquad (1)
$$

Above, ${}^{i}\omega_i$ is the angular velocity of links and ${}^{i}\dot{\omega}_i$ is the angular acceleration of links. Also $(V, \dot{V})$ are the velocity and acceleration of each frame and $(V^c, \dot{V}^c)$ are the velocity and acceleration of center of mass of each link.

"Inertial forces" **F** and "Inertial torques" **N** are then defined for each link, (in terms of the θ_i, $\dot{\theta}_i$, and $\ddot{\theta}_i$);

$$
\begin{aligned}
{}^{i+1}F_{i+1} &= m_{i+1}\,{}^{i+1}\dot{v}^{c}_{i+1} \\
{}^{i+1}N_{i+1} &= {}^{i+1}I^{c}_{i+1}\,{}^{i+1}\dot{\omega}_{i+1} + {}^{i+1}\omega_{i+1} \times {}^{i+1}I^{c}_{i+1}\,{}^{i+1}\omega_{i+1}
\end{aligned}
\qquad (2)
$$

The effect of gravity is included most simply by accelerating the base of the mechanism ${}^{0}\dot{v}_0$ upward by g, and so one does not need to compute the forces and torques due to gravity in the equations above [1].
The acceleration of frame "0" is non-zero in general and is denoted by ${}^{0}\dot{v}_0$ and for ${}^{1}\dot{v}_1$ is modified in the above algorithm as follows.

$$
{}^{0}\dot{\mathbf{v}}_0 = \left[\mathbf{g}\cos\gamma, \quad R\ddot{\theta}_1 - \mathbf{g}\sin\gamma, \quad 0\right] \qquad (3)
$$

Here *R* is the radius of foot, $g\cos\gamma$ and $-g\sin\gamma$ terms result from accelerating the base upwards to account for the effect of gravity, as described above [4].

The Collision Dynamics

As mentioned before, the cycle of step starts after the heel strike and it lasts till the supporting leg converts to the swing leg. The motion equation in the 3 links state includes the walker's motion to the knee strike. Assuming the collision is ideal, which means there's no elastic motion in the collision, completing the motion, the state space should be gained according to the pre-collision parameters, exactly after the knee strike. Also the equations of motion for 2-link state include the walker's motion to the heel strike.
It is assumed that there has happened no elastic motion and also there is no collision between the previous supporting leg and the slope. Thus it is possible to find the state space exactly after the heel strike and the knee strike according to the previous state space parameters, so the cycle will be completed [5].
The state space before the knee strike can be shown as $\left\{\theta^{-}_{st}, \theta^{-}_{th}, \theta^{-}_{sh}, \dot{\theta}^{-}_{st}, \dot{\theta}^{-}_{th}, \dot{\theta}^{-}_{sh}\right\}$. The sign " $^{-}$ " means they are related to the pre-collision moment.
Exactly after the knee strike, the state space including the supporting leg's angle θ^{+}_{st}, the swing leg's angle θ^{+}_{sw} (or θ^{+}_{th}) and their relative velocities.
Although the post-collision state space is unknown, the relation between the angles for the shown model in fig 1 is as below:

$$\theta_{st}^{+} = \theta_{st}^{-}$$
$$\theta_{sw}^{+} = \theta_{th}^{-} = \theta_{th}^{+} \qquad (4)$$

As the angular momentum for the whole walker relative to the contact point of the supporting leg $\left(\mathbf{H}_{tot/cp}^{-} = \mathbf{H}_{tot/cp}^{+}\right)$, and for the swing leg relative to the thigh joint $\left(\mathbf{H}_{sw/h}^{-} = \mathbf{H}_{sw/h}^{+}\right)$ are constant in the collision, $\dot{\theta}_{st}^{+}$ and $\dot{\theta}_{sw}^{+}$ can be easily found, according to obtain the relation between the angular velocities.

Exactly before the heel strike, the state space includes the supporting leg angle θ_{st}^{-}, swing leg angle θ_{sw}^{-} and their relative velocities.

Exactly after the heel strike, the state space includes the supporting leg angle θ_{st}^{+}, swing leg angle θ_{th}^{+} and their relative velocities. The relation between the angles is as below:

$$\theta_{st}^{+} = -\theta_{st}^{-}$$
$$\theta_{th}^{+} = -\theta_{sw}^{-} \qquad (5)$$
$$\theta_{sh}^{+} = -\theta_{sw}^{-} = \theta_{th}^{+}$$

Coming facts are used to find the angular velocities as before:

The angular momentum at the collision for the entire walker relative to the new contact point $\left(H_{tot/cp}^{-} = H_{tot/cp}^{+}\right)$, for the new swing leg relative to the thigh joint $\left(H_{sw/h}^{-} = H_{sw/h}^{+}\right)$ and for the new swing leg relative to the new swing knee $\left(H_{sh/k}^{-} = H_{sh/k}^{+}\right)$ will be constant.

In the 2-D state at the heel strike, 3 equations can be written for the angular momentum where solving them will lead to find the unknown parameters of $\dot{\theta}_{st}^{+}$, $\dot{\theta}_{th}^{+}$, $\dot{\theta}_{sh}^{+}$ in respect to the pre-collision parameters. For this purpose applying the analytic dynamic equations and according to the equations 1 to 5 and the on-line analyzing of it, all the necessary parameters can be found including $\dot{\theta}_{st}^{+}$, $\dot{\theta}_{th}^{+}$, $\dot{\theta}_{sh}^{+}$. The solutions are available in reference [9] in details.

Image Processing

Image processing is applied to obtain $\theta_{st}, \theta_{sh}, \theta_{sw}$. The most important advantage of such process is that the measuring tool has no effect on the measured object. Fast set up and to be independent from any other hardware are some other advantages of the process. Image processing is done with a PC and DSP is not used. The image is taken by a CCD sensor with the resolution of (640x480). The modular design is presented in Figure (6).

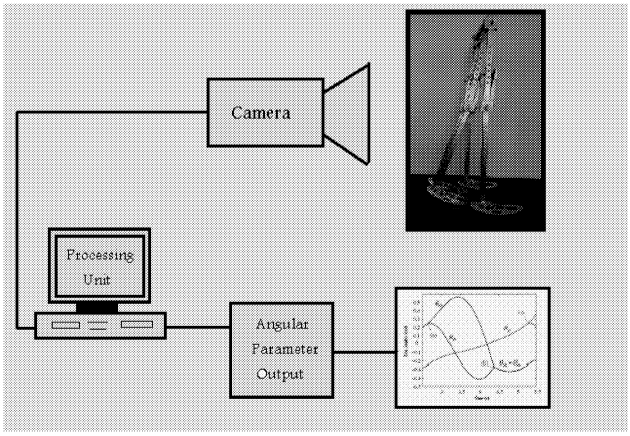


Figure 6 - The modular

The purpose of image processing is to find the position of colored nodes as it is shown in Figure (7). Several frames are taken while the robot is moving and they will be processed later, because the time needed for saving the frames is less than time needed to process them. By this method θ can be found precisely.

Process Algorithm

Each image is consists of 3 position areas and 2 different color areas as it is shown in Figure (7).

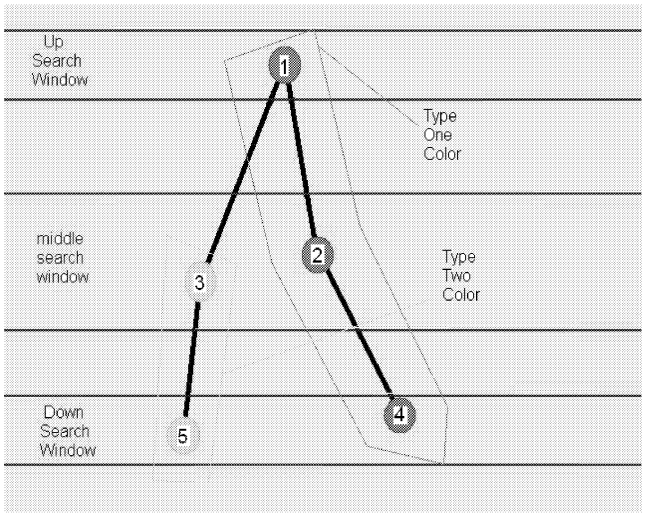


Figure 7 - Image Processing in Detail

The amounts of the position areas are given to the image process program. According to process two different colored areas it is not possible to use hardware filters, so the RGB (Red, Green, Blue) software filters are used.

Each of the amounts of R, G and B for a singular pixel can be in the range of 0 and 255.

In the beginning the program appropriate amounts of the filter parameters and their application must be given to the program according to a sample frame. For example to identify the yellow color the amounts of the R and G must be more than 200 and amount of B must be less than 150. Thus the program has to find the desired color in different

areas. According to the separating method of the position areas and colored areas, it is assured that the obtained point is unique.
As the relation of the distance between the planes - where the nodes are inside them – and the mean distance of them to camera is very close to **1**, all the nodes can be assumed in the same plane. It is due to the effect of immeasurable factors on the transfer function of the pixel positions to the real positions.
Finally, by having the nodes positions and the amount of frames per second it is possible to compute the θ and $\dot{\theta}$.
In cases where the mid-legs is covered by the leg that is faced to the camera - which prevents it to find the nodes- Linear Regression method is used to obtain closest datum to the unknown data. To install the program, Delphi 7.0 is used. The GUI program is shown in Figure (8).

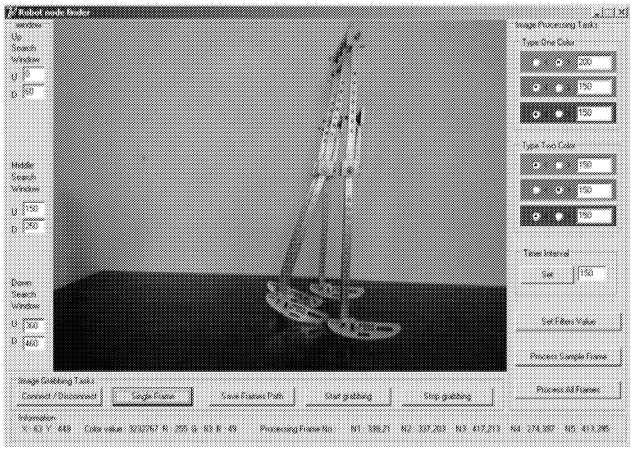


Figure 8 - The GUI Program under Delphi 7.0

Results

To obtain an appropriate cycle of the movement the physical parameters of the model have a certain role. These parameters are shown in Figure (1). Some of this parameter such as the length of the legs and the radius of the sole are constant, but other parameters are changeable. For this purpose, the position of the center of mass and the gyration radius can be changed by increasing the mass.
Applying this method on the manufactured prototype, the appropriate amounts of these parameters found after several tests.
They are presented in Table (1). The physical parameters of the walker are presented in Table (2).

Table 1 - The changed parameters in the test

Model	$m_t,(kg)$	$r_t,(m)$	$c_t,(m)$
A	0.9643	0.14	0.11
B	1.5643	0.102	0.112
C	1.9543	0.0995	0.105
Final	2.4443	0.0995	0.091

Table 2 - The physical Parameters of the model

l_t	0.35m	m_s	1.015m
w_t	0m	r_s	0.197m
m_t	2.4443kg	c_s	0.17m
r_t	0.0995	R	0.2m
c_t	0.091m	γ	0.036rad
l_s	0.45m	g	m/s^2 9.81
w_s	0.03kg	ε_T	0.097rad

The period of each step cycle is obtained 0.8 seconds. The Leg angles are taken from image processor for 5 levels of the movement and are compared with the numerical results which are taken from the numerical simulation. They are presented in the following table.

Table 3 - Numerical and Experimental Leg angles

	Numerical (rad)			Experimental (rad)		
Time (s)	θ_{st}	θ_{th}	θ_{sh}	θ_{st}	θ_{th}	θ_{sh}
0.8	-0.2439	0.2439	0.2439	-0.2439	0.2439	0.2439
1	-0.1654	0.1423	0.4253	-0.1726	0.1521	0.4355
1.2	-0.0895	-0.3325	0.4925	-0.0912	-0.3565	0.4525
1.4	0.0485	-0.2935	-0.2935	0.0435	-0.2785	-0.2785
1.5	0.2439	-0.2439	-0.2439	0.2439	-0.2439	-0.2439

The curves of the angles of $\theta_{st},\theta_{sh},\theta_{sw}$ are plotted in Diagram (1) where they are assumed as functions of time. The confliction points are highlighted in it. (a) is the heel strike point, (b) is Knee strike point and (c) is Heel strike again.

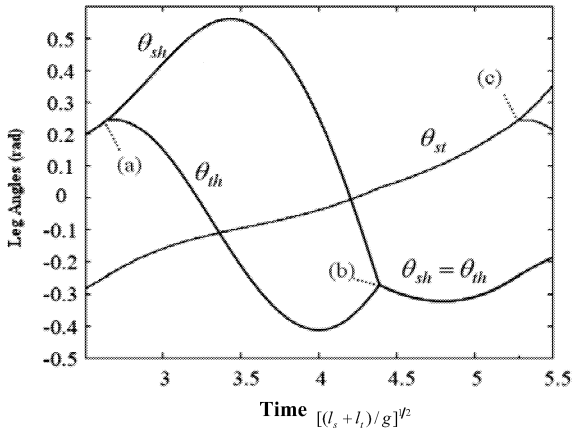


Diagram 1 - The Curves of Foot Angles (Rad) Respect to the Time (Sec)

According to the Diagram and the presented states, the area between (a) and (b) stands for the 3-link mode and (b) to (c) stands for 2-link mode.

Also, as it is expected after the knee strike, the angles of the thigh and the leg will be the same ($\theta_{th} = \theta_{sh}$) and will consist the swing leg's angle (θ_{sw}).
It should be mentioned that the parameters of table (2) are used in simulation and also the initial state space conditions for the best step cycle are as below based on the try and error method:

$$\begin{aligned}
&\theta_{st}(0) = -0.2439 rad \\
&\theta_{th}(0) = 0.2439 rad \\
&\theta_{sh}(0) = 0.2439 rad \\
&\dot{\theta}_{st}(0) = -1.049 rad/s \\
&\dot{\theta}_{th}(0) = 1.049 rad/s \\
&\dot{\theta}_{sh}(0) = 2.23 rad/s
\end{aligned} \qquad (6)$$

Conclusion

In this paper the specifications of the passive biped walkers are presented firstly, and the governing equations of them are explained. Then, by changing the physical parameters, the physical model is analyzed and the appropriate step cycle is obtained. Also the legs' angels for a step cycle simulated as a function of time.
For programming Delphi 7.0 is applied, where SolidWorks and Matlab were used for the designing and numerical simulation, respectively.

References

1. T, McGeer, “Passive walking with knees”, Proceeding of the IEEE Conference on Robotics and Automation, 2:1645-16451, 1997.
2. M, Garcia, A, Ruina, A, Chaterjee, “Efficiency and stability of 2-D passive dynamic kneed walking”, ICRA98, 1998.
3. M, Garcia, A, Ruina, A, Chaterjee, M, Coleman, “The simplest walking model: Stability, complexity, and scaling”, Accepted for publication in the ASME Journal of Biomechanical Engineering, 1997.
4. S, Collins, M, Wisse, A, Ruina, “A Three-Dimensional Passive-Dynamic Walking Robot with Two Legs and Knees”, The International Journal of Robotics Research, Vol. 20, No. 7, pp: 607-615, July 2001.
5. A, Ruina, “Nonlinear-Dynamic study of walking: simulation, analysis, and experiment. A research proposal submitted to the National Science Foundation, 1997.
6. M, Garcia, A, Ruina, A, Chaterjee, “Efficiency, Speed, and Scaling of 2D passive Dynamic Walking”, Submitted to Dynamic and Stability of Systems July 29, 1998.
7. M, Garcia, A, Ruina, A, Chaterjee, “Efficiency, Speed, and Scaling of Small-Slope 2-D passive Dynamic Bipedal Walking”, last revised version of Submitted to Dynamic and Stability of Systems, September, 1999.
8. M, Garcia, “Stability, Scaling and Chaos in Passive-Dynamic Gait Models”, A Dissertation Presented to the Faculty of the Graduated School of Cornell University, January 1999.
9. A. Hajiloo, " Design and Manufacturing of a Passive Biped Walker" , B.Sc. Dissertation, University of Guilan, 2004

Bibliography

Ahmad Bagheri was born in Tehran, Iran, 1966. He received his B.Sc. degree of Mechanical engineering from the University of Mashhad in 1989 and his M.Sc. degree from the University of Tarbiat Modares in 1993 and his Ph. D. degree on Mechatronics at CTU Prague in 1997. He worked four years for Iranian Ministry of Industries. Now he is serving as assistant professor and the head of the mechanical engineering department, University of Guilan. His research interests include biped locomotion, robot modeling and control, Industrial robotics and Mechatronics.

Amir Hajiloo was born in Tehran, Iran, 1981. He received his B.Sc. degree of mechanical engineering from the University of Guilan in 2004. Now he is the student of Master degree at the University of Guilan.

Salar Basiri was born in Tehran, Iran, 1984. He is the student of mechanical engineering at the University of Guilan.

Rahman Yousefi Moghaddam was born in Rasht, Iran, 1980. He received his B.Sc. degree of mechanical engineering from the University of Guilan in 2002. Now he is the student of master degree at the University of Guilan. His research interests include Robotics and Control of the robots, Mechatronics, Nanotechnology, and Manufacturing Systems.

KNOWLEDGE MODELING AND PROCESSING FOR SUPERVISION OF PROCESS FACILITIES

Jörg Krüger
Alexei Lisounkin
Alexander Sabov
Gerhard Schreck
Fraunhofer-Institute for Production Systems and Design Technology
Pascalstrasse 8-9, D-10587, Berlin, Germany
E-mail: { joerg.krueger | alexei.lisounkin | alexander.sabov | gerhard.schreck } @ipk.fraunhofer.de

Michael Pocher
Society for the Promotion of Applied Computer Science (GFaI)
Rudower Chaussee 30
D-12489, Berlin, Germany
E-mail: pocher @gfai.de

KEYWORDS
Cognitive Tasks for Supervision, Process Knowledge, Knowledge Modeling, Temporal Planning

ABSTRACT

Supervision of process facilities is a complex task which objective is to provide an efficient facility operation exploiting tolerances and variations of the process technology as well as workspaces and limitations of the technical components. In its dynamic phases, any technological facility may behave rather like an active and sometimes unpredictable system. That is why advanced process supervision and control systems increasingly include model-based and rule-oriented components to support facility operators in processing of tasks.

This paper discusses the modeling of supervision tasks for technical facilities with a network structure. The facility supervision knowledge is represented in the form of "if-then" objects and operation patterns. For such distribution networks as gas, water, oil, and energy supply, a general data model and simulation-based architecture for a rule-based decision support system was developed.

INTRODUCTION

The complexity of process facilities, control systems and control principles is continually increasing, resulting in new requirements for corresponding operation procedures. Nowadays, control and supervision systems efficiently cover local automation tasks, process data management and archiving, as well as visualization and reporting. Human tasks become concentrated on the supervision of actual system states as well as the planning of the system operation and maintenance for the short and long future.

In its subject, the performing of supervision tasks is mainly focused on analysis of an actual facility state with respect to a defined set of process quality criteria – technical and technological bounds and thresholds, time intervals, and system evolution patterns – and on forecasting the facility behavior for a defined future. The analysis and modeling of role-dependent staff activity, responsibilities, and qualifications and in context of process facilities and power stations were studied by Rasmussen [Rasmussen, J., 1986]. His findings continue to have an influence on the development of decision support systems for technical facilities.

Considering that traditional control and supervision technologies – one-sensor-one-indicator – are being replaced by complex model-based control procedures, the system supervision and decision making has become more complex for the operator, who must reflect on the behavior of the process and the control components as an entire system.

Extensive knowledge about process technology and facility technique has resulted in sophisticated simulation models which can calculate the facility behavior in given environment, under certain conditions. Complementary to this, the decision making process involves analysis of actual application contexts, weighting of different – partially concurrent – targets and objectives, and, finally, elaboration of scaling mechanisms for a quantitative comparison of possible behavior alternatives. Although decision making procedures have been the focus of many researchers, there are deficits of general modeling concepts and treatment procedures due to the high subjectivity of human decision making. Nowadays knowledge related to the decision making process is accumulated and applied mainly in the form of heuristics. These heuristics utilize known situations and reactions to them. This principle is used by many rule-based systems – systems with "if-then" logic.

An additional aspect should be mentioned in consideration of the supervision and operation of technical facilities – their dynamic with respect to technology behavior and technical support routines. This dynamic behavior must be coordinated with decision support procedures.

This paper is devoted to the modeling and implementation of scenario-based supervision techniques for facilities with network structure. The elaboration of supervisory decisions is connected to a simulator for the context dependent facility behavior prediction. The modeling procedure for deci-

sion/action scenarios should to be flexible and easily maintained by facility operation staff during the operation.

COGNITIVE TASK ANALYSIS AND KNOWLEDGE MODELING

Within the facility operation context, the decision elaboration procedure consists of the following steps. First, the decision maker has to detect the need for intervention by means of observing actual process data. For this, the operator has features – data evaluation functions and criteria – which give him an estimation of the optimality and regularity of the system operation or about a malfunction in the system. In regular cases, this information can give the direction for subsequent activities. Based on this evaluation, a target state into which the system should be transferred will be chosen, and the task that the decision maker has to perform will be selected from a review of available resources. When such a task is identified, the proper procedure, i.e., how to do it, can be planned and executed.

If the connection between the data evaluation criteria and the sequence of action to reach the given goals is clear, the supervision procedure is denoted *skill-based* decision. In cases where decision making includes the evaluation of alternatives, the supervision procedure is denoted *rule-based* behavior. The decision making procedure, which involves comparison of models and analysis of goals and depends on the know-how of the operator, is labeled *knowledge-based.* These three levels of decision making procedure are depicted in Figure 1.

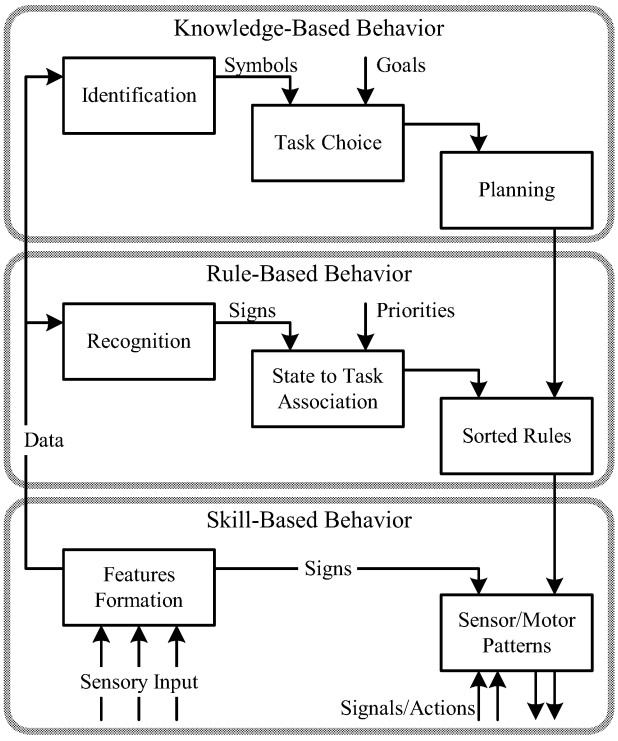


Figure 1. Decision making behavior model.

The main objective of so-called knowledge-based systems is to capture as many regular situations as possible which may appear during facility operation in form of skill-based procedures. This is done in order to ensure a short-time operation and to support a mid-range planning by means of rule-based predicting scenarios.

KNOWLEDGE MODELING FOR FACILITIES WITH NETWORK STRUCTURE

Specific technical realization, technological and administrative requirements, operative performance criteria and long-term planning result in unique supervisory and control schemes for each facility to be controlled. The design, development, and maintenance of such supervisory and control schemes employ a significant amount of human resources for each application. Reasonable methods for modeling a corresponding knowledge base and techniques for the application and maintenance will support high acceptance of such systems and an efficient use.

For technical networks, the design of the knowledge model can closely be connected with the network structure of the facility. It may provide an approach for systematic management and maintenance of knowledge objects.

Network Modeling

Considering network execution for such application fields as water, gas, oil, and energy distribution, the following common structural properties of such facilities can be identified:

- technological steps are encapsulated/concentrated in structurally non-decomposable elements;
- material/energy flows are maintained by pipes/connections;
- material/energy flows are considered to be homogenous when transported between processing elements.

Under these assumptions, network modeling can be provided using the flowchart principle, whereas technological steps are represented by functional blocs and material, energy and signal flow – by links between blocks.

For scalable systems, homogeneous process elements can be put into functional groups with common operational patterns. Additionally, heterogeneous elements may be grouped in order to perform common or collaborated supervision and control tasks. Such groups in the system model will be denoted *hierarchical levels*.

In terms of graphs, technological and structural network elements can be represented by nodes, and material/energy flows in the network can be depicted by arcs. In the following, "elementary", non-decomposable technological operations are called *atomic processes*, and the set of relations among atomic processes is called *facility topology*.

Starting from such commonly used representations of process technology and control structure as flowcharts, com-

plex supervision and control applications can be developed based on the toolbox principle.

Network-Oriented Knowledge Modeling

As discussed above, information about the system behavior will be elaborated by means of processing of system signals and measurements. With each atomic process, a set of variables which uniquely describe actual performance of the process will be defined. Such variables will be denoted *state variables*. The data processing features (see Features Formation in Figure 1) must be developed in order to calculated values of state variables involving sensory input.

For hierarchical groups, where a set of atomic processes has indicators for their joint performance or coordinated operation, these indicators can be taken as state variables as well. Thus, for atomic processes with the same function which is connected in parallel, variables of summarized performance will be used for coordinated control and will be denoted state variables for the group. The analysis of values for state variables may provide information about the operation efficiency for atomic processes, for parts of the system which include several atomic processes, or for the entire system.

Historical data about the evolution of system inputs and state variables describe the operational context for simulation and decision making procedures. Acquisition and classification of these data is a complex process which can be handled by Data Mining methods (see [Lisounkin, 2004]).

When the state variables for the atomic processes have been specified, *features* and *criteria* for the analysis of their values will be constructed. Typically, these features are the upper and lower bounds, deviations from a given value, the accumulation of values over a given time, threshold of time limits, *etc*. Additionally, more complex functions can be defined. The features provide flags with the value "true" if the corresponding criterion achieved and "false" otherwise. The features trigger *process correction actions*. The sequences of process correction actions in connection with corresponding features are denoted *behavioral patterns*.

Summarizing, one can identify that the following principal objects are considered for the construction of a rule-based decision making procedure: process components with their states, process context description, data evaluation features, and correction actions joined in scripts/scenarios.

Decision Making

With regard to supervision systems based on SCADA, the control procedure is based mainly on the following functions:

- keeping given values – operation points – for states of atomic processes by means of local automation components (PLC modules);
- visualization of process data and observation of given bounds for values;
- correction of operation points, switching on/off atomic processes by means of human intervention.

These tasks have been dealt with since the mid 50s. In the meantime, industrial automation systems offer standard modules for performing these tasks. In cases where the controlled process dynamic can be considered linear, small deviations can be compensated by means of standard regulation modules. Currently, such systems have become so sophisticated that they can control even nonlinear behavior when an adequate mathematical model is available.

The need to do analytical modeling using facility engineer and operator expertise results in system design with heuristic components. The implementation of decision making systems is based on rules, in which the facility operation observation and action steps are linked in form of "if-then" objects. The activation of the decision making is based on analysis of data evaluation features on corresponding levels of facility hierarchy. The analysis of the process behavior is performed by a process simulator.

In technical systems constructed from atomic processes, the description of the system operation is usually begun with a definition of functional and behavioral patterns for atomic processes. Subsequently, the operation of grouped elements will be considered.

Keeping in mind that the process and facility have their own dynamics, the control rules must be considered in the context of the system evolution. This is achieved, if the system evolution is simulated under consideration of a given set of decision rules and for a defined time horizon.

KNOWLEDGE OBJECTS

This section is devoted to the modeling of the objects of the proposed decision making system. Here, the model is provided by means of an XML Schema (http://www.w3.org/XML/Schema). For technical networks, an XML-based language for the modeling of supervision and decision making has been developed. In the following, a graphical representation of the knowledge model is given in terms of XMLSPY (Altova GmbH, Austria).

Rule Systems Based on State-Action Relations

We consider the supervision and decision making system which includes the following objects:

- Resource,
- State,
- Condition,
- Rule,
- Action,
- Setpoint.

Object of the type "tResource" represents atomic processes or groups of processes on the corresponding hierarchical level. The entire facility is also an object of the type "tResource". The objects of this type provide the information source and the operation field for the decision making procedure.

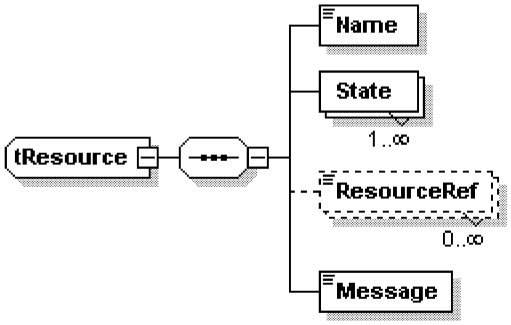


Figure 2: Model of resource objects.

For the decision making procedure, principle components of the resource objects are objects of the type "tState" which collects measurable signals and indicators for the evaluation of facility states and provides an informational starting point for the decision making process.

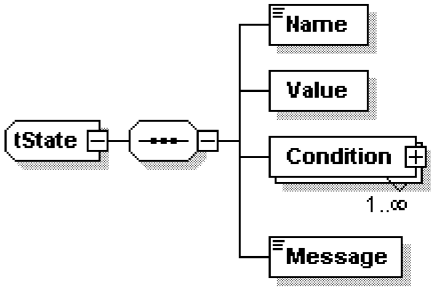


Figure 3: Model of state objects.

For the evaluation of process states, objects of the type "tCondition" are specified. These objects represent data processing features – operators from a predefined set with result value of the type "Boolean". The arguments for these operators are state variables of resource objects.

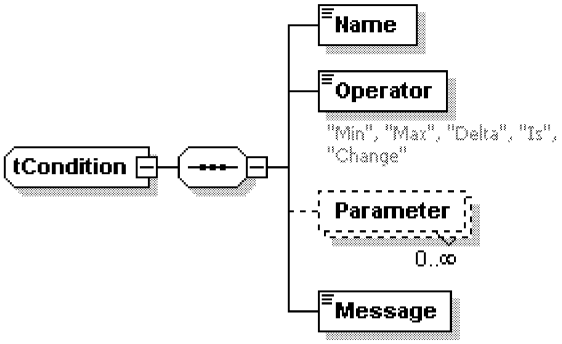


Figure 4: Model of condition objects.

The supervised system allows impacts which influence the future system evolution. The set of impacts correspond to the possibilities of the system operator to change a system behavior in order to achieve aimed results. These impacts are modeled by objects of the type "tAction". The action objects includes objects of the type "tSetpoint", where the reference to a corresponding resource object as well as the type of control impact and its parameters are coded (see Figure 6).

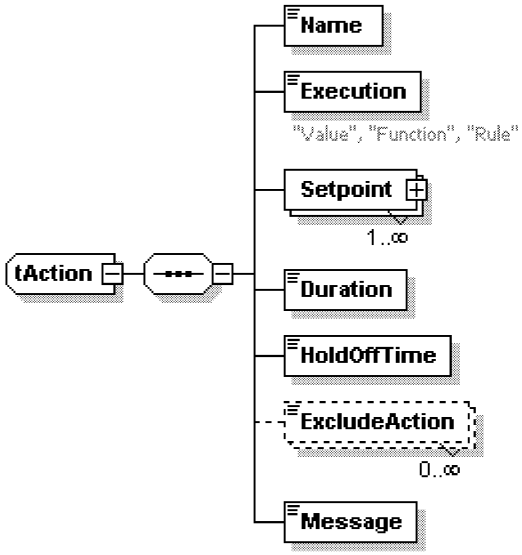


Figure 5: Model of action objects.

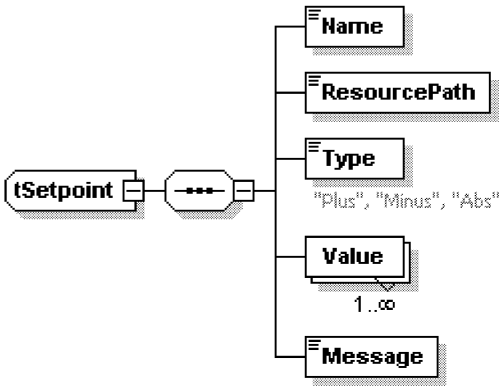


Figure 6: Model of setpoint objects.

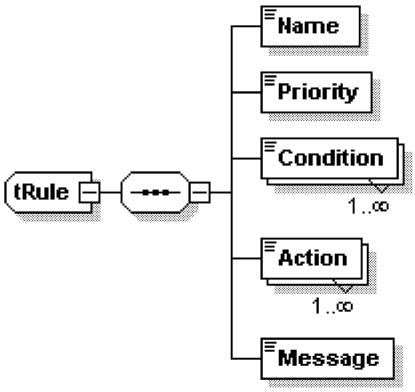


Figure 7: Model of rule objects.

Objects of the type "tRule" link the objects of the type "tCondition" with objects of the type "tAction". If, for some of system states, a condition is failed to meet, a corresponding object of the type "tRule" provides an action which will be performed as reaction.

By the rule objects, the association of system states to control tasks (see Figure 1) is defined. In cases where alternative actions can be identified, the rule objects maintain also priorities of action objects. The model for objects of this type has the following structure:

A set of modeling and visualization environment was developed for the design of the knowledge model in the framework of a technical application (see [Pleßow, 2005]).

INFERENCE SYSTEM

The objective of the inference system is to supervise the compliance of system state values to related conditions over a given time period. The evolution of the process is calculated by means of a process simulator. In cases where some conditions are not met, a sorted list of rules will be elaborated. For actions addressed in these rules, the impact on the functional behavior will be simulated. One can see that the process supervision procedure can be modeled as a tree where alternative rules are represented by branches. This tree is denoted by a *decision tree*.

The model of the decision tree is depicted in Figure 8. One can see that the model has a recursive structure. A proper visualization of the decision tree contributes to better understanding of the system decision scenarios by facility operation staff. These aspects have also been considered by [Pleßow, 2005].

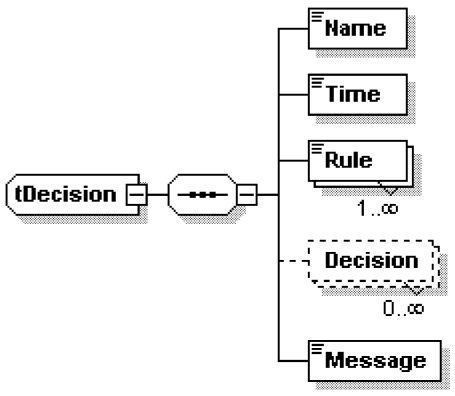


Figure 8: Model of decision tree object.

The function of the inference system was tested and analyzed in a framework of industrial SCADA modules in a control system environment ([Schmidt, 2005]).

CONCLUSION

The idea to perform the modeling of functional and operational behavior patterns in a close relation with a topology model provides systematic and efficient acquisition and processing of rule-based supervision system for technical networks. The knowledge objects are linked to the technological structure of the facility and can be adequately interpreted by facility engineering and operation staff.

Further activities are oriented toward the optimization of the knowledge model structure as well as the elaboration of heuristics for the processing of the supervision rules.

The described approach has been implemented and approved for applications in water supply facilities. Further research will be done in the transfer to other application domains.

ACKNOWLEDGEMENT

The presented knowledge modeling approach was developed in the framework of the R&D project "Akquisition, Management und Integration von Prozesswissen in eine modellbasierte Prozessführung als Repräsentant einer neuen Generation von wissensbasierten Systemen in der Leittechnik," (http://amaryl.ipk.fraunhofer.de) partially funded by the Federal Ministry of Education and Research, Germany. We thank our industrial partner Elpro Prozessindustrie und Energieanlagen GmbH, Berlin, for extensive contacts with companies of the water industry.

REFERENCES

Lisounkin, A. et al. 2004. “Knowledge Acquisition from Historical Data for Case Oriented Supervisory Control.”. In *Emerging Solutions for Future Manufacturing Systems*. Edited by Luis M. Camarinha-Matos, Springer, 499 – 506.

Pleßow, M. et al. 2005. “Tools for Knowledge Acquisition, Modeling and Visualisation applied to Process Supervision”. In *Proceedings of the Industrial Simulation Conference 2005 (ISC-2005, Jun.9-11)*. Berlin, Germany.

Rasmussen, J. 1986. “*Information Processing and Human-Machine Interaction*”. Elsevier Science Publishing Co. Inc., New York.

Schmidt, H.-W. et al. 2005. “Architecture for Simulation and Knowledge Based Supervision for Process Facilities”. In *Proceedings of the Industrial Simulation Conference 2005 (ISC-2005, Jun.9-11)*. Berlin, Germany.

AUTHOR LISTING

AUTHOR LISTING

Abolafia A. **310**
Acosta V. **221**
Agrawal P.K. **367**
Ahlström P. **398**
Alarcón P.P. **13/193**
Angeli C. **216**
Ardizzone E. **53**
Armesto L. **206/236**

Bagheri A. **211/429**
Ball M. **305**
Barbera M. **93**
Basiri S. **429**
Benet G. **68**
Binjie X. **410**
Bolton K. **398**
Bostelmann S. **78**
Brockfeld E. **299**
Bruha I. **37**
Bukvić V. **337**
Burke A. **221**
Busà A. **93**

Cai Y. **405**
Calderón J. **221**
Camacho J **68**
Camahort E. **242**
Carvalho Brito A.E.S. ... **341**
Cepák M. **314**
Chabrol M. **251**
Charfi H. **380**
Chatzinikolaou A. **216**
Chen B.-H. **58**
Chen H. **45**
Cheng C.-H. **101**
Chlaň A. **277/281**
Comelli M. **251**
Correcher C. **68**
Costantino F. **269**
Crosky A. **388**

Dangelmaier W. **150**
Danielsson F. **119**
Das D.B. **348**
de Teixeira J. M.F. **341**
Dehombreux P. **23**
Di Gravio G. **269**
Di Nuovo A.G. **93**
Díaz J. **193**
Douillet P. **422**
Dufour C. **393**
Dvornik J. **28/325**

Earnshaw R.A. **45**
Eickhoff M. **5**
Ender T. **173**
Escriva M. **242**
Esteban J.-L. **193**

Farber K. **371**
Farber P. **371**
Feglar T. **145**
Féniès P. **251**
Fernández-Ramos R. ... **33**
Fleischer J. **173**
Fröhling R. **358**

Gagalowicz A. **380**
Galasso F. **256**
Garbajosa J. **13/193**
Genco A. **231**
German R. **88**
Ghorashi R. **348**
Gokhale M.B. **291**
Gonzalez-Rubio R **129**
Grabot B. **256**
Grekow J. **50**
Groenenboom A. **227**

Hajiloo A. **429**
Hanspal N.S. **331**
Hansson A. Å. **291**
Harneit S. **73**
Hashemi S.M. **211**
Hellgardt K. **331**
Hertkorn G. **285**
Hielscher K.-S. **88**
Hong J.-S. **58**
Hu J. **410**
Huber D. **150**
Huňka F **159**

Joines J.A. **417**

Kale N.A. **201**
Kayakol N. **319**
Kelly D. **388**
Keshavarzi-Roonizi N. .. **348**
Kim B. **393**
Kim Y.K. **405**
Koncar V. **393**
Krajzewicz D. **285/299**

AUTHOR LISTING

Krüger J. **435**
Kudlackova N. **277**
Kühn W. **109**

Lanza G. **173**
Lawrence P. **181**
Legrand X. **388**
Leisen J. **371**
Lennartson B. **119**
Leong S. **137**
Lian Y.-H. **168**
Lin C.-H. **58**
Liporace F.d.S. **115**
Lisounkin A. **358/435**
Lluch J. **242**
Lodewijks G. **187**
Lokshina I. **83**
Lombardo A. **93**
López H. **17**
Lust T. **23**

Mª de Molina R. **68**
Machón I. **17**
Mahajan K. **150**
Marín F.J. **33**
Martı M. **242**
Martin B.R. **417**
Martín J.F. **33**
Märtin M. **371**
McLean C. **137**
McNickle D. **5**
Mebrahtu H. **162**
Mellado M. **68**
Mercé C. **256**
Mikat J. **299**
Milic L. **325**
Milidiu R.L. **115**
Miranda J.C. **221**
Misita M. **337**
Mišković V. **337**
Moghaddam R.Y. **429**
Molina-Conde A.I. **33**
Montés N. **206**
Mortveit H.S. **291**
Mueck B. **150**
Munitic A. **28**
Murty V. **201**

Nagy G. **262**
Nassehi V. **331/348**
Nogueira M.A. **319**
Ottjes J.A. **227**
Ozden M. **101**

Pawlikowski K. **5**
Peri D. **53**
Pleßow M. **353/358**
Pocher M. **358/435**
Puppala A. **310**

Radev D. **83**
Radeva S. **83**
Rentz O. **305**
Reuter M. **73/78**
Rezaee M. **211**
Riane F. **23**
Rijsenbrij J.C. **227**
Ringel J. **285/299**
Ríos F. **33**
Rodehüser E. **371**
Romero-Sánchez J. **33**
Rössel C. **299**
Roux O. **23**

Sabov A. **353/435**
Saffert U. **353**
Safiuddin M. **310**
Sanchez J.M. **242**
Schembra G. **93**
Scherma G. **231**
Schipotinin T. **371**
Schmidt H.-H. **353**
Schreck G. **353/435**
Seyam A.A. **405**
Shams-Nateri A. **376**
Shao G. **137**
Skrifvars M. **398**
Soliman M. **310**
Sorce S. **231**
Šotek K. **281**
Stanojević P. **337**
Stoian V. **63**
Strange C. **162**
Svensson B. **119**

Tchernev N. **251**
Thomassey S. **422**
Thoney K.A. **417**
Thron T. **262**
Tireli E. **28/325**

AUTHOR LISTING

Tornero J. **206/236**
Tripp J.L. **291**
Trivedi R. **201**
Tronci M. **269**
Tuchscheerer W. **299**

Ulgen O.M. **201**

Van Landeghem H. **168**
Veeke H.P.M. **187**
Vivó R. **242**
Vyas K.J. **331**

Wagner I. **88**
Wagner P. **285/299**
Walker R. **162**
Wan T.R. **45**
Wassan N. **262**
Wietschel M. **305**
Williams E.J. **201**
Wösler R. **299**

Xue L. **367**
Yagüe A. **193**
Yeh C-H. **58**

Zaitsev D. **124**
Zubieta L. **129**