

**ESS 2004**

**SCIENTIFIC PROGRAM**



# **PLENARY PAPER**



# Virtual Reality: Computational Modeling and Simulation for Industry

Dietmar P. F. Möller<sup>1,2)</sup>

<sup>1)</sup>University of Hamburg, Faculty of Computer Science, Chair Computer Engineering  
Vogt-Kölln-Str. 30, D-22525 Hamburg, Germany

<sup>2)</sup>College of Engineering, Computer Science, and Technology  
California State University  
Chico, CA95929-0003, USA

## ABSTRACT

This paper outlines the core technologies which underline the principle of virtual reality and the way it is being applied today in industry. In a more general sense virtual reality provides a true 3D interface to a range of computer applications. The essence of virtual reality is immersion, which is the ability to immerse the computer user in a computer generated experience, as an active participant, as opposed to a passive viewer. Hence this paper provides an introduction into the methodology of virtual reality, including its historical background, as well as some basic taxonomy, that are helpful defining the elements of a virtual reality, that are used to create immersive and interactive experience. Moreover this paper report about industrial case study examples of Virtual Reality as an advanced computational method in modeling and simulation of complex dynamic systems.

## 1 INTRODUCTION

Virtual reality (VR) can be described as a synthetic 3D computer generated universe that is a perceived as the real universe. The key technologies behind virtual reality systems (VRS) and virtual environmental systems (VES) are

- Real-time computer graphics
- Colour displays
- Advanced software

Computer graphics techniques have been successfully applied for creating synthetic images, necessary for virtual reality and virtual environments. Creating an image, using computer graphics techniques are used to store 3D objects as geometric descriptions, which can then be converted into an image by specific information of the object, such as colour, position, and orientation in space, and from what location it is to be viewed.

Due to this possibilities computer graphics is now well established and developed as a standard for Computer Aided Design (CAD) systems. With a CAD system e.g. one can render perspective views of the scene under development. The success of CAD has been greatly influenced by the advances of computer architecture due to the microminiaturization of computer chips, and the instruction level parallelism of computer architectures, realized as pipelined instruction processing and superscalar instruction handling in modern microprocessors, that allow to build low cost graphic workstations that can support the real-time intensive manipulation of large graphic databases.

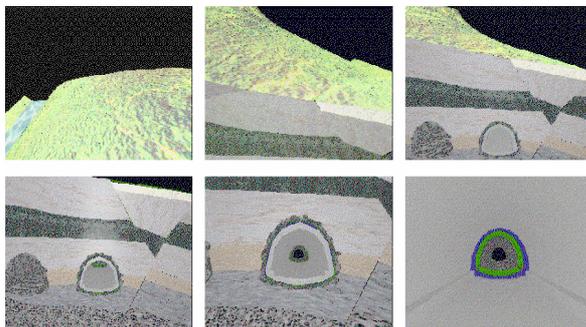
Real-time computer graphics techniques allow the user to react within the time frame of the application domain, which finally results in a more advanced man machine interface, which is the whole rationale for virtual reality systems and virtual environments. This has been possible by overcoming the delays that occur e.g., while rendering of very large database, since the computer power increase and henceforth the graphics performance, due to the instruction level parallelism. An unpipelined processor is characterized by the cycle time and the execution times of the instructions. In case of pipelined execution, instructions processing is interleaved in the pipeline rather than performed sequentially as in non-pipelined processors. Defining the performance of pipeline processing the performance potential of the pipeline equals the number of independent instructions which can be executed in a unit interval of time.

Colour displays are used for displaying the views of the virtual reality and virtual environment universe to provide a visual sensation of the objects from the real-world domain (industrial physical application) into the virtual domain. The colour displays are of great variety, such as monitors fixed to the windows of the simulator cockpit for visual sensation of flying in a flight simulator (see Figure 3), or Head Mounted Displays

which visually isolates the user from the real world. A head mounted display (HMD) can provide the left and right eye with two separate images that include parallax differences, which supplies the users eyes with a stereoscopic view of the computer generated world, which is a realistic stereoscopic sensation.

Advanced software tools are used to support the real-time interactive manipulation of large graphic databases, which can be used to create a virtual environment, that can be anything from 3D objects to abstract data bases. Moreover, 3D modelling and simulation tools are part of the advanced software tools. Hence, a 3D model can be rendered with an illumination level simulating any time of the day, or using OpenGL, a quasi standard for 3D modelling and visualization, one can create geometrical bodies of every shape and size for simulating the different views of geotechnical and geophysical parameters e.g. of a tunnel scenario, as shown in Figures. 1 and 2, which can be moved in size in real time, using advanced simulation software tools, such as NURBS or the MBA algorithm. The image realism can be improved by incorporating real-world textures, shadows, and complex surfaces, etc.

Example 1: Based on OpenGL, a quasi standard for 3D modeling and visualization, one can create geometrical bodies of every shape and size and move them in real time, within a virtual reality simulator, as shown in Figure 1, that shows the sequence of a “flight through a tunnel”. Figure 1 top line from left to right: overall scenic view of the landscape; scenic view and different geological structures; scenic view, different geological structures and tunnel inlets. Figure 1 bottom line from left to right: different geological structures and tunnel portals; different geological structures at the tunnel portal; scenario inside the tunnel with the end of the tube in front of the view.

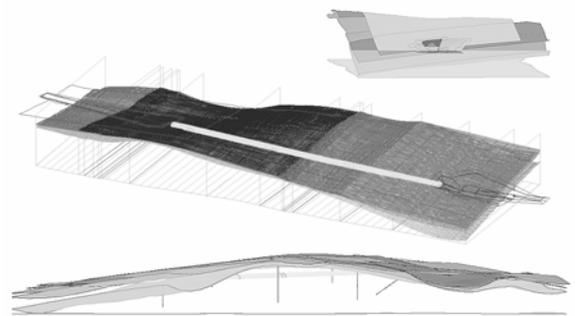


**Fig. 1.** Virtual reality tunnel simulation scenario

Due to intuitive interaction with the virtual reality techniques new scenic presentations are possible, as shown in Figures. 1 and 2, which offers concepts for

modelling and simulation of complex real-world systems with parameterized or non parameterized topologies within a unique framework. This results in rapid prototyping based on flexible modelling tools with concepts for geometry, motion, control, as well as virtual reality components like images, textures, shadowing, rendering, animation, multimedia, etc.

The technical complexity associated with developments in the virtual universe requires the use of metric values, which can then be converted into several important factors that relate to the metric values themselves, especially metric dimensionality, metric attributes, metric types, etc.



**Fig. 2.** 3D model of the virtual reality tunnel simulation scenario of Figure 1.

Top: General 3D NURBS model view  
 Middle: 3D NURBS model view of the tunnel and the geological structure with tunnel portals  
 Bottom: 3D Geological formation model view

The primary goal using virtual reality is to unite the power and flexibility of virtual reality methodology with the insight of ubiquitous computing which can be stated as computation in space and time, based on:

- Image processing, which contains the two sub-topics of image acquisition and image analysis, that are necessary to produce 3D information to generate a useful representation of the environment
- Computer graphic and visualization, which is necessary for modelling virtual environments, creating stereo visions and rendering images, whereby the time to render an image has always been a major issue of computer graphics, especially with animated sequences
- Synthetic scene generation
- 3D modelling based on splines, NURBS and other innovative algorithms such as V-NURBS or VT-NURBS

Henceforth very new scenic presentations are possible, containing branch specific elements and knowledge in

the respective application domain. Moreover the effect of immersion, which means the realization of space depth, allows the user, a fast adaptation to processes in space and time. Virtual reality offer a concept for modelling complex systems with parametric as well as nonparametric topology within a unique frame. This results in rapid prototyping, based on flexible virtual reality modelling tools with concepts for geometry, motion, control, as well as virtual reality components like images, textures, voice, animation, multimedia, video, etc.

The technical complexity associated with developments in the virtual reality domain require for the introduction of characteristics of metric values. In the development of utility metric values, several important factors that relate to the metric values itself must be considered. These areas that are

- Metric dimensionality
- Metric attributes
- Metric types.

A very easy and straightforward approach for realisation of metric valuated dimensions could be found using one-dimensional scaling. Methods of one-dimensional scaling, however, are generally applied only in those cases where there is good reason to believe that one dimension is sufficient.

But metric valuated accuracy and presentation fidelity leading for a multidimensional scale. A multidimensional scale is necessary for an adequate images quality description, if additional information would probably be required. Therefore a multi-dimensional scale must be developed.

Metric valuated attributes are actual quality parameters measured along each quality dimensions, which are:

- Realism: defined as degree of fidelity of representations compared with truth. The degree of realism vary from task to task. To enhance levels of realism often certain kinds of standard features and effects are used.
- Interpretability: corresponds to the level of resolution, which coarsely could be defined as the smallest feature that can be distinguished in an image. This, in turn, directly impacts the level of information that can be interpreted from an image.
- Accuracy: deals with the correctness of objects represented in the scene, and their correct locations. Accuracy depends of source materials used, like geometry derived from imagery, physical models,

etc. as well as the fidelity of the transformations used.

There are a number of possible metric valuated types that could be used for the dimensions of a quality assessment metric. These types are [Yachik, 1998]:

- Criteria based ( on a textual scale which define the levels of the scale)
- Image based (on a synthetic scene where a rating is assigned by identifying the standard image having a subjective quality that is closest to that being rated)
- Physical parametric based (on measured values based on integrated power spectrum, or mensuration error, etc.)

Methodology developers realized that logical constraints or axioms are necessary in order to enhance with mathematical meanings, like [Yachik, 1998]:

- Monotonic  $\Rightarrow$  as items increase in values, the scale number increases
- Continuity  $\Rightarrow$  intermediate scale values have meaning
- Transitivity  $\Rightarrow$  if item a > item b > item c; then item a > item c.

Several scales can be developed that satisfy these axioms to various degrees.

Virtual Reality is a natural domain for collaborative activities because VR allow users doing things they normally cannot do in reality, e.g. being within a molecule, being inside the combustion chamber of an automobile engine, walking through a tunnel in „outer space“ etc.

The big challenge of virtual reality techniques is that it takes us one step closer to virtual objects by making us part of the virtual domain. Computer graphics techniques used in the virtual reality systems of today providing visual images of the virtual universe, but the systems of tomorrow will also create acoustic images of the virtual universe, which can be introduced as sonification or the 5-th dimension of the virtual reality technique – while time is the 4-th dimension – that can stimulate recognizing sounds in virtual environments. One could imagine that more modalities of user interaction, such as tactile and haptic modalities for touching and feeling of virtual objects, can complete the sensation of illusion in virtual worlds, which can be introduced as the 6-th and higher dimension of virtual reality techniques. Moreover, smelling and tasting may also become imaginable in virtual environments, en-

hancing the order of dimension. Henceforth multi-modality, that cover the sensation in virtual worlds, has become one of the major topics in the design of virtual environments. The benefits of the technique of virtual reality are manifold, that is why this technique is so vital to many different application domains, ranging from applications in the automotive and avionics industry, applications in the more advanced military industry, as well as molecular and medical topics, catastrophic management, education and training, etc., and ends up into the different academic research domains. Based on features offered through computer graphics techniques, meaning visualization of highly realistic models, and through the integration of real-time computing, virtual reality enables the user to move around in virtual environments, such as walking in the environment of a tunnel or walking through the through a tunnel, as shown in Figure 1, or to acquire flying skills without involving real airplanes and airports, as realized in virtual training environments for pilots, as shown in Figure 3, etc.

Based on the spatial and temporal geometric description, which can then be converted into an image by specifying the respective information behind, virtual reality techniques can be used as the basic concept for virtual-world simulation, as well as for analysis and prognosis of complex processes in virtual worlds.

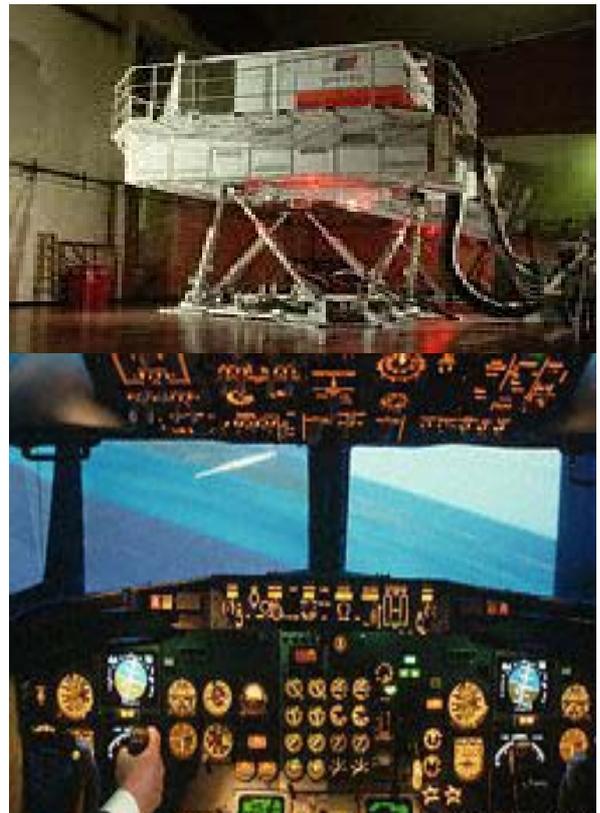
Furthermore, underlying databases in virtual-environments offer the ability to store and retrieve heterogeneous and huge amounts of data for modelling virtual worlds. Hence, virtual reality can be seen as a specific type of a real-time embedded system combining different technological approaches that are integrated within one environmental solution.

In the case of a flight simulator, as shown in Figure 3, the computer graphics techniques are used to create a perspective view of a 3D virtual world, and the view of this world is determined by the orientation of the simulated aircraft. Simulating the complex behaviour of the aircraft requires a sophisticated modelling technique and embedding of several real-time systems, such as engines, hydraulics, flight dynamics, instruments, navigation, etc., as well as weather conditions, and so on, which are components and modes of the flight simulators virtual-environment. The information necessary to feed the flight simulator with real-world data are available from the databases of the aircraft manufacturer and the manufacturer of the aero engines. They describe the dynamic behaviour of the aircraft when taxiing on the ground, or flying in the air, or engine temperature and fuel burn rates, etc. The flight models used in the flight simulator are based on the data obtained from the manufacturer as well as the

data describing the flight controls to simulate the behaviour of the airplane under regular as well as under non-regular flight conditions.

During flight simulation, the pilot – as well as the co-pilot – sit inside a replica cockpit and gaze through the forward-facing and side-facing windows, which are 200° panoramic displays reflecting the computer-generated graphical virtual universe. The flight simulator creates a realistic sensation of being in a real-world plane flying over some 3D landscape, as shown in Figure 3. But today, the flight- simulator panoramic displays do not contain stereoscopic information, the fact that the images are collimated to appear as though they are located at infinity creates a strong sense of being immersed in a 3D world.

Furthermore, immersion can be enhanced by allowing the users head movements to control the gaze direction of the synthetic images that provides the users brain with motion-parallax information to complement other cortical pathways of the visual cues in the brain. This requires tracking the users head in real time, and if the users head movements are not synchronized with the images, the result will be disturbing.



**Fig. 3.** Hydromechatronic flight simulator mock up (top) and cockpit view from inside the flight simulator (bottom)

When visually immersed within a virtual environment there is a natural inquisitive temptation to reach out and touch virtual objects as part of interaction possibilities in the virtual universe, which is impossible, as there is nothing to touch and to feel, when dealing with virtual objects. But, the users sense of immersion can be greatly enhanced by embedding tactile feedback mechanisms in the virtual environment. Embedding tactile feedback needs some specific hardware components, such as data gloves, which enable the user to grasp or to sense real-time hand gestures. Hence data gloves will provide a simple form of touch-and-feel stimulus where small pads along the fingers stimulate a touching and feeling sensation. Thus, if a collision is detected between the users virtual hand – the data glove – and a virtual object, the data glove is activated to stimulate the touch and feel condition. However, the user may not be suddenly aware of the objects mass, as there is no mechanism for engaging the users arm muscles. Therefore, it is necessary to transmit forces from the virtual domain to the user interface, meaning there is a need for embedding articulated manipulators in the virtual environment that could create such forces.

In general for virtual space applications the following main components are available:

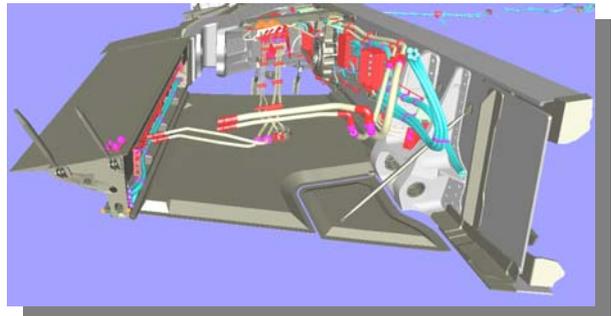
- Space ball and cyber gloves for tactile interaction in virtual space
- In-transparent head mounted devices for visual interaction in virtual space; transparent head mounted devices are used for augmented reality application
- 3D geometric body creation and motion methodology for “virtual space feeling” capability
- 3D visual interactive components for definition, manipulation, animation and performance analysis of geometric bodies
- Object oriented data base for efficient data management
- Objects organization into single-inheritance hierarchies for virtual reality system transparency. When objects are created, they inherit the properties and verbs of their ancestors. Additional verbs and properties as well as specializations of inherited components may be defined to give the new object its unique behavior and appearance
- Computing hardware for computing in space and time

There are many advantages of working in the virtual domain, such as:

- Accuracy due to subject specification: means real world models can be built with great accuracy as they are based upon CAD data of the real objects
- Flexibility: means building virtual representations of anything as well as interacting with this representation due to the virtual reality front ends
- Animated features: means animation of sequences, objects, etc., in space and time

Example 2: Combining these aspects for real-time simulation in virtual environments can be based on the integration of the overall information, but only a few approaches maintain this problem and have been developed like cave automatic virtual-environments, or digital mock up (DMU) in the avionic industry, shown in Figures. 4 and 5, allowing the user a real-time interaction that is not only restricted to the 3D model itself, it also is parameterized, which could lead to a better framework for real-world system analysis, such as

- Statistic and cinematic interference tests
- Development of new methods for DMU application
- Investigation of applicability of new technologies within the virtual product design process



**Fig. 4.** Digital mock up (DMU) of a air planes wing



**Fig. 5.** Digital mock up (DMU) of an air plane wing showing the application of virtual reality to simulate the possibility of a maintenance procedure at the air planes wing

In contrast to Virtual Reality (VR) Augmented Reality (AR) deals with the combination of real and virtual environments. The AR system supports the user, based on semi-transparent output devices, with the relevant computer based information. This means that the images users will see on their AR device show the geometry dependent right perspective in correlation to the real world scenario.

## 2 HISTORICAL DEVELOPMENT OF VIRTUAL REALITY

Like most technologies, Virtual Reality (VR) did not suddenly appear. It emerged in the public domain after a long period of research and development in Industrial, military and academic laboratories. The emergence of VR was closely related to the maturity of other technologies such as real-time computer systems, computer graphics, displays, fibre optics and 3D tracking. When each of these technologies could provide its own individual input, a crude VR working system appeared. But this was a very long way with inventions and discovery. The foundations of today's important technologies used for VR goes back into the twentieth century. The scientific landmarks of which shows the following:

- 1944: Harvard Computation Laboratory completed their automatic general-purpose digital computer
- 1954: Cinerama was invented with 3 cinema screens
- 1956: Morton Heilig invented the Sensorama, a CRT based binocular headgear
- 1960: The Boeing Corporation coined the term Computer Graphics
- 1963: Ivan Sutherland submitted his PhD thesis "SKETCHPAD: A man-machine graphical communication system"
- 1966: NASA commence with the development of a flight simulator
- 1968: Ivan Sutherland published "A Head-mounted 3D Display"
- 1977: Dan Sandin and Richard Sayre invented a Bend-Sensing Glove
- 1981: Tom Furness developed the Virtual Cockpit
- 1984: William Gibson wrote about Cyberspace in the book Neuromancer
- 1989: Jaron Lanier coined the term Virtual Reality
- 1990: Fred Brooks developed the force feedback
- 1991: After founding in 1990, several industries selling their first VR System
- 1993: SGI announced the Reality Engine
- 1994: The Virtual reality was founded

## 3 VIRTUAL REALITY TAXONOMY

Applying the virtual reality methodology to the industrial domain could be stated as combining distributed virtual environments, in order to support collaboration among distance team members developing plans and procedures, doing measurements and data processing, in order to attempt to manage new investigations and organizations collaboratively, as it is needed in global as well as international project development.

One of the most interesting new paradigms in virtual reality methodology in this domain is that 3D representations are not only the lonely possibility of a setting.

Many virtual space applications in today's industry, if not already now, will in future make use of specific graphics. The virtual reality will be visualized in space, which means in terms of 3D, and due to dynamic aspects in time. People that are work in VR projects are able to interact image based within space and time, e.g. fighting a plane, inspecting car crash behavior, interacting with other participants through a graphical user interface, etc. The interweaving of functionality, distribution, efficiency, and openness aspects is very noticeable in computer graphics. The virtual space is graphically visualized flamboyance and for the most part the people work in the outer space application domain will see the same images. In real industrial projects, irrespective of the number of participants, a change of state in the virtual space needs to be communicated to all invented in the project.

Therefore, for virtual space applications, 3D multiuser virtual reality tools have been developed for industrial applications, consisting of the following main components:

- space ball and cyber gloves for tactile interaction in virtual space
- head mounted devices for visual interaction in virtual space
- 3D geometric body creation and motion methodology for "virtual space feeling" capability
- 3D visual interactive system for specification, manipulation, animation and performance analysis of industrial bodies
- object-relational oriented data base for efficient data management in virtual reality applications
- computer hardware for the power of computing in space and time
- objects organization into single-inheritance hierarchies for virtual reality system transparency

When objects are created, they inherit the properties and verbs of their ancestors. Additional verbs and properties as well as specializations of inherited components may be defined to give the new object its unique behavior and appearance.

The presentation of process states is of importance, which has to be realized time dependent, combining real scenarios as well as virtual scenarios, in order to find out optimal geometries, based on a specific mathematical annotation, that are Non Uniform Rational B-Splines (NURBS).

This special kind of B-spline representation is based on a grid of defining points  $P_{i,j}$ , which is approximated through bi-cubic parameterized analytical functions.

$$P_{i,j} = \left\{ \begin{array}{cccc} p_{1,1} & p_{1,2} & \cdots & p_{1,n} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m,1} & p_{m,2} & \cdots & p_{m,n} \end{array} \right\}, p_{i,j} = (x, y, z)$$

$$S(u, v) = \frac{\sum_{i=0}^n \sum_{j=0}^m N_{i,p}(u) N_{j,q}(v) w_{i,j} P_{i,j}}{\sum_{i=0}^n \sum_{j=0}^m N_{i,p}(u) N_{j,q}(v) w_{i,j}}$$

$$0 \leq u, v \leq 1$$

This method allows to calculate the resulting surface or curve points by varying two (surface) or one (curve) parameter values  $u$  and  $v$  of the interval  $[0,1]$ , respectively, and evaluating the corresponding B-Spline basis function  $N_{i,p}$ .

$$N_{i,0}(u) = \begin{cases} 1 & \text{if } u_i \leq u \leq u_{i+1} \\ 0 & \text{otherwise} \end{cases},$$

$$N_{i,p}(u) = \frac{u - u_i}{u_{i+p} - u_i} N_{i,p-1}(u) + \frac{u_{i+p+1} - u}{u_{i+p+1} - u_{i+1}} N_{i+1,p-1}(u)$$

$$U = \{u_0, \dots, u_m\}, u_i \leq u_{i+1}, V \text{ analogous}$$

The parameter values  $u$  and  $v$  can be chosen continuously; the resulting object is mathematically defined in any point thus showing no irregularities or breaks

There are several parameters that approximate given points and thus adapting the view of the described object, and interpolation of all points can be achieved.

Firstly, the polynomial order describes the curvature of the resulting surface or curve, representing the mathematical function at a higher level of flexibility. Secondly, the defined points can be weighted in accordance to their dominance with respect to other control points. Hence a higher weighted point influences the direction of the surface or curve more than a lower weighted one. Furthermore, knot vectors  $U$  and  $V$  define the local or global influence of control points, so that every calculated point is defined by smaller or greater arrays of points, resulting in local or global deformations, respectively, as shown in Figure 6.

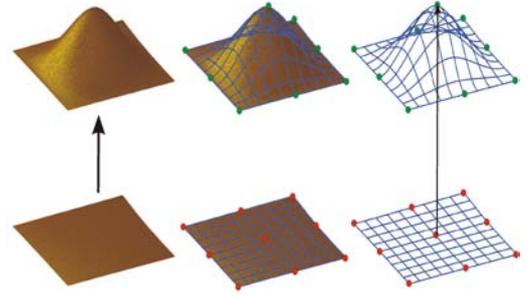


Fig. 6. Modelling and modification of a NURBS surface

NURBS are easy to use, as modelling and especially modifying is achieved by means of control points movement, allowing the user to adjust the object simply by pulling or pushing the respective control points.

Based on these concepts a mathematical methodology is available that allows to interpolate given sets of points, for example the results of scanned data of the human face surface measurements, as shown in Figure 7.

Using multiple levels of surface morphing, this multi level B-spline approximation (MBA) adjusts a predefined surface, i.e. a flat square or a cylinder. Constraints like curvature or direction at special points can be given and are evaluated by the algorithm.

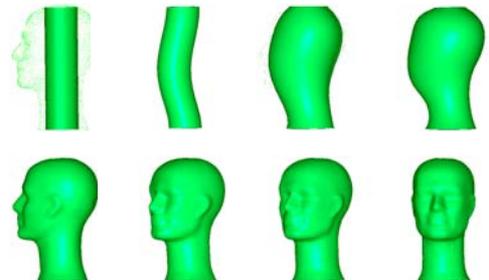


Fig. 7. Morphing, by using multi level B-splines approximation

#### 4 VIRTUAL REALITY SYSTEMS IN INDUSTRY

One of the reasons that VR has attracted so much interest in the military and in the industry is that it offers enormous benefits to so many different application areas.

VR systems firstly have been applied at the US Air Forces Armstrong Medical Research Laboratories developing an advanced fighter cockpit which later has been adapted by the avionic industry as flight simulator for pilot training (see Figure 3) and used by commercial airlines and the military for over twenty years. They are used for training pilots in developing new skills in handling aircraft under usual and unusual operating conditions, and discovering the flight characteristics of a new aircraft.

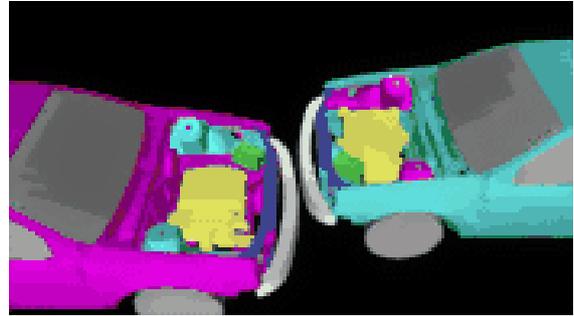
Other applications deals with operations in hazardous or remote environments, where workers could benefit from the use of VR-mediated teleoperation. People working in radioactive, toxic or space environments could be relocated to the safety of a VR environment where they could handle any hazardous materials without any real danger. But telepresence needs further development in tactile and haptic feedback to be truly useful. Today one may find high sophisticated Industrial VR teleoperation applications in the field of nanoscience based on force feedback manipulation, as shown in Figure 8.



**Fig. 8.** Force feedback manipulator

Many areas of design in industry are inherently 3D. Example 3: The design of a car shape needs 3D support while the designer looks for sweeping curves and good aesthetics from every possible view, as well as

placing parts and sub-systems within the car, as shown in Figure 9.



**Fig. 9.** VR design in industry (for details see text)

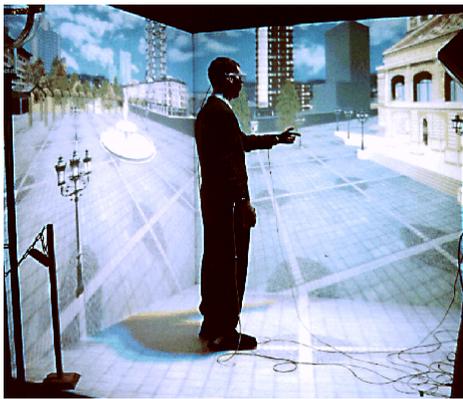
Moreover the expensive physical car crash tests, as shown in Figure 10, have been adapted for use in a virtual car crash simulation environment. This requires the mathematical description of the procedures, a task which is not trivial. In general the real problems arise in designing an interface where such behaviours can be associated with arbitrary geometric databases.



**Fig. 10.** Physical car crash test

Furthermore, shared VR environments will allow possibly remote workers to collaborate on tasks. The VR environment can also provide additional support for the collaboration of industrial designer teams. For

specific application areas, such as land exploration or architecture, caves and domes, as shown in Figure 11, become the most important virtual reality environment.



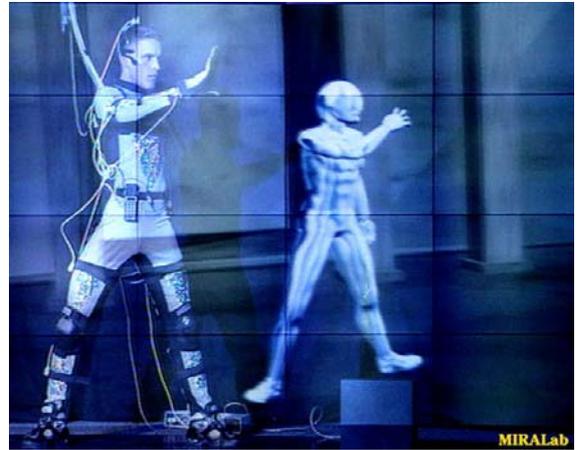
**Fig. 11.** Domes and caves as VR environments (for details see text)

More and more gesture driven interaction as a human factor in virtual environments become of importance. The reason behind is very simple., because if the VR environment is designed to follow some metaphor of interaction in space, and the objects in the environment have behaviour similar to objects in the real world, it can be very intuitive for the user to visit and experience this environments.

One possibility behind gesture driven interaction in VR environments is the possibility using avatars with hand gesture to communicate with deaf people.

The MIRALab at Genova, Switzerland, chaired by Prof. Nadja Thälmann, is one of the pioneers in gesture driven interaction. They build individualized wal-

king models, walking along an arbitrary path, fashion shows, virtual tennis games, and cyber dancing, as shown in Figure 13.



**Fig. 12.** Cyberdance in VR

Moreover VR has become part of entertainment. Current equipment in this domain is used for role playing games. It consists of a number of units networked together in which a team of players stand to play on this roles.

Summarizing the above mentioned VR environments with its specific application In general VR application in industry can be found in:

- Engineering:
  - Aero engine design: where engines are designed to withstand incredible forces during operation, and must function in all types of whether, and over incredible ranges of temperature and atmospheric conditions
  - Submarine design: where VR is used to investigate maintenance and ergonomic issues, as well as model human figures to evaluate how real humans cope with working in the confined spaces associated with submarines
  - Industrial concept design: where collaborative work is used to interact with virtual car concepts, and how such designs can be evaluated for functional correctness, easy assembly and even maintainability
  - Architecture: where VR is a technological progression that brings architects closer to their visual designs
- Entertainment:
  - Computer animation
  - Games systems

- Science:
  - Visualization of electrical fields
  - Molecular modelling
  - Telepresence
- Training:
  - Fire service
  - Flight simulation (see Figure 3)
  - Medicine
  - Military training (see Figure 13)
  - Nuclear industry
  - Accident simulator (see Figure 14)



Fig. 13. Military training using VR



Fig. 14. Accident simulator using VR

## 5 CONCLUSIONS

If we look at the relationships between computers and humans over the past 20 years, we find out that a merge occur bringing both world closer and closer

together. 3D and virtual environments (VEs) meanwhile are around for many years. The simulation industry is also very familiar with modelling, displaying and updating virtual environments in real time. Training simulators for tanks, ships, aircraft and military vehicles use VR and VE as a substitute for a real working environment. But in future the modes of interaction will increase, meaning that the VR and VE systems of tomorrow will allow us beside the interactions of today to sense, to feel, to hear, to smell, etc.

## 6 REFERENCES

- Earnshaw, R. A., Gigante, M. A., Jones, H.: Virtual Reality Systems, Academic Press, San Diego, 1993
- Kesper. B., Möller, D. P. F. : Temporal Database Concepts for Virtual Reality Reconstruction, In: ESS, 2000, pp. 369-376, Ed. D. P. F. Möller, SCS Publ. Ghent, 2000
- Möller, D. P. F.: Virtual Reality and Simulation in Medicine (invited keynote), In: ESM 200, pp. 10-17, Ed. P. Gerill, SCS Publ. Ghent, 2000
- Möller, D. P. F. : Virtual and Augmented Reality: An Advanced Simulation Methodology applied to Geoscience, In: Proceed. 4<sup>th</sup> bMATHMOD, pp. 36-47, Eds.: I. Troch, F. Breitenacker, ARGESIM Report Vol 24, Vienna 2003
- Möller, D. P. F.: Virtual Reality Framework for Surface Reconstruction, In: Networked Simulation and Simuzlated Networks, pp. 428-430, Ed. G. Horton, SCS Publ. Ghent, 2004
- Vince, J.: Virtual Reality Systems, Addison Wesley, Reading, 1995
- Yachik, T. R.: Synthetic Scene Quality Assessment Metrics Development Considerations. In: Proc. VWSIM'98 (Eds.: Landauer, C. and Bellman, K. L.), SCS Publishers, San Diego, 1998, pp. 47-57

## 7 AUTHOR BIOGRAPHIE

**DIETMAR P. F. MÖLLER** is a Professor of Computer Science and Chair of Computer Engineering at the University of Hamburg (UHH), Germany. He is Director of the McLeod Institute of Simulation Sciences at UHH and Adjunct Professor at California State University Chico. His current research interests include modeling and simulation methodology, virtual and augmented reality, embedded computing systems, mobile autonomous systems and robots, data management, e-learning and e-work, as well as nanotechnology and statistics applied to micro array analysis used in cancer research.

# **SIMULATION IN MANUFACTURING AND PRODUCTION**



# d<sup>3</sup>FACT INSIGHT: A SIMULATION-TOOL FOR MULTIREOLUTION MATERIAL FLOW MODELS

Wilhelm Dangelmaier

Bengt Mueck

Christoph Laroque

Kiran Mahajan

Heinz Nixdorf Institute, University of Paderborn

Fürstenallee 11, 33102 Paderborn, Germany

{whd, mueck, laro, kiran}@hni.uni-paderborn.de

## KEYWORDS

Multiresolution Simulation, Material Flow Simulation,  
Virtual Reality, Digital Factory

## ABSTRACT

In a global economy, successful organizations constantly use innovative manufacturing methodologies to stay competitive in business. Among others, simulation is a tool which offers interesting perspectives from the manufacturing system optimization point of view. However, when a simulation model becomes large, and the entire model is simulated at a high level of detail or resolution, computing power tends to become a bottleneck. As a result, if the model is large it is seldom possible to calculate/simulate it in real time. Real time simulation is desirable if an interactive analysis of the manufacturing system is required within virtual environments. Secondly, in a virtual environment a user can view only a part of the simulation model. Hence, in our approach only this part is simulated in high resolution with high effort of computing power. The parts of the model, the user ignored or cannot view, are simulated on a rough level. Consequently, if the user moves, the area which is simulated in high resolution also moves with the user. This way he gets an impression of a simulation in high resolution. This also enables the user to analyze large simulations in real-time.

In this paper, we illustrate a method for detecting the user attention based on modeling approach. These detections will stimulate adjustments in the level of detail. After an adjustment the starting state of the newly activated models have to be generated. Methods to do this are shown. Most of these methods have been implemented in our simulation tool d<sup>3</sup>FACT INSIGHT. Then, a short example of a multiresolution material flow simulation is shown followed by conclusions.

## OBJECTIVES

Simulation of material flow systems is a well-known method to set up new production plants. It is easy to analyze different scenarios and to answer questions like: Will a faster machine raise my throughput? Besides this, processes and dependencies can be detected easily. For such simulations, virtual environments are often

used to give the user a good impression of the model. He can move freely through the model and analyze production processes he is interested in. If he wants to modify or perform interactions with the model, the simulation needs to run simultaneously with the visualization. In this case it is not possible to base the calculation of the visualization on a trace file (Dangelmaier and Mueck 03).

The execution speed of a simulation depends on the size of the model, (which depends on the size of the system being modeled) the detail with which the model is being simulated and the speed of the computer calculating the model. Rougher models leads to a faster calculation. More detailed models need more calculation time. If the models are large and detailed enough, it is not possible to calculate them fast enough and further use them for analysis within interactive virtual environments.

Typically the user can only see a small part of a large scene. We simulate the area, which is surrounding him (and which he can view), with a high level of detail. The areas he is not viewing (or cannot view) are simulated on a low level of detail. If the user moves or turns around, the high detailed area follows the user. So the user gets an impression of a simulation which is calculated completely in high resolution. Because most of the simulation takes place in a rough level of detail, the required calculation time is reduced. As a result, bigger and more detailed models can be analyzed with this approach.

To implement this idea, first we need a representation of models which work in different levels of detail (at the moment these models have to be modeled individually by another modeler). During the simulation/execution one set of models is activated to represent the whole simulation-model. This activation of models has to be identified and if the user moves the activation has to change correspondingly. Besides this, indications of the required level of detail are also required.

For our material flow simulation the state of the system is preserved with tokens and their assignment to objects and an event queue. If the identification stimulates a change in the activation of different models, the state of

the active model has to be generated. Methods will be described later in this article.

## STATE OF THE ART

Some research has been done for modeling and simulating models in different resolutions (e.g. Davis et al. 1998 or Reynolds and Natrajan 97). In these approaches there are models of one object in different levels of detail, between which a switching is possible during the run-time of the simulation – if needed (especially when two partial models want to communicate on different levels of detail). However, if a lot of switches are successively done, inconsistencies can occur. On the basis of this problem, Natrajan proposed to work with only one description of the state and to provide interfaces on each level of detail for interactions with partial models, which are available on different levels of detail. The information which is necessary in the interfaces is only generated by aggregation or disaggregation when needed. Since the interfaces information is only transformed, the description of the state is not affected by these transformations.

An alternative approach is presented by Kim, Lee, Christensen, and Zeigler with the System Entity Structuring and Model Base Management Approach (SES/MB) (cp. Kim, et al. 92). They collect partial models with different levels of detail in a model base for working with different levels of detail. The SES/MB contains a tree comprising possible compositions of partial models for an overall model for the composition of simulation models, which are based on the models of the model base. Models with different levels of detail can be generated from this tree and the MB. The level of detail is determined before the simulation is computed. In other words, in this approach, a level of detail which changes dynamically during run-time is not planned.

These approaches do not use the user’s view as stimulation for changing the resolution of individual objects during the execution time. Secondly, their objective is not to reduce calculation time.

## MODELING

To understand how we model multiresolution models we first take a short look on how we model traditional non-multiresolution models in our approach.

### Non-multiresolution modeling

As mentioned earlier we are using token-event based systems to model material flow systems. So a model consists of positions which represent static resources like machines. The current state of the positions is described with the number of tokens currently assigned to the position (see Figure 1).

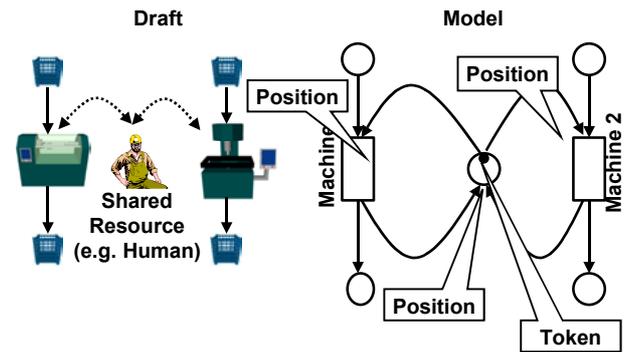


Figure 1: Modeling of production systems with a token based approach.

To change the assignments between tokens and positions we use events. An event occurs only at one specific position. It consists of a time and type. If the simulation time exceeds the time of an event, the event will be executed. A rule, which is part of the model, creates depending on the position, a type of event and the current assignment of tokens to the position a new assignment of tokens to the position and potential new events. So the connection in the material flow is modeled implicitly in the event interpretation rule.

E.g. the event which occurs, might be “Part leaves machine”. The rule now might decrease the number of tokens in the position by one and create a new “Part enters the machine”-event at the following machine at the same time.

### Multiresolution Modeling

To get an hierarchical multiresolution model each position can consist of a whole simulation model (see Figure 2). So if the position is simulated in more detail, the position will be turned off and the simulation model which is part of the position will be switched on. If the positions of the detailed model also contain more detailed models, this will ultimately lead to an hierarchical modeling approach.

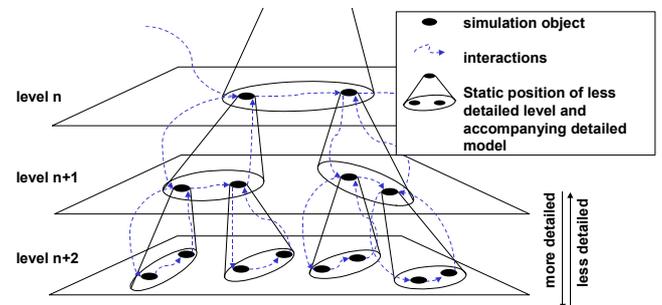


Figure 2: Multi resolution modeling

Events may occur on positions, which are currently not activated. The activation always guarantees that a more detailed model is activated. A transform rule is required

to translate the events into one (or more) events for the more detailed model. If this model is also not activated the rule for this model will direct the event to a further more detailed level.

In our approach the detailed models can view the positions on a rougher level. Hence, they can directly generate events for these positions. A special translation rule is not required for this.

## EXECUTION OF MULTIREOLUTION MODELS

### Activation of the right models

As mentioned earlier, one task is to activate the models which need to be simulated on a different level of detail. To do this, we developed 4 indicators as follows:

- A. Indication by distance: The distance between the user and the object can be used as an indicator. Objects which are far away get low rating.
- B. Indication by the direction of the view: The user can't see objects in his back. So objects with a great angle between the direction of view and the direct connection between the user and the model also get a low rating.
- C. Indication by occlusion: Often the user cannot view the entire scene. Some objects may block the visibility of other objects. Objects the user cannot view see are rated low with this indication. To calculate this indication the 3D-representations of the virtual objects (which might consist of a lot of data) are required. If the user makes small movements the indication can change a lot. This could lead to a lot of aggregation/dissaggregation operations.
- D. Indication by logistic dependencies: If the user takes a close look on one part of the simulation, it is logical that this part of the simulation is logistically connected with other parts of the model. With this indication, the rating of parts which are connected to parts which already have a good rating from other indications, will be increased.

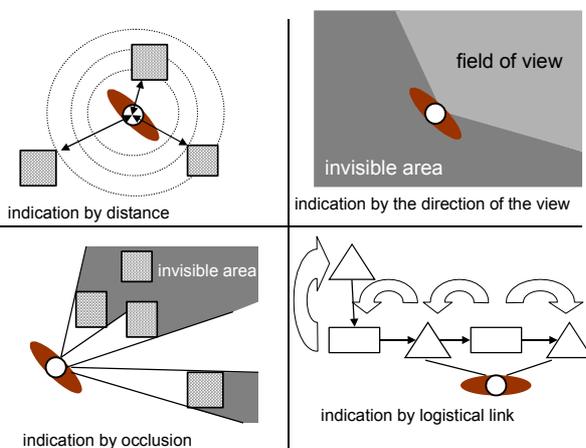


Figure 3: Different indications for the user attention

After calculating the indication of each active position, we decide whether a position has to be simulated in more detail (if a model is available) or a model has to be aggregated (if possible).

### Generating states

When an aggregation or disaggregation operation takes place, a model or a position will be added to the global model. Vice versa a position or a model will be switched off. But the state of the newly added part represents the state when the part was active last (if it was used before). This state might not be the right one for the current simulation time.

A state for the newly activated position or model has to be generated. The first approach is to simulate again all events, which occur on the position/model since the last activation.

If the last activation was done a long time ago, a lot of events have to be simulated afterwards. With the second approach only a part of the past which is sufficient for leveling out of the model will be recalculated. The considered events are selected by the time difference between the actual simulation time and the time assigned to the event. So old events will not be considered. This leads to incorrect states but it can reduce the needed calculation effort enormously.

The re-simulation and the time limited re-simulation are methods for the generation of states which only calculate approximately consistent states. This is due to the fact that potential interaction with other parts of the simulation are not taken into account. If these errors are not acceptable in an application, those methods are not applicable. The models/objects to be activated are calculated without taking the environment into account. Regarding the re-simulation with observation of interactions, the positions to be activated are not calculated without the rest of the simulation. If there are interactions from the model/object to be activated with another position, the model/object is integrated into the calculation of the position. If the position interacts with another position in a different way, this position has to be integrated from this time on, too. Therefore, the number of elements to be simulated increases strictly.

Instead of doing a re-simulation it is also possible to model specific functions, which translates the current state of the position/model, which will be deactivated, into a state for the active position/model. This method requires additional modeling efforts.

### ESTIMATION OF THE REDUCTION OF NEEDED CALCULATION EFFORT

The needed calculation time to execute a model depends on the size of the model. If the user builds a hierarchical

multiresolution model where at least each position of a rougher layer consists out of 2 positions of the more detailed layer, the size of each layer will be the half of the size of the more detailed layer. If the model has got  $l$  layers and  $p$  positions on the most detailed layer, the number of positions of the roughest layer will be less than  $p/2^l$ .

At one point in time only a small part of the simulation should be activated on the most detailed layer. Most of the Simulation be carried out on the roughest layer. Lets assume that the size of the detailed activated small model is  $s_d$ . Then the size of the whole activated model will be less than  $s_d + p/2^l$ . Remember that  $s_d$  should be small. So the activated model is something around  $p/2^l$ . Compared to the model the detailed layer, which has  $p$  positions, the size is decreased by  $2^l$ . If we assume that the needed calculation time depends linear on the size of the model, the execution speed-up will be at least  $2^l$ .

Up to now the calculation didn't take the needed efforts for the calculation of the activation and the generation of states into account. In the following we are showing a rough estimation of the additional needed calculation efforts to give the reader an idea about it.

Only the indication of the activated positions have to be checked. As mentioned before the number of activated positions is much less than the number of the positions on the detailed layer. So the needed calculation efforts are small compared to the efforts of calculating the whole model. The need time to generate the state of a new activated position/model depends on the method, which is used to calculate the state. If the calculation is carried out by methods, which are in the model, the calculation time will depend on these methods. The calculation is a local operation so the calculation time will not depend on the size of the overall model size. If the overall model is huge the calculation time will be small.

## IMPLEMENTATION

### Models represented with XML

Based on the informal description of the chapter "Modeling" we developed a formal description for multiresolution material flow systems. We developed a XML based description language which describes our model. In addition to the simulation model, the description also contains a lot of information for the visualization during the simulation run. Each position has a location and a link to a .x file, which contains a 3D representation (mesh and texture) of the position (e.g. a 3D model of a machine). The code example (Figure 4) gives a brief expression for a position called BL. The first lines describe the localization. Palette.x is the .x file which contains the 3D Model and potential link to a texture file. Delta is the beginning of the rule for processing the events occurring at the position BL.

```
...
<Position Name="BL">
  <PX> -7 </PX>
  <PY> 0</PY>
  <PZ> 0</PZ>
  <F>palette.x</F>
  <Delta>
    <Type>In</Type>
...

```

Figure 4: XML-Sample

More detailed models are directly included in the description. The XML syntax allows to follow the hierarchical multiresolution approach with an recursive notification. E.g. the more detailed model of a position called L consists of two positions called BL and HL. This sub-model can be modeled as a traditional single resolution model and be included in the notification (see Figure 5). If these positions also include more detailed models, an hierarchy will be modeled implicitly.

```
<Position Name="L">
  <PX> 0 </PX>
  <PY> 0 </PY>
  ...
  <Positions>
    <Position Name="BL">
      <PX> -7 </PX>
      <PY> 0</PY>
      ...
    </Position>
    <Position Name="HL">
      <Name>HL</Name>
      <PX> 7 </PX>
      ...
    </Position>
  </Positions>
</Position>

```

Figure 5: Notation of a position with a more detailed model

### Execution of the software

The simulation software d<sup>3</sup>FACT INSIGHT needs an xml description as described above as input. A visual model is generated automatically from the description in the model.

After starting the simulation the user can walk freely through the simulation. The number of tokens assigned to the position is visualized with little red cubes. During the execution of the simulation the assignment is constantly changing. So the user can analyze the simulation within a virtual environment.

For activating more or less detailed models, at the moment, the indication by distance and indication by the direction of the view are used. A mixture of 90% distance and 10% direction leads to good results during tests.

### An Example

Lets assume a production process where parts arrive, put to stock, processed and then dispatched. To model this on this level of detail we need 4 positions. In this example they are called WE (arrival), L (stock), B (processing) and WA (exit). Lets assume the stock (L) and the processing (B) consist of more detailed models. The more detailed model of L should consist of the positions BL and HL and B consists out of ST and P (See Figure 6).

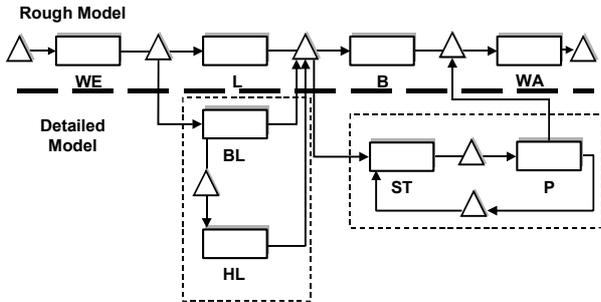


Figure 6: Model of the example process

At the beginning of the simulation run the user is standing far away from the whole model. The rough model is completely activated (see Figure 7).

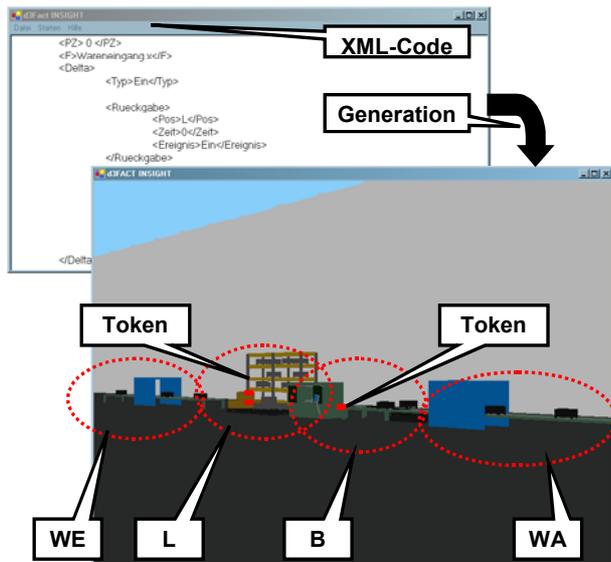


Figure 7: The user is far away. L and B are simulated with a rough model

If the user comes closer, d<sup>3</sup>FACT INSIGHT activates the more detailed models. States for the newly activated positions are generated and the 3d-representives of the more detailed positions are also activated. Figure 8 shows the situation after the activation of the most detailed level.

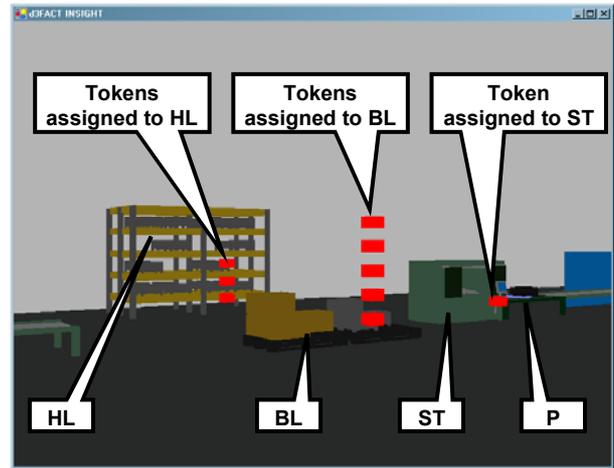


Figure 8: The User is close. The Simulation model works in high resolution.

## FURTHER WORK

In the near future, the missing indications and state generation methods will be implemented. Results about the quality of the generated states are also missing at the moment.

As far as the speed up of the simulation calculation is concerned, some theoretical analysis already exists. But they need some assumptions. Most importantly, measurement with realistic models is also still needed.

## CONCLUSION

Large simulations with a high resolution model cannot be analyzed interactively. But the user also cannot take care of everything at once. He only views small parts of the simulation. In our approach, only the parts the user is viewing are simulated in high resolution. Everything else is simulated with an rough model. If the user changes his interests to other areas, the high resolution area will follow him dynamically. This leads to an experience of a simulation, which is simulated completely in high resolution without the needed calculation efforts. As a compromise the calculated results are not as accurate as a complete detailed simulation.

For this we developed a modeling technique, which allows the modeler to set-up models, which can work in different levels of resolution. Rough models are not generated automatically, but the modeler has to build them. This paper demonstrates methods which can be used to analyze large models interactively in virtual environments.

## REFERENCES

- Dangelmaier, W. und Mueck, B., 2003: Simulation in Business Administration and Management - Simulation of production processes, In: Obaidat, M. S. und Papadimitriou, G. I. (eds.): *Applied System Simulation*, p. 381-396, Kluwer Academic Publishers, Norwell, MA

- Davis, P. K. and Bigelow, J. H., 1998: Experiments in Multiresolution Modeling (MRM). RAND MR-1004-OSD,
- Kim, T. G.; Lee, C.; Christensen, E. R. und Zeigler, B. P., 1992: System Entity Structuring and Model Base Management, In: Davis, P. K. and Hillestad (eds.): *Proceedings of Conference on Variable-Resolution Modeling*, p. 96-101, RAND CF-103-DARPA, Santa Monica, CA
- Reynolds, P. F. Jr. and Natrajan, A., 1997: Consistency Maintenance in Multiresolution Simulations. In: *ACM Transactions on Modeling and Computer Simulation*, 7 p. 368-392, Nr. 3, July. 1997

## AUTHOR BIOGRAPHIES



**Wilhelm Dangelmaier** was director and head of the Department for Cooperative Planning and Control at the Fraunhofer-Institute for Manufacturing. In 1990 he became Professor for Facility Planning and Production Scheduling at the University of Stuttgart. In 1991, Dr. Dangelmaier became Professor for Business Computing at the Heinz Nixdorf Institute; University of Paderborn, Germany. In the year 1996, Prof. Dangelmaier founded the Fraunhofer-Application Center for Logistics Oriented Production. His e-mail address is [whd@hni.uni-paderborn.de](mailto:whd@hni.uni-paderborn.de). The web address of his working group is: [www.hni.upb.de/cim](http://www.hni.upb.de/cim)



**Bengt Mueck** studied computer science at the University of Paderborn, Germany. Since 1999 he is a research assistant at the group of Prof. Dangelmaier, Business Computing, esp. CIM at the Heinz Nixdorf Institute of the University of Paderborn. His main research interests are logistics systems and tools to simulate those systems in different levels of resolution. His e-mail address is [mueck@hni.uni-paderborn.de](mailto:mueck@hni.uni-paderborn.de).



**Christoph Laroque** studied business computing at the University of Paderborn. Since 2003 he is PhD student in the graduate school of dynamic intelligent systems and research assistant at the group of Prof. Dangelmaier, Business Computing, esp. CIM. He is mainly interested in material flow simulation and the “digital factory”. His e-mail address is [laro@hni.uni-paderborn.de](mailto:laro@hni.uni-paderborn.de).



**Kiran Mahajan** studied Mechanical Engineering (Specialization Production Engineering and Organization) at the Delft University of Technology. Since 2004 he is PhD student at the graduate school of dynamic intelligent systems and research assistant at the group of Prof. Dangelmaier, Business Computing, esp. CIM. His research interests include optimization of manufacturing systems using innovative simulation technologies. His e-mail address is [kiran@hni.uni-paderborn.de](mailto:kiran@hni.uni-paderborn.de).

# EVALUATION OF BATCHING STRATEGIES IN A MULTI-PRODUCT WAFERFAB BY DISCRETE-EVENT SIMULATION

Ilka Habenicht, Lars Mönch  
Institute of Information Systems  
Technical University Ilmenau  
Helmholtzplatz 3, D-98694 Ilmenau, Germany  
E-mail: {ilka.habenicht|lars.moench}@tu-ilmenau.de

## KEYWORDS

Semiconductor Manufacturing, Batching, Discrete-Event Simulation, Performance Assessment

## ABSTRACT

In this paper, we present the results of a simulation study for the performance evaluation of certain batching strategies in a multi-product semiconductor wafer fabrication facility (waferfab). Batching in waferfabs means that we can process different lots at the same time on the same machine. As opposite to common scheduling and dispatching decisions in manufacturing beside assignment and sequencing decisions we have to make decisions on the content of a batch. Batching decisions have a large impact on the performance of a waferfab because of very long processing times. In this simulation study, we extend previous work on the evaluation of batching strategies from the two product-case to the case of a larger number of products and different product mix scenarios.

## INTRODUCTION

In semiconductor manufacturing, integrated circuits are produced on silicon wafers. This type of manufacturing is very capital intensive. Lots are the moving entities in a waferfab. Each lot contains a fixed number of wafers.

The process conditions are very complex (Uzsoy et al. 1094, Atherton and Atherton 1995, Schönig and Fowler 2000). We have to deal with parallel machines, different types of processes like batch processes and single wafer processes, sequence dependent setup times, prescribed customer due dates for the lots, and reentrant process flows. Very often, we also have to face with an over time changing product mix including a large number of different product.

Batch machines can process several lots at the same time. However lots of different families cannot be processed together due to the chemical nature of the process. Lots that can be processed together form one family. The processing times of batch operations are usually very long compared to other processes. Therefore batching decisions may effect the performance of the entire waferfab. Especially in the case of a multi-product

waferfab, the dynamic of the waferfab is influenced by the treatment of batches.

Depending on customer requirements lots of different products have to meet different internal and external due dates. Furthermore, based on customer importance some lots can have a higher weight (priority) than other.

Scheduling and dispatching of batching machines is challenging because beside the common assignment and sequencing decisions batch forming decisions are necessary. Due to unequal ready times of the lots on a certain batch machine (or group of batch machines working in parallel) it is sometimes more favorable to form a non-full batch, in other situation it is better to wait for next lot arrivals in order to increase the fullness of a batch.

Batching issues are intensively discussed in the scheduling and industrial engineering literature. We refer to (Mönch and Habenicht 2003) where some related literature is discussed mainly from a deterministic scheduling point of view. Look-ahead strategies for batching are surveyed by Van der Zee 2003.

The authors study the performance of different dispatching and scheduling heuristics for batching tools with respect to minimize due-date oriented performance measures in (Mönch and Habenicht 2003). This work considers only the rather limited two product-case. However, as pointed out for example by Akçali et al. 2000, the performance of batching strategies can be different in multi-product environments. Therefore, we extend our previous work by performing a simulation study for multi-product waferfabs under different product mix situations.

The paper is organized as follows. In the next section, we summarize two batching heuristics from (Mönch and Habenicht 2003) that are used in this study. Then, we describe the simulation model and our experimental design. We present and discuss the results from simulation experiments in the last section of the paper.

## BATCHING ALGORITHMS

In this section, we summarize two batching heuristics from (Mönch and Habenicht 2003) that are relevant for this study. We are interested in two questions. We want to analyze how much we can gain from considering future lot arrivals in a multi-product setting. The second research question is the following one. How does the

number of lot families influences the performance of our batching strategies?

### Notations

We consider one fixed batch tool group. Making a batching decision, we have to decide whether we form a batch only from the set of lots waiting in front of the tool group for processing or to wait for future lot arrivals which means leaving a certain tool idle for a certain period of time. Figure 1 illustrates this issue for a tool group with incompatible families

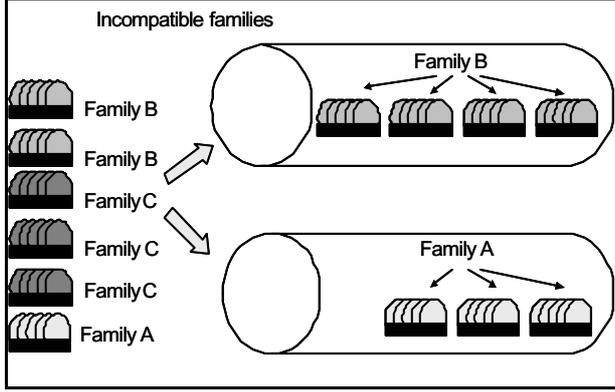


Figure 1: Batching issue with incompatible families

The following notation is used throughout the rest of the paper.

1. Lots belonging to different incompatible families cannot be processed together. There exist  $F$  such families.
2. There are  $n$  lots that have to be scheduled.
3. The fixed tool group contains  $m$  identical machines in parallel. The maximum batch size of the tool group is denoted by  $B$ .
4. There are  $n_j$  lots of family  $j$  to be used for forming and sequencing of batches:

$$\sum_{j=1}^F n_j = n.$$

5. Lot  $i$  of family  $j$  is represented as  $ij$ .
6. The priority weight for lot  $i$  belonging to family  $j$  is represented as  $w_{ij}$ .
7. The ready time of lot  $i$  in family  $j$  is denoted by  $r_{ij}$ .
8. The due date (with respect to the batching tool group) of lot  $i$  of family  $j$  is represented as  $d_{ij}$ .
9. The completion time of lot  $i$  of family  $j$  on the batching tool group is represented as  $c_{ij}$ .
10. The processing time of lots in family  $j$  is denoted by  $p_j$ .

We use the notation  $x^+ = \max(0, x)$  for abbreviation. We frequently refer to the total weighted tardiness as performance measure (with respect to a fixed tool

group) for a given schedule. This measure is defined as follows:

$$TWT := \sum w_{ij} (c_{ij} - d_{ij})^+. \quad (1)$$

### Static Batch Dispatching Heuristic (SBDH)

The first heuristic is a modification of the well-known Apparent Tardiness Cost Dispatching Rule (Vepsäläinen and Morton 1987). In the work of Mason et al. 2002 this rule was adapted to the scheduling of batch machines. We calculate the static batch ATC index  $I_{ij,stat}(t)$  for job  $i$  belonging to family  $j$  at time  $t$

$$I_{ij,stat} = \frac{w_{ij}}{p_j} \exp \left( - \frac{(d_{ij} - p_j - t)^+}{k\bar{p}} \right). \quad (2)$$

The parameter  $k$  is a scaling parameter and  $\bar{p}$  represents the average processing time of the remaining jobs. We sequence the lots of one family waiting in front of the batch tool group in descending order with respect to their  $I_{ij,stat}(t)$  index. We take the first  $B$  lots of this sequence in order to form the batch that has to be processed next for this family. We choose the batch with highest sum of the  $I_{ij,stat}(t)$  indices of the lots of the batch among the families. This batch will be processed next.

For the purpose of finding an optimal  $k$  parameter with respect to the total weighted tardiness of the lot waiting for processing, we repeat the calculation for different values of  $k$  and choose for implementation the schedule that leads to the smallest TWT value.

This heuristic is a full-size batch strategy and does not take any future lot arrivals into account.

### Dynamic Batch Dispatching Heuristic (DBDH)

The second heuristic takes future lot arrivals into account. Therefore, we define a time window  $(t, t + \Delta t)$ . Usually, we chose a fixed portion of the average processing time of the waiting lot as  $\Delta t$ . The set of lots from family  $j$  that are ready for processing at time  $t$  or will arrive inside the given window is denoted by

$$M(j, t, \Delta t) := \{ij / r_{ij} \leq t + \Delta t\}. \quad (3)$$

The elements of  $M(i, t, \Delta t)$  are sorted in descending order with respect to the index

$$I_{ij,dyn}(t) = \frac{w_{ij}}{p_j} \exp \left( - \frac{(d_{ij} - p_j - t + (r_{ij} - t)^+)^+}{k\bar{p}} \right). \quad (4)$$

In the next steps, we only consider the first  $\#lots$  lots of the sorted set  $M(i, t, \Delta t)$ . Here,  $\#lots$  is a fixed number that is a parameter of the heuristic. We build all batch combination using these lots. We calculate the batch index

$$I_{bj}(t) = \frac{w_{bj}}{p_j} \exp\left(-\frac{(sl+rt)^+}{k\bar{p}}\right) \min\left(\frac{n_{bj}}{B}, 1\right) \quad (5)$$

for each formed batch. Therefore we denote by  $d_{bj} := \min_{i \in B_{bj}}(d_{ij})$ : minimum due date among all jobs belonging to the batch,

longing to the batch,

$r_{bj} := \max_{i \in B_j}(r_{ij})$ : maximum ready time of the jobs assigned to the batch,

signed to the batch,

$w_{bj}$ : average weight of the lots contained in the batch,

$n_{bj}$ : number of lots in the batch.

For abbreviation, we use  $sl := d_{bj} - p_j - t$  and  $rt := (r_{bj} - t)^+$ .

This strategy does not necessarily form full batches. Sometimes, it is more profitable to wait for an important lot instead of processing a batch with unimportant lots. From the previous study (Mönch and Habenicht 2003), it is known that DBDH is sensitive to the size of the time window and to the parameter  $\#lots$ .

## EXPERIMENTAL DESIGN

### Framework for Experimentation

We use a discrete-event simulation tool and a simulation model of a waferfab to evaluate SBDH and DBDH. Our basic architecture is described by (Mönch et al. 2002). The center point is a data storage (called data model) which contains all information required to run the dispatching and scheduling algorithms. We extend the data model by additional classes and attributes to adapt it to the two algorithms. The data model connects the manufacturing process emulated by the simulation tool and the proposed production control schemes.

We use the MIMAC test data set 1 (Fowler and Robinson 1995) in a modified version. The original model consists of two different product flows (A, B) with about 200 process steps and more than 80 tool groups. We create new product flows based on product flow A and B to build a multi-product environment.

The simulation model contains 16 batching tool groups. Tool group OXIDE\_1 is bottleneck of the waferfab. In Table 1, information of this batching tool group are provided.

Table 1: Bottleneck Batching Tool Group Information

Tool Group	# tools	$B_{\min}$	$B_{\max}$	$p_{\min}$	$p_{\max}$	Utilization [%]
OXIDE_1	3	2	6	135	1410	84.19

In Table 1, we denote by  $B_{\min}$  the minimum batch size and by  $B_{\max}$  the maximum batch size given in lots. The minimum processing time (measured in minutes) is represented by  $p_{\min}$  and the corresponding maximum processing time by  $p_{\max}$ . The given utilization of the tool group was determined by simulation experiments with the First In First Out (FIFO) dispatching rule. We decided to apply the SBDH and DBDH rule to this tool group.

For the remaining tool groups, we used a slack-based dispatching rule (SLACK). For the calculation of the slack of the lots waiting in front of a certain tool group we calculate a schedule by simply multiplying the processing time of the steps with a dynamic flow factor. For that purpose, we calculate the remaining time of the lot with respect to the due date. Based on this information, we assign a flow factor to each lot (cf. Habenicht and Mönch 2002 for more details). This scheduling method allows us to determine the end dates of the single process steps, in particular future lot arrival information. The end dates serves as internal due dates. We repeat the calculation of the flow factors every 24 hours.

In our experiments we use a moderate workload of the system. Machine failures are exponentially distributed. The model is initialized by using a work in process distribution of the waferfab. The length of a simulation run was 100 days. We take five independent replications of each simulation run in order to obtain statistically significant results.

### Performance Measures

The following performance measures are used:

- Total weighted tardiness (with respect to the entire waferfab) of the lots that are released and finished within the planning horizon under consideration. We define the weighted tardiness of a lot  $i$  as follows:

$$T_i := w_i (C_i^r - d_i)^+, \quad (6)$$

where  $C_i^r$  represents the realized completion time,  $d_i$  the due date and  $w_i$  the weight of the lot  $i$ . In order to calculate the performance measure we sum the  $T_i$  for all lots. We denote this quantity by  $TWT_{total}$ .

- Average cycle time:  $CT$ .
- Throughput of the waferfab (number of completed wafers):  $TP$ .

### Design of Experiments

In (Mönch and Habenicht 2003) we studied the behavior of the batching heuristics under different system condition. We identified different parameters which influence the performance of the batching heuristics. We distinguish two groups of parameters. The first group of parameter characterizes the manufacturing systems:

- number of incompatible families,
- due date settings,

- weight settings.

Parameters that are used for settings in the heuristic, especially the DBDH rule, belong to the second group:

- length of the time window,
- maximum number of lots used for considering all batch combinations,
- setting of the parameter  $k$ .

In (Habenicht and Mönch 2003), we limited the number of families by considering only two products. In this paper, we extend these investigations by considering more products. In our experiments, we vary the time window settings as exclusive parameter of the second group. We fix the other parameters of the heuristics by investigating only the case of optimal  $k$  value setting as described before. The maximum number of lots  $\#lots$  used for considering all batch combinations is ten. The due date is chosen by using a fixed flow factor of two. We consider the case of two, eight, and sixteen different products. An incompatible family is formed by all lots of the a product.

The used factorial design is summarized in Table 2.

Table 2: Factorial Design for this Study

Factor	Level		
	I	II	III
Number of Products	2	8	16
Factor	Level		
	I	II	
Relation of Product Appearance	1:1	2:1	
Weight	With probability: $p_1 = 0.7$ $w_j = 1$ with probability: $p_1 = 0.25$ $w_j = 3$ with probability: $p_1 = 0.05$ $w_j = 10$	With probability: $p_1 = 0.5$ $w_j = 1$ with probability: $p_1 = 0.3$ $w_j = 3$ with probability: $p_1 = 0.2$ $w_j = 10$	
Time Window Size	25% of the average processing time of the lots queuing in front of the batching tools	50% of the average processing time of the lots queuing in front of the batching tools	

For the case of eight products, we copy product flow A and B four times and for sixteen products eight times respectively. Each product flow, created in this way, represents a certain product.

Using DHBH, we derive a new schedule for the tool group every time a lot arrives in front of the tool group. If a batch formed by the time window approach is not full, then we try to increase the fullness of the batch by choosing lots among the waiting lots, but eventually unimportant lots of the same family.

## COMPUTATIONAL RESULTS

In this section, we present the results of simulation experiments with the suggested heuristics. The resulting performance measures are presented in terms of the ratio of the performance measure value of the heuristic and performance measure value obtained by using the SLACK dispatching rule.

Because we do not take future lot arrivals into account for SBDH, we have to consider only the first three factors from Table 2. We use the tuple (dispatching rule, level from factor 1, level from factor 2, level from factor 3) in order to describe the used factor combination.

Table 3 shows the results for the SBDH-based batching strategy for the 2 product-case, Table 4 for the 8 product-case, and Table 5 for the 16 product-case.

Table 3: Results for SBDH for Two Products

Factor Combination	$TWT_{total}$	$CT$	$TP$
SLACK (I-I-I)	1.0000	1.0000	1.0000
SBDH (I-I-I)	0.8409	1.0000	1.0032
SLACK (I-I-II)	1.0000	1.0000	1.0000
SBDH (I-I-II)	0.6312	1.0037	0.9988
SLACK (I-II-I)	1.0000	1.0000	1.0000
SBDH (I-II-I)	1.2190	0.9997	1.0002
SLACK (I-II-II)	1.0000	1.0000	1.0000
SBDH (I-II-II)	1.1052	0.9977	1.0029

Table 4: Results for SBDH for Eight Products

Factor Combination	$TWT_{total}$	$CT$	$TP$
SLACK (II-I-I)	1.0000	1.0000	1.0000
SBDH (II-I-I)	0.9885	1.0034	1.0063
SLACK (II-I-II)	1.0000	1.0000	1.0000
SBDH (II-I-II)	0.9850	0.9995	1.0033
SLACK (II-II-I)	1.0000	1.0000	1.0000
SBDH (II-II-I)	1.0378	1.0077	1.0044
SLACK (II-II-II)	1.0000	1.0000	1.0000
SBDH (II-II-II)	1.1227	1.0194	1.0042

Table 5: Results for SBDH for Sixteen Products

Factor Combination	$TWT_{total}$	$CT$	$TP$
SLACK (III-I-I)	1.0000	1.0000	1.0000
SBDH (III-I-I)	0.9895	1.0106	1.0000
SLACK (III-I-II)	1.0000	1.0000	1.0000
SBDH (III-I-II)	1.0385	0.9953	1.0143
SLACK (III-II-I)	1.0000	1.0000	1.0000
SBDH (III-II-I)	1.0113	0.9968	0.9966
SLACK (III-II-II)	1.0000	1.0000	1.0000
SBDH (III-II-II)	0.9769	1.0032	1.0062

From the experiments with SBDH, we can verify that the batching heuristic outperforms the slack rule only for the case of a homogenous product mix. In the inhomogeneous case, only a small number of lots of the sec-

ond product exists. The SBDH-based batching strategy is a full-batch strategy, which leads to longer waiting times for lots belonging to families with less lots.

There is an even smaller improvement in the 8 product-case and the 16 product-case which confirms this thesis because the number of lots for a single product is even more smaller compared to the 2 product-case.

For the DBDH-based batching strategy the fourth experimental factor (time window size) is also important. The results for DBDH are shown in Table 6 and 7

Table 6: Results for DBDH for Two Products

Factor Combination	$TWT_{total}$	$CT$	$TP$
SLACK (I-I-I)	1.0000	1.0000	1.0000
DBDH (I-I-I-I)	1.1061	1.0033	0.9965
DBDH (I-I-I-II)	1.2252	1.0082	0.9974
SLACK (I-I-II)	1.0000	1.0000	1.0000
DBDH (I-I-II-I)	0.6994	1.0027	1.0002
DBDH (I-I-II-II)	0.7386	1.0039	0.9951
SLACK (I-II-I)	1.0000	1.0000	1.0000
DBDH (I-II-I-I)	0.8551	0.9920	1.0116
DBDH (I-II-I-II)	0.8582	0.9947	1.0004
SLACK (I-II-II)	1.0000	1.0000	1.0000
DBDH (I-II-II-I)	0.6748	0.9917	0.9994
DBDH (I-II-II-II)	0.5956	0.9869	1.0006

Table 7: Results for DBDH for Eight Products

Factor Combination	$TWT_{total}$	$CT$	$TP$
SLACK (II-I-I)	1.0000	1.0000	1.0000
DBDH (II-I-I-I)	0.9621	1.0091	0.9789
DBDH (II-I-I-II)	0.9307	1.0065	0.9753
SLACK (II-I-II)	1.0000	1.0000	1.0000
DBDH (II-I-II-I)	0.9052	0.9952	0.9957
DBDH (II-I-II-II)	0.8858	0.9839	0.9856
SLACK (II-II-I)	1.0000	1.0000	1.0000
DBDH (II-II-I-I)	1.1447	1.0266	0.9982
DBDH (II-II-I-II)	1.2222	1.0412	0.9991
SLACK (II-II-II)	1.0000	1.0000	1.0000
DBDH (II-II-II-I)	1.1568	1.0236	1.0029
DBDH (II-II-II-II)	1.2261	1.0425	0.9914

Both batching strategies, SBDH and DBDH, are sensitive to product mix and weight settings. It is interesting to see that in the 2 product-case the DBDH-based strategy outperforms the other batching strategies only for inhomogeneous product mixes. Considering future lot arrivals allows to decide whether it is advantageous to wait for the next incoming lot of a family with smaller number of lots or to start a non-full batch.

In the 8 product-case, the results are not the same as expected from the 2 product-case. The number of lots for a single product becomes so small for the inhomogeneous product mix that the waiting times for filling a batch are huge. Especially the minimum batch size ( $B_{min} = 2$ ) enforces this effect.

This becomes more clear when looking at the utilization data of the batching tool group. In Table 8, batch utilization (average number of lots that form a batch), tool group utilization and average queue size of the factor combinations II-I-I and II-II-I are shown.

Table 8: Utilization of the batching tool group

OXIDE_1			
Factor Combination	Batch Utilization	Utilization	Average Queue Size
SLACK (II-I-I)	1.0000	1.0000	1.0000
SBDH (II-I-I-)	1.0125	0.9931	0.9652
DBDH (II-I-I-I)	1.3451	0.6125	3.9745
DBDH (II-I-I-II)	1.2982	0.6153	4.5564
SLACK (II-II-I)	1.0000	1.0000	1.0000
SBDH (II-II-I)	0.9994	1.0036	0.9736
DBDH (II-II-I-I)	1.0070	0.8855	1.1583
DBDH (II-II-I-II)	1.0223	0.8262	1.3493

In the case II-I-I, considering future lot arrivals leads to a larger batch utilization and a larger queue size. The tools wait for lots which will arrive during a given time window. In the case II-II-I, the increment of those measures is less. This is caused by the effect that the starting non-full batches is a preferred decision.

The influence of the considered time horizon is pointed out in the results of the 16 product-case shown in Table 9. Especially in situations with a small number of lots per product, the arrival frequency of lots of the same product is very small. Therefore, it is reasonable to enlarge the time horizon for considering future lot arrivals in order to improve batching decisions.

Table 9: Results for DBDH for Sixteen Products

Factor Combination	$TWT_{total}$	$CT$	$TP$
SLACK (III-I-I)	1.0000	1.0000	1.0000
DBDH (III-I-I-I)	1.0079	1.0255	0.9860
DBDH (III-I-I-II)	0.9797	1.0207	0.9689
SLACK (III-I-II)	1.0000	1.0000	1.0000
DBDH (III-I-II-I)	1.0504	1.0186	0.9864
DBDH (III-I-II-II)	1.0196	1.0175	0.9726
SLACK (III-II-I)	1.0000	1.0000	1.0000
DBDH (III-II-I-I)	0.8407	0.9624	0.9342
DBDH (III-II-I-II)	0.8391	0.9566	0.9244
SLACK (III-II-II)	1.0000	1.0000	1.0000
DBDH (III-II-II-I)	0.7843	0.9141	0.9708
DBDH (III-II-II-II)	0.7619	0.8978	0.9589

## SUMMARY

In this paper we evaluated two strategies for batching in a waferfab. We studied the influence of the number of products on the performance of two strategies that we suggested in (Mönch and Habenicht 2003).

The first strategy does not take any future lot arrivals into account. In contrast, we defined a certain time window in which future lot arrivals are considered for the

second strategy. We presented results for a different number of products and different product mixes. The results show that the number of lots in one family is a very important factor for the performance of the strategies. Hence, it is useful to assess the performance of batching strategies in case of product mix changes. The performance of the two heuristics can be improved by determining more meaningful internal due dates and future lot arrival estimates. A finite capacity scheduling algorithm working on an aggregated model (cf Habenicht and Mönch 2002) may lead to more accurate lot arrival information and hence to a better batch decision-making.

## REFERENCES

- Akçali, E., Uzsoy, R., Hiscock, D. G., Moser, A. L., and Teyner, T. J. 2000. Alternative Loading and Dispatching Policies for Furnace Operations in Semiconductor Manufacturing: A Comparison by Simulation. In *Proceedings of the 2000 Winter Simulation Conference*, ed. J. A. Joines, R.R. Barton, K. Kang, and P. A. Fishwick, 1428-1435.
- Atherton, L. F. and R. W. Atherton. 1995. *Wafer Fabrication: Factory Performance and Analysis*. Kluwer Academic Publishers, Boston, Dordrecht, London.
- Fowler, J. W. and J. Robinson. 1995. *Measurement and Improvement of Manufacturing Capacities (MIMAC): Final Report*. Technical Report 95062861A-TR, SEMATECH, Austin, TX.
- Habenicht, I. and L. Mönch. 2002. A Finite-Capacity Beam-Search Algorithm for Production Scheduling in Semiconductor Manufacturing. In *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 1406-1413.
- Mason, S., J. W. Fowler, and W. M. Carlyle. 2002. A Modified Shifting Bottleneck Heuristic for Minimizing Total Weighted Tardiness in Complex Job Shops, *Journal of Scheduling*, 5: 247-262.
- Mönch, L., and I. Habenicht. 2003. Simulation-Based Assessment of Batching Heuristics in Semiconductor Manufacturing. In *Proceedings of the 2003 Winter Simulation Conference*, ed. S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 1338-1345.
- Mönch, L., O. Rose, and R. Sturm. 2002. Framework for Performance Assessment of Shop-Floor Control Systems. In *Proceedings of the 2002 Modeling and Analysis of Semiconductor Manufacturing Conference (MASM 2002)*, ed. J. W. Fowler, J. K. Cochran, 95-100.
- Schömig, A. and J. W. Fowler. 2000. Modelling Semiconductor Manufacturing Operations. In *Proceedings of the 9th ASIM Dedicated Conference Simulation in Production and Logistics*, ed. K. Mertins and M. Rabe, 55-64.
- Uzsoy, R., C.-Y. Lee, and L. A. Martin-Vega. 1992. A Review of Production Planning and Scheduling Models in the Semiconductor Industry, Part I: Sys-

tem Characteristics, Performance Evaluation and Production Planning. *IIE Transactions on Scheduling and Logistics*, 24: 47-61.

- Van der Zee, D.-J. Look-Ahead Strategies for Controlling Batch Operations in Industry – an Overview. *Proceedings of the 2003 Winter Simulation Conference*, ed. S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 1480-1487.
- Vepsäläinen, A. and T. E. Morton. 1987. Priority Rules and Lead Time Estimate for Job Shop Scheduling with Weighted Tardiness Costs. *Management Science*, 33: 1036-1047.

## AUTHOR BIOGRAPHIES

**ILKA HABENICHT** is a Ph.D. student in the Department of Information Systems at the Technical University of Ilmenau, Germany. She received a master's degree in business related engineering from the Technical University of Ilmenau, Germany. Her research interests are in production control of semiconductor wafer fabrication facilities and simulation. Her email address is <Ilka.Habenicht@tu-ilmenau.de>.

**LARS MÖNCH** is an Assistant Professor in the Department of Information Systems at the Technical University of Ilmenau, Germany. He received a master's degree in applied mathematics and a Ph.D. in the same subject from the University of Göttingen, Germany. After receiving his Ph.D. he worked for two years for Softlab GmbH in Munich in the area of software development. His current research interests are in simulation-based production control of semiconductor wafer fabrication facilities, applied optimization and artificial intelligence applications in manufacturing. He is a member of GI (German Chapter of the ACM), GOR (German Operations Research Society), SCS and INFORMS. His email address is <Lars.Moench@tu-ilmenau.de>.

# A NEW METHOD OF FMS SCHEDULING USING OPTIMIZATION AND SIMULATION

Ezedeem Kodeekha  
Department of Production, Informatics, Management and Control  
Faculty of Mechanical Engineering  
Budapest University of Technology and Economics  
E-mail: [ezo12@yahoo.com](mailto:ezo12@yahoo.com)

**KEYWORDS:** FMS scheduling, CIM, Conventional scheduling methods, Break and Build method.

## ABSTRACT

Nowadays, in modern manufacturing the trend is the development of Computer Integrated Manufacturing, CIM technologies which is a computerized integration of manufacturing activities (Design, Planning, Scheduling and Control) that to produce right products right at right time to react quickly to the global competitive market demands. The productivity of CIM is highly depending upon the scheduling of Flexible Manufacturing System (FMS). Shorting the makespan leads to decreasing machines idle time which results improvement CIM productivity. Conventional methods of solving scheduling problems such as heuristic methods based on priority rules still result schedules, sometimes, with significant idle times. To reduce these, the present paper proposes a new high quality scheduling method. This method uses multi-objective optimization and simulation. The method is called “**Break and Build Method**”, **BBM**. The BBM procedure has three stages, in the first **Building stage**; the steps are to build up some schedules using any scheduling methods for example: heuristic ones which are tested by simulation. In the second **Breaking stage**, optimum sizes of batches are determined. In the final **Rebuilding stage**, the most proper schedule is selected using simulation. The goal of use of simulation within manufacturing scheduling is to achieve the two following objectives: first is the visual representation of manufacturing process behavior of a chosen schedule. The second is testing and validation of schedules to select the most proper schedule what can be successfully implemented. There are two-objectives achieved by BBM to the given simple example, one is improved productivity by 31.92% and the other is meeting delivery dates.

The method produces a new direction of manufacturing scheduling using differential calculus, gives a new results and new information for solving simple manufacturing scheduling problem.

## INTRODUCTION

Flexible Manufacturing System (FMS) is an automated manufacturing system which consists of group of automated machine tools, interconnected with an automated material handling and storage system and controlled by computer to produce products according to the right schedule.

Manufacturing scheduling theory is concerned with the right allocation of machines to operations over time.

FMS scheduling is an activity to select the right future operational program and/or diagram of an actual time plan for allocating competitive different demands of different products, delivery dates, and/or sequencing through different machines, operations, and routings that for combination the high flexibility of job shop type with high productivity of flow-shop type and meeting delivery dates.

FMS Scheduling system is one of the most important information-processing subsystems of CIM system. The productivity of CIM is highly depending upon the quality of FMS scheduling. The basic work of scheduler is to design an optimal FMS schedule according to a certain measure of performance, or scheduling criterion. This paper focuses on productivity oriented-makespan criteria. Makespan is the time length from the starting of the first operation of the first demand to the finishing of the last operation of the last demand.

Conventional methods of solving scheduling problems such as heuristic methods based on priority rules (FIFO, SPT, SLACK...) determined the corresponding schedule but usually, still having idle times. To reduce these and improving CIM productivity, this paper presents a new method so called “**Break and Build Method**”, **BBM**. The paper can be classified into forth parts as follow:-First Part: Scheduling using BBM. Second Part: Application of BBM to the simple scheduling problems. Third Part: Conclusion, and References.

## SCHEDULING USING BBM

BBM is a multi-criteria optimization and simulation approach in which the optimum schedule of tasks of High Number of Parts (HNP) are divided into optimum sub-series (batches), then rebuild the schedule again and overlapping production can be realized at certain

condition and tested using one of simulation methods (e.g.: Taylor ED). BBM has two-objectives for this situation, one is a higher productivity and the second is meeting delivery dates.

### BBM Procedure

The BBM procedure is consists of the following three stages:-

#### 1. Building Stage

In the building stage, the steps are to built up an optimum schedule using any scheduling methods such as heuristic method and tested by simulation

### Scheduling Problem

The shop considered in this paper consist of 2-different independent machines  $M_1, M_2$  of load,  $L_1, L_2$  respectively will process 2 demands,  $d_1, d_2$  of units,  $X_1, X_2$ . Each demand processed by 2 operations  $O_1, O_2$  each operation consists of run time  $t$  and set up time  $\delta$  with precedence relationship  $O_1$  precedes  $O_2$  and the processing times are  $P_1, P_2$  respectively, The due date of  $d_1$  and  $d_2$  is  $D$ . Data is summarized at demand table in fig.1. The Objective is to determine the best schedule using productivity criteria.

Table (1) Demand Table.

d	$O_1$	$O_2$	P
$d_1$	$O_1^{11}$	$O_1^{22}$	$P_1$
$d_2$	$O_2^{11}$	$O_2^{22}$	$P_2$
<b>L</b>	<b><math>L_1</math></b>	<b><math>L_2</math></b>	<b><math>S_i</math></b>

### Notations

$O_i^{om}$ : O (Operation time), o (operation number),  
m (machine number), i (demand number),  $t_i^{om}$ : run time,  
r: ready time, s: start time, f: flow time,  $S_i$ : Schedule time,  $\Pi_s$ : Number of schedules, T: Makespan  
 $L_{max}$ : bottleneck machine load,  $\eta$ : Schedule Productivity Index,  $\eta_R$ : Schedule Productivity Rate

### Assumptions

1. No Cancellation. No Breakdown. No Preemption.

2. Operating cost is constant.

3.  $\delta$  is constant,  $r = 0$

Demand chart as in fig. (1), shows how much time required to processing each demand  $P_1, P_2$ , . Load chart as in fig. (2) Shows how much time to be loading each machine  $L_1, L_2$  required to produce the two demands.

### Solution

As in fig. (3), Gantt chart clearly display that the schedule is satisfied according to the precedence relationship but it is infeasible schedule due to the conflict of overload.

### Heuristic Scheduling Methods

A heuristic is a rule of thumb procedure that determines a “good-enough”, satisfactory and feasible solution within certain constraints, but not necessarily guarantees the best or optimal, solution to a problem. A good heuristic is generally within 10% of optimality, the amount of error is not known and degree of optimality is not known. Heuristic methods based on priority rules for job-shop scheduling problem are not a convenience but a necessity for selecting which job is started first on certain machine. Some of the rules used to scheduling problems are FIFO (First In First Out), SPT (Shortest Processing Time) and SLACK.... rules. in this paper the number of schedules to be evaluated is  $\Pi_s = n! = 2$  schedule, where, n: number of demands = 2. The priority rules used in the present paper are FIFO and SPT as following:-

#### a) SPT rule

Table (2) SPT Table

$M_1$			$M_2$		
s	O	f	s	O	f
0	$O_2^{11}$	$O_2^{11}$	$O_2^{11}$	$O_2^{22}$	$f_1$
$O_2^{11}$	$O_1^{11}$	$L_1$	$L_1$	$t_1^{22}$	<b><math>T_1</math></b>

#### b) FIFO rule

Table (3) FIFO Table

$M_1$			$M_2$		
s	O	f	s	O	f
0	$O_1^{11}$	$O_1^{11}$	$O_1^{11}$	$O_1^{22}$	$f_2$
$O_1^{11}$	$O_2^{11}$	$L_1$	$L_1$	$t_2^{22}$	<b><math>T_2</math></b>

### Mathematical Model

The mathematical model for the formulated problem is Objective Function: Minimize

$$T = t_1^{11} + t_1^{22} + t_2^{11} + t_2^{22} + 4\delta \dots (1)$$

$$\text{Subject to } t_1^{11} \rightarrow t_1^{22}, t_2^{11} \rightarrow t_2^{22}$$

$$t_1^{11} \geq t_1^{22} \geq t_2^{11} \geq t_2^{22}, L_1 \leq T \leq D \leq S_i$$

$$T_1 = L_{max} + t_1^{22}, \text{ where } L_{max} = \max(L_1, L_2) = L_1$$

$$T_2 = L_{max} + t_2^{22}$$

$$\text{Since } t_2^{22} \leq t_1^{22}, L_{max} = \text{constant}$$

$$T_2 \leq T_1$$

$$T_2 = T^* = L_{max} + t_2^{22}, L_{max} = O_1^{11} + O_2^{11} \quad O_i^{om} = t_i^{om} + \delta$$

$$T^* = t_1^{11} + t_2^{11} + t_2^{22} + 2\delta \dots (2)$$

The makespan of FIFO ( $T_2$ )  $T_2$  is better than of SPT ( $T_1$ ) but it is not the optimal.

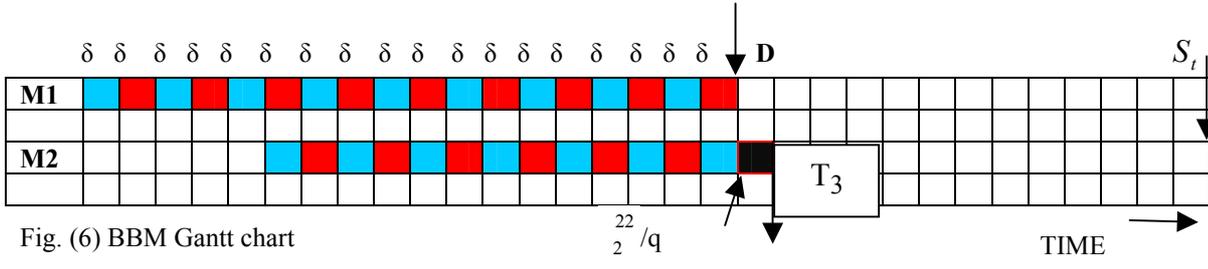
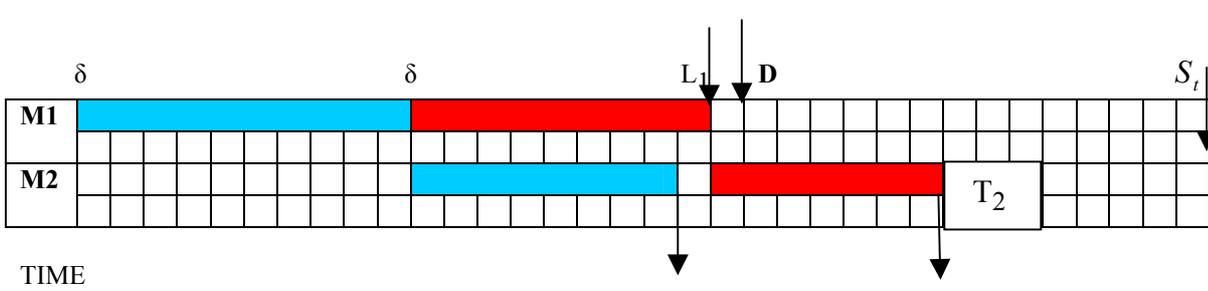
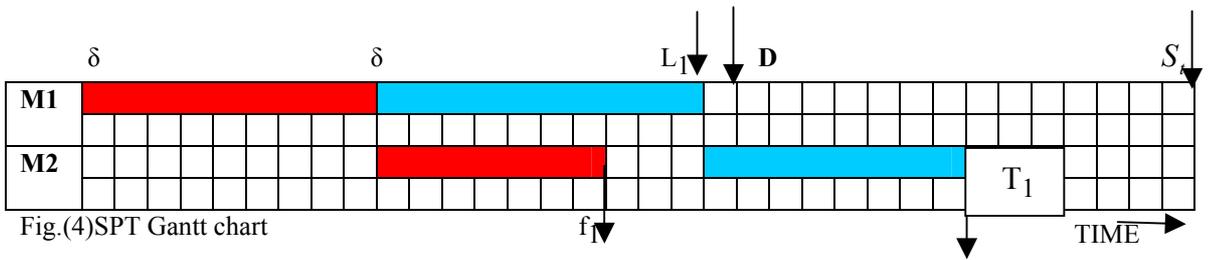
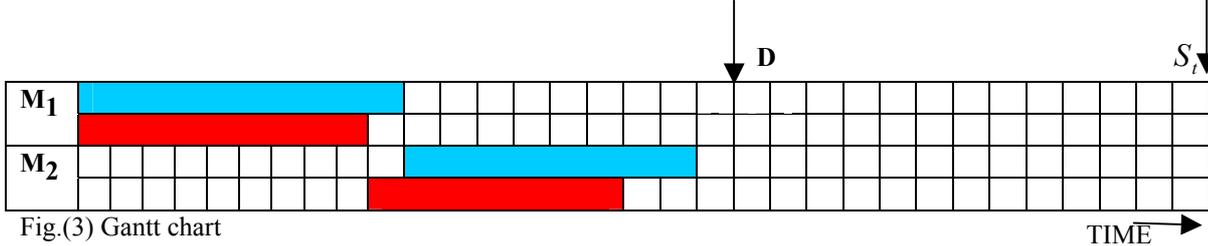
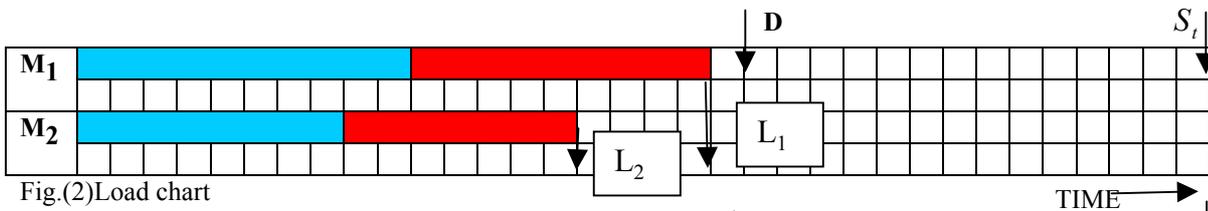
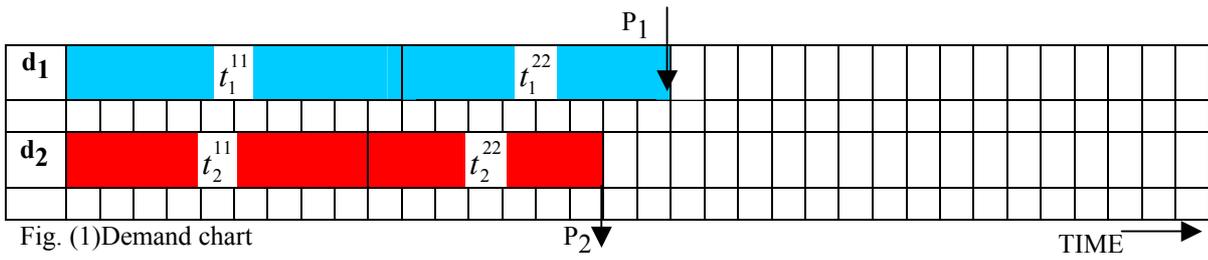


Fig.(5) FIFO Gantt chart  $f_2$

The solution of equation (2) can be tested by one of simulation methods (e.g.: Taylor ED) as shown in fig.(7). Then, build up the proper design of schedule model, but, still there is idle time in machine 2 and also  $T^* \geq D$ .

To minimize (optimize)  $T^*$  and to meet the delivery date, the following breaking stage of productivity criteria oriented-makespan is used.

## 2. Breaking Stage

By dividing the bottleneck machine times ( $L_1$ ) into sub-division of batches of time  $q$  with total set up times of Bottleneck machine 1, ( $q\delta$ ), the last sub-division of batch of time of last operation time at idle machine 2 is the Schedule Black Box ( $t_2^{22}/q$ ), as shown in the Gantt chart fig.(6). The purpose of breaking stage is to determine the schedule breakeven point using breakeven analysis. The **schedule breakeven point** is defined as the optimal sub-division quantity of time at which the total set times ( $q\delta$ ) of Bottleneck machine is equal to the Schedule Black Box ( $t_2^{22}/q$ ). At the schedule breakeven point the makespan is a minimum and the schedule productivity rate is a maximum.

### Determination of Schedule Breakeven Point $B_t^*$

$$T_3 = t_1^{11} + t_2^{11} + q\delta + t_2^{22}/q \dots\dots\dots(3)$$

Since,  $t_1^{11}$  and  $t_2^{11}$  are constant,

$$B_t = T_3 - (t_1^{11} + t_2^{11}) \text{ and called "Schedule Break time"}$$

$$B_t = T_3 - (t_1^{11} + t_2^{11}) = q\delta + t_2^{22}/q, \dots (4)$$

Taking the derivative of  $B_t$  w.r.t  $q$  and equaling zero

$$\frac{\partial B_t}{\partial q} = \delta - \frac{t_2^{22}}{q^2} = 0 \rightarrow q^* = \sqrt{\frac{t_2^{22}}{\delta}} \dots (5)$$

$$B_t^* = 2\sqrt{t_2^{22}\delta} \quad (6)$$

$B_t^*$  is called "Schedule Breakeven Point"

$$T_3 = t_1^{11} + t_2^{11} + 2\sqrt{t_2^{22}\delta}$$

If  $T_3 \leq T_2$  and  $T_3 \leq D$ , then,  $T_3 = T^{**}$

$$T^{**} = t_1^{11} + t_2^{11} + B_t^* \dots (7)$$

$$\eta = (T^* / T^{**}) \rightarrow \eta_R = (\eta - 1) * 100 \dots (8)$$

It concluded that as the number of sub-division of batches  $q$  increases,  $E.(3)$ , the total time( $q\delta$ ) increase, the Schedule black box ( $t_2^{22}/q$ ) decrease.(4), makespan decrease and schedule productivity rate increase.(8) until certain point which is schedule breakeven point  $B_t^*$ ,  $E.(6)$  at which the makespan  $T$  is minimum and schedule productivity rate  $\eta_R$  is maximum.

### Determination of Optimum Units Per Batch $X_i^{om}$

To find out  $X_i^{om}$  (unit/batch) and  $X_{1L}^{22}$  which is the number of units of last batch (unit/batch of time). it must be determined first, the Number of batches of time per operation of demand through machine  $m$ ,  $q_i^{om}$  (batch of time), Length of batch time (h/batch of time)  $\tau_i^{om}$ , and :

$\tau_{1L}^{22}$  : Last batch length (h/batch) also it can be specify the time required to process one unit of batch of demand through certain machine,  $\alpha$  (min/unit). Approximation can be done if required. The following formula are used.

$$q_1^{11} = t_1^{11} * q^* / (t_1^{11} + t_2^{11}), q_2^{11} = t_2^{11} * q^* / (t_1^{11} + t_2^{11})$$

$$\tau_1^{11} = t_1^{11} / q_1^{11}, \tau_2^{11} = t_2^{11} / q_2^{11}, \tau_{1L}^{22} = t_2^{22} / q^*$$

$$X_1^{11} = X_1 / q_1^{11}, X_2^{11} = X_2 / q_2^{11}, X_{1L}^{22} = X_1^{11}$$

$$\alpha_1^{11} = \tau_1^{11} / X_1^{11}, \alpha_2^{11} = \tau_2^{11} / X_2^{11}, \alpha_{1L}^{22} = \tau_{1L}^{22} / X_1^{11}$$

### 3. Rebuilding Stage

In this stage the most proper schedule is selected using simulation. The simulation model rebuild up a gain according to the new condition due to the effect of BBM that to design the final Simulation Model. Corrective actions could be taken if necessary, then, testing and validation of schedules guaranteeing to select the most proper schedule and can be successfully implemented.

### Application of BBM

#### Building Stage

As shown in demand table fig.(4)  $t_1^{11}=1000$  h,  $t_1^{22}=800$  h,  $t_2^{11}=900$  h,  $t_2^{22}=700$  h,  $\delta = 1.75$  h,  $X_1=1200$  unit,  $X_2=540$  unit,  $D=2000$  h

Table (4) Demand Table

d	O <sub>1</sub>	O <sub>2</sub>	P
d <sub>1</sub>	1000	800	<b>1804</b>
d <sub>2</sub>	900	700	<b>1604</b>
<b>L</b>	<b>1904</b>	<b>1504</b>	<b>3408</b>

#### Solution

$$L_{\max} = L_1 = 1903.5h$$

$$\text{SPT: } T_1 = 2703.5 h$$

$$\text{FIFO: } T_2 = 2603.5 h$$

$$T_2 \leq T_1 \rightarrow T_2 = T^* \quad \text{But, } T^* \geq D.$$

so, the following breaking stage must be done.

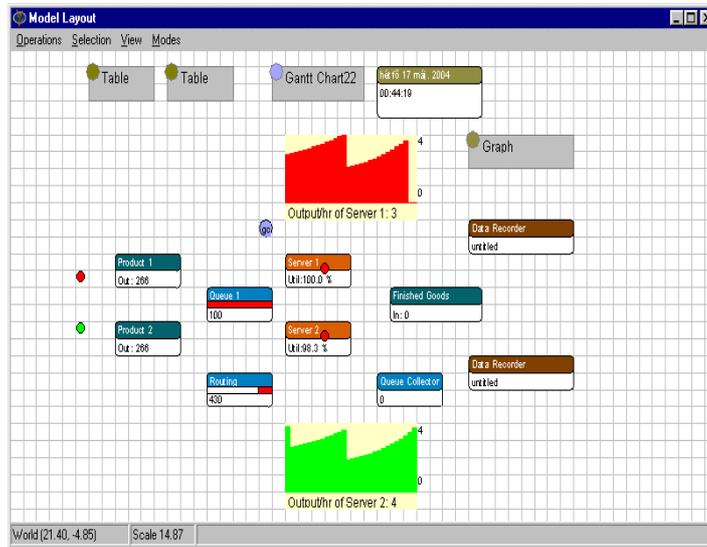


Fig.(7) Simulation Model

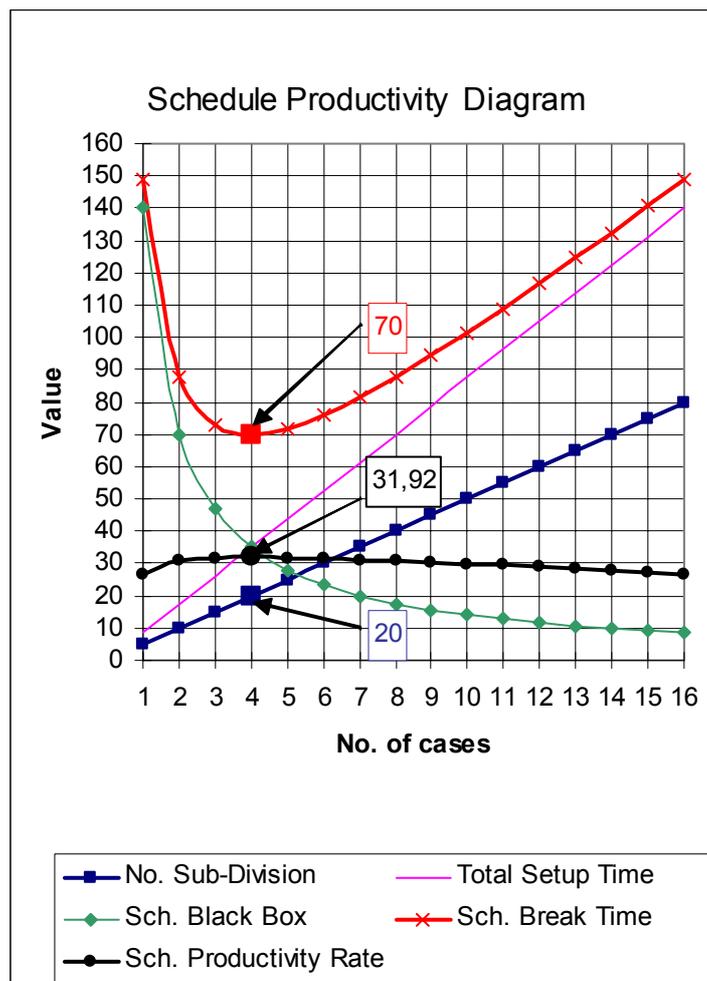


Fig.(8) Schedule Productivity Diagram

## Breaking stage

### Determination of Schedule Productivity Rate

$$q^* = 20 \text{ (batch of time)}, T_3 = 1973.5 \text{ h}$$

$$T_3 \leq T_2, T_3 \leq D T_3 = T^{**} = 1973.5 \text{ h},$$

$$B_t^* = 70 \text{ h} \rightarrow \eta_R = 31.92 \%$$

The schedule breakeven point  $B_t^*$  (70h) at which the Minimum makespan  $T(1973.5\text{h})$  and the maximum schedule productivity rate  $\eta_R$  are (31.92%) as shown in figure (8).

### Determination of Optimum Units Per Batch of Time

$$q_1^{11} = 10 \text{ (batch of time)}, q_2^{11} = 9 \text{ (batch of time)}$$

$$\tau_1^{11} = 100 \text{ h/batch}, \tau_2^{11} = 100 \text{ h/batch}, \tau_{1L}^{22} = 35 \text{ h/batch}$$

$$X_1^{11} = 120 \text{ (unit / batch of time)}, X_2^{11} = 60 \text{ (unit/ batch of time)}, X_{1L}^{22} = 120 \text{ (unit/batch of time)}$$

$$\alpha_1^{11} = 50 \text{ min/unit}, \alpha_2^{11} = 70 \text{ min/unit}, \alpha_{1L}^{22} = 17.52 \text{ min/unit}$$

### Rebuilding Stage

According to the previous figures of the breaking stage, it could be design the optimum schedule of such problem and tested by simulation

## CONCLUSION

The sub-division of batches is a powerful tool for improving the quality of FMS scheduling. In the present paper, for the simplest case (two machine group, two part

types), a new method is proposed to use the above approach. The proposed BBM (Break and Build Method) provide solution for the problem. The results clearly show the effectiveness of the given approach. Future research should be directed to generalize the method to multipart, multi machine group cases.

## REFERENCES

- CARRIE A. 1988. "Simulation of Manufacturing Systems", PP. 418. Wiley
- CHASE, AQUILANO. 1985. "Production and Operation Management": A life cycle approach, PP 853. Irwin
- FRENCH S. 1982. "Sequencing and scheduling: An introduction to the mathematics of the Job-Shop", Wiley, PP. 245.
- GROOVER P., 1987, "Automation, Production Systems, and Computer Integrated Manufacturing", PP. 808. Prentice-Hall
- HAROLD T., JOHN A., OLIVER S., 1975. "Manufacturing Organization and Management", PP. 588. Prentice-Hall
- JACK R, 1992, "The Management Of Operations: Conceptual Emphasis", PP 772. Wiley
- PAUL LOOMBA N., 1978, "Management-A Quantitative Perspective", PP 594. Collier Machillan
- SOMLO J., 2001, "Hybrid Dynamical Approach makes FMS scheduling more Effective", PERIODICA POLYTECHNIC SER. MECH. ENG. VOL.45, NO.2. PP. 175-200. Budapest
- TAKESHI YAMADA, RYOHEI NAKANO, "Job-Shop Scheduling", IEE Control Engineering Series 55, Genetic Algorithms in Engineering Systems, Edited by A.M.S. Zalza and p. j. Fleming, Chapter 7, PP. 134-160.
- THOMAS M., ROBERT A., 1981, "Introduction to Management Science", PP 764. Prentice-Hall
- U. REMBOLD, B.O.NNAJI, A. STORR, 1993, "Computer Integrated Manufacturing and Engineering", PP. 640. Addison-Wesley.

# DYNAMIC CONFIGURATION IN A LARGE SCALE DISTRIBUTED SIMULATION FOR MANUFACTURING SYSTEMS

Koichi Furusawa\*

Kazushi Ohashi†

Mitsubishi Electric Corp. Advanced Technology R&D Center

8-1-1, Tsukaguchi-honmachi

Amagasaki, Hyogo 661-8661, Japan

E-mail: \*Furusawa.Koichi@wrc.melco.co.jp

† Ohashi.Kazushi@wrc.melco.co.jp

## KEYWORDS

Dynamic configuration, Distributed simulation,  
Integrated simulation, Manufacturing, Synchronization

## ABSTRACT

We are developing an integrated simulation environment for manufacturing systems, in which simulators are synchronized in order to guarantee timed consistency among the simulators. We propose a dynamic configuration method in distributed simulation. It automatically configures distributed simulators during integrated simulation and enables efficient execution of large scale integrated simulation. In this paper, we illustrate the proposed dynamic configuration method, and we show its evaluation results under various conditions.

## INTRODUCTION

Manufacturing systems have become increasingly large and complicated, and they consist of various kinds of equipment. We usually use simulation technology to verify their behavior when setting up a new factory or carrying out improvements. High accuracy simulation is necessary to reduce implementation costs.

Many simulators have been already developed in the manufacturing field. They are not easy to combine with each other for an integrated simulation because of the asynchronous execution of the simulators. To solve the problem, several methods of synchronization between simulators are proposed (Carothers et al. 1997, Dahmann et al. 1997, Defense Modeling and Simulation Office 1998). We proposed more efficient synchronization mechanism for manufacturing systems (Furusawa and Yoshikawa 2002).

In simulation for manufacturing systems, various kinds of simulators are integrated. The integrated simulators are so many that they are necessary to be distributed to several PCs in order to reduce the simulation load. The configuration of the distributed simulators are usually decided by the user of the simulators. It is not, however, easy to estimate the PC's load and effectively configure

them in advance. As a result, the total time for the integrated simulation tends to be longer than needed.

In order to solve the above problem, it is useful to dynamically configure the integrated simulators during simulation according to the current situation. From the viewpoint of fault tolerant, the dynamic configuration of distributed simulators is proposed (Welch and Purtilo 1997, Welch and Purtilo 1999). Those purposes are recovering simulators from unexpected troubles. Improvement of the simulation efficiency is not discussed very much. We propose a dynamic configuration method to reduce the total simulation time. In the proposed method, the load of each PC executing simulation and the traffic of data communication between PCs are periodically monitored, and the configuration of the distributed simulators are dynamically and automatically changed during simulation. It enables simulator users to easily simulate a large scale manufacturing system without considering the configuration of the simulators.

## INTEGRATED SIMULATION

Simulators for PLC, CNC, robot, etc. have been developed in the manufacturing field. We can check the control program for the controller using the simulator. However, it is hard to verify the behavior of the whole manufacturing system which consists of lots of machines.

An integrated simulation environment, which can connect many simulators, is useful for testing a whole manufacturing system. Connecting various simulators is possible by communicating control signals and other information between them. Using only data communication does not accurately simulate an actual manufacturing system consisting of many machines, if the simulators are executed asynchronously. The simulation, which does not consider data communication delay between simulators, is not sufficient for system simulation. Moreover, integrated simulators are necessary to be distributed on several PCs, since an integrated simulation for manufacturing systems consists of many simulators. Therefore the

integration of different kinds of simulators involves the following issues:

- Synchronization management
- Communication management
- Configuration management

### Synchronization management

Figure 1 illustrates an example of simulation flow, in which three simulators are integrated asynchronously. The cycle times of the simulators are 100, 70, and 130, respectively. Each simulator executes a cycle of simulation, and exchanges data with connected simulators, then executes next cycle of simulation. In Figure 1, a vertical line is actual time, and an italic figure is logical time of the simulator. A rectangle is execution of simulation, and an arrow is a data flow between simulators.

In Figure 1, when the simulator2 (S2) sends data to the simulator1, 3 (S1, S3) at time 490, the S1 and S3 have already arrived at time 500 and 520, respectively. The simulation does not guarantee the timed consistency, since the S1 and S3 receive a past data.

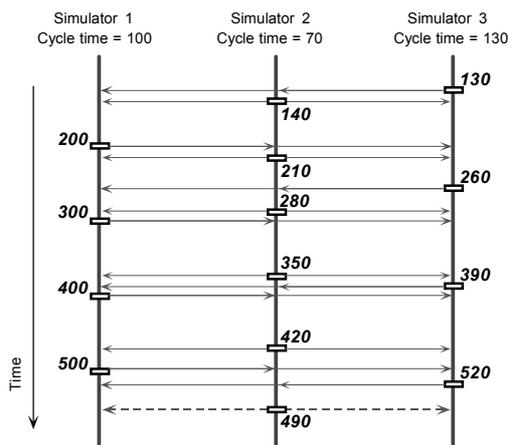


Figure 1: Simulation flow (asynchronous)

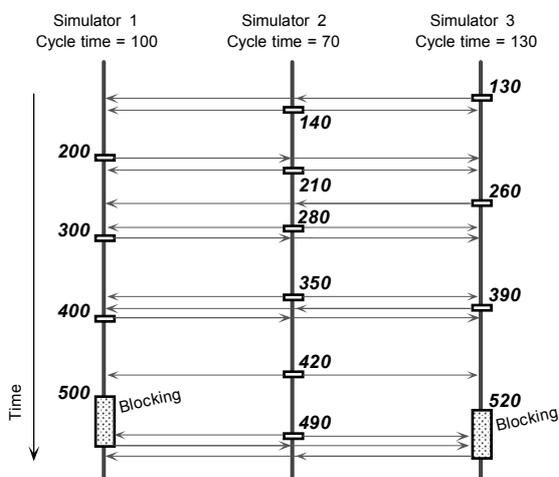


Figure 2: Simulation flow (synchronous)

Figure 2 illustrates a simulation flow in case of synchronous simulation. In this case, the S1 is blocked starting simulation at time 500, since the simulation of the S2 at time 490 is not finished. The S3 is also blocked at time 520 in the same way. It is possible to guarantee the timed consistency by synchronizing connected simulators.

### Communication management

Integrated simulators have their I/O (input and output), and they are connected to others by exchanging data through the I/O. Synchronous data exchange guarantees the timed consistency.

There are many types of communication lines between equipment in manufacturing systems, for example a serial line, Ethernet and so on, and the data transfer speed depends on the line type (Figure 3). The simulation, which does not consider data communication delay according to the lines, is not sufficient for manufacturing systems.

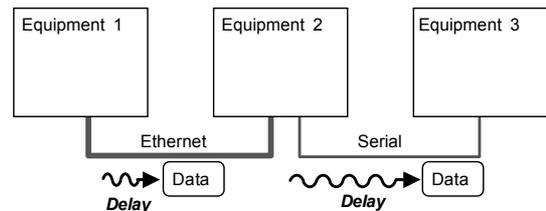


Figure 3: Data communication

### Configuration management

Manufacturing systems consist of many pieces of equipment, and a large number of simulators are integrated and executed simultaneously when they are simulated. It is difficult to execute the all simulators on a PC, since the simulation needs a lot of processing power in general. Consequently, the simulators are necessary to be appropriately distributed on several PCs and be synchronously executed exchanging data for each others (Figure 4). Configuration management is important for a large scale of simulation like manufacturing systems.

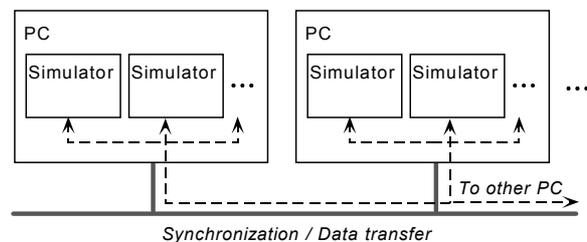


Figure 4: Configuration management of simulators

## CONFIGURATION OF SIMULATORS

We already proposed the method of the synchronization management and the communication management in the

integrated simulation (Furusawa and Yoshikawa 2002). In this section, we discuss the configuration management.

### Configuration in an integrated simulation

In the integrated simulation of manufacturing systems, the various types of simulators are integrated. Some of them require a lot of processing time for their execution. For example, a simulator of a motion controller has very short cycle time, and it usually needs high CPU power to simulate it. If many simulators, which require much processing, are executed on only a PC, its load becomes very high and the execution of the simulation needs a lot of time. As a consequence, the whole simulation does not advance effectively, even if all of the simulators on other PCs finish their simulation cycles.

To solve the above problem, the integrated simulators must be appropriately distributed to several PCs. It is necessary to estimate the capability and number of the PCs corresponding to the type and number of the simulators in advance, and to arrange the simulators to each PC.

### Static configuration

In this section, the static configuration is explained, which defines the configuration of the used PCs and assignment of the simulators to the PCs before executing simulation.

#### Configuration of PCs

Optimizing the efficiency of the integrated simulation using the limited resources involves the following issues:

- Number of PCs

The more simulators the integrated simulation executes, the more PCs it requires.

- Capability of PCs

The more the simulators need the amount of computation, the higher performance PC the integrated simulation requires.

- Network between PCs

The network configuration must be decided considering the location of the PC. For example, in case of executing a high load simulator on a high performance remote PC, the simulator might be connected to others via the internet.

#### Configuration of simulators

The assignment of the simulators to the PCs is necessary to be decided after the configuration of the PCs. The simulators should be appropriately distributed to the PCs in order to uniform their CPU load. In order to decide the desirable configuration, it is necessary to estimate the required CPU load, in advance, corresponding to the types of the simulators.

The communication traffic is another factor to decide the configuration. For example, when some simulators

communicate their large data to each other, the communication traffic becomes very heavy if they are distributed to different PCs (Figure 5). The whole simulation time could be longer. In this case, executing the simulators on the same PC might improve the efficiency. The assignment of the simulators should be, therefore, decided considering the communication traffic between the simulators.

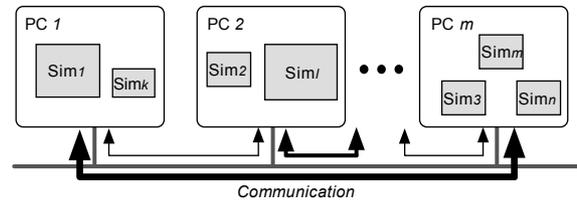


Figure 5: Communication traffic between simulators

### Problems in static configuration

The static configuration is comparatively easy to implement a simulation environment, since it is usually defined before simulation and it is fixed during simulation. From the viewpoint of efficiency, however, there are problems in the static configuration as follows:

- Unbalance of the CPU load

In the simulation for manufacturing systems, there are many types of simulators and the scale tends to be very large. Therefore, it is difficult to estimate the CPU load of the used PCs in advance. They are also changeable according to the condition. Under the static configuration, however, the distributed assignment of the simulators is fixed, and it could cause the unbalanced CPU load.

- Unbalance of the communication traffic

It is difficult to estimate the communication traffic between simulators, and it is changeable as well. The static configuration could cause the unbalanced communication traffic.

- Inefficiency of the simulation execution

The unbalance of the CPU load and communication traffic leads to decreasing the throughput of the PCs. As a result, the required time for simulation becomes longer, and the execution of the whole simulation becomes inefficient.

- Difficulty of deciding the configuration

In case of the static configuration, the simulation user must estimate the specification and number of used PCs, and assign the simulators to them appropriately. It is so heavy a burden, and requires high skill and experience.

### DYNAMIC CONFIGURATION

The static and fixed configuration causes the problems shown in the previous section. Dynamic configuration solves them. We propose a dynamic configuration method, which enables distributed simulators to dynamically change their configuration during the simulation.

The proposed dynamic configuration is realized by gathering the information on the CPU load of the used PCs and the communication traffic between them, and dynamically moving the simulators to the appropriate PC when the load and the traffic exceed a certain level during the integrated simulation.

### Flow of dynamic configuration

In this section, we show the flow of the dynamic configuration in the integrated simulation. The flow is as follows:

- Step1 Gather the information for reconfiguration**  
The information about the CPU load and the communication traffic between simulators is gathered. The information is used in order to judge whether the reconfiguration of the simulator is carried out.
- Step2 Judge the reconfiguration**  
The information gathered in the Step1 is applied to the conditional expression of the reconfiguration, and then the necessity of the reconfiguration is judged.
- Step3 Decide the reconfiguration**  
The new configuration is calculated. It improves the efficiency of the integrated simulation.
- Step4 Stop the integrated simulation**  
The integrated simulation is stopped before its reconfiguration.
- Step5 Execute the reconfiguration**  
In order to change the present configuration to the new one calculated in the Step3, the information on the simulator, for example the current status, is transferred to the reconfigured PC, and the simulator is prepared starting on the new configuration.
- Step6 Restart the integrated simulation**  
The integrated simulation is restarted on the new configuration.

### Conditional expression of reconfiguration

As we explained above, the conditional expression is used to decide the necessity of the reconfiguration. The necessity of the reconfiguration is judged based on the balance of the PC's load and the communication traffic. The conditional expression is lead by the following steps.

The required time for a cycle of simulation on a PC  $P_k$ ,  $TS_k$ , is expressed by the following expression:

$$TS_k = \frac{\sum_{k=k1}^{km_k} L_k}{P_k} \quad (1)$$

where the simulators on the  $P_k$  are  $S_{k1}, S_{k2}, \dots, S_{km_k}$ , and the amount of computation required for the simulation of  $S_k$  is  $L_k$ , and the amount of computation,

which is able to be executed per unit time on the  $P_k$ , is  $P_k$ .

The required time for the data transfer on the  $P_k$  in a cycle of simulation,  $TC_k$ , is expressed by the following expression:

$$TC_k = \sum_{i=k1}^{km_k} \sum_{j=1}^n d_{ij} D_{ij} \quad (2)$$

where the amount of transferred data between  $S_i$  and  $S_j$ , is  $D_{ij}$ , and the required time for transferring a unit data between  $S_i$  and  $S_j$  is  $d_{ij}$ .

The required time for the integrated simulation on the  $P_k$ ,  $T_k$ , is expressed by the sum of Equation (1) and (2).

$$T_k = TS_k + TC_k \quad (3)$$

The standard deviation of the required time on a PC,  $\sigma$ , is expressed by the following expression:

$$\sigma = \sqrt{\frac{1}{m} \sum_{k=1}^m (T_k - \bar{T})^2} \quad (4)$$

where  $P_1, P_2, \dots, P_m$  are PCs used for the integrated simulation, and  $\bar{T}$  is the average of the required time  $T_1, T_2, \dots, T_m$ .

Finally, the conditional expression of the reconfiguration is the following expression:

$$\frac{\sigma}{\bar{T}} \geq r \quad (5)$$

where  $r$  is a judging parameter ( $0 \leq r \leq 1$ ). When the conditional expression (5) is true, the present configuration is judged not to be well-balanced and the reconfiguration is carried out. If the parameter  $r$  is small, the reconfiguration occurs frequently.

### EVALUATION OF THE METHOD

In this section, we validate our dynamic configuration method by simulation.

#### Assumption

In order to validate our dynamic configuration method, we simulated it under various conditions. In the simulation, the number of the integrated simulators is 40 ( $S_1, S_2, \dots, S_{40}$ ), which is supposed to be components of manufacturing systems. The number of used PCs is 10 ( $P_1, P_2, \dots, P_{10}$ ). The method is evaluated changing the following three conditions:

- Performance of PCs
- Amount of computation for each simulator

- Amount of transferred data between simulators

The details of the conditions are shown in Table 1, Table 2 and Table 3. On the performance of PCs, two cases are evaluated as shown in Table 1. The all PCs have the same performance in the one case, and three types of performance in the other case. The required amount of computation for simulators and the amount of transferred data between simulators are randomly fluctuated inside the range shown in Table 2. The interval of their changes is also randomly fluctuated inside the range shown in Table 3.

In Table 1, the performance of PCs means the amount of computation, which can be computed per unit time. In Table 2, the amount of computation means the total amount of computation, which is required for a cycle of simulation of each simulator. All of the integrated simulators are supposed to communicate with each others, and the transferred data means the total amount of data sent and received per cycle of simulation. To simplify the problem, the two conditions for the amount of computation and the transferred data are combined, though they are independent conditions. And the required time for data transfer between two simulators on the different PCs is supposed to be 0.001 sec/KB, and the one between two simulators on the same PC is supposed to be 0.00001 sec/KB. And the cycle time of the every simulator is supposed to be 0.1 sec. The judging parameter  $r$  is 0.1, and the reconfiguration is judged at 60 sec interval. The simulators are initially assigned based on the performance of the PCs.

Table 1: Condition (PC performance)

Condition	Performance of PCs
Cp1	400 [10 PCs]
Cp2	1000 [1 PC], 400 [4 PCs], 200 [5 PCs]

Table 2: Condition (simulation load)

Condition	Amount of computation	transferred data (KB)
Cc1	1	0.01
Cc2	Min 1, Max 2	Min 0.01, Max 0.02
Cc3	Min 1, Max 5	Min 0.01, Max 0.05
Cc4	Min 1, Max 10	Min 0.01, Max 0.10
Cc5	Min 1, Max 20	Min 0.01, Max 0.20
Cc6	Min 1, Max 40	Min 0.01, Max 0.40
Cc7	Min 1, Max 80	Min 0.01, Max 0.80

Table 3: Condition (change interval)

Condition	Change interval (sec)
Ci1 (short)	Min 60, Max 120
Ci2 (long)	Min 480, Max 960

From the view of simulation efficiency, the configuration minimizing  $\max(T)$  in consideration of all PCs and simulators is the optimal one. However, the combination of  $P_1, P_2, \dots, P_m$  and  $S_1, S_2, \dots, S_n$  is  $m^n$ , and it costs too much time to evaluate all patterns and to optimize them. In our evaluation, therefore, we selected two PCs, whose  $\max(T)$  is the largest and the smallest, and then we optimized the assignment of the simulators on these two PCs.

## Evaluation results

We evaluated the required total simulation time under various conditions. In our evaluation, the required total time using our dynamic configuration is compared with one not using it. The required total time includes the time for computation of simulation and transferring data, and the overhead for reconfiguration of simulators, for example moving simulators, restarting them, and so on, is not included. It, however, has not so large influence on the efficiency of the simulation, since the reconfiguration interval is sufficiently long. We show the results of simulation. The duration of the simulation is 10000 sec (almost 3 hours).

Firstly, in the cases using the same performance PCs, the results of the simulation are shown in Figure 6 and Table 4. The case of short fluctuation interval and the case of long one are evaluated. The improvement of the required total simulation time is not confirmed, when the fluctuation range of the amount of computation and the transferred data is small. The larger the range is, however, the larger the reduction of the total simulation time is.

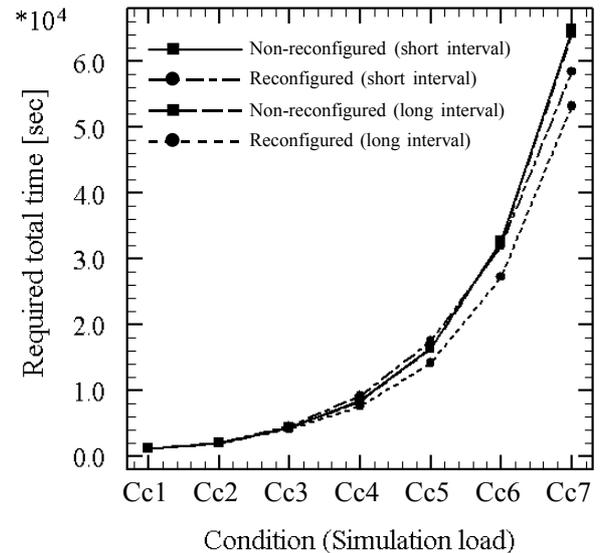


Figure 6: Required total time (condition Cp1)

Secondly, in the cases using the different performance PCs, the results of the simulation are shown in Figure 7 and Table 5. Unlike the first cases, the improvement of

the required total time is confirmed, even when the fluctuation range of the amount of computation for simulating and the amount of transferred data is small. The larger the range is, the larger the reduction of the required total time is, in the same way as the first cases. And the reduction rate is twice as high as the first cases.

Table 4: Reduction rate of total time[%](condition Cp1)

Interval	Cc1	Cc2	Cc3	Cc4	Cc5	Cc6	Cc7
Ci1 (short)	0.0	0.0	-4.2	-8.9	-6.3	1.8	9.9
Ci2 (long)	0.0	0.0	3.8	9.0	13.3	15.7	17.2

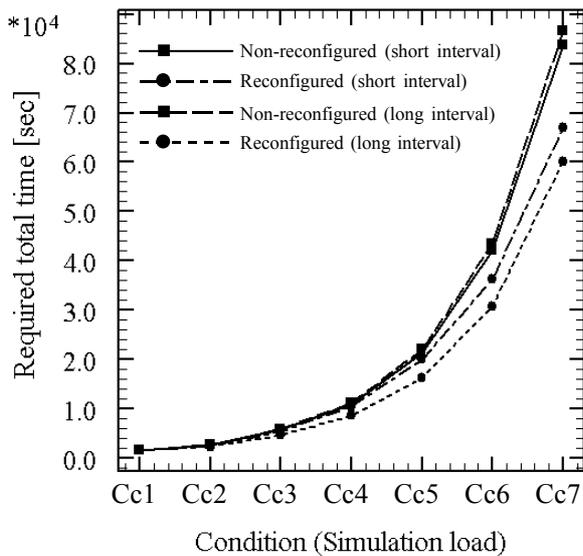


Figure 7: Required total time (condition Cp2)

Table 5: Reduction rate of total time[%](condition Cp2)

Interval	Cc1	Cc2	Cc3	Cc4	Cc5	Cc6	Cc7
Ci1 (short)	0.0	6.5	4.4	4.0	6.4	13.8	20.0
Ci2 (long)	0.0	13.0	20.6	22.7	26.1	29.2	30.6

## CONCLUSION

We proposed the dynamic configuration method of the integrated simulation for manufacturing systems. The proposed method can automatically adjust the configuration of the large scale integrated simulation, and the user of the simulator can efficiently carry out the complicated simulation without considering the configuration, using the limited resources. We evaluated our method and it is confirmed that the required simulation time is drastically reduced, especially when the range of fluctuation of simulation load is large, though the overhead of the reconfiguration is not included in the evaluation.

In this paper, our method is evaluated by only simulation, and the overhead of the reconfiguration is

not discussed. It should be also evaluated in the real situation, in order to examine the effect of the overhead and verify the practicality of our method.

## REFERENCES

- Carothers, C.D.; Fujimoto, R.M.; Weatherly, R.M.; and Wilson A.L. 1997. "Design and Implementation of HLA Time Management in the RTI Version F.0." *In Proceedings of the 1997 Winter Simulation Conference*, 373-380.
- Dahmann, J.S.; Fujimoto, R.M.; and Weatherly, R.M. 1997. "The Department of Defense High Level Architecture." *In Proceedings of the 1997 Winter Simulation Conference*, 142-149.
- Defense Modeling and Simulation Office. 1998. *High Level Architecture Interface Specification, v1.3*.
- Furusawa, K. and Yoshikawa, T. 2002. "Synchronization Mechanism in Integrated Simulation for Manufacturing Systems." *In Proceedings of the MED2002*.
- Welch, D.J. and Purtilo J.M. 1997. "Using Compensating Reconfiguration to Maintain Military Distributed Simulations." *Proceedings of the 1997 Winter Simulation Conference*, 961-967.
- Welch, D.J. and Purtilo J.M. 1999. "Building Self-Reconfiguring Distributed Simulations Using Compensating Reconfiguration." *The Journal of Defense Software Engineering*, 20-23.

## AUTHOR BIOGRAPHIES



**KOICHI FURUSAWA** was born in Osaka, Japan and went to the Osaka University of Japan, where he studied computer science and obtained his B.E. degree and M.E. degree in 1993 and 1995. Then, he works for the Mitsubishi Electric corp. where he is now a research engineer in the controller group of the advanced technology R&D center in the field of a programmable logic controller. His e-mail address is Furusawa.Koichi@wrc.melco.co.jp.



**KAZUSHI OHASHI** was born in Osaka, Japan and went to the Osaka Prefecture University of Japan, where he studied robotics and obtained his B.E. degree in 1988 and went to the Osaka University of Japan, where he obtained M.E. degree in 1990. Then, he works for the Mitsubishi Electric corp. where he is now leading a controller engineering unit in the controller group of the advanced technology R&D center. His e-mail address is Ohashi.Kazushi@wrc.melco.co.jp.

# SIMULATION SUPPORT FOR RESCHEDULING

András Pfeiffer<sup>1</sup>, Botond Kádár<sup>1</sup>, László Monostori<sup>1,2</sup>

<sup>1</sup>Computer and Automation Research Institute Hungarian Academy of Sciences  
Kende u. 13-17, Budapest, H-1111, Hungary

<sup>2</sup>Department of Production Informatics, Management and Control, Faculty of Mechanical Engineering,  
Budapest University of Technology and Economics, Budapest, Hungary  
E-mail: pfeiffer@sztaki.hu

## KEYWORDS

Dynamic scheduling, rescheduling, simulation, stability

## ABSTRACT

The paper discusses the job shop scheduling problem and schedule measurement techniques, especially outlining the methods that can be applied in a dynamic environment. The authors propose a periodic rescheduling method by taking the rescheduling interval and schedule stability factor as input parameters into consideration. The proposed approach is tested on a simulated environment in order to determine the effect of stability parameters on the selected performance measures.

## INTRODUCTION

The broad goal of manufacturing operation management, such as a resource constrained scheduling problem, is to achieve a co-ordinated efficient behaviour of manufacturing in servicing production demands while responding to changes in shop-floors rapidly and in a cost effective manner.

In theory the aim is to minimize or maximize a performance measure. Regarding complexity, the job-shop scheduling problem (and therefore also its extensions), except for some strongly restricted special cases, is an NP-hard optimization problem (Baker 1998; Williamson et al. 1997).

The above mentioned job-shop scheduling is a static case, where all the information is available initially and it does not change over time. Most of the solutions in the literature concerning scheduling concentrate on this static problem. However, in many real systems, this scheduling problem is even more difficult because jobs arrive on a continuous basis, henceforth called dynamic job shop scheduling (DJSS). According to Rangsaritratamee, et al. (2004), previous research on DJSS using classic performance measures like makespan or tardiness concludes that it is highly desirable to construct a new schedule frequently so recently arrived jobs can be integrated into the schedule soon after they arrive.

Scheduling techniques addressing the dynamic – in the current case job shop – scheduling problem are called dynamic scheduling algorithms. These algorithms can be further classified as reactive and proactive

scheduling techniques. Depending on the environment, there may be deviations from the predictive schedule during the schedule execution due to unforeseen disruptions such as machine breakdowns, insufficient raw material, or difference in operator efficiency overriding the predictive schedule. The process of modifying the predictive schedule in the face of execution disruptions is referred to as reactive scheduling or rescheduling (Szelke and Monostori 1999). The reaction to the realised disruption generally takes the form of either modifying the existing predictive schedule, or generating a completely new schedule, which is followed until the next disruption occurs (Kempf et al. 2000).

The practical importance of the decision whether to reschedule or repair has been noted in (Szelke and Kerr 1994), while an additional categorization of scheduling techniques relating to the stochastic or deterministic characteristics of the problem can be found in (Kádár 2002).

It is important to outline, that while rescheduling will optimize efficiency using classic performance measures (makespan or tardiness) the impact of disruptions induced by moving jobs during a rescheduling event is mostly neglected. This impact is frequently called stability (Rangsaritratamee et al. 2004; Cowling and Johansson 2002). In related previous works, the number of times rescheduling takes place was used by Church and Uzsoy (1992) as the measure of stability and it was suggested that a more frequent rescheduling means a less stable schedule. Other approaches defined stability in terms of the deviation of job starting times between the original and revised schedule and the difference of job sequences between the original and revised schedules. One of the shortcomings of these approaches is that they ignore the fact that the impact of changes increases as they are made closer to the current time. Rangsaritratamee, et al. (2004) propose a method which addresses DJSS based on a bicriteria objective function that simultaneously considers efficiency and stability, and so let the decision maker to strike a compromise between improved efficiency and stability. In the approach, two dimensions of stability are modelled. The first captures the deviation of job starting times between two successive schedules and the second reflects how close to the current time changes are made.

Vieira, et al. (2000) presents new analytical models that can predict the performance of rescheduling strategies and quantify the trade-offs between different performance measures. Three rescheduling strategies are studied in a parallel machine system: periodic, event-driven and hybrid, similarly to the work (Church and Uzsoy 1992). They realized that there is a conflict between avoiding setups and reducing flow time, and the rescheduling period affects both objectives significantly, which statement is coincident concluded by Rangsaritratamee, et al. (2004).

In order to decide what action we should take in response to an event, we should have some idea of the value of our current schedule. In the following section evaluation classes of production schedules are introduced.

### **EVALUATION OF PRODUCTION SCHEDULES**

This part of the paper discusses the problem of schedule measurement and especially outlines the techniques which can be applied in a dynamic environment.

The quality of factory scheduling, generally, has a profound effect on the overall factory performance. As stated in (Kempf et al. 2000), an important aspect of the schedule measurement problem is whether an individual schedule or a group of schedules is evaluated. Individual schedules are evaluated to measure its individual performance. For a predictive schedule, the result may determine whether it will be implemented or not. There might be different reasons for evaluating a group of schedules. One of them is to compare the performance of the algorithms with which the different schedules were calculated. The comparison of different schedule instances against different performance measures is an other option in the evaluation of a set of schedules for the same problem.

According to Kempf, et al. (2000), relative comparison assumes that for the same initial factory state two or more schedules are available, and the task is to decide which is better. The task is to decide which one is better from two schedules, or which one is the best from a group of schedules generates additional questions. In a complex manufacturing environment different schedules will probably perform better against different performance measures. Therefore, the selection of the best schedule will always depend on the selected performance measure(s) and thus, on the external constraints posed by the management of the enterprise.

An absolute measurement of schedule quality consists in taking a particular schedule on its own and deciding how „good” it is. This requires some set of criteria or benchmarks against which to measure.

Regarding the predictive schedules, a set of decisions is made on the base of estimates on future events, without knowing the actual realizations of the events in question until they actually occur. Taking this fact into

consideration, Kempf, et al. (2000) differentiate between the static and dynamic measurements of predictive schedules. A static measurement means the evaluation of the schedule independently of the execution environment.

Contrary to static measurement, the dynamic measurement of a predictive schedule is more difficult. In this case, beyond the static quality of the schedule, the robustness of the schedule against uncertainties in the system should also be taken into consideration.

Another aspect in the evaluation of schedules is the state of the manufacturing system after the execution of the schedule. In (Kempf et al. 2000) these parameters are compared as state measurements, which evaluate the end effects of the schedule at the end of the schedule horizon.

Regarding the evaluation classes listed above, a dynamic measurement of individual predictive schedules will be presented in the following sections.

### **SIMULATION IN DYNAMIC SCHEDULING**

Simulation captures those relevant aspects of the production planning and scheduling (PPS) problem, which cannot be represented in a deterministic, constraint-based optimization model. The most important issues in this respect are uncertain availability of resource, uncertain processing times, uncertain quality of raw material, and insertion of conditional operations into the technological routings.

The features provided by the new generation of simulation software facilitate the integration of these tools with the production planning and scheduling systems. Additionally, if the simulation system is combined with the production database of the enterprise it is possible to instantly update the parameters in the model and use the simulation parallel to the real manufacturing system supporting and/or reinforcing the decisions on the shop-floor.

The reason of the intention to connect the scheduler to a discrete event simulator was twofold. On the one hand, it serves as a benchmarking system to evaluate the schedules on a richer model; on the other hand, it covers the non-deterministic character of the real-life production environment. Additionally, in the planning phase it is expected that the statistical analysis of schedules should help to improve the execution and support the scheduler during the calculation of further schedules

In the proposed architecture the simulation model replaces a real production environment, including both the manufacturing execution system and the model of the real factory.

Simulation also generates continuously new orders into the system, while these new orders are scheduled and released by the scheduler.

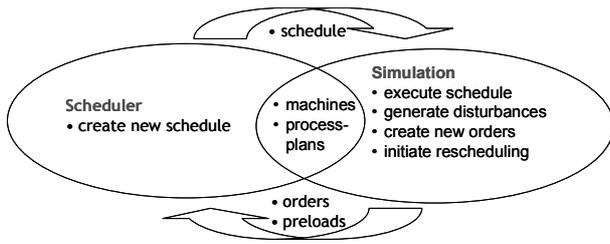


Figure 1. The rescheduling process initiated from the simulation side

The outline of the developed architecture is presented in Figure 1. Rescheduling action can be initiated when an unexpected event occurs or if a main performance measure bypasses a permissible threshold.

The dynamics of the prototype problem have been constructed to preserve realism as closely as possible and make the problem manageable for analysis.

This way simulation is capable for interaction with a specified scheduler, because all the required parameters are available any time for both systems, and so formulating an environment for further analysis on e.g. order pattern or sensitivity on significant parameters.

## PROPOSED METHOD

The study analysis the impact of the rescheduling interval and the rate of schedule modification on classical performance measures as system load, efficiency as well as stability in a single machine prototype system.

From the practical point of view in scheduling it is not possible to create schedules in every minute, however, the (theoretically) best performance of the whole system could be realized if schedule could be able to adapt to any changes, disruptions occurring in real-time. Most industrial planning and scheduling systems create schedules at idle time of the production e.g. at nights, while creating schedules for larger job-shop mostly requires a lot of computational time.

The process of modifying the schedule in the face of execution disruptions is referred to as reactive scheduling or rescheduling – detailed in the previous sections. The reaction to the realized disruption generally takes the form of either modifying the existing (predictive) schedule, or generating a completely new schedule, which is followed until the next disruption or rescheduling event occurs. The first technique is described in (Vieira et al. 2000; Rangsaritratamee et al. 2004), while the second is presented in (Bidot et al. 2003; Cowling and Johansson 2002). The importance of stability is outlined in the selected studies (see “monotonic and non-monotonic approach” in (Bidot et al. 2003), or “2D stability” in (Rangsaritratamee et al. 2004). The most important point is that while scheduling will optimize the efficiency measure, the strategy generates schedules that are often radically different from the previous ones. From the practical point of view the scheduling technique mentioned first seems to be better, while in industrial applications

constructing completely new schedules during schedule execution process must be avoided.

Schedule modification can be executed in given time periods (periodic rescheduling strategy), or related to specified events occurring during schedule execution (event-driven rescheduling strategy). Combining the two methods hybrid rescheduling strategy can be defined under which rescheduling occurs not only periodically but also whenever a disturbance is realized in the system (e.g. machine failures, urgent orders).

Define the time at which a new schedule is constructed as the rescheduling point and the time between two consecutive rescheduling points as the rescheduling interval (RI). At each rescheduling point, all jobs from the previous schedule that remained unprocessed are combined with the jobs that arrived since the previous rescheduling point and a new schedule is built.

In the previous sections the problem of DJSS and related stability measurements were introduced as well as the proposed simulation environment for dynamic rescheduling. In order to prove whether the rescheduling interval and the newly introduced variable schedule stability factor have a significant effect on schedule quality as well as stability an experiment for simulated single machine case was realized.

## Efficiency

The system to be scheduled is a single machine system with continuous job arrivals, but without any due date limitations. According to Baker (1974), the current scheduling problem can be classified as a single machine sequencing case with independent jobs and without due dates. In these situations the time spent by a job in the system can be defined as its flow time and the “rapid turnaround” as the main scheduling objective can be interpreted as minimizing mean flow time. The objective function is calculated as follows:

$$\bar{F} = \frac{1}{n} \sum_{j=1}^n (c_j - r_j) \quad (1)$$

where

$\bar{F}$  is the mean flow time

$n$  is the number of total arrivals

$r_j$  is the point in time at job  $j$  entered the system

$c_j$  is the completion time of job  $j$ , calculated when job  $j$  leaves the system

## Stability

In our study stability is calculated for each available job in the system during schedule calculation by giving penalty (PN) values, using the relation  $penalty = starting\ time\ deviation + actuality\ penalty$ . Starting time deviation is the difference between the start time of the job at the new and previous rescheduling points. Actuality penalty is related to a penalty function associated with deviation of the start time of the job from the current time. Penalty values are only calculated in case starting time deviation is greater than 0. A schedule with less penalty value can be considered as a

more stable schedule. The mean value of stability  $\overline{PN}$  is calculated for all schedules as follows:

$$\overline{PN} = \frac{1}{n_{pn}} \sum_{j \in B} \left[ |t'_j - t_j| + \frac{100}{\sqrt{t_j - T}} \right] \quad (2)$$

where

$B$  is the set of available jobs  $j$  that have not begun processing yet and  $|t'_j - t_j| > 0$   
 $n_{pn}$  is the number of the elements in  $B$   
 $t_j$  is the estimated start time of job  $j$  in the current schedule  
 $t'_j$  is the estimated start time of job  $j$  in the successive schedule  
 $T$  is the current time

### Schedule Stability Factor

When minimizing the objective function Equation (1), in a single machine case the optimal dispatching rule to be selected is SPT (shortest processing time) detailed in (Baker 1974). In the current case we use a truncated shortest processing time (TSPT) rule, in which the schedule stability factor (SF) can be introduced as the measure of the importance of schedule continuity or monotony. SF is the continuity rate of the schedule creation. In case SF equals zero, the new schedule may completely differ from the previous one, in case SF equals 1 the “old” jobs in the successive schedule must have the same position as in the previous one.

### Schedule Creation

SPT based scheduling means, that the priorities of the available activities are calculated by taking only the length of the processing time into consideration. On the other hand, the TSPT rule we introduce – see Equation (3) – generates schedules using SF in order to override the priorities of the activities given by the SPT rule, this way ensuring a more stable schedule. Each priority must have an integer value and it is calculated as follows:

$$prio'_j = (prio_j \times SF + prio_{j,SPT} \times (1 - SF))_{INT} \quad (3)$$

where

$A$  is the set of available jobs  $j$  that remained unprocessed in the previous schedule  
 $prio'_j$  is the modified priority of job  $j$  ( $j \in A$ ) in the successive schedule  
 $prio_j$  is the priority of job  $j$  in the previous schedule  
 $prio_{j,SPT}$  is the temporary priority of job  $j$  calculated using SPT rule

At each rescheduling point the following scheduling procedure is executed:

1. new jobs are added to set  $A$
2. create a priority list of jobs in set  $A$  by using SPT rule

3. compare current and previous priorities for “old” jobs and calculate new priorities by using Equation (3)
4. add remaining priorities to new jobs and sort the list by priority, calculate penalties by using Equation (2)
5. apply successive schedule and continue the schedule execution until the next rescheduling point defined by RI, then return to 1.

## ANALYSIS AND EXPERIMENTAL RESULTS

The above mentioned method was tested on a simulated single machine prototype system in order to measure the characteristics of stability measures in a simple environment.

The simulation system was developed using eM-Plant object oriented, discrete event driven simulation tool, which will be helpful during the extension of the current problem to larger, job shop problems.

In single machine case, minimizing mean flow time we applied SF and RI as input at given shop utilization levels. As output we considered  $\overline{F}$ ,  $n_{pn}$  and total penalty which is the sum of all  $\overline{PN}$  values multiplied by  $n_{pn}$  calculated at the end of each simulation run.

It was experimentally determined that the results from the first 2000 arrivals should be eliminated from computations to remove transient effects. Hence, each simulation run in this study consisted of 12000 arrivals of which the final 10000 were used to compute the performance and stability measurements reported. Each experiment was replicated 10 times to facilitate statistical analysis.

The interarrival time ( $b$ ), i.e. the average time between arrivals for jobs and are generated from exponential distribution with mean calculated using Equation (4):

$$b = \frac{\overline{p} \times n_o}{U \times m} \quad (4)$$

where

$\overline{p}$  is the mean processing time per operation  
 $n_o$  is the number of operations in a job, in the current case equals 1  
 $U$  is shop utilization level  
 $m$  is the number of machines in the system, in the current case equals 1

### Experiment 1

The main goal of Experiment 1 was to analyse the impact of system utilization level on  $\overline{F}$ , where SF was set to 0. Figure 2 shows, that both the system utilization and RI have a significant effect on  $\overline{F}$ . In the following experiment, where stability is examined we would like to use a relatively high utilization level in order to provide as much work-in-process as possible.

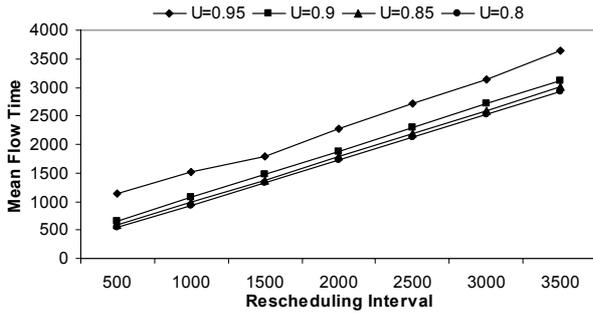


Figure 2. Effect of rescheduling interval and utilization level on mean flow time

As it is expected, extremely high utilization level lead to undesirable system instability, namely increasing the standard deviation of the resulted values and worsening the quality of the experimental results. The maximum acceptable value for  $U$  in the current case is 0.9.

### Experiment 2

The aim of Experiment 2 was to prove the assumption that applying the proposed stability criterion increases the stability of schedule execution however it reduces schedule efficiency. As a second scope of the experiment the effect of schedule stability factor on performance measurements was analysed.

In this experiment  $U = 0.9$  and  $p = 140$  with a triangle distribution  $\{140, 1, 300\}$ , then the mean of  $b$  equals 160. Three rescheduling interval were considered 500, 2000 and 3500 to have results from a wide range of RI. The second group of input parameters was SF, set to 0, 0.25, 0.5, 0.75 and 1.

As we assumed, the lengthening of the rescheduling interval increases stability but decreases the efficiency of the system. Figure 3 shows the illustrative results where SF was set to 0. Efficiency measurement  $\bar{F}$  is represented by the linear increasing dotted line, while the penalty values of the stability measurement are represented by the continuous line having a negative steepness. The penalty values decreased, because a higher number of modification made in the schedule at RI=500 with lower  $\overline{PN}$  values resulted a greater product than the same parameters at RI=3500.

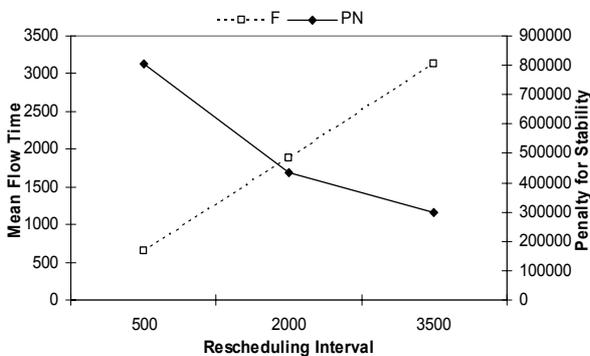


Figure 3. Effect of rescheduling interval on mean flow time and penalty values, in case SF=0

The effect of the parameter SF on penalty values given for stability and efficiency measurement  $\bar{F}$  at different rescheduling intervals are shown in Figure 4 and Figure 5.

Figure 4 shows for all RI curves, that the values increase in a monotonic way, i.e. increasing SF decreases system performance (increasing  $\bar{F}$ ) in each case. Comparing the results to SF=0, in case SF was set to 1, the outcome of the simulation showed an 8% increase of the performance measurement  $\bar{F}$ , in case RI was set to 500. Analyzing the other two cases, when RI was equal to 2000 and 3500, the performance of the system worsened only a few percent. Using these results it can be stated, that the negative effect of a higher SF level on  $\bar{F}$  decreases as the length of the rescheduling interval is growing.

On the other hand, penalty values decreased significantly at each RI (see Figure 5), because the higher SF values reduced the total  $\overline{PN}$  values, i.e. enabled less modification in the schedule.

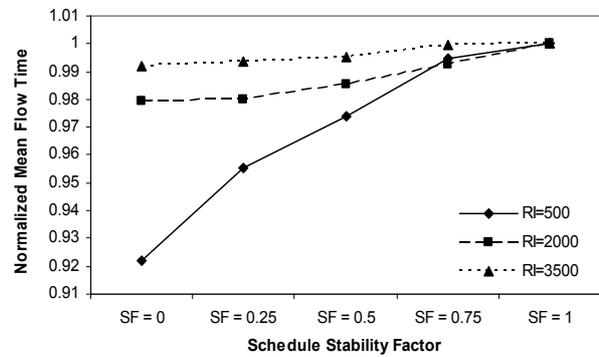


Figure 4. Effect of SF on normalized mean flow time at different rescheduling intervals

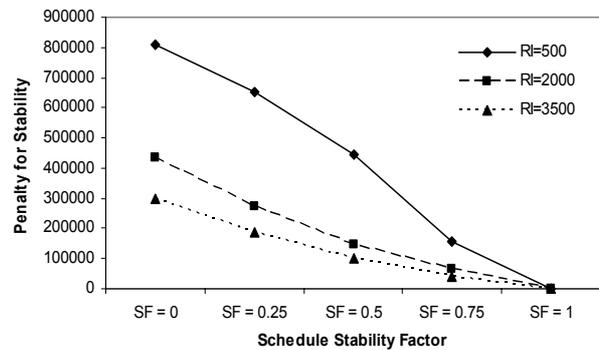


Figure 5. Effect of SF on penalty values at different rescheduling intervals

Comparing  $\overline{PN}$  values at different SF and RI parameter settings, it is interesting, that a penalty value given for SF=0 and RI=3500 is less than a penalty value for SF=0.5 and RI=500, while the efficiency is much better for RI=500.

Applying a limit for penalty values, e.g. let total  $\overline{PN}$  be ab.  $2 \cdot 10^6$ , then the optimal SF values can be selected for the given rescheduling intervals RI=500, 2000 and 3500. These values from Figure 5 are 0.7, 0.4 and 0.25 respectively.

## CONCLUSIONS

The paper discussed the job shop scheduling problem and schedule measurement techniques, especially outlining the methods that can be applied in a dynamic environment. The results of the simulation study based on the proposed architecture showed that both rescheduling interval and the newly introduced variable schedule stability factor have a significant effect on schedule quality as well as stability. In case applying limitations for stability, then for the given rescheduling intervals the optimal SF values can be determined. This significantly improves stability measurements but inconsiderably reduces system performance.

## FUTURE WORK

We would like to extend this experiment to a multi machine job shop system, using the results on stability gathered in this study. We propose a hybrid rescheduling strategy in a dynamic job shop environment defining two types of rescheduling events. The first type is done periodically (e.g. daily or weekly) using RI, releases new orders and involves tasks associated with order release. The second type is done when a disturbance occurs. It does not release new orders but instead reassigns work to off-load a down machine or utilize a newly-available one.

We assume that finding the appropriate schedule stability factor for each given rescheduling situation may result a compromise between the stable schedule execution and schedule quality.

## REFERENCES

- Baker, K. R. 1974. Introduction to sequencing and scheduling, John Wiley & Sons, USA.
- Baker, A. D. 1998. "A Survey of Factory Control Algorithms That Can Be Implemented in a Multi-Agent Hierarchy: Dispatching, Scheduling, and Pull". *Journal of Manufacturing Systems*, Vol. 17, 297-320.
- Bidot, J., P. Laborie, J. C. Beck, T. Vidal. 2003. "Using simulation for execution monitoring and on-line rescheduling with uncertain durations". *Proceedings of the ICAPS'03 Workshop on Plan Execution, Trento, Italy*.
- Church, L. K.; R. Uzsoy. 1992. "Analysis of periodic and event-driven rescheduling policies in dynamic shops". *International Journal of Computer Integrated Manufacturing*, Vol 5(3), 153-163.
- Cowling, P.; M. Johansson. 2002. "Using real time information for effective dynamic scheduling". *European Journal of Operational Research*, Vol 139, 230-244.
- Kádár, B. 2002. Intelligent approaches to manage changes and disturbances in manufacturing systems. PhD thesis. Technical University of Budapest, Hungary.
- Kádár, B.; A. Pfeiffer; L. Monostori. 2004. "Discrete event simulation for supporting production planning and scheduling decisions in digital factories". *Proceedings of the 37th CIRP International Seminar on Manufacturing Systems; Digital enterprises, production networks*, Budapest, Hungary, 444-448.
- Kempf, K.; R. Uzsoy; S. Smith; K. Gary. 2000. "Evaluation and comparison of production schedules". *Computers in Industry*, Vol 42, 203-220.
- Law, A.; D. Kelton. 2000. Simulation modelling and analysis, McGraw-Hill, 669-672.
- Rangaritratsamee R.; W. G. Ferrell Jr.; M. B. Kurz. 2004. "Dynamic rescheduling that simultaneously considers efficiency and stability". *Computers & Industrial Engineering*, Vol 46(1), 1-15.
- Szelke, E.; R.M. Kerr. 1994. „Knowledge based reactive scheduling state-of-the-art”. *Int. Journal of Production Planning and Control*, Vol 5 (March-April), 124-145.
- Szelke, E.; L. Monostori. 1999, Modeling Manufacturing Systems, Chap., Reactive scheduling in real-time production control. Springer, Berlin, Heidelberg, New York, 65-113.
- Vieira G. E., J. W. Herrmann, E. Lin. 2000. „Predicting the performance of rescheduling strategies for parallel machines systems”. *Journal of Manufacturing Systems*, Vol 19(4), 256-266.
- Williamson, D. P.; L.A Hall; J.A. Hooegeveen; C.A. Hurkens; J.K. Lenstra; S.V. Sevastjanov; D.B. Shmoys. 1997. "Short Shop Schedules". *Operations Research*, Vol. 45, 288-294.

## AUTHOR BIOGRAPHIES

**ANDRÁS PFEIFFER** received his M.Sc. degree in mechanical engineering in 2002, from the Technical University of Budapest. Currently he is a Ph.D. student at the Intelligent Manufacturing and Business Processes Research Group in the Computer and Automation Research Institute Hungarian Academy of Sciences (SZTAKI). His current interest includes production planning and scheduling, simulation and emulation of logistic systems. E-mail: pfeiffer@sztaki.hu, webpage: <http://www.sztaki.hu/~pfeiffer>.

**BOTOND KÁDÁR** is senior research fellow at Intelligent Manufacturing and Business Processes Laboratory of Computer and Automation Research Institute, Hungarian Academy of Sciences (SZTAKI). He obtained his M.Eng. and Ph.D degrees from the Technical University of Budapest, Hungary in 1993 and 2002 respectively. Since obtaining his PhD, he has been involved in research and development projects and education at Technical University of Budapest. His current interest includes production control, simulation and multi-agent approaches for production engineering and manufacturing systems. Dr. Botond Kádár is author or co-author of 55 publications. E-mail: kadar@sztaki.hu, webpage: <http://www.sztaki.hu/~kadar>.

**LÁSZLÓ MONOSTORI** has been with the Computer and Automation Institute of the Hungarian Academy of Sciences (SZTAKI) since 1977, now he serves as Deputy Director Research. He is also the head of the Department on Production Informatics, Management and Control at the Budapest University of Technology and Economics. He is an Active Member of the International Institutions for Production Engineering Research (CIRP) and Chairman of the Scientific-Technical Committee on Optimisation of Manufacturing Systems, Vice President of the International Society of Applied Intelligence (ISAI), Chairman of the Technical Committee on Manufacturing Modelling, Management & Control; International Federation of Automatic Control (IFAC). He is associate editor of Computers in Industry, CIRP Journal of Manufacturing Systems, IEEE Transactions on Automation Sciences and Engineering (T-ASE), and member of the editorial boards of some prestigious international scientific periodicals, such as the CIRP Annals. E-mail: laszlo.monostori@sztaki.hu, webpage: <http://www.sztaki.hu/~monostori>.

# BACKWARD SIMULATION IN FOOD INDUSTRY FOR FACILITY PLANNING AND DAILY SCHEDULING

Tom-David Graupner  
Matthias Bornhäuser  
Wilfried Sihn

Fraunhofer Institute for Manufacturing Engineering and Automation IPA  
Nobelstraße 12, 70569 Stuttgart  
Germany  
E-mail: {tdg, mab, whs}@ipa.fraunhofer.de

## KEYWORDS

(Backward) Simulation, Food, Facility Planning, Capacity Planning, Daily Scheduling

## ABSTRACT

In the light of high capital expenditure for manufacturing facilities in food and process industry, an increasing number of variants and frequently changing sales developments, simulation models are becoming more and more important for the support of capacity and production planning.

In the following, the use of a simulation toolset for the support of tactical capacity planning and operative production planning is presented. It was implemented in cooperation with a global player of food industry.

The challenge of the project was the implementation of a backward simulation. Instead of filling food from a tank into cups, the computer simulates the process inversely. The advantage of the chosen procedure is that the provision of all components of a dessert or yogurt product at the time of filling as well as an efficient utilization of resources is ensured.



Figure 1: Picture of a production line

## INTRODUCTION

New products and an increasing number of variants define the requirements to production planning in food industry. In the milk processing industry, production planning and production control are particularly complex as the sequence in which aggregates are provided with different products is highly influential on cleaning effort and therefore on the productivity of the facility. Due to storage life restrictions, pre-defined periods of time for aggregate occupancies have to be respected in the production process. In addition, the products are produced at the latest moment possible in order to meet the retailers' and consumers' demands for fresh products.

Statements whether the facilities can produce the forecast order size on schedule or capacity adaptations are necessary were difficult to make up to now. For production planning, there were no software tools at the orderer's disposal which made it possible to calculate aggregate occupancies on the basis of weekly bottling plans (figure 1) [1, 2].

## THE SIMULATION PROJECT

Our customer decided to approach the problem by representing the existing dessert and yogurt production in a simulation model. Thus, the task was to conceive and develop a flexible simulation model and to train the employees of the company in handling the simulation tool.

The existing facility and its aggregates, facility topology and product-specific facility occupancies was represented in the simulation model. Aggregate- and product-specific cleaning restrictions were included in the model. The aggregates are at the simulation model's command in the form of a library. The advantage of doing so is the possibility to transfer this library into other simulation models. Thereby, development time for new facility topologies of comparable facilities can be significantly reduced.

The simulation model is parametrized by an Access data base. Specific know-how about the simulation software therefore is not necessary for the application as it is controlled via the data base. The master data of the aggregates, products, restrictions and facility topology are stored there and can be modified by the user. In addition, even the strategies for facility occupancy can be influenced by the user via

the data base. The existing control logic, which guarantees an optimum load of the filling lines and a resource-saving aggregate occupancy, are accounted for in the simulation model.

## THE YOGURT AND DESSERT PRODUCTION

The production of fresh milk products is a multi-level process. In the following, this process is presented very roughly only, as the process wanted to be kept secret by our partner.

1. Preparation of mixing = Pumping milk, sugar, vanilla powder from storage containers via pipelines into mixing tanks.
2. Mixing = The ingredients are mixed during a certain period of time.
3. Storing in tanks = The mixed matter is stored until a production line is prepared.
4. Production = The dessert (cream) matter is heated via pipelines and afterwards chilled again.
5. Incubating = The heated yogurt matter is filled into tanks. Subsequently, a bacteria culture is added and the matter has to mature for about 12-36 hours.
6. Chilling = The yogurt matter is chilled at a cooler to slow down bacteria activity.
7. Storing = The finished dessert matter is stored until the filling facility is prepared.
8. Filling = The dessert matter is frothed up with whiskers and then filled into cups.

As it becomes clear, it is a long way from milk to yogurt and dessert. Several aggregates, each with specific properties as throughput, capacity, cleaning time etc., are necessary to run this process.

- Tanks: Filling tanks, incubation tanks, mixing tanks with their properties as mass flows, volume, cleaning restrictions.
- Filling lines with the properties throughput in cups/h, cleaning restrictions, changeover.
- Production lines with their properties throughput in liters/h, sterilization processes, line occupancy.
- The yogurt and dessert matter which is continuously transforming during the process from raw milk to the cup.

Dessert and yogurt products mostly consist of 2-3 components (semi-finished products), which – in different flavours – are filled into cups of different dimensions. Thus, two semi-finished products, e.g. yogurt and strawberry, have to be filled synchronously. Though, the production start of the semi-finished products thereby is totally different.

The challenges in daily planning are the following:

- The lines should never stand still as they are deterministic for the throughput.
- Cleaning is to avoid as this is waste of material and expenditure.
- The number of tank occupancies should be as small as possible in order to economize cleaning effort.
- Simultaneous filling and draining of tanks cuts down throughput time.
- Combining filling orders of the same flavor which are filled into different country-specific packings in the last production step.

## CONTINUOUS PROCESSES IN THE DISCRETE SIMULATOR

In discrete simulation, only defined points of time on a fictive time bar are taken into account. Discrete here means isolated resp. disjointed. When a mobile object enters a resource, the simulator puts a mark shifted of the processing time. The administrator of the incident, who controls all chronological simulation activities, subsequently jumps directly to the next discrete point on the time bar and executes the given orders.

The production considered here is a batch-oriented production. But many decisions go on at discrete points of time so that modelling takes place both batch- and incident-oriented.

**Incident-oriented:** In the receptacles, the semi-finished and finished products are represented as single objects (object with one volume resp. one mass). Advantage: easy batch management in the simulator. Realization of a decentral control.

**Continuous:** In the continuous processes (e.g. production line), there is a flow-oriented representation, i.e. no disintegration of batches into a multitude of objects, but an exact, mathematically correct representation. Advantage: higher simulation speed.

As control and scheduling logic made up for the main complexity in the project, a discrete simulator was used in the presented project and the continuous processes were emulated.

## SCHEDULING LOGIC

In this paragraph, it is shown why a pure forward simulation is not target-oriented. The backward simulation-based approach was realized [3, 4, 5].

### Forward simulation-based approach

By forward simulation, we understand the simulation of production orders in the direction of the material flow. Starting from the start date, the production orders are dispatched to production until they leave production after the processing.

For a better illustration, a production and assembly process is represented in figure 2. A module consists of three components. Each sphere symbolizes a process step. When the three components are completed, assembly can begin.

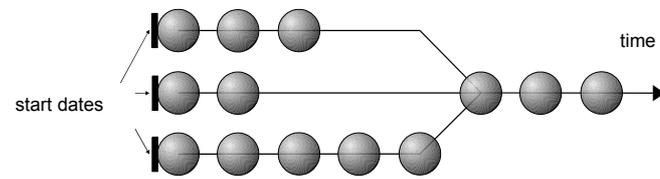


Figure 2: Forward simulation

A weak point of pure forward simulation is that only in a special case all components are completed at the same time. In figure 2 unwanted waiting time arises for the components in the two upper rows.

In order to minimize throughput time, the start dates of the components can be varied. When the processing procedures interact, it is not sufficient to only begin the start times delayed by the waiting times. Such an interaction exists when the processing procedures of different components are executed on the same resources.

A backward simulation is recommended in order to minimize unwanted waiting time.

### Backward simulation-based approach

The backward simulation corresponds to the inversion of forward simulation. The clocks quasi run backwards. The finished products are “decomposed” into their components in the course of time. In figure 3, the “decomposition” of a finished product into its components is depicted. As start date of simulation, the delivery date is chosen.

- For each work step we get the latest start and finish date. Thus, short throughput times can be obtained. When a work step starts later than the corresponding basic date, this unavoidably leads to scheduling delay.
- The completion date of the last work step can be used as start date for forward simulation.

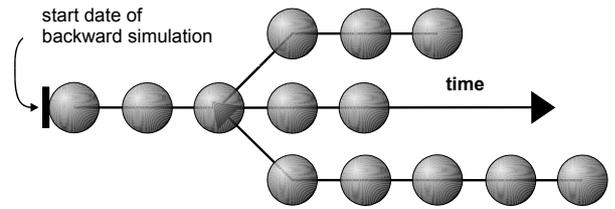


Figure 3: Backward simulation

In some cases, the backward simulation is insufficient for complete scheduling of dessert production as not all processes can be exactly inverted. For instance, many restrictions can only be formulated forwardly. Furthermore, in a pure backward simulation no statements about receptacle occupancies etc. can be made. Thus, subsequent to the backward simulation, a forward simulation is executed.

### Forward simulation (subsequent to backward simulation)

The above-mentioned cognitions show that neither a pure backward simulation nor a pure forward simulation bring about satisfactory results. When executing a forward simulation after a backward simulation and taking over the start dates of the components, the advantages of both simulations are unified (figure 4).

By this combined backward-forward simulation, advantageous start dates of the components are obtained whereby short throughput times result. The modeling discrepancies which may occur because of backward simulation can be compensated by a forward run that can be exactly modeled.

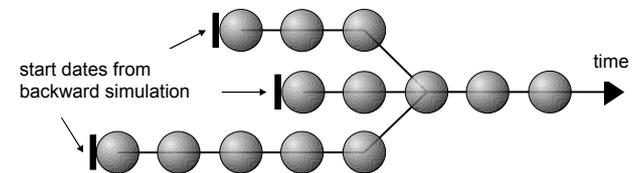


Figure 4: Backward-forward simulation

## REALIZED SCHEDULING LOGIC

The starting point of utilization planning are the delivery dates of customers' orders. Hence, the desired finish dates of a filling order resulting thereof are taken as starting point for the simulation. The decisive question is at which point of time the first process step has to be started at the latest.

This task is solved by undertaking the aggregate occupancies backwards, starting from the last process step "filling" at the desired point of time. The material flow and the time recording comport inversely to reality in the simulation model. A disassembly of yogurt into its basic ingredients can be clearly imagined.

The chosen approach has the advantage that provision of all ingredients of a dessert or yogurt product at the time of filling as well as an efficient resource utilization are ensured. For instance, a classic product of our partner from food industry consists of pudding and cream. It has to be ensured that, at the time of filling, enough pudding and cream are provided in tanks.

The challenge for the developer team in the project therefore always consisted in thinking "backwards" in relation to the production run. After some doubts at the beginning it became clear that rules and restrictions of any kinds in the production run are reversible.

## THE CONFIGURABLE SIMULATION TOOLSET

### Higher planning security in tactical capacity planning

A requirement of our industrial partner was to make reliable statements for tactical capacity planning. The requirement was fulfilled by parametrizing all important master data in the model via the data base. The capacity of an aggregate can be varied, an aggregate blocked and a thitherto in the real facility non-existing "virtual" aggregate can be added. Cleaning restrictions can be abolished or tightened or the facility topology, i.e. the aggregate sequence, can be modified.

The user has the possibility to act out manifold scenarios in the model, e.g.:

- "Can the order load also be produced at the right time with one tank less?",
- "By how many tons can the filling quantity be augmented when investing in one additional tank?" or
- "What influence does easing of a cleaning restriction have on load and productivity?"

So as to answer these questions, the user can dispose of a detailed automatically generated evaluation of a simulation run. For a fast analysis, the user can consult load diagrams of the aggregates generated in HTML (figure 5). For a detailed evaluation, manifold key figures are calculated and the complete occupancy is shown in a Gantt diagram. In this Gantt diagram, the aggregate occupancy in the course of time is

depicted graphically. The aggregate occupancy of an order can be retraced in this depiction.

The model had to be proved shortly after completion for real planning tasks. One question was "What influence do new product variants have on my facility?" resp. "Are the existing capacities sufficient for the forecast sales development?" Furthermore, the consequences of a planned capacity expansion were analyzed.

Within a short period of time, these questions could be answered comprehensibly. Capacity bottlenecks could be identified by the aid of evaluation facilities and the eventual necessary capacity adaptation could be found and proved.

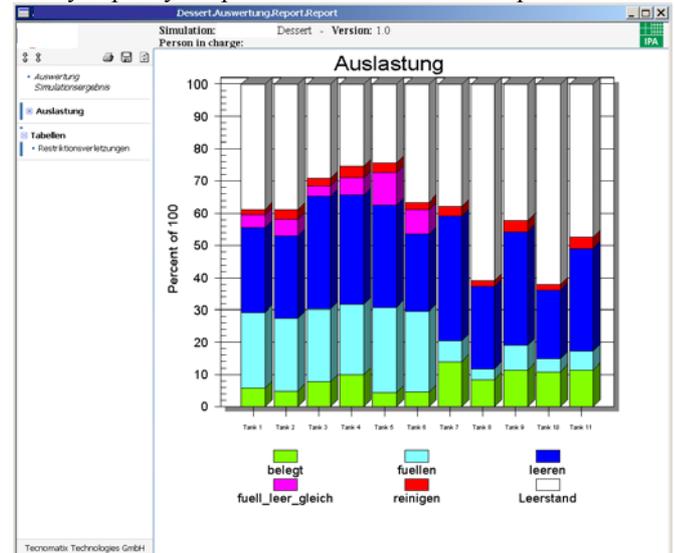


Figure 5: Evaluation of a simulation run – load

### Higher quality and less effort in operative production planning

Another requirement was supporting operative production planning.

Validation of the results showed a discrepancy of the simulation results of less than 5% from reality. The discrepancies are related to the highly fluctuating efficiencies of some aggregates.

This led to high acceptance of the employees in production planning. Another benefit is the short duration of a simulation run. Weekly planning on a standard computer lasts less than three minutes.

Production can be controlled by defining the latest possible start dates for production orders and detailed aggregate occupancy in the Gantt diagram (figure 6).

Furthermore, the user can check the influence on the facility occupancy by exchanging orders and thus, by trying, carry out a sequence optimization of order processing.

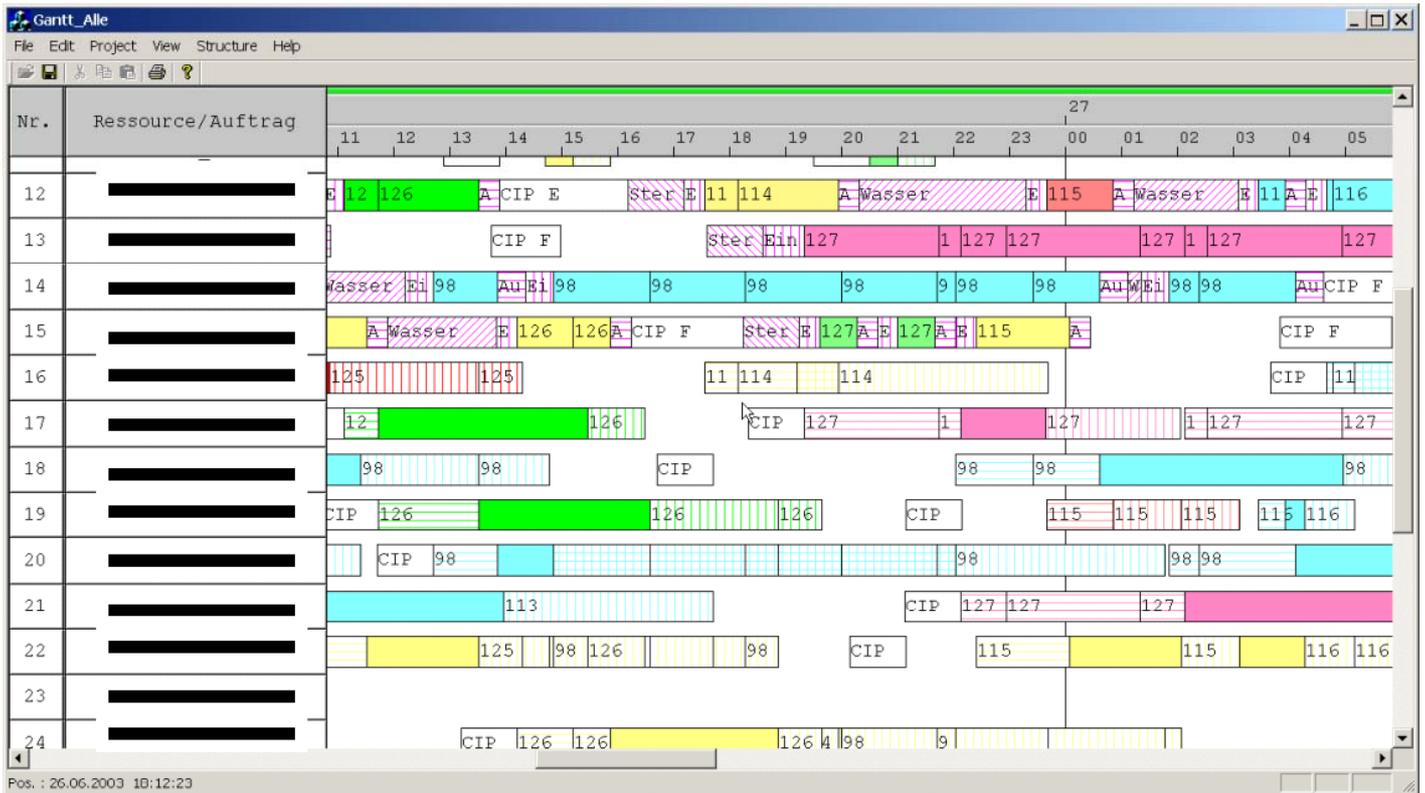


Figure 6: Evaluation of a simulation run – Gantt

## BENEFIT FOR OUR CUSTOMER

A requirement of our industrial partner was to make reliable statements for tactical capacity planning. The user gets the chance to do so by configuring the production, acting out scenarios in the model and analyze them by the aid of detailed evaluations.

Capacity bottlenecks can be identified by means of evaluations and necessary capacity adaptations can be found and proved.

The simulation model has also been used for other plants since it has been completed. Its modular construction makes possible – with little effort only – to picture other plants, too.

The project shows the considerable use of simulation models in automated process industry. It is obtained by higher planning security, quality and speed. Thereby, the ability to quickly react to altered general conditions is increased.

## CONCLUSION

The project shows the considerable use of simulation models in automated food industry. It is achieved by higher planning security, high quality in production planning and an increase in planning speed. The ability to quickly react to altered general conditions is increased.

Withal, development costs of a simulation model mostly only make up for a trickle of the costs caused by a misinvestment.

## REFERENCES

- [1] Sihn, Wilfried; Richter, Hendrik; Graupner, Tom-David:  
Projektierung von Fertigungssystemen durch Konfiguration, Visualisierung und Simulation.  
In: Schulze, Thomas (Hrsg.) u.a.; Otto-von Guericke-Universität Magdeburg / Institut für Simulation und Graphik u.a.: Simulation und Visualisierung 2003: Proceedings der Tagung am Institut für Simulation und Graphik der Otto-von-Guericke-Universität Magdeburg am 6. und 7. März 2003.  
Gent, Belgien: SCS Europe, 2003, S. 9-19
- [2] Graupner, Tom-David:  
Anlagen zeigen schon auf dem Computer ihr ganzes Können.  
In: Metallbearbeitung Deutschland (2003), August, S. 40-41
- [3] Frank, Andreas; Schulte, Jörg:  
Umkehrung von Materialflussmodellen der Vorwärtssimulation für die Rückwärtssimulation zur Terminierung von Produktionsprozessen.  
In: Kampe, Gerald (Hrsg.) u.a.; Gesellschaft für Informatik / Fachausschuss 4.5: Arbeitsgemeinschaft Simulation u.a.: Simulationstechnik (9. Symposium): 9. Symposium in Stuttgart, 10. bis 13. Oktober 1994, Tagungsband.  
Braunschweig; Wiesbaden: Vieweg, 1994, S. 607-612 (Fortschritte in der Simulationstechnik 9)
- [4] Schuler, Klaus; Becker, Bernd-Dietmar; Schulte, Jörg:  
Fertigungsterminierung durch Simulation: ein Anwendungsbeispiel.  
In: Tavangarian, Djamshid (Hrsg.); Gesellschaft für Informatik / Fachausschuss 4.5: Arbeitsgemeinschaft Simulation u.a.: Simulationstechnik (7. Symposium): 7. Symposium in Hagen, September 1991. Braunschweig: Vieweg, 1991, S. 33-44 (Fortschritte in der Simulationstechnik 4)
- [5] Aarts; Lenstra (Editors): Local Search in Combinatorial Optimization. John Wiley & Sons Ltd., 1997

# **SIMULATION IMPROVES SERVICE AND PROFITABILITY OF AN AUTOMOBILE SERVICE GARAGE**

Navin Gupta  
Department of Industrial Engineering  
2157 Manufacturing Engineering Building  
4815 Fourth Street  
Wayne State University  
Detroit, Michigan 48202 U.S.A.

Edward J. Williams  
Management Information Systems & Decision Science  
School of Management  
University of Michigan – Dearborn  
B-14, Fairlane Center South 19000 Hubbard Drive  
Dearborn, Michigan 48128 U.S.A.

## **KEYWORDS**

Service operations, process improvement, process management, discrete-event process simulation

## **ABSTRACT**

Simulation has long been an analytical tool of significant importance and power for process improvement. Historically, the earliest and most widespread uses of simulation were in manufacturing industries; however, it was not long before the power of simulation was applied to improve productivity and assess the relative merits of process change alternatives within various service industry segments such as travel, hotel and restaurant, retail stores, and entertainment venues such as theatres and amusement parks. The study described in this paper describes the successful application of simulation to process management and improvement within a business devoted to aftermarket repair of privately owned automobiles and trucks. We describe the problems encountered by the client and how the simulation study illuminated a pathway to significant improvements in customer service and financial profitability.

## **INTRODUCTION**

Discrete-event process simulation originally proved its worth and power as a process improvement tool within the manufacturing sector of the economy (Miller and Pegden 2000). Somewhat more recently, simulation has likewise become highly respected, and its use widespread, in various service industries (Herbst, Junginger, and Kühn 1997). Indeed, a variety of published results attest to the value of simulation within the service sector of the economy. For example, (Pichitlamken et al. 2003) used simulation to analyze a telephone call center handling both inbound and outbound traffic. (Palacis 2000) described the use of simulation to improve the business processing of accounting transactions within supply chains in the timber industry. (Nanthavanij et al. 1996) described an application in which simulation was used to improve services provided by car-park systems.

In this study, the client was the management of a repair and service shop for privately owned vehicles; this shop provides repair and replacement service for exhaust systems, brakes, steering, suspension, and climate-control vehicle systems. The numerous franchised and licensed shops do business in locations as widely dispersed as Brazil, Morocco, New Zealand, and Spain, in addition to locations within the United States, the country of origin of the business nearly half a century ago. Deteriorating economic conditions among consumers, as have prevailed recently within the United States, provoke postponement of new vehicle purchases and hence increase demand for aftermarket repairs of increasingly used vehicles (Nash 2003).

At the particular franchise location in question, management had noticed significant declines in productivity, efficiency, and profitability accompanied by an increase in operational costs, particularly labor costs. Furthermore, increasing service and waiting times, coupled with deterioration of service quality, were eroding consumer goodwill. As examples, more than 12% of total labor time was expended on rework, the number of customers served had just suffered a 23% year-to-year decline, a large work backlog routinely occupied nearly 75% of the available floor space, and, due to overtime, total labor cost was increasing markedly. The urgency of these problems impelled the managers to seek counsel and recommendations for improvement.

## **PROCESS ANALYSIS**

Spurred by this urgency, client management and the analysts first identified four performance metrics:

1. Overall productivity = labor hours sold ÷ available hours
2. Labor utilization = actual hours worked ÷ available hours
3. Customer satisfaction, as measured by surveys given all departing customers
4. Overtime = hours required to service all customers – regularly scheduled working hours

All these metrics were strongly and directly related to the problems provoking the study.

Next, with extensive help from client management, the analysts constructed a process map. As background, services may be conveniently classified in either of two ways. From the customer's viewpoint, service sought is either periodic preventive maintenance (keep the vehicle in good operating condition) or demand maintenance (restore the vehicle to good operating condition). From the service providers' viewpoint, service is either a minor repair (short duration) or a major repair (long duration). Typically, minor repairs are inspections; major repairs are replacement or restorative work to one or more of the vehicle systems listed previously. The process map comprised seven primary operations:

1. Reception (greet the customer and inquire into the motivation for the customer's visit),
2. Inspection (examine the vehicle to detect all problems meriting attention, possibly extending and/or revising the work deemed necessary at the initial reception),

3. Customer approval (estimating the repair cost and obtaining the customer's approval to undertake the work),
4. Classification of repair (into major or minor; occasionally, both types are required for one vehicle),
5. Performance of repair(s),
6. Final inspection (possibly requiring a short test drive, and certainly requiring cleaning the vehicle of grease or smudge),
7. Invoicing and billing (collecting payment from the customer and returning the vehicle to the customer's custody and use).

Next, the project team (client management and analysts) developed fishbone diagrams (Stevenson 2005) to identify direct causes of the problem. These causes were identified as low productivity and low customer satisfaction, as shown in Figures 1 and 2 below.

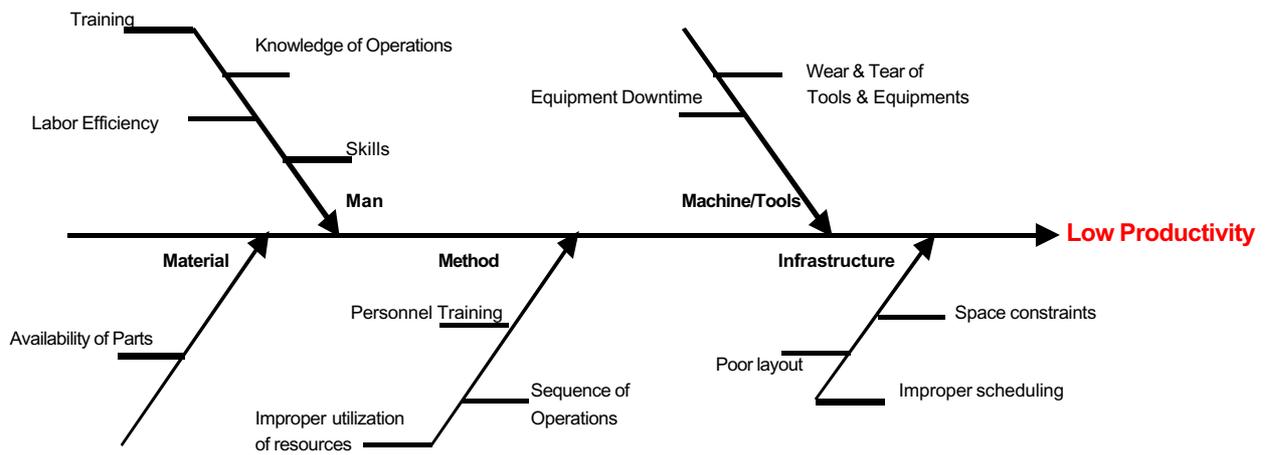


Figure 1. Low Productivity Fishbone Chart

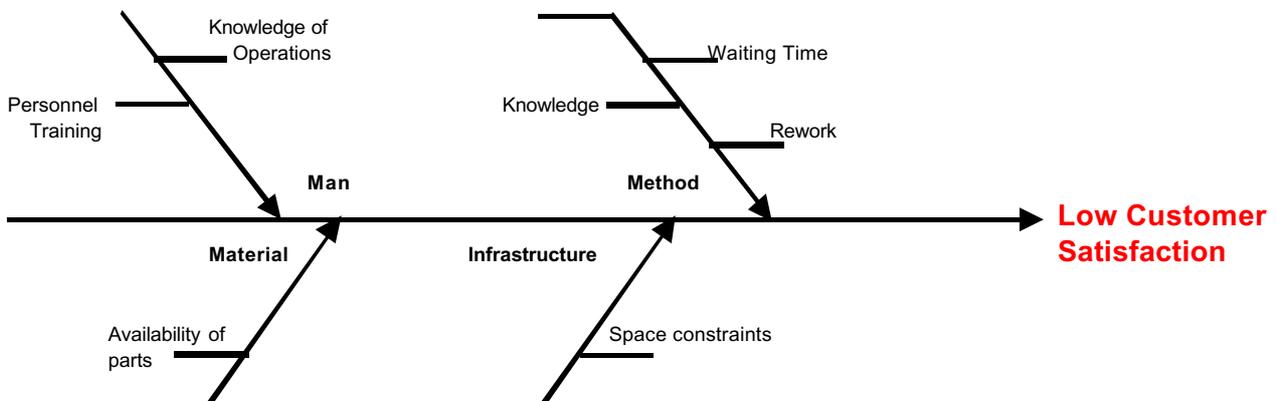


Figure 2. Low Customer Satisfaction Fishbone Chart

After these fishbone diagrams were used to identify these problem causes, preparation of a diagnostic chart (Figure 3, below) helped identify the primary controllable factors. The cause-&-effect mapping was iterated until the primary controllable factors were identified as: inefficient and undisciplined scheduling, inadequate operator training, and sporadic and ineffective equipment maintenance.

## DATA COLLECTION

Extensive data collection efforts examined (a) historical data spanning the six months immediately prior to the beginning of the study, (b) direct observational data covering ten complete workdays (two per week for five weeks), and (c) a Customer Satisfaction Survey personally given to and retrieved from every customer during those ten sampled work days. Therefore, the last two phases of data collection entailed 100% sampling. Process times were collected for all seven phases of the process map above. Data were fitted to closed-form distributions whenever appropriate, using the distribution fitter (“Input Analyzer”) within the Arena® simulation software (Bapat and Sturrock 2003). The Weibull, triangular, uniform, or normal distributions fitted most observational data sets well, with  $p$ -values (for rejection of  $H_0$ : “Data set plausibly comes from fitted distribution”) typically greater than 0.1, and often greater than 0.4. When  $p$ -values fell below 0.1, the observed data were used directly via an empirical distribution.

## THE SIMULATION MODEL AND ANALYSES

The simulation model itself was built using Arena®, a well-known and versatile simulation software package used at both Wayne State University and the University of Michigan – Dearborn for the teaching and practice of discrete-event simulation modeling (Kelton, Sadowski, and Sturrock 2004). This model routinely used standard Arena® modules such as *Create* (customers’ vehicles enter the system), *Dispose* (customers’ vehicles leave the system), *Process* (vehicles undergo evaluation, repair, or inspection), or *Assignment* (attributes such as type of repair, cost of repair, and duration of repair are assigned to a vehicle). Since the model was built, verified, and validated as a team effort, techniques such as structured walkthroughs (Weinberg 1971), tracing of individual entities, deterministic runs in which all distributions were replaced by their means and results compared with spreadsheet computations, degenerate tests, and extreme condition tests were applied early and often (Sargent 2003). The versions of the model considered the base case (current operations) and four alternatives:

1. Use of an appointment system – customers are expected to make appointments in advance.

A “Questionnaire Form” was used to evaluate the customers’ opinions concerning implementation of an appointment system and also to find the preferred time slot(s) for servicing. Implementing an appointment system would presumably guarantee proper workflow of the overall system, which in turn would ensure proper utilization of resources. Hence an appointment system would be expected to improve productivity, efficiency of labor utilization, and customer satisfaction.

2. Additional training of service personnel.  
With the aim of increasing personnel efficiency, a vigorous personnel-training program could be implemented on a periodic basis. Such an implementation presumably would decrease rework and excessive “work in process”. These decreases would then improve utilization of resources, reduce waiting time of the customers, and thereby improve the overall productivity, labor efficiency, and customer satisfaction.

3. Establishment of formal preventive and predictive maintenance procedures.  
In this scenario, the service personnel would be required to follow a daily cleaning schedule of tools and equipment prior to the end of the day. In addition to this expectation, a weekly preventive/predictive maintenance would be scheduled. Implementation of this alternative presumably would decrease equipment downtime and work in process, in turn leading to better utilization of resources, reduced customer waiting time, and increased efficiency and customer satisfaction.

4. Hiring additional service personnel.  
The number of bays cannot be increased due to space constraints, but in this scenario additional employees would be hired on a permanent basis. From a deterministic spreadsheet model, the investigators identified major repairs as the bottleneck; an additional employee already having the required skills to undertake such repairs could be hired. With the help of the additional worker, the process would presumably flow more smoothly than before, and this alternative should also reduce excessive WIP as well as overburden on employees. It could also be expected to improve customer satisfaction by reducing waiting times.

The first three alternatives represented direct responses to the primary controllable factors previously identified during process analysis. Modeling alternative #1 entailed making the arrival rate of customers nearly constant. For alternative #2, mean cycle times for both major and minor repairs were reduced to 90%-92% of their base case values, and rework was likewise decreased by 8%. For alternative

#3, mean times between failures were increased by 30%, mean times to repair were decreased by 25%, and the preventive/predictive maintenance was scheduled to last ½ hour after every 40 hours of operation. For alternative #4, one additional Arena® Resource, representing an employee sufficiently skilled to undertake both major and minor repairs, was added to the bottleneck process, i.e., Major Repair.

The assumptions that alternative #2 (training) would reduce repair times to 90%-92% of their baseline values, and that rework would decline by 8%, was acknowledged as uncertain by both client management and the simulation analysts. These assumptions emerged from the client’s estimates, and those estimates in turn emerged from discreet observations of both fraternal and competitive franchises in the aftermarket vehicle repair business. Certainly it is widely recognized that predictions of training effectiveness are difficult to make and inherently uncertain (Wickens, Gordon, and Liu 1998). Likewise, the predictions that devoting 1¼% (½ hour

of every 80 hours) of potential work time to preventive maintenance would extend mean times to failure by 30% and decrease repair times by 25% were hazy. These estimates also emerged from client estimates, and additionally corresponded reasonably well with case studies cited by (Leemis 1995) extolling the values of preventive maintenance. The second author vividly remembers hearing a wise supervisor and mentor remark years ago (Crabb 1975), “It’s amazing what preventive maintenance will prevent.” Due to the uncertainty surrounding these estimates, all of them were examined in detail via sensitivity analyses, both for verification and validation and to assess the degree to which model results depended on the accuracy of these assumptions (Balci 1998).

The various alternatives called for changes in the ongoing process, organizational structure, infrastructure and technology used in the organization. The anticipated changes with respect to various alternatives are shown in Table 1:

<b>Alternatives</b>	<b>Process</b>	<b>Organizational</b>	<b>Technology / Infrastructure</b>
<b>AS-IS</b>	No Change	No Change	No Change
<b>Appointment System</b>	No Change	An additional responsibility of making all appointments as well as maintaining documents, keeping track of Appointments.	There will be an additional requirement of some kind of scheduling software, which will help in implementing Appointment System.
<b>Training of the Service Personnel</b>	Vigorous Personnel Training program is implemented on a periodic basis.	The Management is required to arrange a Training Program for the service personnel as well hire certain instructor who trains the employees.	Training Manuals/documentations/ visual aids are to be generated to aid in training program.
<b>Preventive &amp; Predictive Maintenance</b>	The Service Personnel are required to follow a daily cleaning schedule of tools and equipments prior to the end of the day. In addition to this a weekly preventive / Predictive Maintenance is Scheduled.	Responsibility is to be given to the Service Personnel to clean the tools & equipments prior to the end of the day.	The Service Station has to invest in the purchase of necessary Tools to perform Preventive / Predictive Maintenance.
<b>Hire Additional Service Personnel</b>	No Change	An additional labor cost is incurred, and also the organization should ensure that the new hired employees are skilled.	No Change

Table 1. Anticipated Process, Organizational, and Technology Changes under Various Scenarios

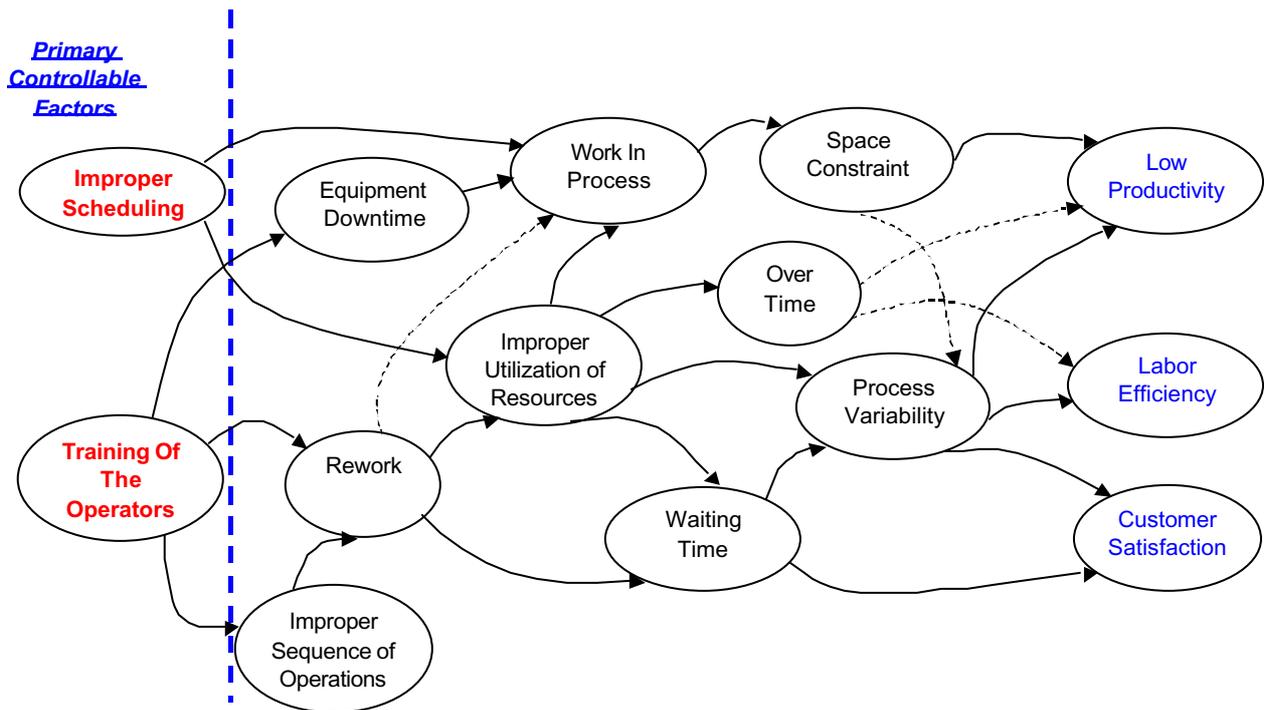


Figure 3. Diagnostic Chart

## RESULTS AND CONCLUSIONS

Since the repair shop runs five eight-hour days per week plus necessary overtime to complete that day's incoming work, it was modeled as a terminating system. All alternatives were run for ten replications, and each replication comprised ten eight-hour days (two calendar weeks), thereby permitting the routine construction of confidence intervals using the Student-*t* distribution inasmuch as results from these replications were pairwise independent.

Mean performance metrics for the base case and the five alternatives were:

Model	Productivity	Overtime (hours/week)	Labor Utilization
Base case	1.38	15.4	0.39
Appointments	1.61	16.6	0.37
Training	2.10	7.8	0.38
Maintenance	1.85	13.1	0.39
New Service Personnel	2.09	11.0	0.45

Table 2. Average Results for Five Scenarios

Various graphs (Figures 4, 5, and 6 below) were plotted to assist client engineers and managers in comparing the different alternatives against the performance metrics of productivity, overtime, and labor utilization.

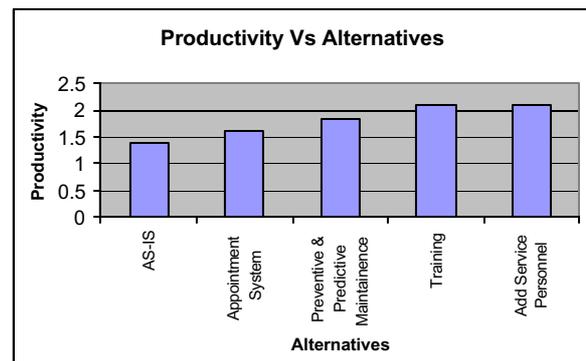


Figure 4. Predicted Productivity Under Various Alternatives

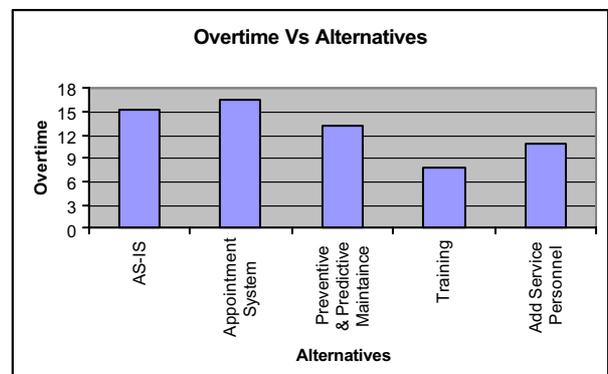


Figure 5. Predicted Overtime Under Various Alternatives

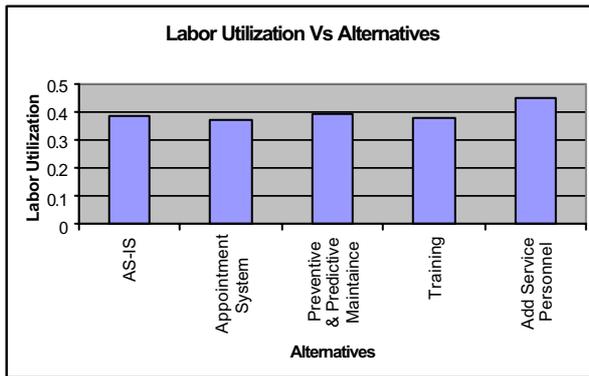


Figure 6. Labor Utilization Under Various Alternatives

It is particularly notable that adding an employee (alternative #4) increased labor utilization; typically, exactly the opposite would be expected. The franchise managers were intrigued to learn from the model that the bottleneck (major repairs) would be sufficiently ameliorated by addition of one worker to increase the utilization of many other workers by reducing their time spent waiting for tasks to arrive, and hence improving overall workload balancing.

Attractive as this improvement in utilization appeared, the client management was even more enticed by the promise of alternative #2 (training), which the analyses predicted would increase productivity as much as adding a new employee, achieve a marked reduction in overtime, and maintain overall labor utilization. The overtime reduction was important to both management (containment of costs) and the workers (greater predictability of their work schedules and hence freedom to schedule activities within their personal and family lives). Both managers and workers realized that the reduced overall amount of overtime could more readily be assigned to those workers more eager to work it – workers' attitudes toward overtime, as is typical in a diverse workforce, varied across the entire spectrum from alacrity to revulsion. From an economic viewpoint, training (as compared to addition of a worker) would increase, rather than decrease, the franchise's ability to adopt to inevitable fluctuations in amounts and types of repair services demanded by the marketplace. Furthermore, almost all workers viewed the training as an enhancement of their long-term employability and as evidence of the willingness of management to invest in them.

Thorough analysis of the criteria must precede making a decision; making the decision then requires consideration of various economic and stochastic scenarios. Cost Benefit Analysis gives us the best economically sound alternative, whereas Sensitivity analysis is used to check the *robustness* of the best alternative.

The factors considered for Cost Benefit Analysis were:

1. Overall Productivity
2. Labor Efficiency
3. Overtime
4. Total Cost

On the basis of Cost Benefit Analysis, it was found that Alternative 2 (Training of Service Personnel) was the best alternative.

Sensitivity analysis of the alternatives was done by increasing & decreasing the values of the factors considered for Cost Benefit Analysis. On the basis of Sensitivity Analysis, it was found that Alternative 2 (Training of Service Personnel) was also the most *robust* alternative.

Therefore, management selected the training alternative for immediate implementation. Data on the four performance metrics were collected again for a six-month period, beginning three months after the training ended (to allow it to achieve full effect). These metrics showed improvements consistent, to within 6%, of those predicted by the simulation analyses of alternative #3. Furthermore, both semi-formal surveys of employee morale and informal observations of it (e.g., noting a reduction in employee turnover) showed improvement, and management felt justified in attributing at least a portion of this improvement to the increased security the employees felt in their careers and the decreased frequency with which unwilling employees had to be dragooned into working unwanted overtime.

In conclusion, we remark that the thorough process analysis and mapping completed before construction of the simulation model, and used to identify alternatives worthy of modeling and analysis, closely match the recommendations of (Eldabi, Lee, and Paul 2003) relative to business process simulation. Certainly nothing inherent in the training alternative precludes the additional alternatives of increasing preventive maintenance and/or hiring an additional worker; therefore, the client's involvement with simulation has extended itself into further investigation of these strategies.

## ACKNOWLEDGMENTS

The first author gratefully acknowledges the mentorship and guidance of Professor Olugbenga Mejabi of Wayne State University and the project assistance of colleagues Saravanan Selveraj, Mayo Santosh, and Asad Sami. The second author likewise gratefully acknowledges the mentorship of Professors Onur M. Ülgen and Benjamin Lev of the University of Michigan – Dearborn. Both authors are pleased to acknowledge the helpful criticism of an anonymous referee, and the assistance of colleague Neelesh Kale in preparing the final manuscript.

## REFERENCES

- Balci, Osman. 1998. "Verification, Validation, and Testing." In *Handbook of Simulation*, ed. Jerry Banks, 335-393. New York, New York: John Wiley & Sons, Incorporated.
- Bapat, Vivek, and David T. Sturrock. 2003. "The Arena Product Family: Enterprise Modeling Solutions." In *Proceedings of the 2003 Winter Simulation Conference*, Volume 1, eds. Stephen E. Chick, Paul J. Sánchez, David Ferrin, and Douglas J. Morrice, 210-217.
- Crabb, Howard C. 1975. Personal oral communication with Edward J. Williams.
- Eldabi, Tillal, Man Wai Lee, and Ray J. Paul. 2003. "A Framework for Business Process Simulation: The Grab and Glue Approach." In *Proceedings of the 15<sup>th</sup> European Simulation Symposium*, eds. Alexander Verbraeck and Vlatka Hlupic, 291-296.
- Herbst, J., S. Junginger, and H. Kühn. 1997. "Simulation in Financial Services with the Business Process Management System ADONIS." In *Proceedings of the 9<sup>th</sup> European Simulation Symposium*, eds. Winfried Hahn and Axel Lehmann, 491-495.
- Kelton, W. David, Randall P. Sadowski, and David T. Sturrock. 2004. *Simulation with Arena*, 3<sup>rd</sup> edition. Boston, Massachusetts: The McGraw-Hill Companies, Incorporated.
- Leemis, Lawrence M. 1995. *Reliability: Probabilistic Models and Statistical Methods*. Englewood Cliffs, New Jersey: Prentice-Hall, Incorporated.
- Miller, Scott, and Dennis Pegden. 2000. "Introduction to Manufacturing Simulation." In *Proceedings of the 2000 Winter Simulation Conference*, Volume 1, eds. Jeffrey A. Joines, Russell R. Barton, Keebom Kang, and Paul A. Fishwick, 63-66.
- Nanthavanij, S., P. Yenradee, V. Ammarapala, and S. Wongtiraorn. 1996. "Performance Analysis of Car Park Systems Using Simulation." In *Proceedings of the 1<sup>st</sup> Annual International Conference on Industrial Engineering Applications and Practice*, eds. Jacob Jen-Gwo Chen and Anil Mital, 726-731.
- Nash, Tom. 2003. "Increased Used-Vehicle Sales Spur Aftermarket Repairs." *Motor* 200(3):58.
- Palacis, Edgars. 2000. "Timber Accounting and Logistics System." In *Proceedings of the Second International Conference Simulation, Gaming, Training and Business Process Reengineering in Operations*, eds. Yuri Merkurjev, Birger Rapp, and Galina Merkurjeva, 231-234.
- Pichitlamken, Jutta, Alexandre Deslauriers, Pierre L'Ecuyer, and Athanassios N. Avramidis. 2003. "Modelling and Simulation of a Telephone Call Center." In *Proceedings of the 2003 Winter Simulation Conference*, Volume 2, eds. Stephen E. Chick, Paul J. Sánchez, David Ferrin, and Douglas J. Morrice, 1805-1812.
- Sargent, Robert G. 2003. "Verification and Validation of Simulation Models." In *Proceedings of the 2003 Winter Simulation Conference*, Volume 1, eds. Stephen E. Chick, Paul J. Sánchez, David Ferrin, and Douglas J. Morrice, 37-48.
- Stevenson, William J. 2005. *Operations Management*, 8<sup>th</sup> edition. Boston, Massachusetts: The McGraw-Hill Companies, Incorporated.
- Weinberg, Gerald M. 1971. *The Psychology of Computer Programming*. New York, New York: Van Nostrand Reinhold Company.
- Wickens, Christopher D., Sallie E. Gordon, and Yili Liu. 1998. *An Introduction to Human Factors Engineering*. New York, New York: Addison Wesley Longman, Incorporated.

## AUTHOR BIOGRAPHIES

**NAVIN GUPTA** received his master's degree in Industrial Engineering at Wayne State University, Detroit, MI in December 2003. He has been deeply involved with the area of Industrial and Quality Engineering. His career in industry began after receipt of his bachelor's degree in Industrial Engineering from Regional Engineering College, India. His first assignment was as an Industrial Engineer at Central Institute of Hand Tools followed by Quality Control Engineer at Sondhi Industries. He is a Certified Quality Engineer and also a member of the American Society of Quality (ASQ). He has been involved in the area of Computer Simulation for the past two years. He did his research in simulation in his master's program; the research was related to an articulation of strategic budgeting with Monte Carlo simulation analysis. His e-mail address is [navin@wayne.edu](mailto:navin@wayne.edu).



**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 classes at the University of Michigan, including both undergraduate and graduate simulation classes using GPSS/H<sup>TM</sup>, SLAM II<sup>TM</sup>, SIMAN<sup>TM</sup>, 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 served as program chair for the 2004 Summer Computer Simulation Conference in San José, California, U.S.A.; and will do likewise for the 2005 Summer Conference in Philadelphia, Pennsylvania, U.S.A. His e-mail address is: [williams@umdsun2.umd.umich.edu](mailto:williams@umdsun2.umd.umich.edu) and his university Web-page can be found at <http://www-personal.umd.umich.edu/~williams>.



**SIMULATION  
METHODOLOGIES,  
METHODS AND  
TECHNIQUES**



# COGNITIVE MAPS BASED ON PLIANT LOGIC

József Dombi and József D. Dombi  
Departure of Informatics  
University of Szeged  
6720, Szeged Árpád tér 2  
E-mail: [dombi@inf.u-szeged.hu](mailto:dombi@inf.u-szeged.hu)

## KEYWORDS

Aggregation operators, continuous valued logic, Cognitive map

## ABSTRACT

In this paper we present a tool for description, and for simulation of dynamic system. Our starting point is the aggregation concept, which was developed for multicriteria decision making. Using a continuous logic operator and proper transformation of sigmoid function we build positive and negative effects. From the input with the aggregation operator we calculate the output effect. This algorithm is comparable with the concept of fuzzy cognitive maps. We show this new technique, which could be much more efficient than the FCM.

## INTRODUCTION

Handling sophisticated systems we face serious difficulties, because we have to approach dynamic systems. Modeling a dynamic system can be hard in a computational sense. In addition formulating a system with mathematical model may be difficult, even impossible.

Developing the model requires effort and specialized knowledge. Usually the system involves complicated causal chains, which may be nonlinear. It should be mentioned to, that numerical data may be hard to get and even uncertain.

Our approach overcomes the above mentioned difficulties. It is qualitative approach, where enough to know rough description of the system and not necessary deep expert knowledge.

First type of this approach proposed by Kosko (Kosko 1986; Kosko 1992; Kosko 1994), and called Fuzzy Cognitive Map (FCM). FCM's are hybrid methods lie in some sense between fuzzy systems and neural networks. Knowledge is represented in symbolic manner, states, processes and events. All type of this information has numerical values.

FCM allows to performing qualitative simulations and experiment with the model. Compared FCM either expert system and neural networks it has good properties as it is relative easy to use for representing structured knowledge and the inference can be computed by numeric matrix operation instead of applying rules.

## BASIC CONCEPT OF PLIANT COGNITIVE MAP (PCM)

In this paper we make closer the FCM concept to the real world modeling. We will use the cognitive maps as a formal way of representing knowledge and modeling decision making, which was introduced by Axelrod (Axelrod 1976). Kosko used fuzzy values and matrix multiplication to calculate the next stage of the concepts representing by cognitive map. Instead of values we will use time dependent functions like an impulse function representing the positive and negative influences.

Another improvement is dropping the matrix multiplication concept, because on one hand it not works using continuous logic (or fuzzy logic), where the truth value has 1 and the false is 0 as usually used and the negative effect build by negation. The other hand more general operators could be more effective.

Logic and the cognitive map model correspond each other, and much more easy for the expert to build up and construct the system and from the identified system extract the knowledge.

Combining cognitive maps with logic can avoid many of knowledge extraction problems instead of rule based system. The classical knowledge representation in expert system is made through a decision tree. This form of knowledge presentation can not model the dynamic behavior of the world.

The cognitive map describes the whole system by a graph showing the cause effects along concepts. It is a directed graph with feedback, that the world as a collection of concepts and causal influences between the concepts. From logic point of view the causal concepts are unary operators of a continuous valued logic, which may negation operators in the case of inhibition effects. The value of node reflects the degree of the activity of the system at a particular time. Concept values are expressed on a normal range denoting a degree of activation rather than an exact quantities value. The inverse of the normalization could express the values coming from the real world. In spite of Fuzzy Cognitive Maps we do not use thresholds to force the values between 0 and 1. The mapping is a variation of the "fuzzification" process in fuzzy logic, but this destroys getting the possibility of quantitative results. In pliant logic we use only continuously strict monotonously increasing functions, and the inverse function always corresponds the real world values with the logical values.

In FCM the causal relationship are expressed by either positive or negative signs ordered by different weights. As we mentioned this will be replaced by unary operators.

Let  $\{C_1, \dots, C_m\}$  be concepts. Let define over the concepts a directed graph. A directed edge  $w_{ij}$  from concept  $C_i$  to concept  $C_j$  measures how much  $C_i$  causes  $C_j$ .  $w_{ij} \in [0, 1]$  where  $\frac{1}{2}$  is the neutral value, 0 is maximum negative and 1 is maximal positive influence or causality (In FCM  $w_{ij} \in [-1, 0, 1]$ ):

- $w_{ij} > \frac{1}{2}$  indicates direct (positive) causality between concepts  $C_i$  and  $C_j$ . That is the increase (decrease) in the value of  $C_i$  leads to increase (decrease) on the value of  $C_j$ .
- $w_{ij} < \frac{1}{2}$  indicates inverse (negative) causality between concepts  $C_i$  and  $C_j$ . That is the increase (decrease) in the value of  $C_i$  leads to decrease (increase) on the value of  $C_j$ .
- $w_{ij} = \frac{1}{2}$  indicates no relationship between  $C_i$  and  $C_j$ .

In pliant case  $w_{ij}$  depends on time(t) i.e.  $w(t) = (w_{ij}(t))_{n \times n}$ . The activation level  $a_i$  of concept  $C_i$  calculated by an iteration process. In FCM is

$$a_i^n = f\left(\sum_{i=1}^n w_{ij} a_i^o\right) \text{ where } a_i^n \text{ is the new activation}$$

level of concept  $C_i$  at time t+1,  $a_i^o$  is the activation level of concept  $C_i$  at time t and f is threshold function.

FCM has the advantage that we get the new state vector a by multiplying the previous state vector a by the edge matrix W showing the effect of the change in the activation level of one concept to another concept. In the pliant concept we aggregate the influences instead of summing up the values. The result is always remaining between 0 and 1, so we can avoid to normalization as an artificial step. The aggregation is pliant logic is general operation, which contain the conjunctive operators and disjunctive operators as well. Depends on the parameter – called neutral value – of aggregation operator we can build logical operators (Dombi operators). Using PCM (Pliant Cognitive Maps) can be used to answer “what if” question based on an initial scenario. Let  $a_0$  the initial

state vector. Repeatedly calculate with the aggregation operator the new state until the system convergence (i.e.  $|a_i^o - a_i^n| < \varepsilon$ ). We get the resulting equilibrium vector, which provides the answer to the “what-if” question. The PCM can be used all the areas covered by FCM.

## AGGREGATION AND ITS PROPERTIES

Beside the developed logical operators in fuzzy theory appears a non logical operator. The reason was insufficiency of using either conjunctive or disjunction operators for real world situation [Zimmermann]. General class of the fuzzy operators is called t-norm and t-conorm (disjunctive case). Denoting by  $c(x, y)$  the conjunctive operator and  $d(x, y)$  the disjunctive operator then

$$c(x, y) \leq \min(x, y)$$

$$d(x, y) \geq \max(x, y)$$

In the real world situation often occur that an aggregation value  $a(x, y)$  is

$$\min(x, y) \leq a(x, y) \leq \max(x, y)$$

The rational form of an aggregation operator is (Dombi 1982a):

$$a(x_1, \dots, x_n) = \frac{1}{1 + \left(\frac{1-v_0}{v_0}\right) \left(\frac{v}{1-v}\right)^n \prod_{i=1}^n \left(\frac{1-x_i}{x_i}\right)}$$

or

$$a(x_1, \dots, x_n) = \frac{1}{1 + \left(\frac{v_*}{1-v_*}\right)^{n-1} \prod_{i=1}^n \left(\frac{1-x_i}{x_i}\right)}$$

Where  $v$  is the neutral value or previous value of the node (see later).

The corresponding negation function is

$$n_v(x) = \frac{1}{1 + \frac{1-v_0}{v_0} \cdot \frac{1-v}{v} \cdot \frac{x}{1-x}}$$

$$n_{v_*}(x) = \frac{1}{1 + \left(\frac{1-v_*}{v_*}\right)^2 \cdot \frac{x}{1-x}}$$

The aggregation operator is axiomatically based and it has several good properties as

1. defined on (0,1) and the values are also in (0,1)
2. associativity

3. continuously
4. strictly monotonously increasing
5. continuous on  $[0,1)$  interval
6.  $a(0,0)=0$  and  $a(1,1)=1$

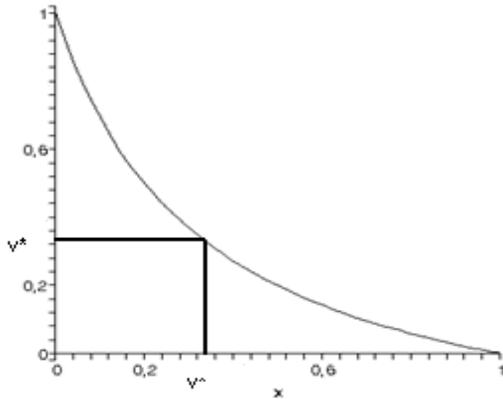
1.  $n(a(x, y)) = a(n(x), n(y))$
2.  $a(x, n(x)) = v_0$
3.  $a(x, v_0) = x$

Aggregation is connected to negation operators. The proportions of that negation are

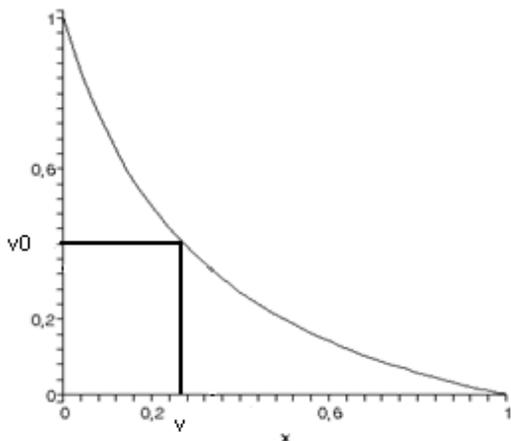
1. defined on  $(0,1)$  and the values are also in  $(0,1)$
2.  $n(0)=1$
3.  $n(1)=0$
4. continuous
5. strictly decreasing function
6. involutiv (the double negation is the identity)  
 $n(n(x))=x$

We can ordered to a negation a  $v_*$  fixed point such  $n(v_*) = v_*$  or defined  $v_0$  the so called common threshold and neutral value than  $n(v) = v_0$ .

Usually  $v_0 = \frac{1}{2}$



Figures 1: Negative function width  $v_*$  threshold



Figures 2: Negative function width  $v$  neutral value and  $v_0$  threshold

The negative function and aggregation operator is closely related. It can be seen easily that

The properties of the aggregation are natural: 1, aggregating positive values and negating it is the same if we aggregating negative values; 2, aggregating positive and negative values we get the neutral values back; 3, aggregating  $x$  with the neutral value we get back  $x$ .

The property ensures that we can replace the sum function with the aggregation function of the pliant logic. The neutral value here is  $v_0$  instead of 0 using in FCM. The neutral value corresponds the aggregation width the logical connectives to.

It can be proved that

1. if  $x, y \leq v$  than  $a(x, y) \leq \min(x, y)$
2. if  $x, y \geq v$  than  $a(x, y) \geq \max(x, y)$
3. if  $x \leq v \leq y$  than  
 $\min(x, y) = x \leq a(x, y) \leq y = \max(x, y)$

First means that if the values are less then neutral value then the aggregation is conjunction.

Second means that if the values are larger then neutral value then the aggregation is disjunction.

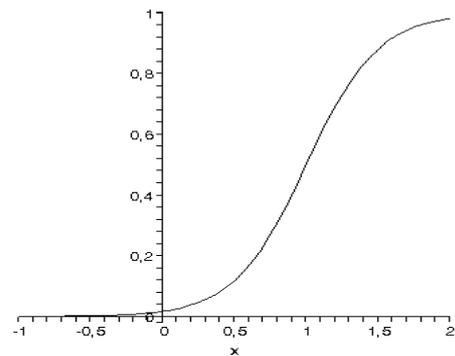
Third means that the aggregation of positive and negative values is always between the two values.

We can model conjunctive and disjunctive operator with the aggregation operator. If  $v$  is close to 0 then operation has disjunctive character and if  $v$  is close to 1 then the operation has conjunctive character. From this property it can be seen, that using aggregation we have much more possibilities instead of using the sum function in FCM. Changing in the nodes the neutral value different operation can be carried out.

## PRODUCING INPUT INFLUENCES

Our starting point is the sigmoid function.

$$\delta_a^t(t) = \frac{1}{1 + e^{-\lambda(t-a)}}$$



Figures 3: Sigmoid function

It is easy to see:

1.  $\delta_a^{(\lambda)}(a) = \frac{1}{2} (= v_0)$
2.  $\delta'_a(a) = \lambda$
3.  $\delta_a^{(-\lambda)}(t) = 1 - \delta_a^{(\lambda)}(t)$

The sigmoid function natural way maps the values to the (0,1) interval. Positive (Negative) influences can be build with  $\delta_a^{\lambda_1}(t)$ ,  $\delta_b^{\lambda_2}(t)$  and conjunctive operator where  $\lambda_1 > 0$ ,  $\lambda_2 < 0$  and  $a < b$  ( $\lambda_1 > 0, \lambda_2 < 0$  and  $b < a$ ).

Using the Dombi operator (Dombi 1982b) and sigmoid

$$c(x_1, w_1, \dots, x_n, w_n) = \frac{1}{1 + \left( \sum_{i=1}^n w_i \left( \frac{1-x_i}{x_i} \right)^\alpha \right)^{\frac{1}{\alpha}}}$$

function with proper weights we can get

$$\delta_{a,b}^{\lambda_1, \lambda_2}(t) = \frac{1}{1 + \frac{\lambda_2}{\lambda_1 + \lambda_2} \cdot e^{-\lambda_1(t-a)} + \frac{\lambda_1}{\lambda_1 + \lambda_2} \cdot e^{-\lambda_1(t-b)}}$$

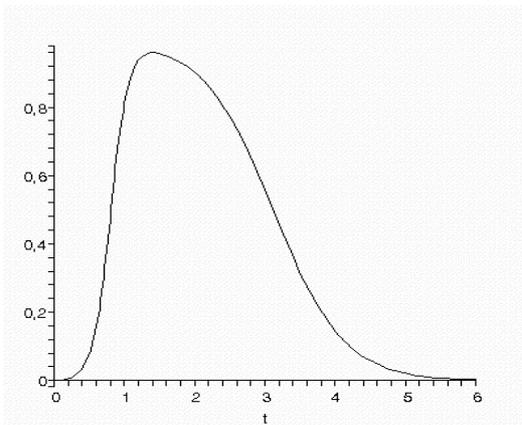
where  $\lambda_i > 0$  and  $a < b$ .

For the aggregation we have to transform  $P(t)$  the

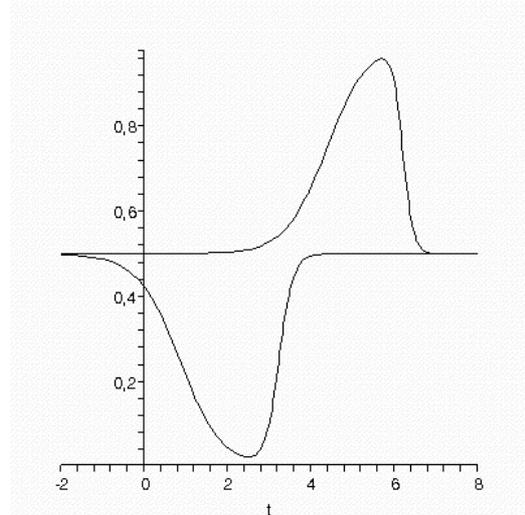
positive influence into  $\left[ \frac{1}{2}, 1 \right]$  interval and  $N(t)$  the

negative influence into  $\left[ \frac{1}{2}, 0 \right]$  interval:

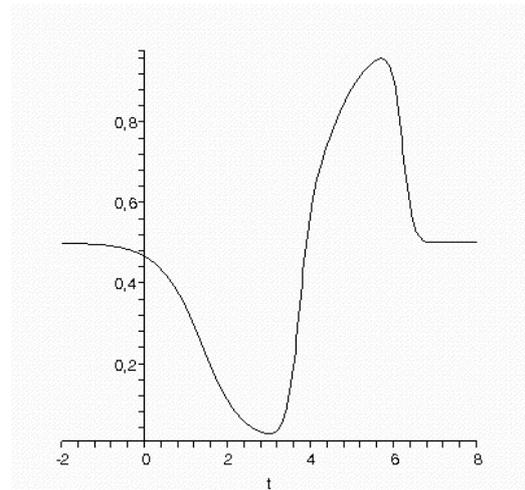
- $P(t) = \frac{1}{2} \left( 1 + \delta_{a,b}^{\lambda_1, \lambda_2}(t) \right)$
- $N(t) = \frac{1}{2} \left( 1 - \delta_{a,b}^{\lambda_1, \lambda_2}(t) \right)$



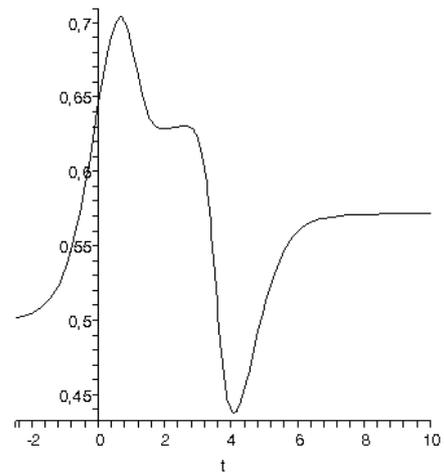
Figures 4: Asymmetrical influence on [0,1]  
 $a = 1, b = 3.5, \lambda_1 = 10, \lambda_2 = 2$



Figures 5: Transformation of two influences  
 $a^p = 4, b^p = 6, \lambda_1^p = 2, \lambda_2^p = 8$   
 $a^n = 1, b^n = 3, \lambda_1^n = 2, \lambda_2^n = 6$



Figures 6: Aggregation of transformed positive and negative input influences



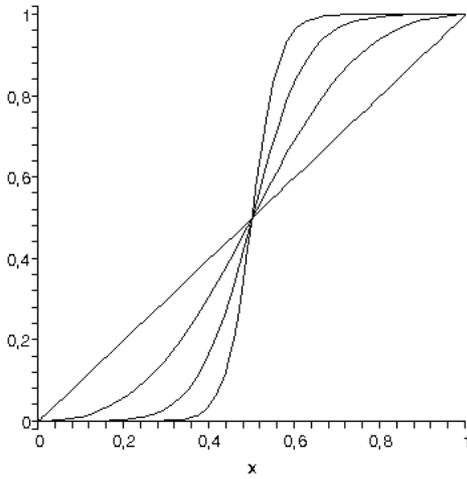
Figures 7: Aggregation influences

## VALUE TRANSFORMATION

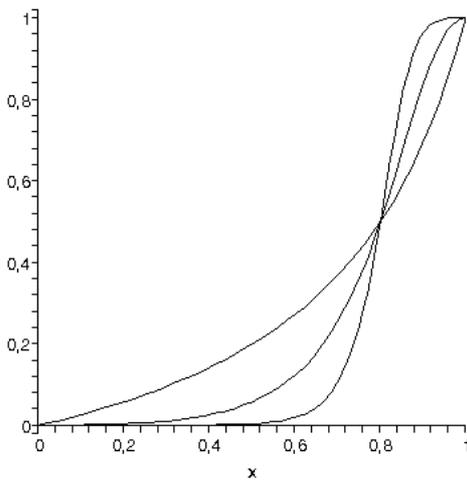
To complete the Pliant Cognitive Map we have to give unary transformation on the values produced by the aggregation. In the pliant logic the general form of the modification operators are:

$$\kappa_{\nu_{ij}}^{\lambda_{ij}}(x) = \frac{1}{1 + \left( \frac{1 - \nu_{ij}}{\nu_{ij}} \cdot \frac{1 - x}{x} \right)^{\lambda_{ij}}}$$

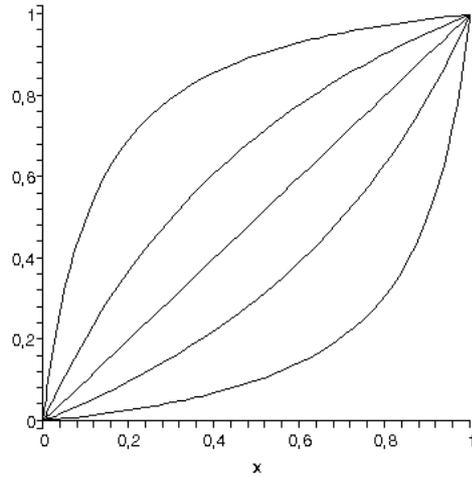
The sign of  $\lambda_{ij}$  means that it is a positive or negative influence, the value of  $\lambda$  mean the sharpness and  $\nu_{ij}$  is the actual value of  $C_j$ .



Figures 8: Modification operator with the parameters  $\nu = 0.5, \lambda = 1, 2, 4, 8$



Figures 9: Modification operator with the parameters  $\nu = 0.2, \lambda = 1, 2, 4, 8$



Figures 10: Modification operator with the parameters  $\lambda = 1, \nu = 0.1, 0.3, 0.5, 0.7, 0.8$

## CONSTRUCTION PCM

It is easy to build PCM. The following steps should be carries out:

1. Collect the concepts
2. Define the expectation values of the nodes (i.e. threshold values of the aggregations)
3. Build a cognitive map (i.e. draw a directed graph between the concepts)
4. Define the influences (i.e. are they positive or negative)

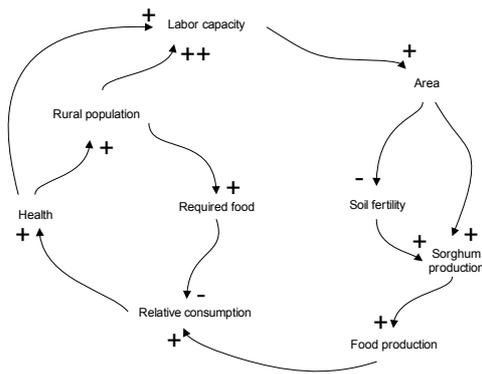
The iterative method:

1. Use the proper function to the input nodes  $\delta_{a,b}^{\lambda_1, \lambda_2}$
2. Build the  $P(t)$  and  $N(t)$  for every nodes
3. Aggregating in the nodes the positive and negative influences, where the  $\nu_*$  value is the previous value of  $C_j$

The system now ready to make simulation test. We are developing a program to test the system. First we are studying artificial situation, and this shows that the system is very flexible, and easy to adapt to various situation. For the real world application we invent learning process finding the best parameter. This lead to a nonlinear problem.

## APPLICATION

Our model consists of four interacting sector: population, food consumption, sorghum production, and animal production. As the model contains approximately 200 variables, it is impossible to discuss it in detail. Figure 11 presents a feedback diagram of the part of the model structure.



Figures 11: Feedback diagram of model structure

We generate several input function, and calculate it's effects of the nodes. The model works as it was expected. In the future we will work on an identification process i.e. based on historical data. We would like to find the proper parameters of the modification operators.

## CONCLUSION

We propose a new type of numerical calculus modeling complex systems based on positive and negative influences. This concept is similar to FCM, but the functions and the aggregation procedures are quite different. It is based on a continuous valued logic and all the parameters have semantic meaning. We are working a real world application and on an effective learning of the parameters of the system.

## REFERENCES

- Axelrod R., 1976. "Structured of Decision: the cognitive maps of political elites", Princeton University press, New Jersey
- Bazara, M.S.; Sherali, H.D.; Shetti, C.M.; 1993. "Nonlinear programming: Theory and Algorithms", John Willey&Sons, New York, 325p
- Dombi J., 1982a, "Basic Concept for a theory of evaluation: the aggregation operator", European Journal of Operations Research, 10:282-293
- Dombi J., 1982b. "A general class of fuzzy operators, the De Morgan class of fuzzy operators and fuzziness measures induced by fuzzy operators" Fuzzy Sets and Systems, 8:149-163
- Dombi J. 1987. *Membership function as an evaluation*, ITRC report, Bristol,
- Fodor J., (1996). "A new look at fuzzy connectives ", Fuzzy Sets and Systems, 57:141-148
- Kosko B. 1986. "Fuzzy Cognitive Maps", Int. Journal of Man-Machine Studies, Vol. 24, pp. 65-75
- Kosko B. 1992. *Neural Networks and Fuzzy systems, A dynamic system approach to machine intelligence*, Prentice Hall, New Jersey
- Kosko B., Dickerson J. 1994. "Fuzzy virtual worlds". AI Expert, pp. 25-31

Kosko B. 1997. *Fuzzy Engineering*, Prince-Hall, New Jersey

Kosko B., 1998. "Global Stability of Generalized Additive Fuzzy Systems", IEEE Transactions on Systems, Man and Cybernetics- Part C: Applications and Reviews, Vol. 28, No 3

Tjark. E. Struif Bontkes 1993. "Dynamics of rural development in southern Sudan", System Dynamics Review, Vol. 9, No 1, Winter, 1-21p

## AUTHOR BIOGRAPHIES



**JÓZSEF DOMBI** was born in Zalaegerszeg, Hungary, and went to the University of Szeged, where he studied mathematics. His scientific degree: M. Sc. degree mathematician (1972), Ph.D. degree (1977, Summa cum laude on Hydrogen transfer reaction), CSC degree Candidate of the mathematical science (1994, The structure of the operators of fuzzy sets in the respect of multiple criteria decisions).

His research area is Intelligence, Operation Research, Optimization of logistic process. He has 70 publications. He won sum scientific prizes: including European Information Technology prize in Brussels (1997), the best software prize on the COMDEX in Las Vegas (1999) and Kalmár László prize (1998).

His e-mail address is: [dombi@inf.u-szeged.hu](mailto:dombi@inf.u-szeged.hu) and his Web-page can be found at <http://www.inf.u-szeged.hu/~dombi>.



**JÓZSEF D. DOMBI** was born in Szeged, Hungary, and went to the University of Szeged, where he studied informatics. He is a student. He several times successfully attends on the Hungarian Scientific Conference. He's research interest is modeling dynamic systems and artificial intelligence. His e-mail address is: [d.j.dombi@dopti.hu](mailto:d.j.dombi@dopti.hu).

# FUZZY CONTROL OF COMBUSTION WITH GENETIC LEARNING AUTOMATA

Zoltán Himer<sup>1</sup>, Géza Dévényi<sup>2</sup>, Jenő Kovács<sup>1</sup>, Urpo Kortela<sup>1</sup>

<sup>1</sup>University of Oulu, Systems Engineering Laboratory,  
P.O. Box 4300, FIN-90014 University of Oulu, Finland

Fax: +358-8-553-2439, email: [himi@paju.oulu.fi](mailto:himi@paju.oulu.fi)

<sup>2</sup>Technical University of Budapest, Department of Power Engineering  
Egry József u. 18, Budapest, H-1111 Hungary email: [dgeza@freemail.hu](mailto:dgeza@freemail.hu)

## Abstract

It is difficult to achieve effective control of time variable and nonlinear plants such a fluidized bed boiler. A method of designing a nonlinear fuzzy controller is presented. However, its early application relied on trial and error in selecting either the fuzzy membership functions or the fuzzy rules. This made it heavily dependent on expert knowledge, which may not always available. Hence, an adaptive fuzzy logic controller such as Adaptive Neuro-Fuzzy Inference System (ANFIS) removes this stringent requirement.

This paper demonstrates the application of ANFIS a nonlinear Multi Input Single Output fuel feeding and combustion system and a fuzzy controller design for the system with optimization with Genetic Learning Automata (GLA).

An ANFIS model has been developed to determine the exact amount of fuel fed to a combustion chamber. This property is impossible to measure directly, but it is required for improving combustion control.

The control of the combustion base on two Takagi-Sugeno type controllers, which were optimized by GLA. The control system has been validated on experiment data obtained in a case-study power plant. The results have shown that the system is able to capture the nonlinear feature of the fuel feeding system.

## Key words

Combustion control, non-linear systems, ANFIS, Combustion control, Genetic Learning Automata

## 1. Introduction

In the last decade, the interest of burning multifuel has arisen in Finland using mainly fluidisation technology. The multifuels are usually mixtures of different bio fuels (peat, woodchips, sawdust, and bark) but in some case, coal and municipal wastes are burned with. The more intensive use of multifuels can be explained by: a) the increasing demand of using domestic fuels (e.g. peat), b) the thermal utilisation of the high caloric-value

paper-industry by-products (wood chips, sawdust, and bark) which would be waste and c) diverting municipal solid wastes from landfill.

Beside the economical and environmental advantages, there are several difficulties with burning bio fuels and municipal wastes. The combustion of those fuels or fuel-mixtures has different properties compared to the conventional fuels (coal, gas, and oil). Bio fuels and municipal wastes are very inhomogeneous. The properties (heat value, moisture content, homogeneity, density, mix ability) may vary in a large range. It causes non-steady, agitated combustion conditions; even if steady fuel feed volume is maintained, leading to increase in the emission level and variation of the generated heat flow. Those property variations are not predictable or directly measurable, only their effects on the combustion, on the steam generation and on the power production can be observed through the O<sub>2</sub> content of the flue gas. This paper presents an ANFIS system, which determines the amount of fuel fed to the combustion chamber. Combined with a stoichiometric model, it predicts the flue gas properties, including the O<sub>2</sub> content.

## 2. Description of the neuro-fuzzy controller

Fuzzy Logic Controllers (FLC) has played an important role in the design and enhancement of a vast number of applications. The proper selection of the number, the type and the parameter of the fuzzy membership functions and rules is crucial for achieving the desired performance and in most situations, it is difficult. Yet, it has been done in many applications through trial and error. This fact highlights the significance of tuning fuzzy system.

Adaptive Neuro-Fuzzy Inference Systems are fuzzy Sugeno models put in the framework of adaptive systems to facilitate learning and adaptation [1]. Such framework makes FLC more systematic and less relying on expert knowledge.[2],[3] To present the ANFIS architecture, let us consider two-fuzzy rules based on a first order Sugeno model:

Rule 1: if (x is A<sub>1</sub>) and (y is B<sub>1</sub>) then (k<sub>1</sub> = p<sub>1</sub>)

Rule 2: if (x is A<sub>2</sub>) and (y is B<sub>2</sub>) then (k<sub>2</sub> = p<sub>2</sub>)

One possible ANFIS architecture to implement these two rules is shown in Fig. 1. Note that a circle indicates a fixed node whereas a square indicates an adaptive node (the parameters are changed during training). In the following presentation  $O_{Li}$  denotes the output of node  $i$  in a layer  $L$ .

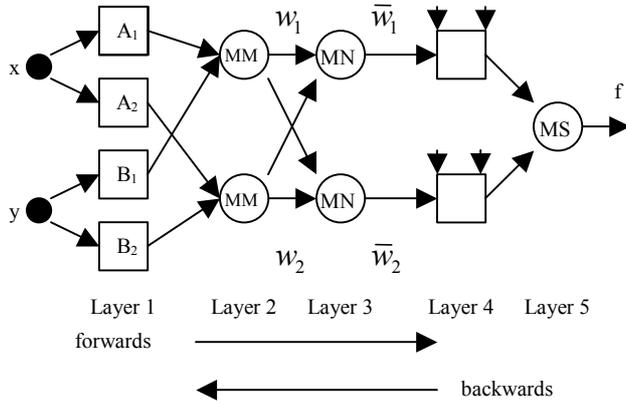


Fig. 1 Construct of ANFIS controller

Layer 1: All the nodes in this layer are adaptive nodes,  $i$  is the degree of the membership of the input to the fuzzy membership function (MF) represented by the node:

$$O_{1,i} = \mu_{A_i}(x) \quad i = 1, 2 \quad (1)$$

$$O_{1,i} = \mu_{B_{i-2}}(y) \quad i = 3, 4$$

$A_i$  and  $B_i$  can be any appropriate fuzzy sets in parameter form. For example, if bell MF is used then,

$$\mu_{A_i}(x) = \frac{1}{1 + \left[ \left( \frac{x - c_i}{a_i} \right)^2 \right]^{b_i}} \quad i=1,2 \quad (2)$$

where  $a_i, b_i$  and  $c_i$  are the parameters for the MF.

Layer 2: The nodes in this layer are fixed (not adaptive). These are labelled  $M$  to indicate that they play the role of a simple multiplier. The outputs of these nodes are given by:

$$O_{2,i} = w_i = \mu_{A_i}(x) \mu_{B_i}(y) \quad i=1,2 \quad (3)$$

The output of each node in this layer represents the firing strength of the rule.

Layer 3: Nodes in this layer are also fixed nodes. These are labelled  $N$  to indicate that these perform a normalization of the firing strength from previous layer. The output of each node in this layer is given by:

$$O_{3,i} = \bar{w}_i = \frac{w_i}{w_1 + w_2} \quad i=1,2 \quad (4)$$

Layer 4: The output of each node is simply constant:

$$O_{4,i} = \bar{w}_i k_i \quad i=1,2 \quad (5)$$

where  $k_i$  is design parameter

Layer 5: This layer has only one node labelled  $S$  to indicate that it performs the function of a simple summer. The output of this single node is given by:

$$O_{i,5} = \sum_i \bar{w}_i k_i = \frac{\sum_i w_i k_i}{\sum_i w_i} \quad i=1,2 \quad (6)$$

The ANFIS architecture is not unique. Some layers can be combined and still produce the same output. In this ANFIS architecture, there are two adaptive layers (1, 4). Layer 1 has three modifiable parameters ( $a_i, b_i$  and  $c_i$ ) pertaining to the input MFs [4]. These parameters are called *premise* parameters. Layer 4 has one modifiable parameters ( $k_i$ ). That parameter called *consequent* parameter.

### 3. Optimization of the Fuzzy controller using Genetic Learning Automata

Standard genetic or genetic searching algorithms are used for numerical parameter optimization and are based on the principles of evolutionary genetics and the natural selection process [5].

A general genetic algorithm contains, usually, the next three procedures: selection, crossover and mutation. These procedures are responsible for the “global” search minimization function without testing all the solutions. Selection corresponds to keeping the best members of the population to the next generation to preserve the individual with good performance (elite individuals) in fitness function. Crossover originates new members for the population, by a process of mixing genetic information from both parents, depending of the selected parents the growing of the fitness of the population is faster or lower. Among many other solutions, the parent selection can be done with the roulette method, by tournament, random and elitist [6]. Mutation is a process by which a percentage of the genes are selected in a random fashion and changed. The population of the bit string chromosome in genetic algorithms is replaced by a corresponding string of binary-action learning probabilities. The value at the  $i^{th}$  position of each member of the population defines the probability of the allele value in the

corresponding bit string of being 1 at the position. The probabilities are initialized to  $p_i(0)=0.5$  for all  $i$ , so there is equal probability a 1 or 0 being selected at each position. The system therefore has a very high degree of randomness for the initial generation. A population of the bit strings that directly determines the phenotype is generated stochastically at each generation by sampling the probability distribution of the population.

The probabilities at each position are regarded as the action probabilities of a binary-action discrete stochastic learning automation. The two actions of the learning automata are generating a 0 and generating a 1 at the corresponding position in the phenotype string in each generation. Since there are two actions, only the probability of one of the actions is required. In this paper we have defined the probabilities stored in the population as being the probability of selecting a 1.

Probabilities are updated at each generation on the with the Linear Reward/Penalty algorithm. [7] The probability  $p_i$  is the probability of a 1 being the action generated at the  $i^{th}$  position of the bit string. This is updated at each generation by the following

if the  $i^{th}$  position is 1 at generation  $n$

$$p_i(n+1) = p_i(n) + \Theta B(n)(1 - p_i(n)) \quad (7)$$

if the  $i^{th}$  position is 0 at generation  $n$

$$p_i(n+1) = p_i(n) - \Theta B(n)(p_i(n)) \quad (8)$$

where  $\Theta$  is the learning rate parameter and  $B(n)$  is generated by adjusting the raw fitness in the current generation. The value of  $B(n)$  for  $j^{th}$  string at generation  $n$  is given by

$$B_j(n) = \frac{f_j(n) - \min(f)}{\max(f) - \min(f)} \quad (9)$$

Here is  $\min(f)$  and  $\max(f)$  refer to the minimum and maximum raw fitness in the current population and  $f_j(n)$  is the raw fitness of the  $j^{th}$  string.

The fitness function in our case is

$$f_j(n) = \frac{\sum_1^N (y_{Comb} - \hat{y}_{Comb})}{\sqrt{N}} + m * \frac{\sum_1^N (y_{O2} - \hat{y}_{O2})}{\sqrt{N}} \quad (10)$$

where the  $m$  is a weighting factor. In our case  $m=2$  to emphasize the importance of oxygen content which is directly related to the flue gas emissions.

In the implemented algorithm a population of 60 individuals, an elitism of 6 individuals was used, the crossover of one site splicing is performed and all the members are subjected to mutation except the elite. The mutation operator is a binary mask generated randomly according to a selected rate that is superposed to the existing binary codification of the population changing some of the bits.[8] Crossover is performed over half of the population, always including the elite. The individuals are randomly selected with equal opportunity to create the new population.

Dynamic crossover and mutation probability rate was used in the GLA operation, as they provide faster convergence when compared to constant probability rate [9].

#### 4. Model of combustion

The role of the combustion process is to produce the required heat energy for steam generation at the possible highest combustion efficiency. The efficiency depends on the completeness of burning and the waste heat taken away in the flue gas by the excess air flow. The higher the burning rate and smaller the waste heat is the higher efficiency. However, excess air is required for ensuring complete burning. The  $O_2$  content of the flue gas is directly related to the amount of excess air. The aim of the combustion control, from the efficiency point of view, is to keep the  $O_2$  content around 3-5 % [10]. In multi-fuel fired fluidised bed power plants (see Fig. 2), it is a difficult task due to the inhomogeneous properties of the fuel.

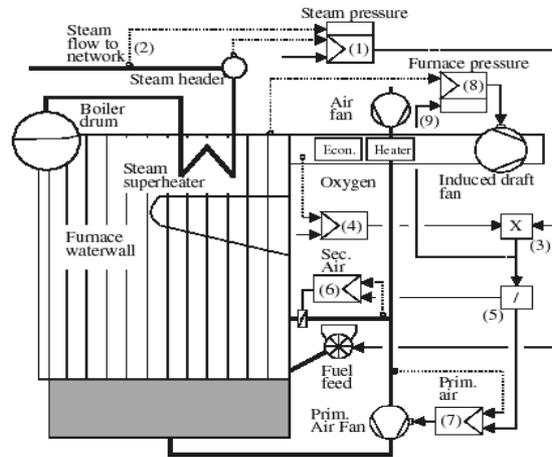


Fig. 2 Fluidized bed power plant

The combustion model, utilising the ANFIS structure based on [11], calculates the combustion power ( $P_{comb}$ ) and flue gas components ( $C_f$ ), including the oxygen content, from the fuel screw  $Q_{Hz}$ , signal primary airflow  $F_p$ , secondary airflow  $F_s$ . (see Fig. 3)

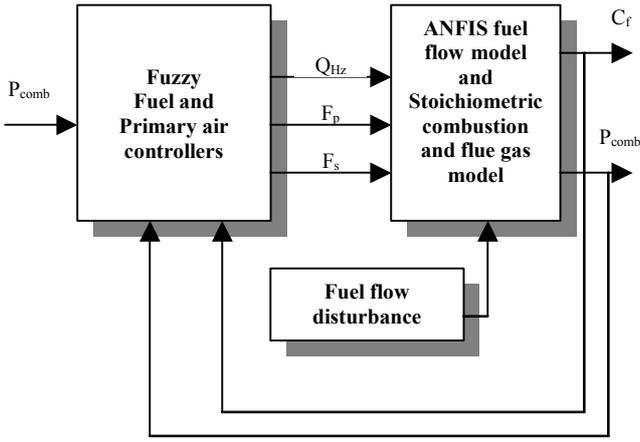


Fig. 3 Control system of combustion process.

The fuel and primary air fuzzy controller (see Fig.3) consists of two parallel Fuzzy controllers. The error signal from the oxygen content drives the fuzzy controller of the primary airflow, while combustion power is controlled by the flue screw signal.

The reference signals for the fuel screw  $Q_{Hz}$ , primary airflow  $F_p$  and secondary airflow  $F_s$  signals are calculated by the linearization model as a function of the reference of the combustion power such as:

$$\begin{aligned} Q_{Hz} &= 0.2663P_{comb} - 9.7207 \\ F_p &= 0.0737P_{comb} + 10.912 \\ F_s &= 0.2663P_{comb} - 4.0049 \end{aligned} \quad (11)$$

The error signal for the controllers divide in three region low, middle and high. The membership functions are shown by fig. 4 and 5

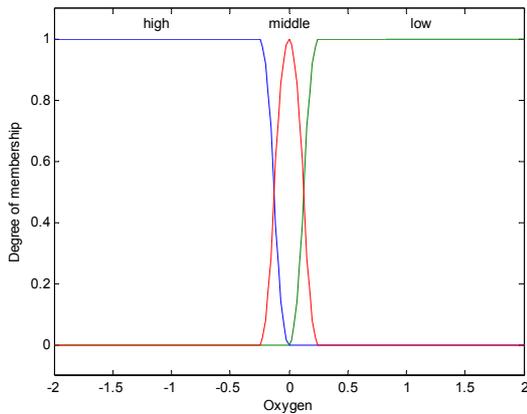


Fig. 4 Membership function of the oxygen error

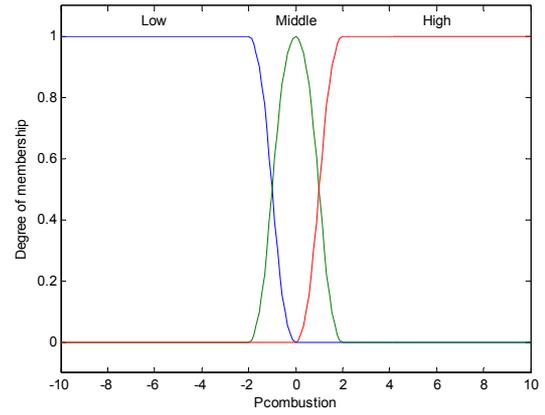


Fig. 5 Membership function of the Combustion power error

The outputs MFs of the controllers are constants, which mean in our case six parameters.

## 5. Experimental results

The system was optimized for a power level change and the fuel flow disturbance.

After 392 generation the optimal parameter for the fuzzy controller was found.

$$\begin{aligned} KP_{High} &= 0.8129 \\ KP_{Middle} &= 0.4681 \\ KP_{Low} &= 0.7814 \\ KO_{High} &= 0.5682 \\ KO_{Middle} &= 0.2568 \\ KO_{Low} &= 0.5182 \end{aligned} \quad (12)$$

The fitness function result

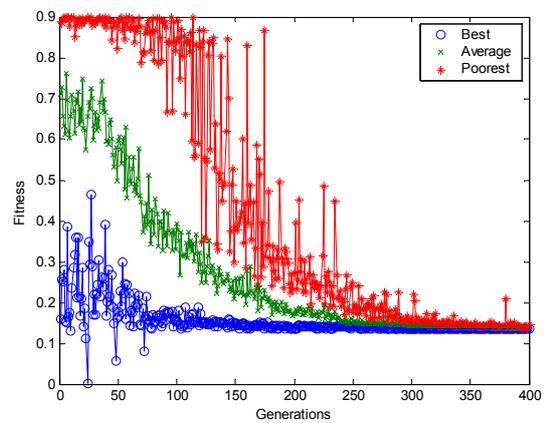


Fig. 6 Fitness function by the generation of the GLA

The model have limitation on each inputs, the combustion power change has also limitation. The set point function and the fuel flow disturbance for the optimization were the follow:

Time [s]	500	1000	1500	2000	2500
Set point [MW]	102	102	115	115	115
Fuel disturb [kg]	-5	+5		+5	-5

The result is compared by the self-tuned PI and a Genetic Algorithm tuned PI controller [11], [13]

	PI	PI with GA	Fuzzy GLA
RMSE	0.2812	0.2412	0.1324
Comp. time		32 hours	78 hours

The table shows the improvement of the control by the RMSE value, the drawback is the optimization time. The optimization was running 78 hour on a Pentium 2.4 GHz computer.

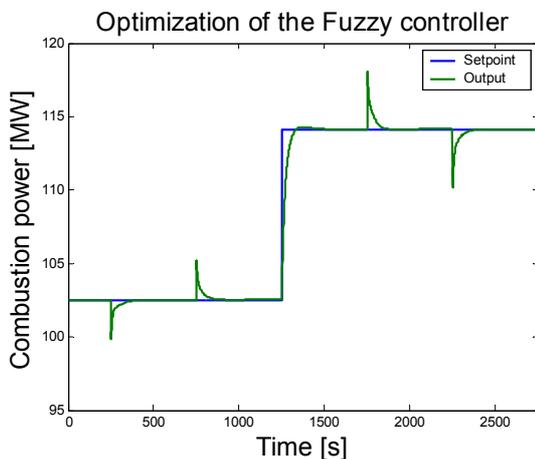


Fig. 5 Fuzzy combustion power controller optimization with GLA

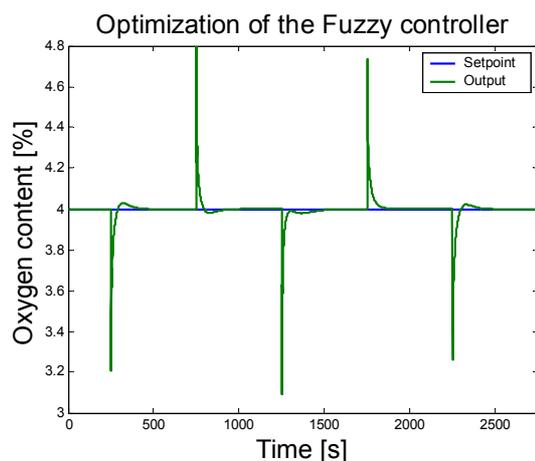


Fig. 6 Fuzzy Oxygen content controller optimization with GLA

In the following, the performance of the new controller based on the ANFIS model will be compared to the performance of the real process. The reference signal

for the combustion power is taken from the measurement data. The simulation shows that by applying the new controller structure together with the ANFIS model, much smaller deviation in the oxygen content can be achieved while satisfying the same demand for combustion power.

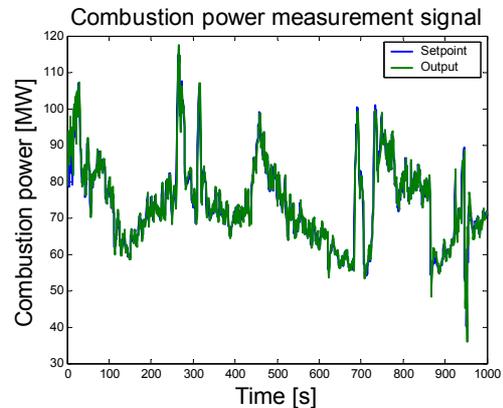


Fig. 7 Combustion power response: comparison of the achievement in real process and in the simulated control system.

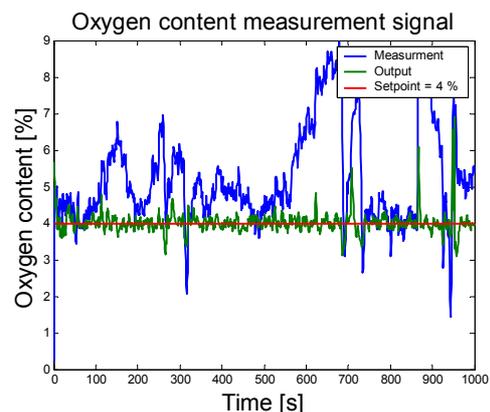


Fig. 8 Oxygen content response: comparison of the achievement in real process and in the simulated control system.

## 6. Conclusion

In this paper, ANFIS neuro-fuzzy controller was studied via optimization by Genetic Learning Automata. Neuro-fuzzy controller combines the theory of artificial neural networks and fuzzy systems. GLA providing successful parameter optimization for the ANFIS controller. The drawback of the method is the time consuming computation. Simulation result revealed that neuro fuzzy model was capable of closely reproducing the optimal performance.

## References

- [1] R. Jang, C. Sun, E. Mizutani, *Neuro-fuzzy and soft computation* (Prentice Hall, NJ,1997)
- [2] Fahd. A. Alturki, Abel Ben Abdennour, Neuro-fuzzy control of a steam boiler turbine unit, *Proceeding of the 1999 IEEE, International Conference on Control Applications*, Hawaii, USA 1999 pp 1050-1055
- [3] E. Ikonen, K.Najim, Fuzzy neural networks and application to the FBC process, *IEE Proc.-Control Theory Appl.* Vol. 143, May 1996 pp 259-269
- [4] S. H. Kim, Y. H. Kim, K. B. Sim, H. T Jeon, On Developing an adaptive neural-fuzzy control system, *Proc. IEEE/RSJ Conference on intelligent robots and systems* Yokohama, Japan, July 1993 pp 950-957
- [5] J. H. Holland, *Adaptation in natural and Artificial System*, MIT Press, 1992
- [6] J.S.Yang & M.L. West, A Case Study of PID Contoller tuning by Genetic Algorithm *Proceedings of IASTED International Conference on Modelling and Control*, Innsbruck,2001
- [7] M.Howell Genetic Learning Automata, *Internal report Loughborough University*,2000
- [8] J. Vieria, A. Mota, Water Gas Heater Nonlinear Physical Model: Oprimization with Genetic Algorithms *Proceedings of IASTED International Conference on Modelling Identification and Control*, Grindelwald, Switzerland 2004
- [9] T. L. Seng, M. B. Khalid, Tunning of a Neuro-Fuzzy Contoller By Genetic Algorithm *IEEE Transaction on Systems, Man and Cybernetics vol.29 no.2* 1999
- [10] K. Leppäkoski & J. Kovács, Hybrid model of oxygen content in flue gas. *Proc. IASTED International Conference on Applied Modelling and Simulation*, Nov, 2002, Cambridge, MA, USA, pp 341- 346
- [11] ] Z. Hímer, V. Wertz, J. Kovács, U. Kortela Neuro-fuzzy model of flue gas oxygen content *Proceedings of IASTED International Conference on Modelling Identification and Control*, Grindelwald, Switzerland 2004
- [12] Z. Hímer, G. Dévényi, J. Kovács, U. Kortela, Control of Combustion based on Neuro-fuzzy model *Proceedings of IASTED International Conference on Applied Simulation and Modelling*,Rhodos Greece 2004
- [13] H. Ghezelayagh, K.Y.Lee, Traning Neuro-fuzzy boiler identifier with Genetic Algorithm and error-back-propagation, *IEEE* 1999

## AUTHOR BIOGRAPHY



**ZOLTÁN HÍMER** (M.Sc. 2001 Budapest, Hungary) is a Ph.D. student since 2001 at the Systems Engineering Laboratory, University of Oulu, Finland. His research interests include fuzzy-neuro modelling and fuzzy control, genetic algorithms, Controller optimisation and their application to energy systems and power plant control problems.



**GÉZA DÉVÉNYI** (M.Sc. 2000 Budapest, Hungary) is a Ph.D. student since 2002 at the Power Engineering Department, Technical University of Budapest, Hungary. His research interests include power plant atomization, finite element calculation, high-voltage switch gear design and their application to energy systems.



**JENŐ KOVÁCS** (M.Sc. 1991 Budapest, Hungary, Ph.D. 1998 Oulu, Finland) is a senior assistant at the Systems Engineering Laboratory, University of Oulu, Finland. His research interests include adaptive control, constrained control, advanced modelling and their application to energy systems and power plant control problems.



**URPO KORTELA**, born in Finland, 1945, is the head professor of the Systems Engineering Laboratory, University of Oulu, Finland. He graduated as M.Sc. in Technical Physics in 1970 at the University of Oulu, Finland. He received the Licentiate of Technology in 1973 at the University of Oulu and the Doctor of Technology in 1981 at the University of Helsinki, Finland. His interest lies in the research in control engineering and system theory: state and parameter estimation and advanced control methods. The application field consists of power plant modeling and control, control and fault diagnosis of pulp and paper processes, and field bus technology.

# Industrial Maintenance Metrics based on Simulation and Fuzzy Logic

*Agostino G. Bruzzone*

*MISS - DIP*

Via Opera Pia 15, 16145 Genova, Italy

Email agostino@itim.unige.it - URL st.itim.unige.it

*Chiara Briano*

*BRBStudio,*

Genoa Voltri Terminal –Office Tower 16158 Genoa, Italy

Email chiara.briano@liophant.org - URL www.brbstudio.com

*Simone Simeoni*

*Liophant Simulation*

Via Molinero 1, 17100 Savona, Italy

Email simone.simeoni@liophant.org - URL www.liophant.org

## KEY-WORDS

M&S, Fuzzy Logic, VV&A; RCM, AI.

## ABSTRACT

This paper is focused on the development of a model for applying Reliability Centred Maintenance (RCM) on complex system by integrating simulation and artificial intelligence (AI); the application is related to maintenance planning of groups of gas turbine power plants. This paper will explain the reasons, the criticalities and the lessons learned during the development of the research, leading to the realisation of a Fuzzy Logic Schedule Evaluator (FUSE).

## INTRODUCTION

Complex industrial systems are subjected to the impact of many stochastic components especially in maintenance procedures; in effect the market evolution during the last years provided a growing opportunity in development of maintenance services. It was evident during the last 10 years that complex objects, such as planes, helicopters, gas turbines, and even big vehicles could provide interesting opportunities in service improvements. Improving the maintenance it becomes possible to increase reliability and availability with direct impact on final user costs and savings; for instance a SAR (Search and Rescue) fleet of helicopters could be drastically reduced if the availability is improved by a more efficient maintenance planning, in this case some component have a “restoring time” of over 18 months, so reducing lead time of some of these element to 12 months (i.e. by a better inventory planning and spare parts use) could provides the same quality of service reducing 30% the aeromobile fleet and saving several million dollars per helicopter. Similar problem impact bus fleets of a mass transportation company, obviously with different impact of availabilities, lead times and costs, but very interesting saving on the large fleet of vehicles.

These cases are related to complex entities, where a predefined preplanning of operations is requested in order to guarantee the safety and efficiency; obviously in these cases, where maintenance is mostly preventive, it becomes possible to have high margins by a smart organization respect to cases where the breakdowns are mostly pure stochastic events with very large standard deviation respect mean expectations.

To provide an additional case it is interesting to compare the problem of maintenance in Fossil Power Plants and Gas Turbines; the first kind of plants are usually very high tailored, and pre-planned maintenance have a different impact respect Gas turbines, where the high solicitation requires very detailed pre-planned actions at precise time events. In this framework to use models in order to improve the overall efficiency it is quite critical and could provide significant savings. This paper present a case study just related to a group of Gas Turbines distributed over 11 plants; it is interesting to note that during the last 10 years the Service Business Unit was growing to a leading positing in term of profits in all major Power Plant providers, guaranteeing usually 50% of the profits with just 10% of the personnel; this is a confirmation of the investments of the Power Plant users in this sector and of the interesting opportunities for providing additional improvements in the sector, guarantee a durable benefit. In most case, advanced Power Plant users, today requires special contracts supporting not only turn-key, but operation and management and even more; the case presented is related to an user interested in establishing a set of procedures and tools for guaranteeing a on-line control of Service performances and actions for continuous improvements.

## APPLICATION FRAMEWORK

If we consider the problem to manage effectively the service of Gas turbine power plant it is important to define the boundaries of this framework; first of all each of these plants produce several thousand dollar of power each day and so each additional day of “non-

production” introduce very high costs. In addition the dynamic components of turbine and generators requires continuous special inspections due to high mechanical/thermal solicitations, where the intervals are technically predefined and related to the use of the different machines. It is evident that Gas Turbine use is strongly dependent on a large quantity of stochastic components, first among the others the behaviour of the demand and the market evolution; based on these parameters the request could require to work intensively, to put the machines “out of line” or to operate at “110%” with additional stress; the consumption of Equivalent Hours (the technical parameter for use of these components) is highly variable for each machine. In addition during each inspection the quantities of components to be substituted is variable based on a very unpredictable set of technological aspects. In term of management the lead time to have the spare parts available, due to their “construction complexity” and high technologies involved, is sometime pretty long (over 6 months) and involve sensible investments; these lead times are obviously stochastic as well as the related costs (due to the market situation and to the urgency of the requests, introducing expensive expedition activities). Similar problems affect the refurbishment/revamping of components (i.e. blades) that due to the nature of the processes (i.e. recoating) involve stochastic times, quantities and quality. All these factors combined provide a very complex framework with many interactions among different aspects. In addition the preventive planning of the power plants, especially if the number of turbines is quite high, introduce strong correlations among the different events: for instance two inspections in the same site could not be feasible due to available space for mounting/dismounting the components, while technical constraints could request to stop a turbine before to overpass some limit threshold. The Power Plant users have additional needs, often stated in a contract, to guarantee to have minimum production capability during the different part of the days/week/year, and/or concentrating/distributing maintenance operations. In addition the relationship among the different site considering the planning horizons of about 5/10/15 years with 10 power plants it is evident that a delay/shift in an inspection affects all the following events with possible critical results. This brief description provide an evident framework of the necessity to move forward to extensive use of modelling and simulation as support for developing an effective planning also in the simple case of a single user single provider with a limited number of power plants. In effect the authors developed in the past several models for being applied to this sector, involving case with multi users scenarios, geographical service clustering operation, new warehousing solutions etc.; in effect the research presented year is based on a case developed in cooperation between BRB, DIP Genoa University for a Power Plant Provider, on a group of about 11 combined cycles of a common user.

These power plants planned to be completed in 2003-2004 have a related service agreement introducing the necessity to apply maintenance procedures in accord with the RCM theory; this explicit request from the user justified the development of an ad hoc model for analysing the planning and simulating the inventory level of inspection kits.

## PROJECT OUTLINE

It is interesting to provide an overview of this project timeline in order to report the evolution of this M&S (Modelling & Simulation) project and the impact of reusability, portability and other aspects that affects these developments. As already mentioned the authors had experience in cooperation with major Power Plant Providers in developing models in the past to face different aspects (development of new regional maintenance organisations, introduction of outsourcing, optimal planning etc.); around the beginning of the new millennium the authors developed for a Construction & Engineering company, with a Power Division, a solution for being integrated with ERP (Enterprise Resource Planning) for supporting management of service division; this solution had to face a wide set of open issues:

- Large Set of Information Dbase and Tools
- Overlaps among Management Systems
- Miss-Integration in IT systems
- Management Gaps in Covering Service Processes
- “Early Preliminary Phase” of new ERP Implementation Project
- New Logistics and Organisational Changes
- Limited Historical Experience in Gas Turbines
- Short Delivery Time

Due to these scenario the author proposed a solution based on the development of new system to be integrated with ERP based on well defined business procedure, but able to support decision making process from managers also stand alone.

The Solution integrated a Simulator, as core engine for providing estimations, risk analysis and time/cost/quality figures, a new common database, a decision-making module, and some optimisation units for inventory management and inspection scheduling.

This solution was applied to some industrial cases with success, however it was requesting significant investments to be acquired and implemented.

The origins of this project were to develop a support for a specific project (service contract for 11 power plants) imposing a reduced budget and development time for providing a solution to this case; the application case was very similar to the previous mentioned however the final power plant user provided just guidelines and fixed budget for guaranteeing application of RCM policies to the acquired power plants.

The Power Plant Provider requested a specific ad hoc solution that was able to provide real benefits to the final user and eventually to be extended to other cases,

but obviously confirmed the constraints in term of time and budget and look around for possible providers to the different requested lines (i.e. FMECA analysis).

The authors presented a modular proposal and started to negotiate the details; it was immediately evident that the full integrated solution (even if this company was using an ERP already supported for integration) was not suitable due to the resources involved in this specific project. During the contacts it emerged the interest for a module able to provide a smart evaluation/optimisation of the gas turbine planning; in this way it was possible to provide a convenient proposal based on the reuse of conceptual models developed by the authors for the integrated solutions, to be implemented in a light tool.

The modules originally were two fuzzy evaluators, one focusing on inventory management (estimating the critical level of the items/components combining costs, lead time, critical level on inspection etc) and another one focusing on scheduling issues; both modules was applying fuzzy rules in order to consider the stochastic nature of the parameters and were integrated with the simulator in order to have an iterative automated optimisation/validation of the results.

The critical arising point was that the simulation model to be used in this case was not including many parameters fundamentals for some optimisation module (i.e. details of the item logistics and material flows due to the fact it is much less detailed) so it was difficult to reuse the optimisers within this project. Vice-versa the schedule evaluator was much more reusable and it was possible to be implemented by available data as support for the critical phases of settings and validation. So the project was outlined, however due to the fact that a significant amount of resources was allocated to traditional FMECA analysis to another provider, it was decided in this phase to develop just the schedule fuzzy evaluator without optimisation. The fuzzy optimiser in effect was using a simulator in close loop for iterating the alternatives and evolving to overall planning improvements; the solution defined was to develop the fuzzy evaluator and to keep open space for further developments of the optimiser. The technical requirements were fixed and the conceptual model was verified in order to guarantee the consistency, introducing changes and updates based on the specific case; for instance the original model was much generic, emphasising multi-customers/users, while in this case the single user allowed to a different tailoring with more attention on single power plant site reports. Reporting and Key Performance indexes were redefined based on the user needs as well as user interface and database integration. In the middle of the project, after the completion of conceptual model definition, during early phase of implementation a specific management problem show up. The Power Plant Provider had the necessity to evaluate different kit configuration to serve these power plants, based on some preliminary analysis carried out; however the evaluation was required in pretty short term and without increasing the budget, so a part of the development phase, that was proceeding

quite efficiently, was moved to adapt a Monte-Carlo simulator for inventory evaluation.

The authors decided to provide quickly two alternative models for evaluating these phenomena: an analysis based on queue theory and a Monte-Carlo simulator adaptation from a similar case already available. This approach was very satisfactory because allowed in very short time, whit the resource available, to support the decision, providing a comparison between the preliminary independent estimations developed by the company and the queue model, but also correcting these quantities by the more comprehensive Monte-Carlo model. The results obtained allowed to support the relation between power plant user and provider in defining the kits to be used in the inventory.

In the meantime the project evolved and the implementation was completed; the first critical issue was the model setting, originally based on similar case experience from the authors.

However during the accreditation procedure with the model users, it arises that some of the hypotheses from the developers were too general and don't guarantee to provide a common baseline with the traditional reference of the company; so it was decided to proceed to a supervised setting & testing in order to guarantee the model consistency; this approach required some additional time, due to the necessity to coordinate developers and users, however it was necessary to guarantee the accreditation. After this phase common tests were developed and the results were evaluated as fully satisfactory from the evaluation model.

After the completion of the project, during the final phase of period of guarantee, the team of the Power Plant Company, more directly involved in overall planning of inspections was interested to review the models and the simulation; this provided an opportunity to define possible developments in order to cover additional aspects with special attention to a stochastic more detailed simulation of planning and inventory levels. Currently the authors are finalising a proposal for proceeding in this direction by the development, in multi steps, of different modules for simulating others groups of power plants with high details and introducing evolutionary optimisers (in close loop with this new simulator) for inventory and schedule; while the final phases of the project expect to have a full integration with global "industrial optimisation" of the planning. These new modules will be integrated with original fuzzy evaluator for providing support to the hierarchical estimation of the critical components of a planning provided by the optimiser.

## **FUZZY MODEL INTRODUCTION**

The idea of realising a Fuzzy Schedule Evaluator was born from a real case to be analysed. The biggest Italian company specialised in Power Plants realisation and maintenance, had to build for an major customer a group of 11 Combined Cycle Plants. The agreement

between the company and its customer included the maintenance service for all the plants following the principles of RCM (Reliability Centred Maintenance).

The previous fruitful cooperation among the company and the authors, led to think to a possible advanced research to be presented to the final customer as a value added service. In effect the Authors had previous experiences in developing ERP-Integrated set of tools for maintenance and stock level control, applied in different real industrial contexts, including software for Failure Modes and Effect Analysis (FMEA/FMECA), a scheduler and a simulator integrating Artificial Intelligence (AI) techniques such as Genetic Algorithms (GAs). The interest on the subject, led the research to be developed on the subject of scheduling maintenance actions on the plants. Combined Power plants in fact have various kind of maintenance interventions, the most important of them are Minor Inspections and Major Inspections, that are made respectively each year and each three years of activity of the plant, depending on the Equivalent Operative Hours (EOH) of work, typical of each turbine. Different kind of maintenance requires also different kind of maintenance kits, composed by spare parts characterised by utilisation probability coefficients and time for refurbishment. This is the reason why another theme on which the study in object had focused was the determination of the criticality of each different kit of spare parts. Usually, the starting point for the definition of maintenance planning is the fixing of the start-up date for each plant, called PAC date. PAC dates of all plants are usually distributed on a timeframe in such a way to optimise resources and successive maintenance plans, in order to have a homogeneous distribution of workloads. Starting from the settled dates, the maintenance interventions could be planned based on a series of limits and constraints. Policies and constraints in a maintenance plan for power plants could be of various kinds, but the most important are:

- Strong constraints (e.g. contractual the customer does want the stop for maintenance in a specific month of the year)
- Regular constraints (e.g. the distance between two minor inspections, where the turbine producer provide just recommendations of the review interval)
- Weak constraints (e.g. requests of preference. not specified in the contract framework)

These and other rules are the basis of the logical processes introduced in the FUSE model. All the constraints have been introduced in the schedule evaluation system following the rules of Fuzzy Logic.

## THE FUZZY EVALUATION MODEL

The goal of FUSE research was to determine an evaluation system based on advanced techniques in order to provide a measurement of performance for a maintenance plan determined by Subject Matter Experts (SME), and able to identify where are located

criticalities, so that experts can correct punctually the main weaknesses of the planning.

The rules on which the evaluation schedule model is based are the rules of maintenance for Power Plants, plus specific rules introduced by the company for a better management of the resources. Every rule has been introduced in the model according to Fuzzy Logic. This technique allows considering a wider range of possibilities than traditional techniques. Introducing a different weight for different kind of constraints and considering more cases than simply *true* or *false* for the respect of each condition, the Fuzzy Logic allows to analyse in a more precise way a problem like the evaluation of maintenance planning of power plants, that is typically very complex, non-repetitive and so not very provided with historical data.

The complexity of the problem is increased by the fact that power plants to be maintained can be property of different customers, or located on different sites, or both conditions. These conditions have been considered in the logic of the model because are fundamental for the definition of some of the rules that allow to evaluate planning. An additional problem is connected to the availability of resources for maintenance in terms of spare parts. As previously mentioned, different kits are used for different inspections, and inside kits there are different spare parts with their own usage/consumption probability and lead-time for refurbishment. In order to determine the need of spare parts and their availability in relation to the maintenance plan, a further analysis on the kit criticalities has been made in the framework of the cooperation. The problem was in effect to define if the forecast number of kits for each kind needed was sufficient for the basic hypothesis on which the plan was structured. As mentioned, the PAC dates of each plant will determine more or less the period of the years in the timeframe in which the different maintenance inspections of each plant should be made. So, defining the occurrence in the same time of same kind of inspections on different plants, it is possible to define how many kits for each kind are necessary to respect the sequence of inspections, considering contemporaneous or too near interventions. The Company divided the plants into groups considering them as they were belonging to different customers, based on an homogeneous distance among different PAC dates, but mostly on the analysis of different kits criticalities. The groups of customers cycle among their plants the different kits they have, so the planning and the kit availability are very strictly connected.

## THE LOGIC RULES AND CLASSES

Fuzzy Logic principles need to follow a process based on Fuzzyfication and Defuzzyfication of data. For this reason the authors followed the below mentioned approach. First of all, the model considers some parameters that are typical of each plant, such as EOH

conversion coefficient (in hours), last minor inspection date, last major inspection date, customer name and site. Plants information is one of the input files of the model, containing all constraints, while a second input file with maintenance planning completes the data set. The data available include among the others:

- Group id. Number and Name,
- Location Site (different plants could be on the same site),
- Customer name (the same customer can hold various groups on different sites),
- Last information update instant (if not yet working, it's the forecasted start up date),
- Working hours in the last update (delta from the last major inspection)
- EOH factor for considering strain and stress of components
- Last minor inspection date (this includes also major dates due to the fact that major inspection includes always a minor one)
- Last major inspection dates
- Interval between two minor inspections, same group
- Interval between two major inspections, same group
- Weight for the constraint on interval between two major same group
- Interval between a minor inspection and a major inspection, same group
- Weight for the constraint on interval between a major inspection and a minor inspection same group
- Interval between two minor inspections on different groups on same Site
- Weight for the constraint on interval between two minor inspections on different groups on same Site
- Interval between two major inspections on different groups on same Site
- Weight for the constraint on interval between two major inspections on different groups on same Site

The scenario to be analyzed is really complex, in fact it was necessary to consider intervals and constraints for each planned maintenance. In order to do this it was implemented a Fuzzy Logic Module inside. Also to carry out an estimation of the maintenance planning considering all the intervals and the constraints it was applied Fuzzy Logic.

MINOR INSPECTION CONSTRAINT				
				8900
TV		0	4150	0.00
V	0	4150	8300	0.00
G	4150	8300	10375	0.71
L	8300	10375	12450	0.29
TL	10375	12450		0.00

After this session the model estimate if it is possible stop the group in the month for the planned maintenance. It was possible to define five different classes to convert this constrain in Fuzzy Values where TV means Too Close, V means Close, G means Perfect, L means Far and TL means Too Far. In the example the interval between the next Planned Maintenance and the previous activate the G Class with a value of 0.71 and the L Class with a value of 0.29. In the same way it was defined the class for the interval between two major

inspection, a minor and a major on the group the site and the customer. Due the importance to respect of the constraint between two major inspections in the definition of the classes it's necessary to assign a specific weight to the respect of the different intervals.

### Schedule Performance Metrics

The origins of these initiatives are related to the market evolution of power plant service, with special attention to Gas Turbine, obviously in this area the target functions are related to some aspects:

- Quality of the service*
  - Power Plant Availability
  - Delays in Operations
- Cost of the Services*
  - Inventory Costs
  - Expedition Extra Costs
  - Direct & Indirect Costs
- Constraints Respect*
  - Contractual Terms
  - Technical Requirements
  - Other Aspects

The authors decided to use for measuring the overall performance to combine together the different terms by applying the overall fuzzy model by a hierarchical approach that was able also to consider the time distance in the future of the event versus an horizon fuzzy factor, defined by the user, to have a smooth estimation of impact of far future problems (i.e. overlapping of two major inspection in 15 years in the future is much less critical than the same event in the next 12 months, because in the first case it could be easily correct in advance, while the second don't leave degree of freedom). The Model attributes excellence to the planning with a reading key, which enables to understand where the most critical points are. The simulator uses a hierarchical approach to identify the excellence of the input planning. The final excellence is the results of the excellence of the planned maintenance of the each group of the scenario. The planned maintenance quality is defined using Fuzzy Rules in order to estimate the respect of the constraints. For each planned maintenance the model calculates the intervals between each planned maintenance in order to verify the compatibility with the input constrains and these are converted in Fuzzy measures using Triangular Membership Classes with 50% overlapping.

### Simulation Model Metrics

The simulator was estimating delays on the inspections due to the unavailability of kits; the different kits with their lead times and with the necessity to refurbish part of the relative inventory was defined as stochastic components; obviously these delays represent usually just shift from the predefined planning, however they could introduces additional problems as well as power

plant stoppage with very high costs. The authors developed a risk level factor as metrics, this was the probability to overpass a cumulative number of delays during a timeframe over a set of turbines; to this risk level corresponded an expected mean and standard deviation on the delays on major and minor inspections (aspects with different impact). By the simulation it was possible to obtain an estimation about the current situation, and the obtainable situation with a different number of kits; these results provided a significant improvement in term of delay reduction. Currently the results obtained are confirming the expectations and the statistical database of effective performance is growing.

## CONCLUSIONS

Based on these developments it was possible to complete in quite short term, with very reduced resources a challenging problem. The stronghold in this development was an extensive development of the expertises and experience in this field: it was evident that previous conceptual models were well defined and guaranteed possibility to be tailored. Vice-versa the implementation, data integration, and target functions were requesting a significant cooperation developers/users. It was confirmed that management criteria are heavily related to the case study, however the flexibility of the original model architecture allowed to proceed successfully in retuning the model to new policies, providing satisfactory reference baseline.

## REFERENCES

1. Boyce M.P. (2001) "Gas Turbine Engineering Handbook", Butterworth-Heinemann
2. Bruzzone A.G. (2004) "Power Plant Service Evaluation Based On Advanced Fuzzy Logic Architecture", Proc. of SCSC2004, San Jose, Julye
3. Bruzzone A.G., Kerckhoffs (1996) "Simulation in Industry", Genoa, Italy, October, Vol. I & II, ISBN 1-56555-099-4
4. Bruzzone A.G., Colla G., Mosca R., Scavotti A. (1998) "Simulation as Verification Support for Qualitative Risk Analysis in Industrial Facilities based on Fuzzy Logic", Proceedings of Emergency Management Conference, Boston, 4-9 April
5. Bruzzone A.G., Giribone P., Revetria R., Solinas F., Schena F. (1998) "ANN as a Support for the Forecasts in the Maintenance Planning", Proceedings of Neurap98, Marseilles, 11-13 March
6. Bruzzone A.G., Colla G., Mosca R., Scavotti A. (1998) "Simulation as Verification Support for Qualitative Risk Analysis in Industrial Facilities based on Fuzzy Logic", Proceedings of Emergency Management Conference, Boston, 4-9 April
7. Bruzzone A.G., Mosca R., Pozzi Cotto S., Simeoni S. (2000) "Advanced Systems for Supporting Process Plant Service", Proceedings of ESS2000, Hamburg, Germany, October
8. Bruzzone A.G., Mosca R., Simeoni S., Pozzi Cotto S., Fracchia E. (2000) "Simulation Systems for Supporting Gas Turbine Service Worldwide", Proceedings of HMS2000, Portofino, October 5-7
9. Bruzzone A.G., Simeoni S. (2002) "Cougar Concept and New Approach to Service Management by Using Simulation", Proceedings of ESM2002, Darmstad Germany June 3-5
10. Bruzzone A.G., Mosca R. (2002) "Simulation And Fuzzy Logic Decision Support System As An Integrated Approach For Distributed Planning Of Production", Proceedings of FAIM2002, Dresden, July 15-17
11. Bruzzone A.G., Giribone R., Revetria R. (2002) "Integrating Small & Medium Enterprise in an Eprocurement using Java Applet Technology", Proceeding of SCI2002, Orlando, July
12. Bruzzone A.G. (2002) "Supply Chain Management", Simulation, Volume 78, No.5, May, 2002 pp 283-337 ISSN 0037-5497
13. Bruzzone A.G., Revetria R., Briano E. (2003) "Design of Experiments and Montecarlo Simulation as Support for GAS Turbine Power Plant Availability Estimation", Proceedings of MIC2003, Innsbruck, February 10-13
14. Cox E. (1994) "The Fuzzy System Handbook", AP Professional, Chestnut Hill, MA
15. Elliott T.C., Chen K., Swankamp R., (2002) "Standard Handbook of Powerplant Engineering", McGraw-Hill, NYC
16. Giribone P., Bruzzone A.G. & Tenti M. (1996) "Local Area Service System (LASS): Simulation Based Power Plant Service Engineering & Management", Proc.of XIII Simulators International Conference, New Orleans LA, April
17. Giribone P., Bruzzone A.G. (1997) "Design of a Study to use Neural Networks and Fuzzy Logic in Maintenance Planning", Proceedings of Simulators International XIV, SMC'97, Atlanta, Georgia, April
18. Giribone P., Bruzzone A.G. (1998) "Development of Innovative Maintenance Support Techniques", Proceedings of Applied Informatics'98, Garmisch-Partenkirchen, Germany, February 23-25
19. Levitt J. (2002) "Complete Guide to Preventive Maintenance", American Management Association
20. Montgomery D.C. (1997) "Design and Analysis of Experiments", Wiley & Sons, NYC
21. Robinson C.J., Ginder A.P. (1995) "Implementing Tpm: The North American Experience", Productivity Press Inc
22. Wang L.X. (1997) "A Course in Fuzzy Systems and Control", Prentice Hall, Upper Saddle River, NJ
23. Wiereman T. (1999) "Developing Performance Indicators for Managing Maintenance", Industrial Press, NYC
24. Zadeh, L. (1965) "Fuzzy Logic for the Management of Uncertainty" Janusz Kacprzyk Editor

# THE MODELING TECHNOLOGIES EVOLUTION FOR FOSSIL POWER PLANT SIMULATORS

Alexander Rubashkin, president of “Power Plants Simulators”  
Vladimir A. Rubashkin, technical director of “Power Plants Simulators”  
Russian Federation, Moscow, Semenovskiy per., 15, office 224  
[pps@edunet.ru](mailto:pps@edunet.ru), [www.fpps.ru](http://www.fpps.ru)

## KEYWORDS

Simulation technologies, fossil power plants.

## ABSTRACT

One of the main thesis of different simulator vendors is that all good simulators are arranged internally more or less in the same way. It is not the truth. The article describes and compares 3 generations of simulation technologies for fossil power plants.

## INTRODUCTION

This article is devoted to the consideration of different technologies for developing the models of fossil power units used by development engineers in the world. Depending on the used technologies for developing these models the authors distinguish the three generations of simulators for fossil power plants. The main features of models of each generations are described below.

One of the main thesis, which, as a rule, the development engineers of different companies uphold, consists of the fact that all good simulators are arranged more or less in the same way. Therefore the question, where to place the order for a simulator, consists for the Customer mainly of the price and the personal preference.

## SIMULATORS WE ARE SPEAKING ABOUT

First of all, we must mention, which simulators are involved here.

The persons of different specialties are engaged in the electricity production. It is clear that all of them must be trained, and the different specialists must be trained in the different way and with the different means of training.

In this article it is a question of training means for the boiler and turbine operators.

The professional skill of operators includes as minimum the two main components:

- Theoretical knowledge, for example, the knowledge of maintenance manuals
- Skill (or practical skills) to control the power unit.

In this article we don't concern the theoretical knowledge, while we discuss only the simulators, the main goal of which is the training of practical skills.

Now the following question is natural: which practical skills of operators should be trained? It is evident, for

example, that the operators need the practical skills including the motor ones and for working with control system installed at the unit – it can be either the traditional automatic control system (ACS) with operating board or the modern DCS. In this case, even if ACS covers completely all the regimes of equipment operation, while this completeness is realized far from always, and even if the ACS utilization factor is high, the situations appear frequently in the real operation, when the operator must make decisions on controlling the power unit and to carry out them in operation under conditions of rigid limit of time. This skill must be trained.

It should be noted beforehand that many development engineers of simulators, which can be used in the best case for training the operators to work with ACS, try to persuade the potential customers that if the simulator can be used for training to work with ACS, it can be used the more so for training the operators to control the power unit in the complicated technological situations. But it is not the same in reality.

## WHAT DETERMINES THE INSTANTANEOUS VALUES OF PARAMETERS IN A REAL LIFE

Before describing the different technologies for simulating the dynamic processes it is necessary to give to the reader, which is far from dynamics problems, the explanation, what determines the instantaneous values of different physical parameters in the real physical system.

We take for it the example. Let us have a cold room, where we have just switch on a heating device. What determines the air temperature in this room in 1 minute? In 5 minutes? In other time?

It is clear that the more powerful are the heating device, the warmer is in the room in 1 minute and in 5 minutes. Thus, a temperature in the room depends in each moment on the “consumption” of coming heat. We put intentionally the word “consumption” relatively to heat in quotation marks, because we don't speak in such way about heat in our everyday life, though just this word defines the essence of matter.

And what is when a small hinged window pane or even the window are open in the room. It is evident that the temperature in room will be in 1 minute less than in the first case, but by how much less, it depends on the fact, what is open – the small hinged window pane or the window itself.

Thus, really the temperature in the room depends in each moment of time on imbalance of the flows of coming and leaving heat.

On what else does the temperature in the room depend in 1 minute after beginning of heating from? That is evident that it depends on the initial temperature in room. On what else? From the sizes of the room itself – than less is the size of room, the quicker it is heated with the same heating devices. It means that the less is the size of room, the warmer will be in it in 1 minute after beginning of heating with the same other conditions.

Thus, if to use the more formal language, the air temperature in a room is an integral of the flows of coming and leaving heat. The integration rate depends on the room sizes in the reversed proportion and more formally – on the general heat capacity of the room.

We considered the question of temperature. The same judgments and conclusions are applied to pressures, but in the case flows are a consumption of substance. The pressure in some point of real physical system is the integral of imbalance of coming and leaving substance. The coefficient at integral is inversely proportional to the inner volume (capacity) of the given point.

It will be shown furthermore that the application of these ideas to the development of power units models was the serious step forward in the technology of simulating fossil power plants.

## GENERATIONS OF SIMULATORS

In this Chapter three generations of simulators are determined on the basis of approach used for the creation of object models, and their characteristics are given.

### Simulators That Include A Model Directly Reproducing The Known Processes Of The Object

For simulators of the first generation (SIG) the models are constructed on the basis of known static and dynamic characteristics of the object. In the most cases the experimental data obtained directly at the operating object are the source of these characteristics. The off-line calculation methods can be sometimes used to obtain some characteristics. For example, in the past for some simulator projects in Russia the curves of transient response were first calculated on the basis of so called “Normative method of calculating the dynamic characteristics of one-through boilers”, while then the model was developed on the basis of these transient response curves. On the whole the model is constructed as a software system, which reproduces the known regimes and processes of the object. In this case the model structure reproduces primarily the structure of channels, which link the input effects with the outlet variables. In essence the object becomes the black box: its technological structure and design characteristics, which lie in the basis of how it works, remain outside the model frameworks. The dynamic characteristics, on the reproduction of which the model is based, are usually treated as the linear ones. In the very insidious cases the coefficients of transfer functions approximating them put in a dependence on some parameter that is considered to be a decisive one, for example, on the load. It is natural that the principle of superposition is used for calculating the reactions on a combination of inlet effects.

This principle can apply to any linear system. The intermediate regimes and processes, for which the experimental (or off-line calculated) data are absent, are realized by means of an interpolation. The regimes and processes outside the region of known data are obtained by using an extrapolation.

The essential advantages of such models are:

- a problem of securing the solution stability is practically absent, because the number of feedback's is minimal; it allows to carry out the calculations with the relatively large steps in time, i.e. with the small expenditure of computer time, and it reduces substantially the demands of calculating capacity of computers used for simulation;
- a possibility to separate strictly the work on development models between the specialists of different professions: some of them determine the characteristics and construct the scheme of channels, while the other ones reproduce these scheme and characteristics in the computer; the latter specialist – mathematicians and programmers – must not at all understand the technology and physics of the processes.

On the other hand this approach has some essential shortcomings:

- low accuracy being the result, first of all, of the fact that the substantially nonlinear object (the power unit is a perfectly such object) is reproduced as a linear one and, secondly, due to the fact that any initial characteristics obtained in experiments on a real object are known to have not a high accuracy;
- low reliability of the processes, which can be reproduced with such model for the intermediate (interpolation) regimes and especially those coming outside the frameworks of experimental data (extrapolation), in particular, for the start-ups;
- such model can't be constructed for the object, for which the experimental characteristics haven't been determined, i.e. for the object being in the stage of design, construction or mounting.

The result of low model's accuracy and reliability consists of the fact that the thermal and mass balances are quite often not fulfilled in them, and the users first of all define this shortcoming.

### Simulators That Include A Model Based On Conservation Equations With Coefficients Obtained From An Experiment

The recognition of fact that the models must be constructed directly on the basis of physical laws, which define the functioning of real object, became a significant step ahead in the area of developing the models of power units. First of all, they are the laws of conservation of energy (heat), mass and momentum. For example, the application of laws of conservation for developing the models has obtained in the USA the name of application of «the main principles». They are the same laws, which we spoke about in the previous chapter.

The laws of conservation are mathematically written as the differential balance equations. The heat balance is described by an equation, where the time derivative for temperature or heat content is proportional to the difference between the consumption's of heat supplied and

removed from the working medium. The temperature itself, or heat content, is calculated by integrating this difference (imbalance).

The heat balance equations are written for all components and working media under consideration, for example, for steam flowing through the superheater bank, flue gases given the heat to this bank, metal of bank tubes and so on. The mass balance can be described by an equation, where the time derivative of pressure is proportional to the imbalance of flow rates of supplied or removed working medium (steam and/or water, flue gases etc.). The pressure itself is calculated by integrating this imbalance.

The object is already not a black box in the model based on the set of differential balance equations. As the equations are written for the interconnected object's components, the structure of balance equations and their interconnection reflect the structure of object's components.

The principal questions are: how to determine the heat consumption or flow rate of working medium (steam, water, gases) being present in the balance equations as well as to determine the coefficients of derivatives, from which the dynamics of processes in simulator depends.

We will consider it on the example of heat balance equation for the metal of bank tubes of platen superheater of a boiler (not taking into account the metal distribution along the wall length and thickness):

$$M \cdot c_m \frac{dt_m}{d\tau} = \alpha_{out} S_{out} (T_g - T_m) - \alpha_{in} S_{in} (t_m - t)$$

where  $M$  is a metal mass of bank tubes,

$c_m$  is a specific heat of metal

$\alpha_{out}$  is a coefficient of outside heat transfer (from gases to metal)

$S_{out}$  is a heating surface on the gas side

$\alpha_{in}$  is a coefficient of inside heat transfer (from metal to steam)

$S_{in}$  is a heating surface from the steam side

$T_g$  is a temperature of flue gases (average)

$T_m$  is a metal temperature (average)

$t$  is a steam temperature (average)

$\tau$  is a time.

The difference between the heat flow from gases to the outer metal side (usually that is a heat supplied to metal) and the heat flow from inner metal side to steam (usually that is the heat removed from metal) is written in the right side of equation. The flows link this equation with balance equations of other components:

- balance of gases heat in the area of platens, where the same flow that is outer for metal is the heat removed from gases;
- heat balance of steam, where the heat flow that is inner for metal is the heat supplied to steam.

The calculation of these flows during simulation represents some difficulties. First of all, it concerns the determination of heat transfer coefficients and especially the coefficient of heat transfer from gases to metal. The problem consists of the fact that in the area of platens there are both heat transfers by radiation (due to the high gases temperature) and convection (due to the velocity of gases motion). The heat transfer coefficient for each kind of these heat transfers depends on the composition of flue gases, their temperature, the geometric characteristics of platens and gas duct etc.

The accurate calculation of these heat rates for all operating regimes of boiler could be fulfilled on the basis of formulae and recommendations of the widely known in Russia "Normative method of thermal calculation of boiler plants", which contains for it all necessary recommendations. However, this procedure is very complicated and demands a large number of initial design data.

It is much easier to determine these rates from the experimental data obtained at the real object. It is possible to see that these two heat flows for metal (supplied and removed ones) are equal in a steady mode of power unit operation. It is quite simple to calculate the heat flow to steam for a steady mode of the real unit, if the steam flow rate and pressure as well as the steam temperature before and after platens are known. All these parameters are measured, as a rule, on all power units. If a development engineer has the information for some steady modes, he can try to invent an approximating function for calculation of heat flows from gases to metal and from metal to gas depending on some parameters really measured at the power unit. For example, it is possible to construct the approximation function depending on the gases temperature in some point of gas path, where the gases temperature is measured at the real power unit. In our case it is not obligatory if this gases temperature were the gases temperature in the platens area, because there is a correlation between the gases' temperatures in the different points of gas path. At last it is possible to approximate not the heat flows themselves, but to make the approximation of heat transfer coefficients from gas to metal and from metal to steam on their base. Then it should be necessary to calculate in simulator the heat flow by multiplying the "supposed" value of heat transfer coefficient obtained on the basis of approximation procedure by the heat transfer surface and the temperature difference.

The same approach can be also used for another components of boiler and turbine.

It should be specially spoken about the coefficient of derivative, on which the dynamic properties of metal temperature in the area of platens (in our example) in the model will depend. In its physical sense this coefficient represents the general heat capacity of platen metal. The basic meaning of this heat capacity, which should be adjusted in the dynamic calculations depending on the current metal temperature, can be determined on the basis of detailed analysis of initial design and constructive data.

Another method consists of the simple using as adjusting coefficients the coefficients of derivatives in the balance equation (not only for the heat and not only for the metal). For any particular transient process (unloading, shut-down, cold start etc.) by changing and adjusting these coefficients you can realize the dynamic properties of power unit, which correspond to Customer's understanding or interpretation of how the power unit works. Of course the customer's understanding is based upon his experience. I.e. again the coefficients of model equation are determined on the basis of experimental data.

The coefficient of derivative can be tuned to satisfy the customer's interpretation if some static parameters (heat transfer coefficients, temperature differences, heat fluxes themselves) are calculated very accurately. Some authors call it the separation of static and dynamic problems

(while the nature “solves” these problems jointly – as the only problem).

We call the simulators using the models of such type as the second generation simulators (S2G). They have the substantial advantages before S1G, first of all because they reproduce in much more details the object’s structure and the physics of processes. The balance of heat and mass in all components and in the model on the whole is brought together in principle. It is easier to obtain the initial information from the real power plant, which the development engineer of S2G is using in the development, than for S1G, for which the much higher accuracy is required. It is here mainly a question of information on the static (steady state) and dynamic regimes of equipment operation, which can be obtained in some cases without the special tests of equipment – by the method of passive experiment.

However the same shortcomings are inherent in these models as in the models S1G:

- uncertainties in the accuracy and reliability of reproducing the unsteady regimes, for which the experimental data are absent (for start-ups, for not expected modes etc.)
- difficulties in developing such model for the object, which has no working prototype
- etc

These shortcomings are the result of the fact that in developing such models as well as the models S1G, though to a lesser degree, the development engineers come «from processes to processes» - from the processes that in one way or the other are fixed at the real object, to the processes, which are implemented on the model. Such models are quite convenient for the simulator development engineers, because in developing them they are responsible not for the model adequacy in any processes and regimes of objects, but for its adequacy only in the regimes, the data for which were received from Customer. The main goal is to satisfy the Customer’s expectations. There appears to be the main reason, why the majority of simulators’ development engineers not only in Russia, but in the world are still developing S2G.

### **Simulators That Include A Model Based On Balance Equations With Coefficients Obtained From Design Data**

The static and dynamic characteristics of real power unit are determined by a large number of factors, which can be arbitrarily divided in the following groups:

- main design parameters of equipment that are chosen in the design stage and are subject to the accurate evaluation such as:
  - values of heating surfaces in the different zones of boiler,
  - cross section for gas passing through the different zones of boiler,
  - amount and design parameters of high-pressure heaters and low-pressure heaters,
  - metal mass of separate components of equipment
  - etc
- Parameters generalizing some set of made design decisions, which are at the design stage not subject to

the accurate calculation for the concrete equipment; however basing on the designs the values of such parameters can be preliminarily evaluated on the basis of statistical data generalization for the similar equipment; the preliminarily estimates can be later adjusted after the completion of mounting and the putting the equipment into operation; for example, such parameters include:

- rate of use for different heating surfaces
- thermal resistance of insulation
- height of flame in furnace
- etc
- Outside factors not depending on the design parameters of power unit such as:
  - Composition of fuel coming at the present moment (for example, in some moment of time the fuel may have the elaborated moisture content or the boiler can operate with the mixture of different fuels)
  - Ambient air temperature
  - Temperature of cooling water
  - etc
- Factors depending on the distinctions of power unit maintenance such as:
  - Degree of heating surface contamination in boilers, of tubes in condenser etc
  - Value of air suction in the different boiler elements and in the turbine condenser,
  - etc

The operator interprets the properties of power unit through automatic control system installed on the object. Therefore the additional factors effecting on the perception of power unit properties by a man are:

- properties of measuring transducers and special features of their mounting,
- Properties of DCS

To take into account all these factors is so difficult, when the simulator is developed. It is also one of the reasons, why the technology of developing the models S2G was for a long time the main technology for developing the power unit models.

However the technologies has to evaluate, and it was necessary to make the next principle step. The result of this step was the technology of developing the models for simulators of the third generation (S3G).

The main features of modeling technology S3G are the following:

1. The modeling is based on the so called main principles:
  - laws of balancing heat, mass and momentum,
  - equations of water, steam and gaseous mixtures equations,
  - criteria equations of heat transfer
2. The united system of differential and algebraic equations, which describes its behavior in all operating regimes (from the cold start-up initial state till the nominal state of unit operation with full load) is constructed for the power unit being modeled.
3. **All coefficients** of this system of equations are directly or indirectly determined on the bases of **design data** of modeled object.

4. The values of a main amount of coefficients in these equations (no less than 95% from them in accordance with our estimation), which depend on the design equipment parameter that are subject to the accurate estimation, can be precisely calculated. These values are determined at the initial stage of simulator development and are the final ones. It means that the development engineer doesn't change the values of these coefficients in the process of adjustment and testing. A case represents the exception, when the mistake is found in determining some coefficient.
5. The values of statistically estimated and generalized parameters are evaluated. The coefficients in equations, which depend on them, are calculated by means of these values. In Russia the statistical evaluation of generalized parameters for boiler plants are taken from "Normative method for calculating boiler plants". For example, there are the statistical estimates of heat loss in environment, air flows in flue gases in the regenerative air heaters, rate of use for the heating surfaces in gas ducts of the boilers of different type and so on.
6. The method of statistical evaluation is also used for the factors depending on the special features of power unit operation.
7. If necessary, the values of statistically estimated parameters are furthermore corrected.
8. The external factors effecting on the equipment behavior must be for the models S3G the boundary conditions, which can be effectively changed in the process of operation. It means, for example, that a start-up of unit can be begun at S3G with the usual fuel and finished with a wet fuel.

An example of statistically estimated coefficient is the coefficient connecting the heat loss in environment with the difference of current metal temperature and ambient air temperature. If it will be so in future that within the frameworks of simulator and with these coefficients the power unit is cooled quicker or slower than in reality, the coefficients are corrected in the corresponding side - and nothing more.

The technology S3G makes the following important steps in comparison with S2G:

- the calculations of heat flows, water and steam flows in all regimes of modeled equipment operation are carried out on the basis of accurate formulas, and no approximation is used
- the heat capacities and inside volumes of all elements of power unit are taken correctly into account; in this case their values are not used as the adjusting coefficients for achievement of the necessary dynamic characteristics; the object model works directly with the initially calculated specific quantities of metal and the inside volumes of all power unit elements

Due to all above mentioned, S3G has the following important consumers' properties:

- sufficiently accurate reproduction of **any** static regime of equipment operation; for example, a special testing of a few S3G installed in training center of Moscow 26-th power plant showed that the discrepancy of values of the main static parameters of power unit operation at the simulators and at the similar re-

gimes of real equipment operation falls in the measurement error

- sufficiently accurate reproduction of **any** dynamic regime of equipment operation
- there is a possibility to train the personnel to power unit start-ups from **any** thermal conditions
- the models S3G allow to Customer to pay attention on the potential problems of measuring devices of real objects (transducers, thermocouples) or real DCS, because if a parameter of a steady mode in simulator and real object are not coincided, in many cases just the parameter in simulator is correct, while there is a problem at the real power unit either with the DCS or with the measurement devices

It is reasonably safe to say that S3G go not from the processes to the processes, as S1G and S2G do, but they go from the design data to the processes. The task of S3G development engineer consists of the fact to model correctly on the basis of design data the static and dynamic properties of separate elements, which form the power unit (furnace, platen superheater, pipeline, turbine valve, condenser, high- and low pressure heaters etc.); and then any variant of start-up, correct or not correct, can be reproduced in simulator from any thermal condition by natural way without the additional adjustments of simulator.

In fact the S3G is more than just a tool for training of beginners. In addition it is a tool for increasing the skill of the most experienced and skilled operators. They can test here any situations, which rarely occur. An adequate and explainable reaction will be in the result.

There is a sufficiently simple method for Customer to understand, either the development engineer creates S3G or not. As a result of development of simulator the Customer will receive in the best case S2G, if the development engineer:

- asks as the initial data for modeling from Customer the information on dynamic properties of real object, but not only the design data
- drag the specialists of Customer to the development of simulator before the moment of beginning its tests

## COMPARISON OF THE SIMULATORS OF DIFFERENT GENERATIONS

The models S1G are still used by the developer engineers, for which the creation of simulators is not the main professional activity. These models can be used for the following purposes:

- connection with a model of real DCS for the debugging of relatively simple DCS components and training of operator for working with these components
- initial training of the beginners, for which it is still early to fulfill at simulator the complex regimes and which can not yet evaluate the adequacy of simulator model

The S2G is a serious step ahead to the quality of modeling in comparison with S1G. As a rule the developers of S2G simulators are professionals in the field.

At present the most part of commercially developed simulators are the S2G simulators. The S2G allows to achieve the acceptable quality of technological process

modeling with its careful implementation in the case, if the main goal of training is formulated in such way:

- the training is limited by some multitude of beforehand defined operations
- the training is begun from the beforehand coordinated initial conditions (for example, the unloading of power unit from 100 to 70%, the power unit start-up from the beforehand agreed initial conditions)
- the training is carried out by the beforehand known scenarios (the unloading of power unit from 100 to 70% must be fulfilled only by the beforehand determined way)

Many development engineers of S2G are proud by the fact that they drag the Customer's specialists to the simulator development at the early stages of the development and that they use the data from the real object. They declare that it is the only way to approach the simulator to the real power unit.

Reading this description of simulators' generations, somebody from S2G development engineers can assert the following:

- All leading development engineers develop S3G. Only the procedure of determining the same coefficients in the same equations is different.
- The practice showed that there is a possibility to determine the same coefficients on the basis of static and dynamic properties of the power units.

Is it so? The equations of S3G models operates on the concept of metal mass (and by means of mass the heat capacity of this metal is calculated for the current temperature of metal) of not heated boiler surfaces: headers, by-pass tubes etc. The concrete values of mass for all such elements (for example, the by-pass tubes before the 1<sup>st</sup> injection, the supply headers, a header of the 1<sup>st</sup> injection, the outlet header and so on) are calculated in S3G by development engineers on the basis of design data, and they are placed in the model. Due to the distribution of object in the space, the not heated element has an independent meaning, and its mass can't be simply added to the mass of heated element. It is impossible to determine separately the mass of heated and not heated surfaces on the basis of experimental dynamic properties of boiler: unlike the direct problem of determining the sum by the items the reverse problem of division of sum to the items has no solution. Therefore, if a simulator development engineer is using in the process of model development the dynamic data from object or if he attracts the Customer's specialists to the development of simulators before the tests, it means that such simulators use the *other* equations in comparison with S3G. For example, these *other* equations unlike the S3G equations don't take into account the distribution of specific amount of metal along the spatial coordinate. To our knowledge, practically nobody from the leading world producers of simulators for fossil power plants even don't ask the Customer to present the detailed data on the not heated surface of boiler. It means that their equations don't take into account the heat storage in metal of these surfaces.

## THE PRESENT STATE OF THE ART IN THE WORLD SIMULATOR ENGINEERING

The Russian company "Power plant simulators" takes the leading positions in Russia in the area of development of simulators for training the operators of boilers and turbines due to the development of S3G.

In January 2004 the specialists of company took part in the annual conference "Fossil Simulation and Training", which was organized within the frameworks of "2004 Western Simulation MultiConference" by International Society for Computer Simulation (SCS). A lot of leading world simulator vendors participated in the conference.

The Russian company "Power plant simulators" has demonstrated at this Conference the abilities of its simulators and made a plenary report.

The main subject of Russian representatives' report was the technology of developing models for S3G, which is successfully used by this Russia company at the Russian market already for more that 10 years. More than 20 simulators were developed by this technology and are used successfully in Russian electric power industry. What is more, it is exactly due to the high quality of models the simulators of this Russian company were chosen as the basic software for organizing the International competitions of professional skills of fossil power plants operators - Cyberthon. These competitions have been twice successfully organized in the beginning in Republic of South Africa and then – in Russia. It is planned to carry out in October of 2004 the next international competitions of professional skills of fossil power plants operators - Cyberthon-2004.

At the discussion, which took place at the Conference "Fossil Simulation and Training"-2004 after the report of Russian representatives, it was evident that no one companies participated in the Conference is ready today to develop a simulator for fossil power plant on the basis of design data as it was done by company "Power plant simulators". All of them demand the intensive participation of Customer's specialists in the process of developing simulator from the beginning. In fact it means that all of them are developing only S2G.

Why does the simulator engineering stop at the boundary of S2G? Is it possible that the leading world producers were not able to create the simulation technologies for S3G? Certainly they could do it. But the conditions in the simulator engineering industry were developed in such way that the technology of creating S2G was extremely convenient for the development engineers including as well from the commercial point of view. And the Customer knew nothing that was better. It was explained to Customer that there was no other way for developing the qualitative simulator. IAEA (International Atomic Energy Agency) helped unintentionally to the simulator development engineers in this problem: IAEA made a decision that each nuclear power plant has to have s simulator. It signified that the power plant has not only to spend money for simulator, but it has in addition to be pleased with this simulator (if the simulator is not suitable for training the personnel, it is formally impossible to maintenance the nuclear power plant). The development engineers explained to Customer that if the situation has been formed, when the Customer was interested essentially in

the quality of simulator more than the development engineer itself, the Customer has to help to the development engineer to create a good product. So the Customer really became the coauthor of simulator. Using it the development engineer shifted to the Customer a responsibility for the simulator quality: the Customer provides its own engineers and presents the regimes, while the development engineer makes in essence “all you wish”.

Gradually such approach has migrated from the nuclear power plants to the fossil ones, because the Customers from fossil power plants were constantly “educated” that the experience of developing the simulators for nuclear power plants was the most advanced in the world.

In fact the transfer to S3G means for development engineer that he takes all the responsibility for himself. In this case the Customer keeps away from the development, and its natural role of outside critic of simulator quality is returned to him. Who needs in such responsibility, if the Customer doesn't require it?

About 10 years ago, when the favorable scientific-research contacts between Russia and West were begun, the “advanced” technologies of developing S2G prevailing in the world came in Russia and mainly in the nuclear power industry. The situation is natural in the world, when the same companies that specialized in simulators for nuclear power plants receive the large contracts and develop also the simulators for the fossil power plants. There is another situation in Russia: the leading Russian producers of simulators for NPP have yet not managed to develop no one serious simulator for a fossil power unit or power plant. One of the reasons consists of the high prices, with which these development engineers got used to work (if a nuclear power plant is obliged to have a simulator in any case, it is possibly to increase the price). However we believe that the question concerns not only the price. First S3G began to appear in Russia approximately in the same time, and from year to year the more and more power engineers know that these simulators are of very high quality.

## CONCLUSIONS

To get a next contract for simulator, the development engineers bring forward often as the evidence of their successful activity the examples of other “successful” projects claiming that the Customer was pleased with its simulator. The situation in simulator engineering, when “the Customer is pleased”, convinces little somebody. There are many reasons, why a Customer can be pleased with simulator. The different people can be pleased on the side of the same Customer. For example, a chief manager of a power plant can be pleased in public with simulator, while the ordinary operators can be not pleased. The chief manager can be pleased in public, because he paid money for the simulator and, he can't confess that these expenditures didn't justify the expectations. Another reason, why somebody can be pleased with the available S2G, simply because he was not familiar with a S3G.

The conclusion that S3G use the **different equations** and the **different technology** for simulating the power unit in comparison with S2G, which was substantiated above in this article, can explain many things:

- why the situation is practically unbelievable, when an operator of high class, which is well acquainted with some S3G, can be not pleased with it and at the same time to be pleased with some S2G
- why it is possible successfully to carry out with S3G the competitions of professional skills of fossil power plants operators including the international ones and suggesting to the participants to fulfill the most complicated tests continuing for hours; the different teams of operators fulfill in these tests the hundreds of different control operation in the different sequence, while the simulator reacts adequately on all their actions
- etc

To order of a S2G instead of a S3G can be justified from the viewpoint of Customer only in a case, if the S2G simulator is in few times cheaper than the S3G, because the customers' qualities of S3G and S2G simply cannot be compared.

At present according to our understanding of the situation, no one leading world simulator vendor for fossil power plants except of Russian company “Power plant simulators” has not even declared that he has been developing a S3G. Even if somebody will start today to change his own technology of simulation to S3G, he needs a few years for creating the necessary software and for development his first S3G.

## REFERENCES

1. A.Rubashkin, V.Rubashkin, “Simulation technologies for fossil power plants used in Russia” // 2003 Western Simulation MultiConference, Orlando FL, 2003
2. A.Rubashkin, V.Rubashkin, T.Shuk “Objective simulation of fossil power plants” // 2004 Western Simulation MultiConference, San Diego CA, 2004
3. “Heat calculation of boiler units (normative method)”. Moscow, Energy, 1973

## AUTHOR BIOGRAPHIES

**ALEXANDER S. RUBASHKIN** was born in 1936 in Moscow. He graduated from the Moscow Energy Institute. He has a PhD degree in the “power plant automation” field. For more then 20 years he worked in the world well known Russian engineering company that specializes on the fossil power plants - ORGRES. He started research and development in the field of fossil power plant simulation in the middle of 70<sup>th</sup>. In 1992 he founded and started up the “Power plant simulation” company.

**VLADIMIR A. RUBASHKIN** was born in 1963 in Moscow. He graduated from the Moscow institute of railway transport with “computer science” specialization. After the graduation for 6 years he had been working in the Moscow academic institute of control problems. From 1992 he has been working in the “Power plant simulation” company on the technical director position.

# GENERIC BI-LAYERED NET, AS THE NATURAL COMPUTATIONAL MODEL OF THE CONSERVATION AND INFORMATION PROCESSES

Béla Csukás

Gyöngyi Bánkuti

Institute of Mathematics and Information Technology, University of Kaposvár

Guba S. u. 10. Kaposvár, 7400, Hungary

[csukas@mail.atk.u-kaposvar.hu](mailto:csukas@mail.atk.u-kaposvar.hu), [bankuti@mail.atk.u-kaposvar.hu](mailto:bankuti@mail.atk.u-kaposvar.hu)

**KEYWORDS:** generic modeling, dynamic simulation, conservation and information processes, direct computer mapping, Generic Bi-layered Net

## ABSTRACT

In Direct Computer Mapping the simple building blocks of the conservational and informational processes are mapped onto the generic “active” and “passive” elements of an executable program. The recently developed Generic Bi-layered Net model provides a common framework for the simulation of the hybrid (continuous and discrete, quantitative and qualitative) balance-based and rule-based processes. The common features of the process models are represented by a bi-layered net of variable structure that also determines the network (ring) structures of the influence routes and flux routes, as well as the Gantt Chart view of the process.

## INTRODUCTION

The computer modeling of the continuous and discrete processes have been evolving in three different ways. The processes are usually described by a set of algebraic, differential and/or integral equations IPDAE (Pantelides, 2001). The ‘a priori’ (white box) models are derived from the simple first principle primitives, and then they are transformed into various sophisticated mathematical constructs. The ‘a posteriori’ (black box) models differ only in the origin, but the numerical solution of the identified mathematical equations is similar.

Artificial Intelligence developed various knowledge-based methods without the explicit consideration of domain specific structures and of fundamental conservation laws. The attempts to bridge this gap, with various kinds of qualitative models were not successful enough, because the qualitative knowledge representation evolved on its own, without effective connection to the quantitative modeling. The execution of the hybrid, discrete / continuous models is a difficult question, because the usual integrators do not tolerate the discrete events, while the usual representation of the continuous processes cannot be embedded into the discrete models conveniently.

The general formal models of the systems had be developed before the powerful computers appeared.

According to the Kalman’s approach (Kalman et al. 1969), the state space model of the continuous processes

$$\langle U, X, Y, \underline{x}(0), f, g, t \rangle \quad (1)$$

is described by the transition (f) and output (g) functions in the continuous time t:

$$\begin{aligned} \dot{\underline{x}}(t) &= f(\underline{x}(t), \underline{u}(t), t) \\ \underline{y}(t) &= g(\underline{x}(t), \underline{u}(t), t) \end{aligned} \quad (2)$$

where,

$\underline{u}(t) \in U$  = the input variables,

$\underline{x}(t) \in X$  = the state variables and

$\underline{y}(t) \in Y$  = the output variables

of the process. The abstract automaton representation of the discrete processes describes the same in the discrete time k.

The General Net Theory (Brauer 1980) describes a generalized net model for the description of the structures. Many net models, like the early appeared and very innovative Petri Net (Petri 1962), as well as the various State-Transition Nets belong to the above family.

Many recently used engineering methods had been established before the onset of powerful computers. Modeling starts either from the consideration of changes in characteristic measures, or from the rules and signs. Next this is transformed into mathematical construct. It usually cannot be solved, should be discretized, and finally, the computer executes simple arithmetical steps.

In *Direct Computer Mapping DCM* (Csukás and Bánkuti, 2003a), we can map the simple building blocks of the conservational and informational processes onto the generic “active” and “passive” elements of an executable program (see Fig.1.). The balance elements and the signs, as well as the elementary transitions and the rules can be described by brief uniform programs, executed by the same kernel algorithm. Direct Computer Mapping of process models allows the computer to know explicitly about the very structures and bounds of the physical world. In this knowledge representation, the model is organized rather by the transitions, than by the state. ***The key issue is that the computational software (and hardware) can copy the natural structure and building elements of the investigated problem.***

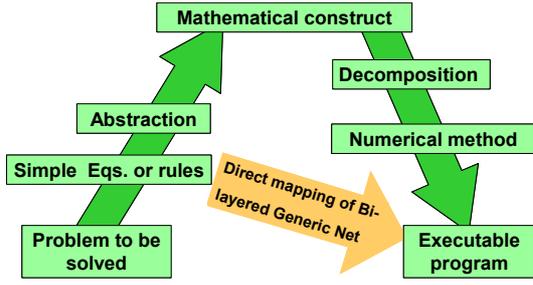


Figure 1: The idea of the Direct Computer Mapping

## GENERIC BI-LAYERED NET MODEL OF COMPLEX PROCESSES

The recently developed *Generic Bi-layered Net* model (Csukás and Bánkuti, 2003b) is a theoretically established and practically validated powerful realization of the Direct Computer Mapping. It is a special case of the General Net Theory on the one hand, as well as an explicitly structured, generic combination of the state space model and of the abstract automaton.

The *generic, bi-layered net model* can be defined by the ten-tuplet of

$$\langle P, A, B, G, X, Y, \Phi, \Psi, \underline{r}, t \rangle \quad (6)$$

where  $\langle P, A, B \cup G \rangle$  is a net. The communication channels  $B$  and  $G$  determine the passive  $\rightarrow$  active

$$B(\tau) \subset P(\tau) \times A(\tau) \quad (7)$$

$${}_j b_i(\tau) = \langle p_j(\tau), a_i(\tau) \rangle \in B(\tau) \quad (8)$$

$$\exists_j b(\tau) \mid \forall_i \langle p_j(\tau), a_i(\tau) \rangle \in {}_j b(\tau) \quad (9)$$

$$\exists_j b(\tau) \mid \forall_i \langle p_j(\tau), a_i(\tau) \rangle \in {}_j b(\tau) \quad (10)$$

and active  $\rightarrow$  passive data flows

$$G(\tau) \subset A(\tau) \times P(\tau) \quad (11)$$

$${}_i g_j(\tau) = \langle a_i(\tau), p_j(\tau) \rangle \in G(\tau) \quad (12)$$

$$\exists_i g(\tau) \mid \forall_j \langle a_i(\tau), p_j(\tau) \rangle \in {}_i g(\tau) \quad (13)$$

$$\exists_i g(\tau) \mid \forall_j \langle a_i(\tau), p_j(\tau) \rangle \in {}_i g(\tau) \quad (14)$$

respectively. Index  $j$  designates the ordered sets of the existing output ( ${}_j b$  (9)) and input ( ${}_j g$  (14)) connections for the  $j$ -th passive element. Similarly, index  $i$  defines the ordered sets of the existing output ( ${}_i g$  (18)) and input ( ${}_i b$  (10)) connections for the  $i$ -th active element. Variable  $\tau$  denotes the optional points or intervals of the continuous or discrete time  $t$ , declaring the existence of the respective elements and relations.

The *passive elements*  $P$  are associated with state variables  $X_j$  and with an operator, describing the change of the state:

$$\forall_j p_j \rightarrow X_j \in X; \psi_j \in \Psi; \psi_i = \begin{bmatrix} \underline{y}_j \mid \underline{g}_j \downarrow \\ \uparrow \mid \underline{x}_j \mid \underline{b}_j \end{bmatrix} \quad (15)$$

where  $X_j$  contains any structured data set, and operator  $\psi_j$  describes how  $\underline{y}_j$  changes the state via the channels  $\underline{g}_j$ . The *active elements*  $A$  are characterized by the operator  $\varphi_i$ , providing a mapping. This determines how the output changes  ${}_i y_j \in {}_i \underline{y}$ , carried by  ${}_i \underline{g}$  are calculated from the coordinated input readings  ${}_j x_i \in {}_j \underline{x}_i$  that comes from the passive elements through the channels  ${}_j b_i$ :

$$\forall_i a_i \rightarrow \varphi_i \in \Phi; \varphi_i = \begin{bmatrix} \underline{x}_i \mid \underline{b}_i \\ \uparrow \mid \underline{y}_i \mid \underline{g}_i \end{bmatrix} \quad (16)$$

The operators  $\psi_i \in \Psi$  and  $\varphi_i \in \Phi$  *may be anything from a simple input/output mapping to a brief program*, calculating the elementary process or the rule.

Variable  $\underline{r}$  designates the geometrical and/or the property coordinates of the distributed systems and/or population balances. As an example, the GBN implementation of a simple hybrid automaton (of the Single Switch Server problem) is shown in Fig. 2.

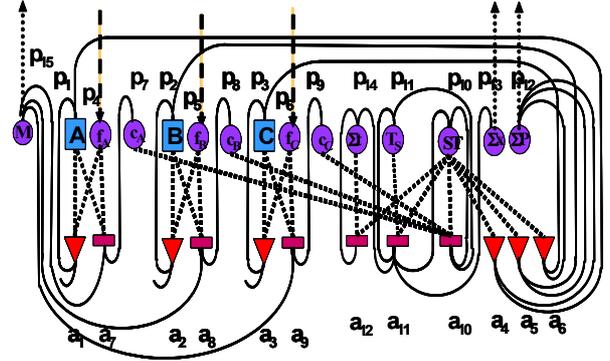


Figure 2: The network view of a simple hybrid model

Two examples for the passive elements and one example for an active element are the followings:

$$p_1 \rightarrow \Psi_1 : \begin{bmatrix} \Delta M_{A1} = 1Y_1 & \Delta M_{A4} = 4Y_1 \\ 1X_1 = M_A & 1X_7 = M_A \end{bmatrix} \quad (17)$$

$$\Delta M_A := M_A + \sum_i {}_i Y_1$$

$$p_{10} \rightarrow \Psi_{10} : \begin{bmatrix} St = 11Y_{10} & St = 10Y_{10} \\ \forall_i {}_{10} X_i = St \end{bmatrix}$$

$$St = \text{Prod } A \vee \text{Prod } B \vee \text{Prod } C \vee \text{Setup} \vee \text{Wait} \quad (18)$$

In the network view of the net we can interpret the alternating, connected, ordered set of the communication channels

$$\{ {}_j b_i, {}_i g_j, {}_j b_{i_2}, {}_{i_2} g_{j_3}, {}_{j_n} b_{i_n}, {}_{i_n} g_{j_{n+1}} \} \quad (19)$$

They are called *influence routes*, which determine a special network structure. The influence routes carry the influence. E.g. the perturbation of the content  $X_{j_1}$  of the element  $p_{j_1}$  affects the content  $X_{j_{n+1}}$  of the element  $p_{j_{n+1}}$ , according to the influence:

$$\{j_1 \Delta x_{i_1}(t_1), i_1 \Delta y_{j_2}, j_2 \Delta x(t_2)_{i_2}, \dots, i_n \Delta y_{j_{n+1}}\} \quad (20)$$

where  $j \Delta x_i$  and  $i \Delta y_j$  refer to the perturbation of the state and the change, respectively. The sensitivity and its special forms, such as observability and controllability can be studied by means of the influence route network. The minimal (generating) influence routes are the basic edges. The maximal influence routes are the transferring routes and the complete loops. The simplified structure of the influence routes is a special ring, where the two algebraic operations are the concatenation and the common part.

As we have emphasized, the existence of elements A, P, B and G, as well as the contents X and Y of the communication channels depend on the time while  $\tau$  denotes the well-defined points or intervals of the continuous or discrete time  $t$ , when the given element, channel or sign does exist. This temporal behavior of the system of variable structure can be seen from the Gantt Chart view (see Fig. 3).

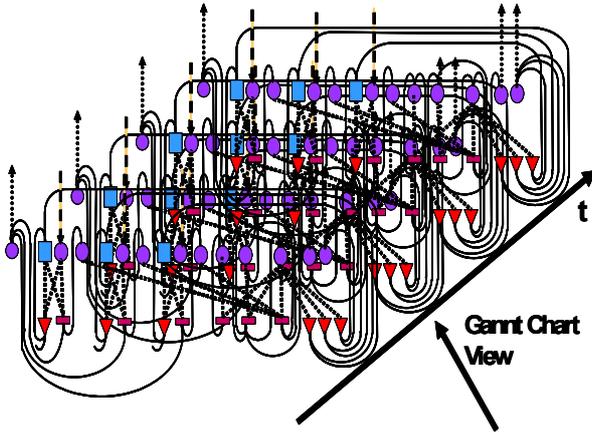


Figure 3: The Gantt Chart view of a hybrid model

## CONSERVATIONAL PROCESSES

An important special case of net (6) is the class of **balance processes**, where the basic part of the state  $X_j$  is a measure and, the operator  $\psi_j$  summarizes the simultaneous rates. Depending on the discrete or continuous time, operator  $\psi_j$  generates also the appropriate

$$\psi_j(\underline{y}_j) = \frac{dX_j[p_j]}{dt} \approx \frac{\Delta X_j[p_j]}{\Delta t} = \sum_i i y_j \quad (21)$$

difference or differential equations, called balance equations. In the balance model the descendent of the mappings,  $\varphi_i$  can be divided into two disjunct parts, corresponding to the increases (+) and decreases (-) of the characteristic measures:

$$i \underline{y}[i \underline{g}] = i \underline{y}^+[i \underline{g}^+] \cup i \underline{y}^-[i \underline{g}^-] \quad (22)$$

The active elements of the balance process models describe the various transportations and transformations.

Conservational process is a special case of the balance process, where there are constant measures  $\underline{C}$  determined by the model specific conservation laws. Simultaneously all of the measures  $\underline{M}$  can be combined from these constant measures, according to the respective stoichiometry  $\underline{S}$ , i.e.:

$$\exists \underline{C} \exists \underline{S} \mid \underline{M} = \underline{S} \cdot \underline{C} \quad (23)$$

$$\underline{0} = \underline{\Gamma} \cdot \underline{M} = \underline{\Gamma} \cdot \underline{S} \cdot \underline{C} \quad (24)$$

where  $\underline{\Gamma}$  is the process rate matrix.

If the operator  $\varphi_i$  can be determined by a well-defined single rate  $v_i$ , then the change of the conservational measures the expression of

$$\dot{\underline{M}}(t) = \underline{S}_{\Gamma}^T \cdot \underline{v}(t) \cdot \underline{V} \quad (25)$$

can be written for (where  $\underline{V}$  is the reference measure, e.g. the volume).

As an example consider in Fig. 4 the GBN model of a simple enzyme reaction, where the active elements correspond to the elementary processes of the a1:E+S  $\rightarrow$  ES; a2:ES  $\rightarrow$  E+S; a3:ES+I  $\rightarrow$  ESI; a4:ESI  $\rightarrow$  EW; a5:EW  $\rightarrow$  E+W and a6:E+W  $\rightarrow$  EW reactions, as well as of the a7, a8, a9 = transportations.

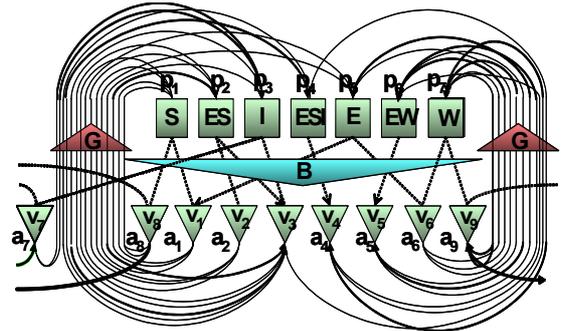


Figure 4. GBN representation of an enzyme reaction

where e.g.

$$p_1 \rightarrow S; \psi_1: \begin{bmatrix} 8Y_1, 1Y_1, 2Y_1 \\ 1X_8 = c_S, 1X_1 = c_S \end{bmatrix}$$

$$\Delta M_S = 8Y_1 + 1Y_1 + 2Y_1; c_S = \frac{\Delta M_S}{V}$$

$$a_3 \rightarrow v_3; \varphi_3: \begin{bmatrix} 2X_3 = c_{ES}, 3X_3 = c_I, 5X_3 = k_3 \\ 3Y_2 = -v_3, 3Y_3 = v_3, 3Y_4 = v_3 \end{bmatrix} \quad (26)$$

$$v_3 = k_3 \cdot c_S \cdot c_I \cdot V \cdot \Delta t$$

The respective measure vectors and stoichiometric matrices are the followings:

$$\underline{C} = \begin{bmatrix} C \\ H \\ O \\ E \end{bmatrix} \quad \underline{M} = \begin{bmatrix} S \\ ES \\ I \\ ESI \\ E \\ EW \\ W \end{bmatrix} \quad \underline{S} = \begin{bmatrix} 4 & 4 & 4 & 0 \\ 4 & 4 & 4 & 0 \\ 0 & 2 & 1 & 0 \\ 4 & 6 & 5 & 1 \\ 0 & 0 & 0 & 1 \\ 4 & 6 & 5 & 1 \\ 4 & 6 & 5 & 0 \end{bmatrix} \quad \underline{S}_{\Gamma}^T = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & -1 & 1 & -1 \end{bmatrix} \quad (27)$$

In the balance process models the alternating, connected, ordered set of the communication channels

$$\left\{ j_1 g_{i_1}^-, i_1 g_{j_2}^+, j_2 g_{i_2}^-, i_2 g_{j_3}^+, \dots, j_n g_{i_n}^-, i_n g_{j_{n+1}}^+ \right\} \quad (28)$$

is called flux route. Flux routes determine another network structure, which carry the constant and conservational measures. For example, if we modify the value of measure  $X_{j_1}$  in element  $p_{j_1}$ , then this change effects on measure  $X_{j_{n+1}}$  of the element  $p_{j_{n+1}}$ , i.e.:

$$\left\{ j_1 \Delta y_{i_1}^-(t_1), i_1 \Delta y_{j_2}^+(t_2), j_2 \Delta y_{i_2}^-(t_1), \dots, i_{n-1} \Delta y_{j_n}^+(t_n) \right\} \quad (29)$$

The  $\Delta y_{j_i}^{+/-}$  values refer to the dispersion of the changes in the rate of subsequent processes (multiplied by the stoichiometric coefficients). The minimal (generating) flux routes are the basic edges. The maximal flux routes are transferring routes and the complete loops. The simplified structure of the flux routes is a special ring, where the two algebraic operations are the concatenation and the common part.

### INFORMATIONAL PROCESS, AS A SPECIAL PART OF CONSERVATIONAL PROCESS

It is to be noted that a part of the above conservational model, responsible for the enzymatic control, can be replaced for a simplified model of rules and signs, respectively. This is illustrated in Fig. 5. where the signs are symbolized by circles and the rules are represented by bar nodes.

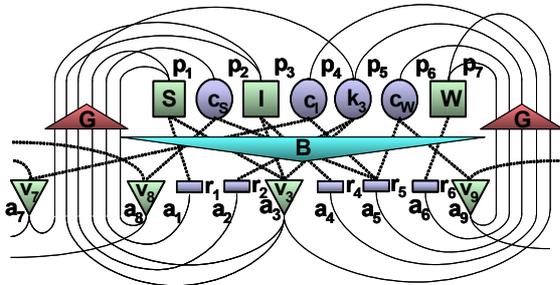


Figure 5: An example for the conservation based information process

Examples for the  $\psi$  and  $\phi$  mappings of a sign ( $p_6$ , evaluation of  $W$ ), and of a rule ( $a_5$ , calculation of the reaction rate) are as follows:

$$\begin{aligned} p_6 \rightarrow c_W; \psi_6 : \begin{cases} 6Y_6 \\ 6X_5 = c_W \end{cases} \\ c_W = M_W / V \\ a_5 \rightarrow r_5; \phi_5 : \begin{cases} 2X_5 = c_S, 4X_5 = c_I, 6X_5 = c_W \\ 5Y_5 = k_3 \end{cases} \\ k_3 = F(c_S, c_I, c_W) \end{aligned} \quad (30)$$

The human made “artificial” processes often do not have the above described self-determined control, but they can be supplied by an informational process,

determining the control signs and rules. Fig. 6 shows an example for the GBN model of a controlled heat exchanger. In the Figures V and W, as well as H and Q refer to the volume and the enthalpy of the inside and outside liquid, respectively. The inlet and outlet flows of the inside and outside agents are signed by the symbols  $V_b$  and  $V_k$  as well as  $W_b$  and  $W_k$  respectively. The heat transfer is symbolized by the elementary process  $Ht$ .

The temperature  $T$  is measured by the thermometer  $m$ , and it is compared with the set point  $a$ . With the knowledge of this difference, the PID controller calculates the control action  $u$ . In the right hand side the Generic Bi-layered Net model of the heat exchanger and the controller are represented by a connected pair of a conservational and an informational process. In the practical realization the informational process is carried out by another physical (electronic / electric, hydraulic or pneumatic) process (i.e. another conservational process). The apparently paradox, but meaningful notion of conservation based informational processes is illustrated in Fig. 7 more plausibly. Here the liquid flow through the jacket of the heat exchanger is controlled by another liquid flow through another vessel. By decreasing the size of the heat exchanger and by the simultaneous increase of the upper right hand side

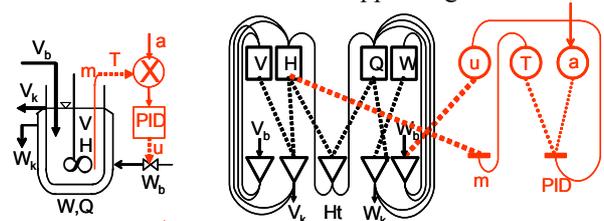


Figure 6: A conservational / informational process

volume, the exciting question appears whether the parts changed their relative position. It means beyond a certain point we recognize that the heat exchanger controls the other unit. From theoretical point of views, this results a new interpretation of the informational process. Accordingly, a given part of the conservational process behaves as an informational process with respect to its complementary part, if this special part consumes and produces significantly less conservational measures, than the complementary process, while, along the feedback influence loops and transferring influence routes the informational process exerts more influence on the complement, than the completing part on it.

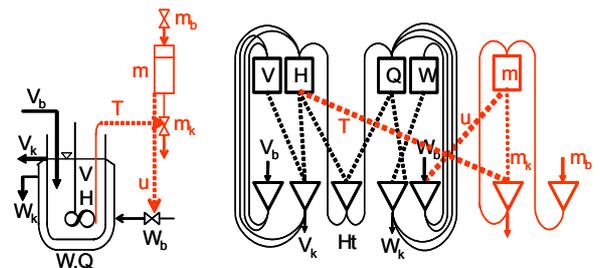


Figure 7: Plausible visualization of a conservation based information process

The informational process can be a special part of the self-determined natural processes (e.g. neural system, enzyme regulation), or it can be a supplied part of the non-self-determined artificial one (e.g. control systems). The essential feature of the informational process is that it

- transports negligible amount of conservational measures with the complementing part and with the environment,

- while it has a greater influence on the operation of the complementing part than *vice versa*.

If the above criteria are fulfilled, then it is not necessary to describe the conservational processes for this special subsystem. Instead, we can read, calculate and overwrite the appropriate signs simply. Accordingly, we neglect the conservational process carrying these signs, and deal only with the informational process carried by the vehicle conservation process.

### CLASSIFICATION OF THE PROCESS MODELS

The above described relation between the conservational and informational processes can be overviewed by the classification of the processes according to Fig. 8.

The set of balance processes is a subset of the Generic Bi-layered Net processes. Conservational processes are in a subset of the balance processes. Both of the balance and conservational processes might have a special part that consumes and produces less additive measures, but exerts more influence on the completing part. These special parts can be transformed into the respective informational processes. Marginally the whole balance or conservational process can be mapped into an informational process. Another case is, when the balance or conservational process is supplied with an informational process. In addition there are also primary informational processes.

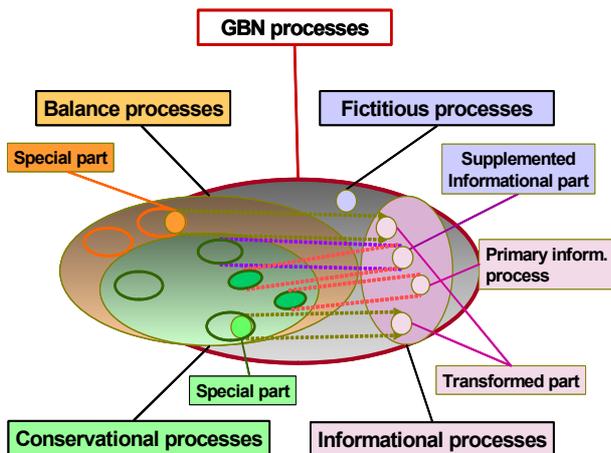


Figure 8: The classification of the GBN processes

On the other hand, all of the above described informational processes must have a vehicle conservational process that carries the signs and

executes the rules. The brain and the computers can work as conservational processes, themselves. It is to be noted that there might be also fictitious processes, outside of the set of the informational processes.

### SOFTWARE / HARDWARE IMPLEMENTATION

The generic elements of the proposed hybrid simulator are brief programs, organized in a bi-layered architecture. The programs in the passive and active layers, as well as the communication between them have different functionalities, which makes possible to execute the continuous changes and the sequential events with the same kernel. The brief programs  $\Phi$  associated with the “active” elements **A**, with the knowledge of the output of the “passive” ones **P**, calculate the changes or rules, and then modify the input data of the “passive” layer. Finally the brief programs  $\Psi$  associated with the “passive” elements **P** are executed.

When building a simulator for a completely new class of problems, only the respective new executing prototypes of the active and passive elements have to be supplied in the mappings’ database. The various applications use the same data structures, core algorithms and interfaces; that is why the individual simulators of various abilities can easily be integrated with each other. In GBN models the passive **P** and active **A** elements are described by the passive

$p(\text{Identifiers, Kind, Time, Location, Content, Call\_for\_Operators, Others});$

and active

$a(\text{Identifiers, Kind, Time, Location, Inputlist, Call\_for\_Operators, Outputlist, Others})$

dynamic database or program partitions. These dynamic partitions are executed by a general and optionally extendable kernel. Accordingly the execution consists of four, cyclically repeated consecutive steps, as follows:

- (1) active elements read the Content (**X**) from the associated passive elements according to the Inputlist (**B**);
- (2) operators  $\Phi$  calculate the changes (**Y**),
- (3) passive elements are changed according to the Outputlist (**G**)
- (4) operators  $\Psi$  calculate the new state.

The method offers robust solution for the hard, hybrid, multidimensional and non-linear problems, as well as supports parallel programming.

The hardware implementation of the Generic Bi-layered Net can be solved by a hypothetical multiprocessor computer (see Fig. 9), consisting of two classes of small Neumann machines. Programming means the configuration of processor network, the declaration of the initial conditions and the distribution of the brief elementary programs amongst the units. Execution is

based on the wired or wireless communication between the two layers of processors, having their own unique broadcast and multiple receiving frequencies. The four steps of the run are: the P→A broadcasting, the execution of the ‘active’ units, the A→P broadcasting and the execution of the ‘passive’ units, respectively.

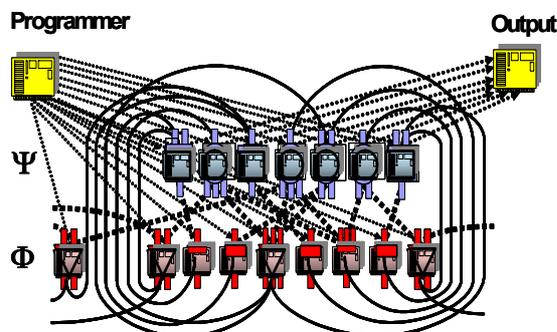


Figure 9: Software implementation of the GBN

## PRACTICAL APPLICATIONS

The methodology has been applying for the solution of various difficult practical problems. A couple of typical examples are the followings: Simulated Moving Bed preparative chromatography with cyclically changing initial and boundary conditions (Temesvari et al. 2004); batch, controlled co-polymerization of directed structure in partially mixed volume, with discrete feeds (Csukás and Balogh 1998); macro level simulation of a chicken poultry (common use of quantitative and qualitative knowledge); metabolic networks: underdetermined systems with roughly estimated model parameters (Csukas and al. 2003); planning and scheduling of an agricultural farm (including cost analysis and year long cash-flow calculation); quantitative health risk analysis of a multi-product plant (changing allocation of contaminant sources and workers).

## CONCLUSIONS

In Direct Computer Mapping we map the simple building blocks of the conservational and informational processes onto the generic “active” and “passive” elements of an executable program, directly. Direct Computer Mapping of process models allows the computer to know explicitly about the very structures and bounds of the physical world. The recently developed Generic Bi-layered Net model is a theoretically established realization of the Direct Computer Mapping. It is a special case of the General Net Theory on the one hand, as well as an explicitly structured, generic combination of the state space model and of the abstract automaton. The Generic Bi-layered Net model provides a common framework for the simulation of the hybrid (continuous and discrete, quantitative and qualitative) balance-based and rule-based processes. The common features of the process models are represented by a bi-layered net of variable structure that also determines the network (ring)

structures of the influence routes and flux routes. The advantage of the new methodology is that the computational model is specified by the very structures and building elements of the process to be modeled.

## References

- Brauer, W. Ed. 1980. “Net Theory and Applications”. Springer Lecture Notes in Computer Science, (84).
- Csukás B. and Bánkuti Gy. 2003a. “Direct computer mapping of process models”. In: *Foundations of Computer Assisted Process Operations*, I. E. Grossmann and C. M. MacDonald, Eds., AIChE INFORMS, pp. 577-581.
- Csukás B. and Balogh S.: “Combining Genetic Programming with Generic Simulation Models in Evolutionary Synthesis”. *Computers in Industry*, 36, 181-197.
- Csukás B. and Bánkuti Gy. 2003b. “Generic Bi-layered Net model of conservational and informational processes”. In: C. H. Dagli, A. et al. Eds. *Intelligent Engineering Systems through Artificial Neural Networks*, 13, ASME Press, New York, pp. 769-774.
- Csukás B., Debelak, K. A., Prokop, A., Balcarcel, R. R., Tanner, R. D., Bánkuti Gy, Balogh S. 2003. “Generic Bi-layered Net Model Based Discrimination of Chemical and Biological Warfare Agents”, AIChE Annual Meeting, San Francisco, Manuscript 474f.
- Kalman, R., Falb, P. and Arbib, M.. 1969. “Topics in Mathematical System Theory”, McGraw Hill.
- Pantelides, C. C. 2001. “New Challenges and Opportunities in Process Modeling”. Proceedings of ESCAPE-11, Copenhagen, Elsevier.
- Petri, C. A. 1962. “Kommunikation mit Automaten”, *Schriften des Institut für Instrumentelle Mathematik*, 2, Bonn.
- Temesvári K., Aranyi A., Csukás B., and Balogh S. 2004. “Simulated Moving Bed Separation of a Two Components Steroid Mixture”. *Chromatographia*,

**Acknowledgement:** the work was supported in part by Hungarian Scientific Research Fund T 037-297.

**BÉLA CSUKÁS** was born in Keszthely, Hungary. He studied chemical engineering and process control at the University of Veszprém and obtained CSc/PhD degree in 1985. He worked for research institutes, industrial R&D and universities. Now he is Professor of Information Technology, leading the Institute of Mathematics and Information Technology at the University of Kaposvar, Hungary.

**GYÖNGYI BANKUTI** was born in Nagybjom, Hungary. She studied mathematics and mechanical engineering at the Technical University of Budapest and obtained PhD degree in 1990. She worked for various firms before moving to the University of Kaposvár, Hungary, where she is Associate Professor of Applied Mathematics and Chair of Department of Applied Mathematics and Physics.

# VALIDATION OF COMPUTER FLUID DYNAMIC SIMULATION FOR DISPLACEMENT VENTILATION

Edit Stevensné Száday

Department of Building Services Engineering, Faculty of Mechanical Engineering  
Budapest University of Technology and Economics (BUTE)

Műgyetem rkp. 3., Budapest, H-1111, Hungary

e-mail: szaday@rit.bme.hu

## KEYWORDS

Displacement ventilation, CFD simulation, measurement, analytical method

## ABSTRACT

Computer simulations are used to determine information that current design regulations do not take into account and that would be either impossible or uneconomical to discover through direct measurement. At the same time, these simulations depend on data provided by the design regulations. In this article, I describe how I used Computer Fluid Dynamics (CFD) to design a simulation of the occupied zone, compared the results with those of direct measurement, and applied an analytical method to verify the results. By entering measured values for the inlet velocity, the inlet temperature, the outlet temperature, and the radiator average surface temperature into equations and running 290 points of iteration, my method yielded field distributions of the air temperature and velocity in part of the occupied zone. The analytical method I used was based on REHVA (Federation of European Heating and Air-conditioning Associations) guidelines.

## INTRODUCTION

For the user of the premises, the occupied zone is the most important element of the air conditioning system. The evaluation of the entire system depends on whether the draft and thermal comfort criteria are met there, as well as whether the contaminant requirements are satisfied. The current design regulations and standards are based on average values which do not provide information concerning the fields for air temperature, air velocity, and contaminant concentration. Therefore we cannot reach any conclusions regarding the potential recirculations and the thoroughness of the ventilation within the occupied zone. Experimental methods or computer simulations can yield such information.

Prior to the 1980's, direct measurement was the only method available to determine the fields for air temperature, air velocity, and contaminant concentration. With the arrival of Computational Fluid Dynamics codes, however, came the possibility of a comprehensive analysis of the occupied zone. To solve the partial differential equations, describing the chosen

model, we need the initial and boundary conditions including, among others, supply airflow rate and temperature—both of which could be derived from the design procedure. The current design guidelines (in the absence of sufficient computer capacity and software) are still very much in practice.

## MEASUREMENT SYSTEM

The measurements were taken at the Ventilation Laboratory of our department on the air distribution measuring system.

The examination chamber is inside our Laboratory so the “outside” temperature was constant during the measurements. We disabled any radiation from outside by putting shade on the windows. The measurement started after the temperature became steady which we could monitor with the help of the gradient measuring pole.

The collection of the temperature and velocity data happened in three surfaces perpendicular to the inlet face (see Fig. 1.). Data were collected on all three surfaces at 8 heights in 51 points each through a computer and evaluated with excel.

A radiator served as our heat source with 51°C forward and 42°C return warm water temperatures.

The measured air flow rate was 0.22m<sup>3</sup>/s which entered the chamber through a low velocity air terminal device. The outlet was in the middle of the ceiling.



Fig. 1: Measurement setup

## COMPUTER FLUID DYNAMIC SIMULATION

The computer simulation solves a set of partial differential equations with a numerical method. These partial differential equations are the conservation of mass, momentum, energy, and contaminant concentrations.

A three-dimensional steady-state numerical simulation has been performed to examine the displacement ventilation in cooling conditions according to the measurement setup.

The simulations have been implemented using the commercial code FLUENT. A computational grid of 152460 cells has been chosen, after having verified the grid independence of the results. The reference calculation hypotheses are:

- fixed temperature boundary conditions at the radiator surface,
- other walls are adiabatic,
- standard k- $\epsilon$  turbulence model,
- standard wall functions,
- velocity inlet on the round face of the inlet unit
- pressure outlet on the ceiling.

As a result this method provides the field distributions of the air temperature and velocity (Fig. 2 and Fig. 3).

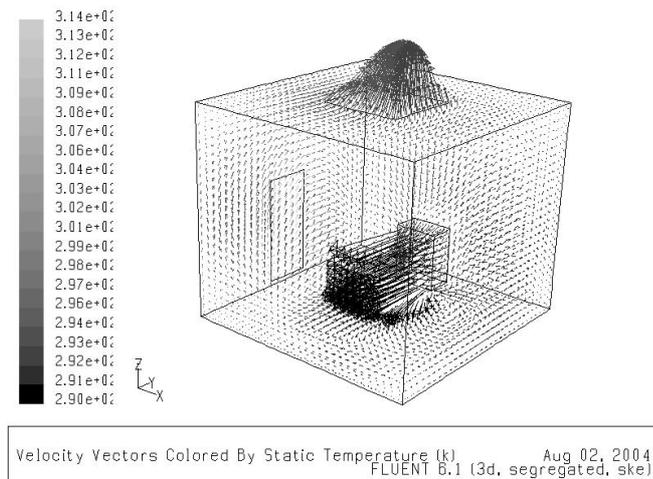


Fig. 2: Velocity vectors colored by temperature

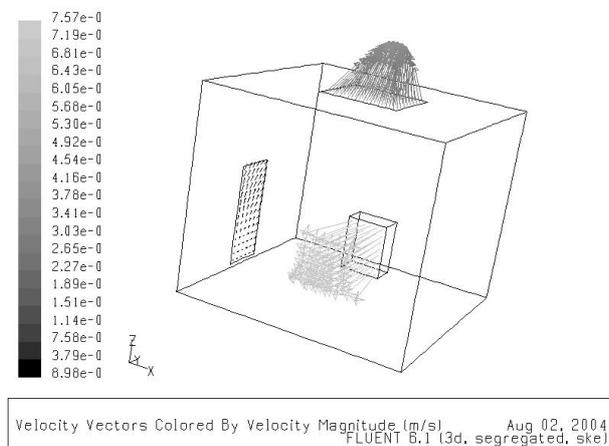


Fig. 3: Velocity vectors colored by velocity

## COMPARISON BETWEEN MEASURED DATA AND SIMULATION RESULTS

With CFD simulation there are countless small details which need to be worked out in order to get the most reliable model, and thus the solution closest to the real case. The CFD simulation method provides the most detailed information about the entire room. However, validation of this method is necessary. On the other hand relying only on measurement results is not sufficient due to the multiple ways in which errors can occur (instrumental, human, recording, etc.).

There is always a question of how to compare the different results. In my work I have created the mesh for the CFD simulation but the points I got after the iteration were too many to handle. I had to create certain surfaces (on which data were collected during the measurements) in my model. (Fig. 4.)

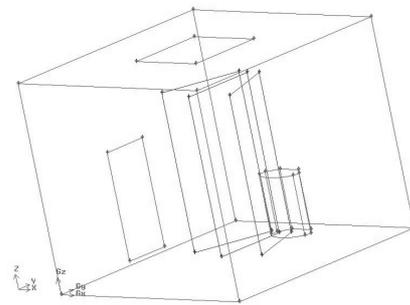


Fig. 4: Grid of the numerical method

After that I created and listed only the points of interest, where the measurements were taken with coordinates, temperature, and velocity values. Only then was I able to compare the two results and derive the necessary consequences. Fig.5 and Fig.6 show the velocities from the measurement and the simulation. As can be seen from these figures the CFD model follows the real experimental setup so the measured values are within an acceptable range.

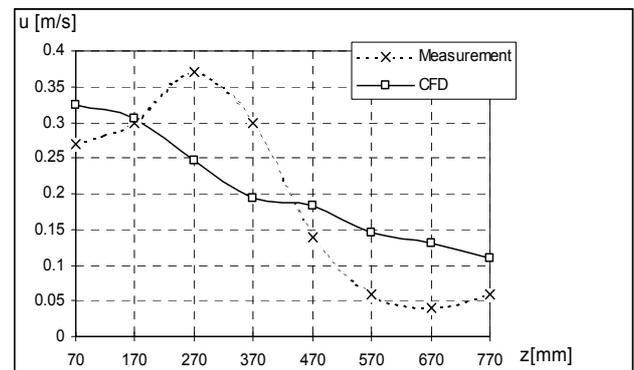


Fig. 5: Measured (dotted line) and calculated (CFD, continuous line) velocity magnitudes at different height (distance from the inlet unit was 1015mm)

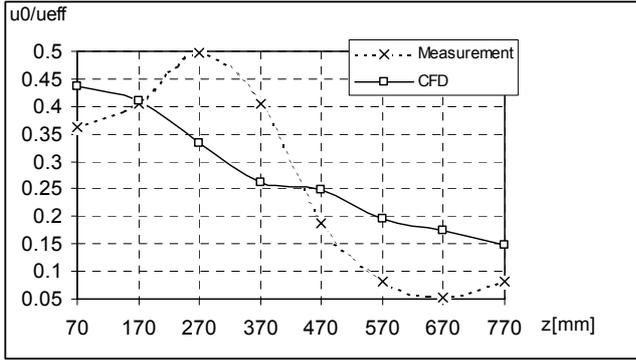


Fig. 6: Measured (dotted line) and calculated (CFD, continuous line) dimensionless velocity magnitudes at different height (distance from the inlet unit was 1015mm)

## CHECKING RESULTS WITH ANALYTICAL METHOD

The design procedure used for validation was developed by REHVA.

The determination of the supply airflow rate depends on the goal of the air conditioning (or ventilation). On the basis of this the procedure distinguishes between design criteria for *contaminant stratification* and for *excess heat removal*.

As our subject is non-industrial premises, the contaminant is CO<sub>2</sub>. The indicator for the process taking place in the room is:  $\Delta h/\Delta x = +\infty$ .

### ❖ Design criteria for contaminant stratification

#### 1. Determination of the Input data

The input data required for both design criteria are the following:

- room size
- location and number of people, type of human activity
- location, number and specification of other heat and contaminant sources
- requirements for the occupied zone: design air temperature of the occupied zone (at the height of 1.1 m for sedentary, 1.7 m for standing occupants), acceptable maximum air velocity near the floor, acceptable maximum contaminant concentration at inhalation level, air temperature difference between the head (1.1 or 1.7 m) and ankle level (0.1 m), and the required air flow rate.

#### 2. Selection of stratification height

The height of the lower stratification layer must be set slightly above the height of the inhalation level.

3. The condition of contaminant stratification is that the density of the contaminant is less than the density of the air surrounding it. With the help of appropriate literature, we determine the convection air flow rate around the various heat sources at a given height on the basis of convective heat emission, location and characteristic measurements. (M is the number of the heat sources.)

4. It is necessary to ensure contaminant concentrations lower than the allowed value in the occupied zone, so the supply air flow rate ( $\dot{V}_s$ ; [m<sup>3</sup>/s]) above the inhalation level (1.1 or 1.7 m) must be kept in balance with the sum of ascending air flow rates from the heat sources, minus descending air flow rates from the cooler surfaces ( $\sum_{i=1}^M \dot{V}_{conv,i}$ ; [m<sup>3</sup>/s]).

$$\dot{V}_s = \sum_{i=1}^M \dot{V}_{conv,i} \quad (1)$$

5. Calculation of the exhaust contaminant concentration ( $c_e$ ; [mg/m<sup>3</sup>])

In light of the following conditions, the contaminant concentration of the exhaust air can be calculated with the help of equation (2).

Conditions for equation (2):

- the air-conditioning is continuous ( $\dot{V}_s = \text{constant}$ )
- the contaminant concentration of supply air ( $c_s$ ; [mg/m<sup>3</sup>]) is constant
- the indoor air is uniform,
- there is no local exhaust in the room ( $\dot{V}_s = \dot{V}_e$ ),
- the source of contamination is constant ( $\dot{C}$ ; [mg/s])

$$c_e = c_s + \frac{\dot{C}}{\dot{V}_s} \quad (2)$$

6. Evaluation of the contaminant concentration of the inhaled air ( $c_{exp}$ ; [mg/m<sup>3</sup>])

As a result of human heat sources, fresh air replaces the ascending air. At the point of inhalation the air quality is higher (in our case CO<sub>2</sub> level is lower) than measured with no person at that point. This process can be expressed numerically with the help of the Personal Exposure Index ( $\varepsilon_{exp}$ ):

$$\varepsilon_{exp} = \frac{c_e - c_s}{c_{exp} - c_s} \quad (3)$$

From equation (3) the inhalation contaminant concentration:

$$c_{exp} = \frac{1}{\varepsilon_{exp}} \cdot (c_e - c_s) + c_s \quad (4)$$

where  $\varepsilon_{exp}$  can be derived from the literature as the function of the supply air flow rate.

Another way to determine the inhalation contaminant concentration is to assume that the inhalation and the supply contaminant concentration difference is 0.5-0.7 times the exhaust and the supply contaminant concentration difference. Assuming a value of 0.5, the inhalation contaminant concentration:

$$c_{exp} = 0,5 \cdot (c_e - c_s) + c_s \quad (5)$$

In the event that the inhaled contaminant concentration determined by either equation (4) or (5) is above the acceptable level the increase of the supply air flow is necessary

❖ Design criteria for excess heat removal

1. Determination of the Input data
2. Calculation of the cooling load

The calculation of the cooling load can happen according to standards or cooling load programs.

3. Calculation of the maximum temperature increase from supply to exhaust air

The design procedure assumes constant vertical temperature gradient in the room. The temperature of the supply air along the floor increases from  $T_s$  to  $T_f$ . According to the so called "50% rule"  $T_f$  temperature is the arithmetic mean of the supply air and exhaust air temperature ( $T_e$ ) (see equation 6) and can be calculated with the design air temperature ( $T_{oz}$ ) and the maximum acceptable temperature gradient  $\left(\left(\frac{\Delta T}{H}\right)_{\max}; [K/m]\right)$  (see

equation 7). The air temperature rises from  $T_f$  to  $T_e$ . The desired temperature difference, then, can be calculated with the aid of the maximum acceptable temperature gradient (see equation 8).

$$T_f = \frac{T_s + T_e}{2} \quad (6)$$

$$T_f = T_{oz} - \left(\frac{\Delta T}{H}\right)_{\max} \cdot z, \quad z=1,1m \quad (7)$$

$$T_e - T_s = 2 \cdot \left(\frac{\Delta T}{H}\right)_{\max} \cdot H, \quad (8)$$

where  $H; [m]$  is the interior height

4. Determination of the supply and exhaust air temperature

The supply air temperature from equations 6 and 8:

$$T_s = T_f - \left(\frac{\Delta T}{H}\right)_{\max} \cdot H \quad (9)$$

$$T_e = 2 \cdot \left(\frac{\Delta T}{H}\right)_{\max} \cdot H + T_s \quad (10)$$

5. Determination of the supply air flow rate

The supply air flow rate can be calculated from the heat removed from the space according to equation 11.

$$\dot{V}_s = \frac{\dot{Q}_t}{\rho \cdot c_p \cdot (T_e - T_s)} \quad (11)$$

6. Recalculation of the temperature increase along the floor ( $T_f - T_s$ )

The temperature increase along the floor according to equation 12 [2,3]:

$$T_f - T_s = \frac{1}{\rho \cdot c_p \cdot \dot{V}_s \cdot \left(\frac{1}{\alpha_{\text{rad}}} + \frac{1}{\alpha_{\text{conv}}}\right) + 1} \cdot (T_e - T_s) \quad (12)$$

❖ Common last steps for both criteria

7. Verification of the calculated supply air flow rate against codes and standards

8. Selection of the supply air flow rate

The supply air flow rate for the system ( $\dot{V}_{s,s}$ ) is chosen as the biggest among the calculated air flow rates from equations 1 and 11, and the required air flow rate according to the regulations.

9. Recalculation of the vertical temperature distribution in the room and estimation of the pollutant stratification height

The temperature difference between the exhaust and supply air can be calculated from equation 11 as follows in (13):

$$T_e - T_s = \frac{\dot{Q}_t}{\rho \cdot c_p \cdot \dot{V}_{s,s}} \quad (13)$$

$$\left(\frac{\Delta T}{H}\right) = \frac{(T_e - T_s)}{2} \cdot \frac{1}{H} \quad (14)$$

$$T_f = T_{oz} - \left(\frac{\Delta T}{H}\right) \cdot z, \quad z=1,1m \quad (15)$$

$$T_s = T_f - \left(\frac{\Delta T}{H}\right) \cdot H \quad (16)$$

$$T_e = 2 \cdot \left(\frac{\Delta T}{H}\right) \cdot H + T_s \quad (17)$$

In the event of  $\left(\frac{\Delta T}{H}\right) > \left(\frac{\Delta T}{H}\right)_{\max}$  the increase of the supply

air flow rate is necessary, with the increased air flow rate the vertical temperature distribution in the room needs to be recalculated with equations 13-17.

The pollutant stratification height can be determined with iteration from the sum of the convection air flows around the various heat sources which is balanced with the supply air flow rate for the system ( $\dot{V}_{s,s}$ ).

10. Selection of diffusers, verification of the adjacent zones

A suitable diffuser needs to be chosen to achieve the required performance. It is strongly recommended to use diffusers from manufacturers who supply their products with reliable documentation. The calculation of the adjacent zone depends on, among others, the discharge angle and the type of the diffuser.

With the help of this method the airflow rate, the supply air temperature, and the exhaust air temperature can be calculated. The results of the calculation can be seen on Fig. 7. The inlet velocity can be computed from the airflow rate with the help of the effective area given by the inlet catalog.

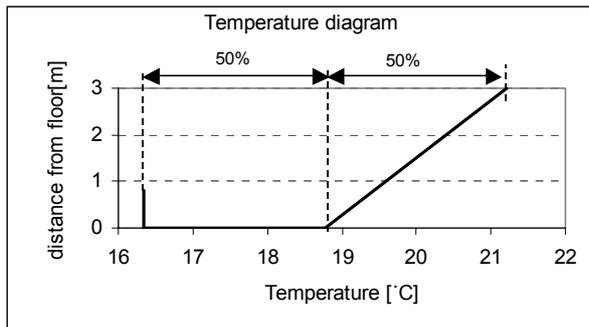


Fig. 7: Temperature diagram

## CONSEQUENCES

Checking the results with an analytical method is relatively quick and it is the key in validating the CFD simulation. The measuring equipment is very expensive, and often it is not even possible to measure the desired parameters for instance when the building does not even exist yet. Doing the calculations right and comparing them with the measured data would give us a powerful tool which together with the CFD simulation creates a fast way to evaluate the desired system.

## REFERENCES

- Skistad, H. (1994) "Displacement Ventilation" ISBN 0 86380 147 1
- Kis Piroska Z.; (1994) "Examining Ventilation of Offices" MSc Thesis, Budapest University of Technology and Economics (Hungarian)
- Eberhardt P.; (1995) "Room Ventilation with Displacement Ventilation" MSc Thesis, Budapest University of Technology and Economics (Hungarian)
- Mundt E.; (1995) "Displacement Ventilation Systems – Convection Flows and Temperature Gradients", Building and Environment, Vol. 30, No. 1, 129-133
- Tóth A.; (1998) "Examining Occupied Zone Measurements for Displacement Ventilation" MSc Thesis, Budapest University of Technology and Economics (Hungarian)
- Yuan X., Chen Q., Glicksman, L.R. (1998) "A Critical Review of Displacement Ventilation", ASHRAE Transactions, 104(1A), pp. 78-91 7.
- Yuan X., Chen Q., Glicksman, L.R. (1999 A) "Models for Prediction of Temperature Difference and Ventilation Effectiveness with Displacement Ventilation", ASHRAE Transactions, 105(1), pp. 353-367
- Yuan X., Chen Q., Glicksman, L.R. (1999 B) "Performance Evaluation and Development of Design Guidelines for Displacement Ventilation", ASHRAE Transactions, 105(1), pp. 298-309
- Yuan X., Chen Q., Glicksman, L.R. Hu Y., Yang X. (1999 C) "Measurements and Computations of Room Airflow with Displacement Ventilation", ASHRAE Transactions, 105(1), pp. 340-352
- Chen Q., et al. (1999) "Performance Evaluation and Development of Design Guidelines for Displacement Ventilation" ASHRAE Research Project –RP-949
- Skistad H., Mundt E., Nielsen P. V., Hagström K., Railio J.; (2002) "Displacement ventilation in non-industrial premises" Rehva, Guidebook No 1.

- Stevensné Száday E., Magyar T.; (2004) "Design Procedures for Displacement Ventilation (part 1.)", Magyar Épületgépészet, LIII. 2004/No.4. pp. 3-7. (Hungarian)
- Stevensné Száday E.; (2004), "Design Procedure developed by REHVA for Displacement Ventilation." Conference on Mechanical Engineering 2004, Volume 2 pp. 321-325
- Stevensné Száday E., Magyar T.; (2004) "Design Procedures for Displacement Ventilation (part 2.)", Magyar Épületgépészet, LIII. 2004/No.8. pp. 3-7. (Hungarian)

## AUTHOR BIOGRAPHY



**EDIT STEVENSÉ-SZÁDAY** received her MSc in Mechanical Engineering at BUTE, in 1994. She has been working at the Faculty of Mechanical Engineering of BUTE since 1994, recently as an assistant researcher. Her main research interest is Displacement ventilation and simulation in which she is completing her dissertation. Her teaching areas include ventilation and air-conditioning systems.

# REFINING SPECTRAL ANALYSIS FOR CONFIDENCE INTERVAL ESTIMATION IN SEQUENTIAL SIMULATION

Don McNickle  
Department of Management  
University of Canterbury  
Christchurch, New Zealand  
Email:Don.McNickle@canterbury.ac.nz

Gregory C. Ewing,  
Krzysztof Pawlikowski  
Computer Science and Software Engineering  
University of Canterbury  
Christchurch, New Zealand

## KEYWORDS

Sequential simulation, confidence intervals, spectral analysis

## ABSTRACT

The method of Spectral Analysis proposed by Heidelberger and Welch (SA/HW) is an effective and efficient way of calculating the error of a sample mean in sequential simulation. A simple modification to the method improves the coverage of the resulting estimators in the case of sequential simulation.

## INTRODUCTION

Sequential stochastic discrete-event simulation, i.e. stochastic simulation with on-line analysis of output data, is generally accepted as the most effective way to secure representativeness of samples of observations collected during simulation (Law and Kelton 2000). In this scenario, a simulation experiment is stopped when the statistical error of the estimate(s) reaches a required (low) level.

The method of Spectral Analysis proposed by Heidelberger and Welch (1981), which we abbreviate to SA/HW, has proved to be an effective and efficient way of calculating the statistical error. While more sophisticated spectral methods have been proposed (e.g. Lada, Wilson and Steiger, 2003), SA/HW is the only currently known method of sequential estimation of steady-state mean values in which designers have large freedom for deciding about the granularity of sequential data analysis, since SA/HW can be applied after grouping data in blocks of arbitrary size. This makes it an attractive choice for parallel simulation executed under the Multiple Replications in Parallel (MRIP) scenario (see Ewing, Pawlikowski, and McNickle, 2002.) In this paper we consider a simple modification of the SA/HW algorithm which improves the coverage of the estimators still further.

## THE SPECTRAL ANALYSIS METHOD

Simulation output often consists of highly correlated sequences of observations, for example waiting times of successive customers in a queue. Estimating the error in the mean waiting time thus requires techniques that account for this correlation. The best known method is that of Batch Means, where the means of batches of observations chosen large enough to be almost

independent are used to construct a confidence interval. Two problems with Batch Means are: in sequential simulation the granularity imposed by the batch length means that runs may be longer than needed; and finding an easy algorithm to reliably determine the size of the batch length is difficult. Figure 1 illustrates this problem for a particular algorithm. Here Batch Means has been used to produce supposedly 95% confidence intervals (the horizontal line.) However the actual coverage (see Section 2) of the confidence intervals drops off as the traffic intensity (and hence correlation of waiting times) in an M/M/1 queue increases.

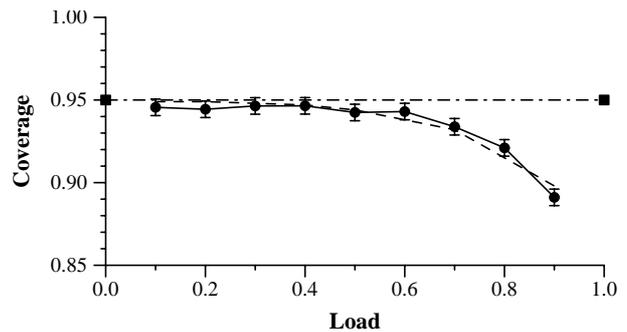


Figure 1: Coverage produced by an automated Batch Means Method. M/M/1 queue

The reduction in coverage turns out to be almost entirely due to the fact that the algorithm for determining batch length has produced batches that are too short. Daley (1968) gives formulas for the serial correlation of M/M/1 waiting times. Law (1977) outlines the steps needed to calculate the serial correlations between the batch means from these correlations. Using this method we can estimate the expected coverage (plotted as a dashed line) from the average batch lengths that the algorithm has produced. Thus almost all of the reduction in coverage appears able to be explained by the fact that the batches are too short. Since reliably estimating small correlations is difficult, the method of Batch Means always carries this risk: that the batches will be too short and hence batch means will remain significantly correlated.

On the other hand the Spectral Analysis method of estimation of the variance of a steady-state mean from a correlated sequence of observations  $x_0, x_1, \dots$  explicitly takes account of correlation between the observations. It was originally proposed as a simulation output analysis

method in Heidelberger and Welch, (1981). The variance is obtained as the value of the periodogram  $P(f)$  (of the analysed sequence of observations) at frequency  $f=0$ . Because of high variability of a typical periodogram at low frequencies, in SA/HW its value at  $f=0$  is obtained through a regression fit to the logarithm of the averaged periodogram, where fitting is done using a polynomial of degree  $d$  (typically  $d = 1$  or  $2$ ). The fitting is done using  $K$  fixed points of the periodogram. As shown in Heidelberger and Welch (1981), if  $d=2$ , then the confidence interval of the sample mean can be obtained using quantiles of the Student t-distribution with 7 degrees of freedom (if  $K=25$ ). The periodogram can be calculated either over the sequence of individual observations or over the sequence of their batch means. Thus SA/HW can be also applied to sequences of batch means of arbitrary size, instead of individual observations, greatly reducing storage and processing costs. In a subsequent paper, Heidelberger and Welch (1981b) considered a range of alternative values for  $d$ , and adaptive smoothing techniques. However they concluded that for both fixed-length and sequential simulation, the modifications offered no substantial improvement over their original recommendation of  $d=2$  and  $K=25$  points.

### COVERAGE ANALYSIS

Coverage analysis is widely used for assessing the quality of different methods used for constructing confidence intervals on the basis of simulation output data. By performing a large number of experiments we estimate the fraction of the generated confidence intervals which actually contain the true value of the parameter. If the method is accurate then when the theoretical confidence level has been set for example to 95% this fraction should also be close to 95%.

We performed sequential analysis of coverage, using the methodology described in Pawlikowski, Ewing and McNickle (1998), to produce coverage of SA/HW estimates with a relative precision of 0.01 at the 95% confidence level. It is worth noting that for each setting of the parameters of the reference models, getting coverage results with the statistical accuracy required meant that up to 14,000 separate experiments were needed.

Experiments were conducted for a number of reference models: M/M/1, M/D/1 and M/H<sub>2</sub>/1 and some simple network models. Here we give only the results for the queueing models, with traffic intensities ranging from 0.1 to 0.9. Figure 2 shows the coverage produced by the original SA/HW algorithm in sequential simulation for estimating the mean waiting time in the queue, plotted against the load.

There are two effects that can be noted. The first is that the coverage becomes poorer as the models become more variable. And the second effect is that the coverage reduces slightly, but steadily as the load in each of the queues increases.

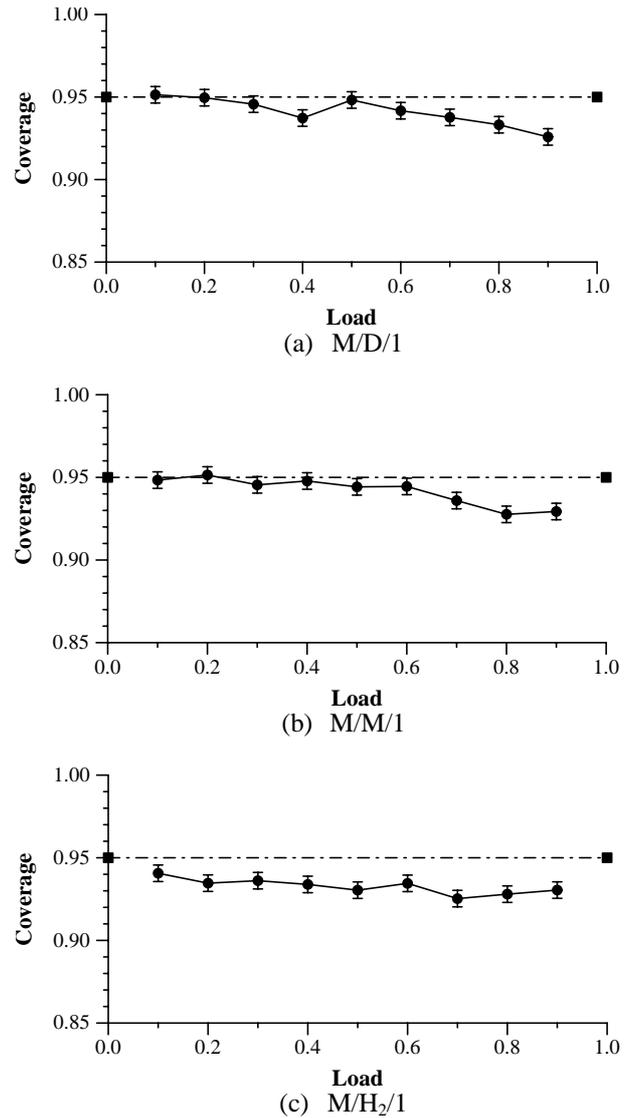


Figure 2: Coverage for SA/HW, M/D/1, M/M/1, M/H<sub>2</sub>/1 Queues

We are using the sequential version of SA/HW described in Pawlikowski (1990) in which the observations are grouped into a number of batches, and only the batch means are used as data. Our hypothesis is that that this fall-off in coverage is due to the increased run lengths required for more variable models, or as the traffic intensity increases, which in turn have resulted in batches of larger size. Large batches, we claim, may result in an inappropriate but easily fixed shape of the fitting polynomial.

### A MODIFICATION TO SA/HW

As mentioned previously, one attraction of the method is that it can be applied to grouped data, with essentially no change in the algorithm. Grouping the data reduces storage and network costs, so this is an attractive option. As the batch length increases the spectrum becomes flatter, tending towards the constant needed to estimate the variance of the overall mean. Heidelberger and Welch recommend approximating the log of the

periodogram by a low order polynomial, preferably of order  $d=2$ , in order to estimate the log of the periodogram at zero. For problems where the acceptable relative error is fairly high (e.g. greater than 10%) we have found that this works reasonably well, because the spectrum does have a shape that decays away from zero frequency. However where a very small degree of relative error is required we have found that the simulation can stop too early, with coverage well below the specified level.

The reason for this appears to be that the fitted polynomial is often convex upward when the simulation stops. In fact over the range of queueing models we have observed that about 90% of the simulations using SA/HW with  $d=2$  stopped with a convex upward quadratic.

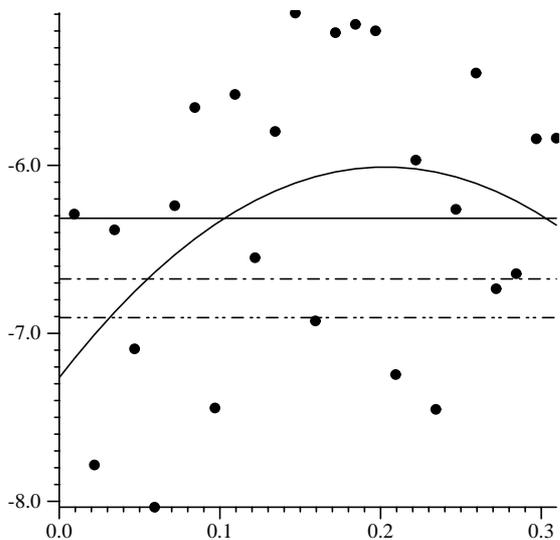


Figure 3: Typical Quadratic and Average Fits to the Log of the Averaged Periodogram at Stopping Time

For example Figure 3 shows the average and quadratic fits to grouped data, with a batch length of 1024, of waiting times for an M/M/1 queue with a traffic intensity of 0.8, at the time when the simulation stopped with an estimated relative error of 0.05, for a 95% confidence interval.

The upper and lower dashed lines show the values that must be reached for the simulation to stop for  $d=2$ , and  $d=0$  respectively. Thus this simulation will stop if a quadratic fit is used, but will not stop if  $d=0$ .

Since the stopping criterion is satisfied when the y intercept falls below a prespecified level it is clear why this form of fitting polynomial is most likely to occur at the stopping time. However a quadratic ( $d=2$ ) with a positive slope at zero is unrealistic, since the periodogram from simulation output should be a reducing function of frequency, especially after batching. Heidelberg and Welch (1981b) commented on the relative values of  $d = 0, 1$  and  $2$ , and suggested three adaptive methods for picking or altering the degree of the polynomial during the run. They concluded that for both fixed-length and sequential simulation, they offered no substantial improvement

over their original recommendation. However the fraction of sequential simulations which actually stop with a convex upward quadratic suggest a simpler approach which appears to work well.

Since grouping has reduced the periodogram to close to that of an independent process, an obvious modification is to replace the polynomial by simply averaging the values in order to estimate the intercept, in cases where an inappropriate (i.e. increasing at zero) polynomial occurs. This is equivalent to fitting a polynomial of degree zero.

Thus if:  $d=2$  and the slope of the quadratic at  $f=0$  is positive, we use the average of the periodogram points as the estimate of  $P(0)$ . The Heidelberg and Welch method requires two constants:  $C1(K,d)$  to produce an unbiased estimate of  $P(0)$ , and  $C2(K,d)$  to give the approximate degrees of freedom of the t-distribution. For  $d=0$ , the values, which were not included in the original paper, are:

Table 1: Constants for the Average Fit

K	D	C1(K,d)	C2(K,d)
25	0	.987	76
50	0	.994	154

Thus if the stopping criterion appears to have been met and the slope at  $f=0$  is positive, we use the average value and the parameters in Table 1 to re-estimate the variance. The simulation only stops if this estimate of the error is small enough.

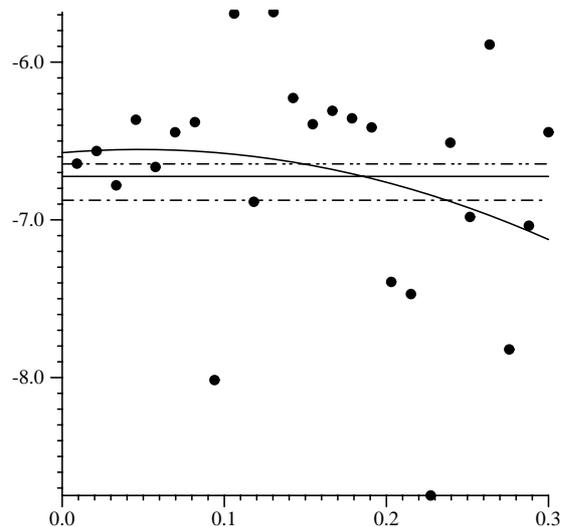


Figure 4: A Case where the Average Fit results in a Reduced Variance Estimate

It should be noted that if the quadratic fit has a positive slope at  $f=0$  this does not guarantee that the average will produce a larger variance estimate, as Figure 4 shows. In this example the upper dashed line is the

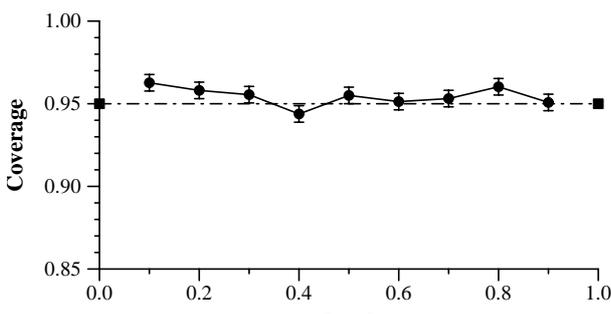
stopping criterion for  $d=0$ , while the lower line is that for  $d=2$ . Thus in this case the simulation will stop if an average is used, but will continue if we use the quadratic fit, in spite of the quadratic having a positive slope at  $f=0$ .

Thus we consider two versions of the modification: if the slope of the quadratic at  $f=0$  is positive, we use the average unconditionally to re-calculate the variance, (“Slope Protection”), and using the average only if it provides a larger estimate of variance than the quadratic (“Conditional Slope Protection”).

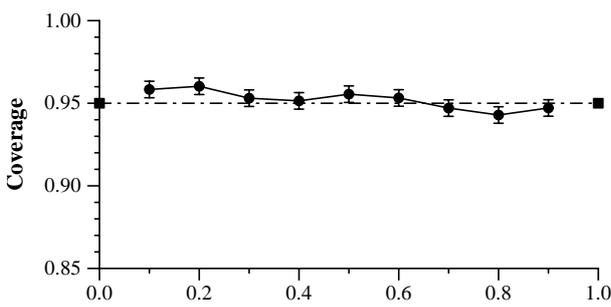
## RESULTS

The simulation were carried out using the Akaroa2 Simulation package (Ewing, Pawlikowski, and McNickle, 1999.) The implementation of SA/HW, except for the modification as above, is as described in Pawlikowski, (1990)

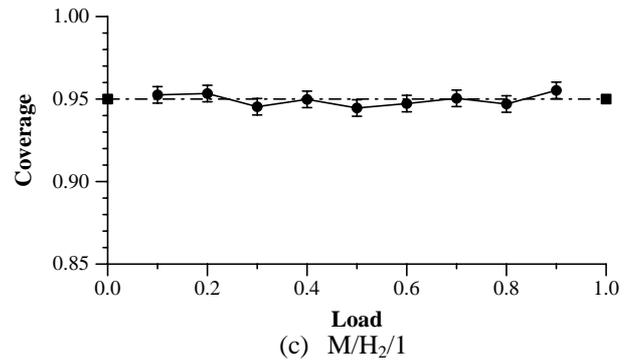
Figures 5 and 6 show the effects of these two schemes on the three queueing models. The coverage is uniformly increased, with the larger increase coming from the conditional scheme. The results for other reference models were consistent with those for simple queueing models presented here.



(a) M/D/1

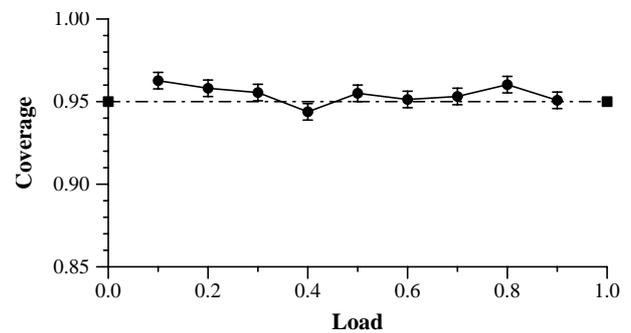


(b) M/M/1

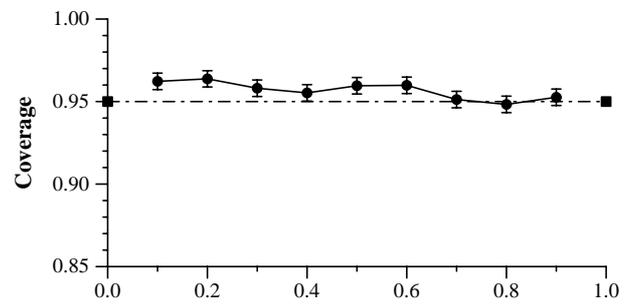


(c) M/H<sub>2</sub>/1

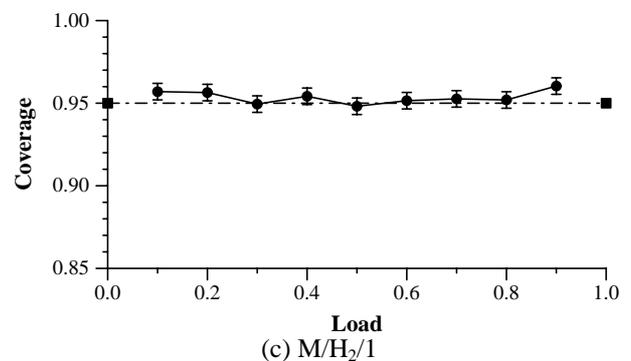
Figure 5: SA/HW with Slope Protection



(a) M/D/1



(b) M/M/1



(c) M/H<sub>2</sub>/1

Figure 6: SA/HW with Conditional Slope Protection

## CONCLUSIONS

The method of SA/HW has been found experimentally to produce coverage values which agree well with those expected. Further improvements in coverage of SA/HW in sequential simulation can be obtained by adding a simple extra step to the calculation of the stopping criterion to check if the fitted quadratic is increasing at zero. When this happens using the average value of the periodogram instead of a fitted quadratic to estimate the variance of the mean provided coverage levels that were almost exactly those required. The conditional use of the average only if it gave a larger estimate of the error, produced results which were typically slightly above the specified level of coverage and could be considered as providing an additional margin of accuracy.

## REFERENCES

- Daley, D. J. 1968 "The serial correlation coefficients of waiting times in a stationary single server queue." *Journal of the Australian Mathematical Society*, vol. 8, 683-699.
- Ewing, G., K. Pawlikowski, and D. McNickle. 1999. "Akaroa2: Exploiting Network Computing by Distributing Stochastic Simulation". *Proceedings of the European Simulation Multiconference ESM'99, Warsaw*. International Society for Computer Simulation. 175-181
- Ewing, G., Pawlikowski, K. and McNickle, D. 2002. "Spectral Analysis for Confidence Interval Estimation under Multiple Replications in Parallel". *Dresden, Germany: Proceedings of the 14th European Simulation Symposium, ESS'2002*. 52-55 & 61. October 2002.
- Heidelberger, P. and P. D. Welch. 1981. "A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations." *Communications of the ACM*, vol. 24, no. 4 (April), pp. 233-245
- Heidelberger, P. and P. D. Welch. 1981b. "Adaptive Spectral Methods for Simulation Output Analysis." *IBM Journal of Research and Development*, vol. 25, 860-876
- Lada, E. K., J. R. Wilson, and N.M. Steiger. 2003. "A Wavelet-based Spectral Method for Steady-state Simulation Analysis." *Proceeding of the Winter Simulation Conference*, 422-430.
- Law, A. M. 1977. "Confidence intervals in discrete event simulation: a comparison of replication and batch means." *Naval Research. Logistics Quarterly*, vol. 24, 667-678.
- Law, A. M. 1983. "Statistical Analysis of Simulation Output Data". *Operations Research*, vol. 31, no. 6, 983-1029
- Law, A. M. and Kelton, W. D. 2000. *Simulation Modelling and Analysis*, 3<sup>rd</sup> Edition. New York: McGraw-Hill.
- Pawlikowski, K. 1990. "Steady State Simulation of Queueing Processes: A Survey of Problems and Solutions". *ACM Computing Surveys*, vol. 22, no. 2, 123-170
- Pawlikowski, K., G. Ewing and D. McNickle. 1998. "Coverage of Confidence Intervals in Sequential Steady-State Simulation". *Journal of Simulation Practise and Theory*, vol. 6, 255-267

## BIOGRAPHIES

**DONALD MCNICKLE** is an Associate Professor of Management Science in the Department of Management 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. His email address is <don.mcnickle@canterbury.ac.nz>.

**GREG EWING** is a research associate in the Department of Computer Science and Software Engineering at Canterbury; where received a Ph.D. His research interests include simulation; distributed systems; programming languages, 3D graphics and graphical user interfaces. He has made contributions to the Python programming language; and has recently been nominated for membership of the Python Software Foundation. His email address is <greg@cosc.canterbury.ac.nz>

**KRZYSZTOF PAWLIKOWSKI** is a Professor of Computer Science at the University of Canterbury. His research interests include quantitative stochastic simulation; and performance modelling of telecommunication networks. He received a PhD in Computer Engineering from the Technical University of Gdansk; Poland. He is a Senior Member of IEEE and a member of SCS and ACM. His email address is <krys@cosc.canterbury.ac.nz> and his web page is <<http://www.cosc.canterbury.ac.nz/~krys/>>.

# NUMERICAL APPROXIMATION OF TAYLOR COEFFICIENTS FOR SOLVING FIRST ORDER ODEs

István SeleK

Department of Informatics, Faculty of Mechanical Engineering  
Budapest University of Technology and Economics  
H-1111 Budapest, Múgyetem rkp. 5. Phone: (1)-463-1170  
E-mail: selek@rit.bme.hu

## KEYWORDS

Ordinary differential equations, Taylor coefficients, numerical methods, approximation.

## INTRODUCTION

It is well-known, that differential equations are used to describe the processes of engineering, economics...etc. It is often necessary to solve a system of N first order ordinary differential equations (ODEs) of the following format:

$$y'(x) = f(y(x), x), \quad y(x_0) = y_0 \quad (1)$$

The majority of differential equations does not have an analytical solution, only a limited number of ODEs (see (1)) have an exact analytical solution. It is often the case, that even quite simple systems might have a complicated nonlinear differential equation representation, or it could be very difficult or even impossible to give the differential equations exactly. With the advent of digital computers, complex equations or systems of equations could powerfully and exactly be solved with various numerical methods. Computers can handle large amounts of data easily and quickly. Probably the most serious drawback of numerical methods is that they can only approximate the continuous solution with a series of discrete points. A large number of formulas were developed to solve these kinds of equations. Adams-Bashfort and Runge-Kutta methods are used fairly extensively nowadays. Both methods use discrete points to approximate the integral of functions from  $x_i$  to  $x_{i+1}$  where  $x$  is the variable of function.

In this paper an algorithm is proposed to approximate the Taylor series of the solution of equation (1) in given points. The proposed method can be useful for solving ODEs with continuous methods and for evaluating numerical derivatives.

## THE ALGORITHM

Let us consider the differential equation (1) with its initial condition. Let us presume that the values of  $y'(x)$  and  $f(y(x), x)$  in  $x_i$  are known, and  $f(y(x), x)$  satisfy the Lipschitz condition in every  $x$  and  $y_1, y_2$ :

$$|f(x, y_1) - f(x, y_2)| \leq L_f |y_1 - y_2| \quad (2)$$

Where  $L_f$  is a scalar number ( $L_f \in R$ ). Let us introduce a new variable  $\xi$  in the vicinity of  $x_i$ .

$$x = x_i + \xi \quad (3)$$

The Taylor series of  $y'(x)$  in the neighbourhood of  $x_i$  is given in the following format:

$$y'(x_i) = \sum_{p=0}^{\infty} \frac{\xi^p}{p!} y^{(p+1)}(x_i) \quad (4)$$

When integrating equation (4) the approximate value of  $y(x)$  around  $x_i$  results. Predictor methods (like Adams-Bashfort) can be based on this formula because it traces the progress in variable  $x$ . The increment of function  $y(x)$  is determined by

$$\Delta y = \int_{x_i}^{x_{i+1}} y'(x) dx = \sum_{p=0}^{\infty} \frac{y^{(p+1)}(x_i)}{(p+1)!} \xi^{p+1} \quad (5)$$

On the other hand the increment of the function can be approximated with another formula, with the use of a discrete Runge-Kutta method.

$$\Delta y = \sum_{p=1}^{\infty} \alpha_p k_{i,p}(\xi) \quad (6)$$

Coefficients  $k_{i,p}(\xi)$  of the previous formula are calculated with

$$\begin{aligned} k_{i,1} &= \xi f(y(x_i), x_i) \\ k_{i,2} &= \xi f(x_i + A_2 \xi, y(x_i) + A_2 k_{i,1}) \\ &\vdots \\ k_{i,p} &= \xi f(x_i + A_p \xi, y(x_i) + \sum_{q=1}^{p-1} A_{p,q} k_{i,q}) \end{aligned} \quad (7)$$

where  $\alpha_p$  and  $A_{p,q}$  are constants. By using the formulas of partial differentiation, let us expand the  $n^{\text{th}}$  order differentials of  $y(x)$  according to the format in (4). The results are

$$\begin{aligned} y'(x_i) &= f(y(x_i), x_i) \\ y''(x_i) &= \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} f \Big|_{x=x_i} \\ &\vdots \end{aligned} \quad (8)$$

The Taylor series of  $f(y(x), x)$  around  $x_i$  is determined by the following equation:

$$f(y(x_i) + k(\xi), x_i + \xi) = f(y_i, x_i) + \frac{\partial f(x, y)}{\partial y} \Big|_{x_i} k(\xi) + \frac{\partial f(x, y)}{\partial x} \Big|_{x_i} \xi + \dots \quad (9)$$

Coefficients of equation (6) can finally be reproduced like Runge-Kutta method. After that there are two formulas available to determine function  $\Delta y(x)$ . Let us assume that the two formulas are equal.

$$\sum_{p=1}^{\infty} \frac{y^{(p)}(x_i)}{p!} \xi^p = \sum_{p=1}^{\infty} \alpha_p k_{i,p}(\xi) \quad (10)$$

By using this equation the Taylor coefficients of  $y(x)$  can be taken. However, a problem arises: The formula in the left side of (10) is determined by  $p^{\text{th}}$  order functions of  $\xi$ , while the right side is determined by a first order function of  $\xi$ . As the coefficients are calculated  $k_{i,p}$  has to be written as a function of  $\xi^p$ . By using equation (7), let us consider a new function:

$$K_{i,p}(\xi) = f(x_i + A_p \xi, y(x_i) + \sum_{q=1}^{p-1} A_{p,q} k_{i,q}(\xi)) \quad (11)$$

Let us suppose that the Taylor coefficients up to  $z^{\text{th}}$  order ( $z \in \mathbb{Z}$ ) are supposed to be calculated. First the Runge-Kutta coefficients  $k_{i,p}(\xi)$ ,  $\alpha_p$  and  $A_{p,q}$  have to be taken ( $i = 1 \dots z$ ,  $z^{\text{th}}$  order Runge-Kutta method). As  $K_{i,p}$  is determined as the polynomial of variable  $\xi$ , the series expansion depending has to be given on  $\xi$  of  $f(x_i + A_p \xi, y(x_i) + \sum_{q=1}^{p-1} A_{p,q} k_{i,q}(\xi))$  around  $(x_0, y_0)$ . It can be done numerically. Let us consider  $h$  as a step of variable  $x$ , ( $x_i + h = x_{i+1}$ ) and determine the value of  $K_i(\xi)$  in points  $\xi_t = th$  where  $t = 0, 1, 2, \dots, N$  and  $N \geq z$ , ( $N \in \mathbb{Z}$ ). After that a set of points  $\{\{\xi_t, K_{p,t}\}_{t=0}^N\}_{p=1}^z$  result. Now, let us find function  $K_i(\xi)$  in the following format:

$$K_{i,p}(\xi) \approx \sum_{t=0}^N a_{i,p,t} \xi^t \quad (12)$$

where  $a_{i,p,t}$  are unknown constant coefficients. Values of  $a_{p,t}$  are determined with the help of the least squares method. After replacing the coefficients into equation (10) and equalizing the identical indices of  $\xi$ , the approximation of Taylor coefficients in  $x_i$  are calculated.

$$y'(x_i) = f(y(x_i), x_i) \quad (13)$$

$$y''(x_i) = 2 \sum_{m=1}^{z-1} \alpha_m a_{i,m,1} \quad (14)$$

$$y^{(n)}(x_i) = n! \sum_{m=1}^{z-1} \alpha_m a_{i,m,n-1} \quad (15)$$

Along the given order of Taylor coefficients, the precision of the proposed method can be extended, when the number of approximate points  $N$  are increased, however it is not necessary to calculate coefficients  $K_{i,p}(\xi)$  up to the  $z^{\text{th}}$  order.

## NUMERICAL RESULTS

In this section the performance of the method is analyzed. Two application examples are presented. The first example is the simpler one, where  $K_i(\xi)$  could be written as a polynomial of variable  $\xi$ , and there is no need to use the least squares method. In the second example a more general differential equation with  $f(y(x), x)$  function is examined. After the Taylor coefficients of the solution of ODEs are calculated, let us compare the results with the coefficients that were determined by the analytical solution in both examples. Let the problem be the same in both cases: the Taylor coefficients of the solution of given differential equations up to fourth order have to be calculated, with the consideration of this initial condition  $y(x_0 = 0) = y_0$ . Let us consider the following equation as the first example

$$y'(x) = y(x)x \quad y(0) = 1 \quad (16)$$

First let us determine the analytical solution of equation (16) by using the separation of variables method. After the conversion the result is

$$\int_{y_0}^y \frac{1}{y(s)} ds = \frac{x^2}{2} \quad (17)$$

After the expansion of equation (17) the analytical solution of Eq. (16) and its derivatives are determined by the following formulas:

$$\begin{aligned} y(x) &= \exp\left(\frac{x^2}{2}\right) \\ y'(x) &= x \exp\left(\frac{x^2}{2}\right) \\ y''(x) &= (1 + x^2) \exp\left(\frac{x^2}{2}\right) \\ y'''(x) &= (3x + x^3) \exp\left(\frac{x^2}{2}\right) \\ y^{(4)}(x) &= (3 + 6x^2 + x^4) \exp\left(\frac{x^2}{2}\right) \end{aligned} \quad (18)$$

By using formulas (3), (5) and (18) the increment of  $y(x)$  around  $x = 0$  has the following format:

$$\Delta y = \frac{1}{2} \xi^2 + \frac{3}{24} \xi^4 + \dots + \mathcal{O}(\xi^5) \quad (19)$$

Now the Taylor coefficients are calculated numerically and  $\Delta y$  is written. As the Taylor series up to

fourth order is to be calculated by using formula (11) the Runge-Kutta equation has the following form:

$$\Delta y = \frac{1}{6}k_{0,1} + \frac{2}{6}k_{0,2} + \frac{2}{6}k_{0,3} + \frac{1}{6}k_{0,4} + \dots + \mathcal{O}(\xi^5) \quad (20)$$

where,

$$\begin{aligned} k_{0,1} &= \xi K_{0,1}(\xi) = \xi f(x_0, y_0) \\ k_{0,2} &= \xi K_{0,2}(\xi) = \xi f\left(x_0 + \frac{\xi}{2}, y_0 + \frac{k_{0,1}}{2}\right) \\ k_{0,3} &= \xi K_{0,3}(\xi) = \xi f\left(x_0 + \frac{\xi}{2}, y_0 + \frac{k_{0,2}}{2}\right) \\ k_{0,4} &= \xi K_{0,4}(\xi) = \xi f\left(x_0 + \xi, y_0 + k_{0,3}\right) \end{aligned}$$

and  $f(y, x) = xy$ . By using the values of  $x_0$  and  $y_0$  and the form of  $k_{0,p}$ ,  $K_{0,p}(\xi)$  functions could be calculated. The formulas of these functions are

$$\begin{aligned} K_{0,1}(\xi) &= 0 \cdot 1 = 0 \quad , \quad k_{0,1} = 0 \\ K_{0,2}(\xi) &= \frac{\xi}{2} \cdot 1 = \frac{\xi}{2} \quad , \quad k_{0,2} = \frac{\xi^2}{2} \\ K_{0,3}(\xi) &= \frac{\xi}{2} \left(1 + \frac{\xi^2}{4}\right) \quad , \quad k_{0,3} = \frac{\xi^2}{2} + \frac{\xi^4}{8} \\ K_{0,4}(\xi) &= \xi \left(1 + \frac{\xi^2}{2} + \frac{\xi^4}{8}\right) \quad , \quad k_{0,4} = \xi^2 + \frac{\xi^4}{2} + \frac{\xi^6}{8} \end{aligned}$$

By using formula (20) the numerical approximation of  $\Delta y$  is determined by a following equation:

$$\Delta y = \frac{1}{2}\xi^2 + \frac{3}{24}\xi^4 + \dots + \mathcal{O}(\xi^5) \quad (21)$$

Comparing the given result to the analytical form of  $\Delta y$  it is evident, that the two expressions are equal. It was easy to see that how to write  $K_{0,p}$  as a polynomial of variable  $\xi$  and the basis of the method. In the next example let us consider the following equation:

$$y'(x) = y(x) \cos(x) \quad y(0) = 1 \quad (22)$$

The analytical solution of equation (22) is determined by the following form:

$$y(x) = C \exp(\sin(x)) \quad (23)$$

By using the initial conditions  $C = 1$  results. Let us

write the derivatives of  $y(x)$  in the point  $x = 0$  :

$$\begin{aligned} y'(0) &= \left[ \cos(x) \exp(\sin(x)) \right]_{x=0} = 1 \\ y''(0) &= \left[ \left( \sin(x) + \cos(x)^2 \right) \exp(\sin(x)) \right]_{x=0} = 1 \\ y^{(3)}(0) &= \left[ \left( -\cos(x) - 3 \sin(x) \cos(x) + \right. \right. \\ &\quad \left. \left. + \cos(x)^3 \right) \exp(\sin(x)) \right]_{x=0} = 0 \\ y^{(4)}(0) &= \left[ \left( \sin(x) + 3 \sin(x)^2 - 4 \cos(x)^2 + \right. \right. \\ &\quad \left. \left. - 6 \sin(x) \cos(x)^2 + \cos(x)^4 \right) \right. \\ &\quad \left. \exp(\sin(x)) \right]_{x=0} = -3 \end{aligned}$$

By using these results the Taylor series of  $y(x)$  around  $x = 0$  has the following form:

$$y(x) = 1 + \xi + \frac{1}{2}\xi^2 - \frac{3}{24}\xi^4 + \dots + \mathcal{O}(\xi^5) \quad (24)$$

As the Taylor coefficients are calculated numerically, the  $K_{0,p}$ , ( $p = 1, 2, 3, 4$ ) coefficients in the Runge-Kutta method have to be written as a polynomial of variable  $\xi$ . In the first example it is easy to write  $K_{0,1}$ . It is known that  $f(x, y) = y \cos(x)$ , by using this the first coefficient can be calculated.

$$K_{0,1} = 1 \cos(0) = 1 \quad , \quad k_{0,1} = \xi \quad (25)$$

Let us say that  $K_{0,1}$  does not depend on variable  $\xi$ . In the next step let us consider a stepsize  $h = 0.1$  and the number of interpolation points  $N = 6$ . Let us determine  $K_{0,2}(\xi)$  in given points (See in Table 1), where

$$K_{0,2} = f\left(x_0 + \frac{\xi}{2}, y_0 + \frac{k_{0,1}}{2}\right) \quad (26)$$

and the values of  $K_{0,2}(\xi)$  :

Table 1: The values of  $K_{0,2}(\xi)$  in given points

$\xi$	$K_{0,2}(\xi)$
0.0	1
0.1	1.04868
0.2	1.09450
0.3	1.13708
0.4	1.17608
0.5	1.21114

To determine the polynomial the points have to be interpolated and the method of least squares has to be used. The format of the approximate polynomial is the same as in the previous section in equation (11). According to the method of least squares the coefficients of the approximate polynomial of  $K_{0,2}(\xi)$  have the following format (The values of coefficients up to fifth order are shown in Table 3):

$$K_{0,2} \approx 1 + 0.5\xi - 0.125\xi^2 + 0.0625\xi^3 + \dots \quad (27)$$

By using this result up to the third order the value of  $k_{0,2}$  is obtained.

$$k_{0,2} = \xi + 0.5\xi^2 - 0.125\xi^3 + 0.0625\xi^4 \quad (28)$$

Substituting the given  $k_{0,2}$  to  $K_{0,3} = f(x_0 + \frac{\xi}{2}, y_0 + \frac{k_{0,2}}{2})$  and resumming the method  $k_{0,3}$  is

$$\begin{aligned} K_{0,3}(\xi) &\approx 1 + 0.5\xi - 0.1249\xi^2 + 0.1241\xi^3 + \dots \\ k_{0,3}(\xi) &= \xi + 0.5\xi^2 + 0.1249\xi^3 - 0.1241\xi^4 \end{aligned}$$

and  $k_{0,4}$

$$\begin{aligned} K_{0,4}(\xi) &\approx 1 + 1.0001\xi - 0.0022\xi^2 + 0.3572\xi^3 + \dots \\ k_{0,4}(\xi) &= \xi + 1.0001\xi^2 - 0.0022\xi^3 + 0.3572\xi^4 \end{aligned}$$

as a polynomial of variable  $\xi$ . (Table 2 shows the values of  $K_{0,3}(\xi)$  and  $K_{0,4}(\xi)$ )

Table 2: The values of  $K_{0,3}(\xi)$  and  $K_{0,4}(\xi)$  in given points

$\xi$	$K_{0,3}(\xi)$	$K_{0,4}(\xi)$
0.0	1	1
0.1	1.05111	1.09959
0.2	1.10390	1.19646
0.3	1.15741	1.28718
0.4	1.21057	1.36760
0.5	1.26223	1.43296

Figure 1 shows the values of  $K_{i,p}$  in discrete points, and their numerical approximations with a polynomial of  $\xi$ . The forms of approximate polynomials can be seen in Table 3.

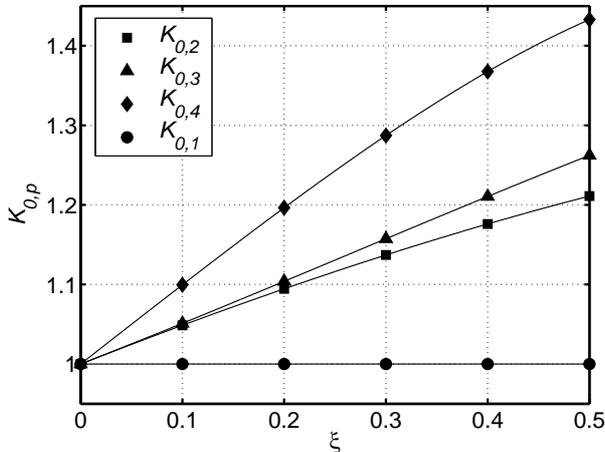


Figure 1: The numerical approximation of  $K_{i,p}(\xi)$  functions

By using formula (14) the numerical approximation of Taylor coefficients is determined by the following forms:

$$\begin{aligned} y'(0) &\approx \left(\frac{1}{6} + \frac{2}{6} + \frac{2}{6} + \frac{1}{6}\right) = 1 \\ y''(0) &\approx 2 \left(\frac{2}{6} \cdot 0.5 + \frac{2}{6} \cdot 0.5 + \frac{1}{6} \cdot 1.0001\right) = \\ &= 2 \cdot 0.50001 = 1.0002 \\ y'''(0) &\approx 6 \left(-\frac{2}{6} \cdot 0.125 + \frac{2}{6} \cdot 0.1249 - \frac{1}{6} \cdot 0.0022\right) = \\ &= 6 \cdot -0.0004 = -0.0024 \\ y^{(4)} &\approx 24 \left(-\frac{2}{6} \cdot 0.0625 - \frac{2}{6} \cdot 1.241 - \frac{1}{6} \cdot 0.3572\right) = \\ &= 24 \cdot -0.1217 = -2.9208 \end{aligned}$$

The analytical and numerical solutions are compared, see Table 4, 5 and 6. These tables contain the Taylor coefficients along with the value of  $h$ . The tables show that when decreasing the stepsize  $h$  the error of Taylor coefficients is also decreasing along given conditions.

Table 3: The forms of approximate polynomials

	$\xi^0$	$\xi^1$	$\xi^2$	$\xi^3$	$\xi^4$	$\xi^5$
$K_{0,1}$	1	0	0	0	0	0
$K_{0,2}$	1	0.5	-0.125	-0.0625	0.0026	0.0013
$K_{0,3}$	1	0.5	0.1249	-0.1241	-0.063	0.0156
$K_{0,4}$	1	1.0	-0.002	-0.3572	-0.400	0.1006

Table 4:  $h = 0.5$

	$y'(0)$	$y''(0)$	$y'''(0)$	$y_{(4)}(0)$
Analytical	1	1	0	-3
Numerical	1	0.9644	0.4141	-4,8992

Table 5:  $h = 0.1$

	$y'(0)$	$y''(0)$	$y'''(0)$	$y_{(4)}(0)$
Analytical	1	1	0	-3
Numerical	1	1.0002	-0.0024	-2.9208

Table 6:  $h = 0.01$

	$y'(0)$	$y''(0)$	$y'''(0)$	$y_{(4)}(0)$
Analytical	1	1	0	-3
Numerical	1	1	$-5 \cdot 10^{-7}$	-3

## CONCLUSIONS

A general algorithm designed for the approximation of Taylor coefficients of the solution of first order ODEs has been presented. This method gives a continuous approximation, it can approximate the solution with a  $z^{\text{th}}$  order polynomial in any points where

the function of solution is interpretable. The presented method based on the standard Runge-Kutta method, but it takes the constants used by RK as a function of a local variable. Therefore the main difference between this method to other standard methods for example discrete Runge-Kutta, Adams-Bashfort, Adams-Moulton etc. while the standard methods give a discrete approximation point by point this method approximate the function of solution by a continuous polynom in given points it is show more information the behavior of the solution in the neighborhood of approximated points. This method only the first step to the simulation it can take as a predictor method so it can be decisive the integration with variable stepsize to determine the value of the stepsize or the order of approximation. The method can be useful in determining the derived functions of solution up to  $z^{\text{th}}$  order.

## REFERENCES

- A. Coddington; N. Levinson. 1955. *Theory of Ordinary Differential Equations*. McGraw-Hill, New York.
- E. Hairer; S.P. Norsett; G.Wanner. 1987. *Solving Ordinary Differential Equations, I. Nonstiff Problems*. Springer, Berlin.
- Halász Gábor; Huba Antal. 2003. *Műszaki mérések*. Műegyetemi Kiadó, Budapest, ISBN 963-420-744-8
- P. Henrici. 1962. *Discrete Variable Methods in Ordinary Differential equations*, Wiley, New York.
- Peter Henrici. 1985. *Numerikus Analízis*. Műszaki könyvkiadó, Budapest, ISBN 963-10-6419-0
- P. Davis. 1961. *Interpolation and approximation*. Blaisdell, Waltham.
- P.N. Brown, G.D. Byrne, A:C Hindmarsh, VODE. 1989. "A variable coefficient ODE solver." *SIAM J. Sci. Stat. Comp.* **10**, 1038-1051
- Stoher József; Bulirsch Roland. 2002. *Introduction to numerical analysis*. Springer, New York. ISBN 0-387-95452-X
- Stoyan Gisbert; Takó Galina. 1995. *Numerikus Módszerek II*. ELTE-TypoTEX, Budapest. ISBN 963-7546-53-7
- W.H. Enright, K.R. Jackson, S.P. Norsett and P.G. Thomsen. 1986. "Interpolants for Runge-Kutta formulae." *ACM Trans. Math. Software* **12**, 193-218

## AUTHOR BIOGRAPHY

**ISTVÁN SELEK** received the M.Sc degree in mechanical engineering with a major in applied mechanics and machine design from the Faculty of Mechanical Engineering Budapest University of Technology and Economics (BUTE) in 2003. In 2003 he began his Ph.D. studies in BUTE. His primary research interest is the simulation of highly nonlinear processes with time delays, by using Soft Computing methods.

# THE IMPORTANCE OF THE OBJECTIVE FUNCTION DEFINITION AND EVALUATION IN THE PERFORMANCE OF THE SEARCH ALGORITHMS

Javier Otamendi  
E-mail: jotamendi\_30@yahoo.com

## KEYWORDS

Simulation and optimization, search methods, capability index, cost functions.

## ABSTRACT

A qualitative and quantitative comparison of simulation optimization methodologies is presented to specifically study the importance of the objective function on the election of the algorithm to generate the alternatives to be simulated. If a capability index is used in conjunction with a powerful rejection algorithm and a efficient metaheuristic, the average number of repetitions per simulated alternative is to be decreased significantly, facilitating the simulation of more alternatives. The probability of the convergence of the algorithm increases then considerably. The credibility in the solution is also raised since the decision is made both in terms of average behavior and variability.

## INTRODUCTION/PROBLEM DEFINITION

The problem is to optimize a simulation vector of responses  $\Psi$  over a region  $\Theta \subset \mathcal{R}^p$  with respect to an ordered p-tuple of input factor settings  $X = (x_1, x_2, \dots, x_p, \dots, x_p) \in \Theta$  (Jacobson and Schruben 1989). In other words, a p-tuple (or alternative or combination) must be selected among a full set in terms of several responses (or criteria or fitness).

The difficulty to solve the problem is triple. First, each of the J individual responses  $Y_j \in \Psi$  is a random variable  $Y_j = f(X)$ , which must be estimated. Second, the individual responses must be aggregated in an overall response or objective function  $Y = f(\Psi)$ , which is then also a random variable. Third, the total number of p-tuples to evaluate, I, might be very large.

The solution comes in the form of a simulation optimization methodology (SOM) which combines a generation algorithm of a subset of p-tuples to be simulated and the evaluation of an objective function.

Figure 1 summarizes a simulation optimization methodology within the framework of a generic optimization methodology, particularizing it for a simulation study. The system is first represented in a simulation model and then an iterative procedure is followed: "simulations are normally used in a scenario-by-scenario base, with the designer generating a

solution and subsequently having the computer evaluating it" (Gama and Norford 2002).

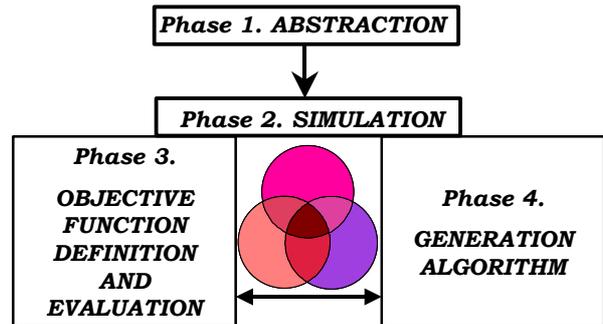


Figure 1. Simulation Optimization Methodology

The origin of any SOM resides in the impossibility of analyzing the huge set of alternatives that are under study. The size of the subset of the alternatives that a methodology is able to simulate basically depends on the calculation speed of the computer.

If the number of alternatives is not excessive and each and every one of them might be simulated in the required time, the WHOLE SET should be analyzed and no generation algorithm is needed. The problem is reduced to the evaluation of each alternative.

If, as it is usually the case, the set is too large because the execution of one repetition is too slow or because there is a continuous variable under consideration, then a SUBSET is to be analyzed. The generation algorithm is crucial and will be dependent on the shape  $\Psi$  over the region  $\Theta$ , which depends on the objective function that has been defined. "The success of the optimization procedure depends on the choice of the objective function and its functional relationship to the control parameters" (Hilgers and Boersma 2001).

What follows is the qualitative and quantitative study of the available SOMs, grouped not only in terms of their different formulations but also on their desirable characteristics. The study is performed first in terms just of the objective function to explain the way it is defined and evaluated. Then, the focus is in the combination of an objective function and generation algorithms and how they affect the final solution.

## DESIRABLE CHARACTERISTICS OF THE SIMULATION METHODOLOGY

Many SOMs have been developed. They all try to find a satisficing alternative, with a fitness value close to that of the optimum alternative, without having to analyze all the available alternatives, that is, in a reasonable execution time.

A desirable search methodology must accelerate then the selection of satisficing alternatives. The optimum simulation search methodology will be one in which the defined multicriteria objective function allows for the simulation of only a few alternatives, among which the subpar alternatives are run only one time or not simulated at all, and the real candidates are run a small, preset number of times ( $R_{max}$ ). In summary, and since the response surface  $\Psi$  is a random variable, the objectives of a SOM are to “maximize statistical power and psychological validity” (Wager and Nichols 2003).

Let's separate the above objectives into three characteristics, namely, efficacy, efficiency and credibility (Law and Kelton 1991).

### Efficacy and credibility

The two characteristics are closely related. If efficacy is a measure of the distance to the ideal objective function value, the credibility is the psychological measure that the solution obtained is the optimum one.

Efficacy is to be calculated as follows:

$$EFFICACY = 100\% * \left( 1 - \frac{|Y - Y^*|}{|Y^*|} \right)$$

where:

- $Y$  = objective function value for a particular combination or p-tuple
- $Y^*$  = objective function value for the optimum combination, which is not known, unless all of the combinations are analyzed.

Credibility is the confidence of the decision maker in the solution obtained, that is, the degree of belief that the optimum solution has been obtained. It will be quantified as a number between 0 and 100.

Efficacy and credibility are affected by two factors. The first is the percentage of simulated alternatives. The more alternatives are run, the higher the credibility will be. If the whole set is evaluated, efficacy of the selected alternative is 100%, since  $Y^*$  is known and  $Y_{selected} = Y^*$ . If only a subset is evaluated, there is no certainty that the selected alternative is the optimum one. That is why a good methodology should search for  $P(Y_{selected} = Y^*) \rightarrow 1$ .

The second is the number of runs made per alternative ( $R$ ). The more runs performed, the higher the credibility, since the estimation of the individual objective functions is improved, unless a small number guarantees that the alternative is not optimal but subpar, and therefore is subject to be rejected.

### Efficiency or the total number of runs, NTOT

Efficiency is the time spent to select an alternative among those included in the available set.

Quantitatively, the efficiency is measured as the total number of simulation runs ( $NTOT$ ), which might be split into the following factors:

$$NTOT = \frac{\text{Runs}}{\text{Alternative}} * \text{Alternatives}$$

It is possible to improve the efficiency by reducing either factor of the expression. Depending on the objective function to estimate, the number of runs might be different (first factor) and also the method to select the subset of alternatives to simulate (second factor).

The above formula might be further specified if the number of runs per alternative depends on the objective function of an alternative. If the estimation of the value of the objective function makes an alternative feasible, more repetitions should be made than if the estimation makes an alternative clearly infeasible or subpar.

Then, the distribution of the repetitions per alternative,  $\Omega$ , might be defined as:

$$p(\Omega = R) = p(R) \quad \forall R \in [0, R_{max}]$$

where:

- $R$  is the number of repetitions that are to be performed per alternative
- $R_{max}$  is the maximum number of repetitions
- $p(R) = I_R/I$ ; probability of running an alternative  $R$  times
- $I$  = total number of alternatives
- $I_R$  = number of alternatives that are run  $R$  times.

Therefore, the total number of runs is the sum of the total number of runs for each  $R$ :

$$NTOT = \sum_{R=0}^{R_{max}} NTOT_R$$

which is the number of alternatives times the number of runs for each  $R$ :

$$NTOT = \sum_{R=0}^{R_{max}} R I_R$$

which might be calculated also as the total number of alternatives times the average number of runs per alternative:

$$NTOT = \sum_{R=0}^{R_{max}} R I p(R) = I \sum_{R=0}^{R_{max}} R p(R) = I * R_{mean}$$

NTOT might be lowered then by either reducing  $I$  (simulating only a small subset), decreasing  $R_{max}$  per evaluated feasible alternative, or altering  $\Omega$  (shifting the weight towards the lower values of  $R$ ). Any of the factors are dependent on the objective function used.

## OBJECTIVE FUNCTION DEFINITION AND EVALUATION

An important step of the SOM is to define the objective function that is going to evaluate the fitness of an alternative, that is, the capability of an alternative to meet requirements. This objective function is an aggregation of several individual criteria, and must include an idea of risk.

It is necessary, first, to define a fitness value per criterion that allows for a satisfaction and aspiration analysis following the multicriteria decision making theory (Zeleny 1982). An alternative will be satisfactory if it fulfills certain satisfaction requirements and it will be ideal if it reaches certain aspiration levels. The fitness value must include the information provided for each criterion (satisfaction limits - or worst permissible values - and aspiration level - or ideal value) and the random variables  $Y_j$  obtained by the repetitive execution of simulation models.

The fitness value must quantitatively reflect the degree of fulfillment of requirements. Therefore, it must include in the analysis the whole output distribution per criterion, and not only its mean behavior. The proposed alternatives must simultaneously generate values inside the satisfaction limits, avoiding adverse situations, and a mean behavior close to the ideal value or aspiration level.

Then, a multicriteria fitness value must be defined to aggregate individual fitness value, in order to perform an optimization or search process with a single objective and select among the alternatives that are within the satisficing subset. This multicriteria fitness value summarizes the degree of satisfaction of all the individual criteria.

### Individual criterion

The main characteristics of one criterion must be:

1. Subjective: comes from someone's mind.
2. Give an idea of preference: alternatives are to be compared and ordered.

Each criterion must be optimized, either by:

1. Maximizing its value, for example, service level.
2. Minimizing its value, for example cost.
3. Searching for a target value, for example budget.

In all three cases, both an aspiration level  $T_j$  and upper and lower satisfaction levels ( $LSL_j$  and  $USL_j$ ) are subjectively set (Barba-Romero and Pomerol 1997), levels that define the acceptability or feasibility of a given alternative.

Then, it is necessary to estimate the performance of criterion  $j$  for each alternative  $i$  in a measure  $v_{ij} = f(Y_{ij})$  and compare it with the desirable values. The probability distribution of  $v_{ij}$  is to be estimated from  $R$  repetitions of the simulation model.

### Aggregated criterion

It is not trivial to develop a multicriteria measure  $v_i = f(v_{ij})$  that aggregates the individual criteria. The individual measure for each criterion has to be selected and then combined all together into a single measure. The bigger problem is to put all the measures in the same unit.

Even in the same units, priorities or ranking between the criteria might be set. Therefore weights that combine priorities and change of units are to be determined.

### Setting the subjective values

As already mentioned, the determination of the objective function is subjective in nature, since the weights, the targets and the specification limits are subjectively assigned.

In terms of the weights, several procedures are available (for example, Analytic Hierarchy Process (AHP) (Saaty 1980)) to give some sense to the assignment process.

Regarding the other parameters ( $USL$ ,  $T$ ,  $LSL$ ), it is usually possible to correctly assign those values since the decision maker knows the system in hand, especially for the target.

To conclude, it should be mentioned that it is usually easier to set the specification limits than the weights (Barba-Romero and Pomerol 1997), since each criterion is studied independently.

## QUALITATIVE ANALYSIS OF OBJECTIVE FUNCTIONS

In this section, the two main types of multiple objective functions are introduced.

### Average

Usually, the individual average value for any measure,  $\bar{v}_{ij}$ , is used, and its confidence interval given to provide an idea of variation.

The multicriteria fitness value is generally a linear combination of the averages for each criterion, where the weights are used to change units and measure the relative importance of each criterion:

$$\bar{v}_i = \sum_j w_j \bar{v}_{ij}$$

The weaknesses of this fitness value is that it only includes mean behaviour and that several criteria with different units are aggregated.

The following are two possible aggregations and how they are optimized.

### Economic

Both individual criteria must be converted to the same monetary units. If a criterion is not monetary, it is incorporated as a penalty for not fulfilling requirements (Hilgers and Boersma 2001):

$$Economic = Profit - Penalty$$

This objective value is to be maximized.

### Deviation

Each objective function  $j$  has its own target value,  $T_j$ . The deviation  $z_j$  is calculated as the difference between that target and the average value for the criterion,  $\bar{Y}_j$ :

$$z_j = -|\bar{Y}_j - T_j|$$

If subjective numeric weights are assigned to each criterion,  $w_j$ , the aggregated objective function is then:

$$Deviation = \sum_{j=1}^J w_j z_j$$

This objective value is to be minimized.

### Process capability index (PCI)

The second option comes from the field of quality control. The whole distribution  $Y_{ij}$  is compared with both the specifications and the target and a summary

measure estimated. This measure is called Process Capability Index (PCI).

There are several capability indices (Kotz and Lovelace 1998). However, two have come up as the ones more commonly used in industry. One is  $C_{pk}$ , whose value is above 1 if the whole distribution is between the specification limits regardless of where the average or the median lie with respect to the target, and  $C_{pm}$ , which aggregate the satisfaction of the limits and the closeness to the target.

The  $C_{pmF}$  index has also been proposed (Otamendi 2001) to combine the two previous indices in to a single measure:

$$C_{pmF_{ij}} = \begin{cases} C_{pm_{ij}} & \forall MCpk_{ij} \geq 1 \\ 0 & \forall MCpk_{ij} < 1 \end{cases}$$

where:

$$C_{pm_{ij}} = \frac{USL_j - LSL_j}{6\sqrt{s_{ij}^2 + (\bar{Y}_{ij} - T_j)^2}}$$

$$MCpk_{ij} = \min \left[ \frac{USL_j - P50_{ij}}{P99.865_{ij} - P50_{ij}}, \frac{P50_{ij} - LSL_j}{P50_{ij} - P0.135_{ij}} \right]$$

$USL_j$  = upper specification limit (criterion  $j$ )

$LSL_j$  = lower specification limit (criterion  $j$ )

$T_j$  = target value (criterion  $j$ )

$\bar{Y}_{ij}$  = average of the data (alternative  $i$ , criterion  $j$ )

$s_{ij}$  = standard deviation of the data

$P_{k_{ij}}$  =  $k\%$  percentile of the data

and  $MCpk$  is just the generalization of the  $C_{pk}$  index for non-normal distributions.

The  $C_{pmF}$  index takes a value of 0 if the alternative does not fulfill requirements, and  $C_{pm}$  if all the requirements are met. In other words, subpar alternatives are given a value of 0 and the rest a value which combines the distance to the ideal and the requirements. This objective value is to be maximized.

In the multicriteria case, besides the linear combination option presented for the averages, there is also a possibility that makes sense to develop the MPCl (Multicriteria PCI). Since the capability indices are unitless, and the philosophy is to fulfill requirements, a conservative option is to assign the fitness value of the worst criterion to the alternative:

$$C_{pmF_i} = \min (C_{pmF_{ij}}) \quad \forall j = 1 \dots J$$

The strength of this type of aggregation is that each criterion is first evaluated on its own, and then combined into the unique measure. Each and every criterion must be fulfilled for the alternative to be valid and the aggregation performed.

The weakness of the capability indices is that extreme percentiles are used for their calculation. These  $P_k$  are estimated with more repetitions than the ones needed to estimate average behaviour.

### QUANTITATIVE COMPARISON OF OBJECTIVE FUNCTIONS

A case study is presented to quantitatively compare the objective functions mentioned in the previous section. This case study will also be used later in the article to compare SOMs that include a search algorithm.

#### The system

The inventory system under study has been adapted from (Taha 1988), by changing the given random variables to non-normal, asymmetric distributions. Four products are stored and sent to customers.

Stock level is revised periodically (*review interval, RI*). If it falls below the *reorder point (RO)*, a call is made to the supplier to ask for the needed quantity to reach the *maximum inventory level (ML)*.

The possible values for each factor are:

- RI  $\in$  {2, 3, 4, 5, 6, 7}                      6 levels
- RO  $\in$  {250, 275, 300, ..., 725, 750}        21 levels
- ML  $\in$  {3500, 3600, ..., 5400, 5500}        21 levels

for a total of  $21 \times 21 \times 6 = 2646$  possible combinations or alternatives (*RI, RO, ML*) to choose from.

The decider looks to minimize cost without reducing the level of service. The target for the cost is 7% with an upper limit of 12%. For the service level, the target is 99.5% and the lower limit is 95%.

The estimation of the aggregated objective function for each alternative is based on  $R_{max} = 100$ , so the total number of runs performed for the WHOLE SET are:

$$NTOT = R_{max} * I \\ = 100 * 2646 = 264600 \text{ runs.}$$

#### Economic

It is calculated as:

$$Economic = Profit - Penalty$$

where the profit is a monetary value in currency units:

$$Profit = Gross \text{ income} - Costs$$

and penalty is a subjective monetary value assigned to the lack of service, for example, 10000 m.u. (monetary units) per day of lateness:

$$Penalty = 10000 * \text{Late Days}$$

Table 2 includes the results for the best 20 alternatives.

Table 2. Results for *Economic* Objective Function

FACTORS / PARAMETERS			OBJ. FUNCTION	
REVIEW INTERVAL	REORDER POINT	MAXIMUM LEVEL	Economic	
			Fitness	Efficacy
3	675	3500	373.44	100.00%
4	700	3500	373.31	99.96%
3	700	3500	373.30	99.96%
4	725	3500	373.16	99.93%
3	725	3500	373.16	99.93%
4	750	3500	373.00	99.88%
3	750	3500	372.94	99.87%
2	675	3600	372.59	99.77%
2	700	3600	372.08	99.64%
2	725	3600	371.90	99.59%
2	750	3600	371.69	99.53%
2	750	3700	371.60	99.51%
2	750	3800	371.24	99.41%
2	700	3900	370.82	99.30%
2	425	4000	370.82	99.30%
2	450	4000	370.73	99.28%
2	725	3900	370.71	99.27%
2	675	3900	370.70	99.27%
2	750	3900	370.70	99.27%
2	475	4000	370.66	99.26%
<b>Optimum</b>			<b>373.44</b>	

The results are very similar for all of the first 20 p-tuples, with the best alternative presenting a review interval of 3. For comparison purposes, in the last column, a measure of efficacy is included.

#### Deviation

Table 3 includes the results for the best 20 alternatives, for subjective weights of 90% and 10% for the service level and cost, respectively.

Table 3. Results for *Deviation* Objective Function

FACTORS / PARAMETERS			OBJ. FUNCTION	
REVIEW INTERVAL	REORDER POINT	MAXIMUM LEVEL	Deviation	
			Fitness	Efficacy
4	575	3600	0.13	100.00%
2	600	3900	0.14	88.17%
4	575	3500	0.15	87.36%
4	625	3600	0.15	86.36%
2	650	3900	0.15	81.94%
4	700	3500	0.16	77.39%
2	625	3800	0.16	75.65%
3	700	3500	0.16	75.20%
4	625	3500	0.16	74.53%
4	675	3700	0.16	74.36%
3	625	3500	0.16	73.76%
2	575	3800	0.16	73.11%
2	650	3800	0.17	69.94%
2	600	4000	0.17	69.83%
2	700	3900	0.17	69.64%
4	675	3600	0.17	68.79%
2	675	3900	0.17	68.66%
2	600	3800	0.17	68.11%
2	650	4000	0.17	67.65%
4	675	3500	0.17	66.74%
<b>Optimum</b>			<b>0.13</b>	

For this objective function, the best alternative is one in which the review interval is set to 4, and a low 66.74% efficacy for the 20<sup>th</sup> alternative is obtained.

#### Process capability index $C_{pm}F$

Table 4 includes the results for the best 20 alternatives.

Table 4. Results for  $C_{pm}F$  Objective Function

FACTORS / PARAMETERS			OBJ. FUNCTION	
REVIEW INTERVAL	REORDER POINT	MAXIMUM LEVEL	CpmF	
			Fitness	Efficacy
2	500	3800	0.65	100.00%
4	575	3600	0.64	99.77%
2	550	3700	0.60	92.52%
2	600	3700	0.58	90.04%
2	600	3900	0.58	89.13%
4	575	3500	0.57	88.59%
4	625	3600	0.57	87.71%
2	600	3800	0.56	86.64%
2	500	3900	0.56	86.48%
2	575	3800	0.56	86.09%
2	550	3900	0.55	85.90%
2	650	3900	0.54	84.43%
2	475	3800	0.54	83.29%
2	650	3700	0.53	82.84%
2	550	3800	0.53	82.83%
4	700	3500	0.52	81.35%
2	525	3800	0.52	80.88%
2	625	3800	0.52	80.18%
3	700	3500	0.52	79.98%
2	575	3700	0.52	79.93%
<b>Optimum</b>			<b>0.65</b>	

The review interval is found to be 2 for this measure.

**Comparison**

Table 5 includes a comparison of the results obtained for each objective function, for the best 20 alternatives found feasible according to the  $C_{pm}F$  index.

Table 5. Comparison

FACTORS / PARAMETERS			OBJECTIVE FUNCTION					
REVIEW INTERVAL	REORDER POINT	MAXIMUM LEVEL	Economic		Deviation		CpmF	
			Fitness	Efficacy	Fitness	Efficacy	Fitness	Efficacy
2	500	3800	319.99	85.69%	0.47	-162.90%	0.65	100.00%
4	575	3600	-4.40	-1.18%	0.13	100.00%	0.64	99.77%
2	550	3700	297.93	79.80%	0.63	-286.86%	0.60	92.52%
2	600	3700	309.78	82.95%	0.17	66.46%	0.58	90.04%
2	600	3900	363.64	97.38%	0.14	88.17%	0.58	89.13%
4	575	3500	-105.52	-28.26%	0.15	87.36%	0.57	88.59%
4	625	3600	-4.82	-1.29%	0.15	86.36%	0.57	87.71%
2	600	3800	349.23	93.52%	0.17	68.11%	0.56	86.64%
2	500	3900	364.27	97.54%	0.37	-84.57%	0.56	86.48%
2	575	3800	342.90	91.82%	0.16	73.11%	0.56	86.09%
2	550	3900	363.98	97.47%	0.19	51.87%	0.55	85.90%
2	650	3900	366.11	98.04%	0.15	81.94%	0.54	84.43%
2	475	3800	264.69	70.88%	0.21	40.46%	0.54	83.29%
2	650	3700	336.57	90.13%	0.18	62.60%	0.53	82.84%
2	550	3800	342.99	91.85%	0.38	-95.17%	0.53	82.83%
4	700	3500	373.31	99.96%	0.16	77.39%	0.52	81.35%
2	525	3800	334.32	89.52%	0.21	34.94%	0.52	80.88%
2	625	3800	351.05	94.01%	0.16	75.65%	0.52	80.18%
3	700	3500	373.30	99.96%	0.16	75.20%	0.52	79.98%
2	575	3700	297.84	79.76%	0.20	44.66%	0.52	79.93%
<b>Optimum</b>			<b>373.44</b>		<b>0.13</b>		<b>0.65</b>	

It is surprising that the best alternative produces a very low value for *Deviation* and the second a low value for *Economic*. It is clear then that the final decision depends on the objective function used. And each decider might choose a very different response measure.

Even the choice of subjective values for the parameters within the three objective functions might vary the final solution. To study the robustness of the final decision, a variation in the subjective values of the monetary penalty, the criteria targets and specification limits is tested. For the monetary value, a smaller value of 5000 m.u. is proposed instead of the original 10000 m.u. For the cost criterion, the target value is kept at an almost achievable value of 7% whereas the upper satisficing value lowered from 12% to 9%. For the service criterion, the target is set at 100% and the lower satisficing limit raised to 98%. Again, it looks like it is easier to set the values for the targets and limits than the one for the penalty. The results obtained for the

$C_{pm}F$  index for the best 20 alternatives are included in Table 6.

Table 6. Sensitivity Analysis

FACTORS / PARAMETERS			OBJECTIVE FUNCTION					
REVIEW INTERVAL	REORDER POINT	MAXIMUM LEVEL	Economic		Deviation		CpmF	
			Fitness	Efficacy	Fitness	Efficacy	Fitness	Efficacy
4	575	3600	184.40	49.38%	0.13	100.00%	0.26	100.00%
2	600	3900	367.04	98.29%	0.14	88.17%	0.23	89.34%
4	575	3500	134.28	35.96%	0.15	87.36%	0.23	88.79%
4	625	3600	183.98	49.27%	0.15	86.36%	0.23	87.91%
2	600	3800	360.43	96.52%	0.17	68.11%	0.22	86.84%
2	575	3800	357.30	95.68%	0.16	73.11%	0.22	86.29%
2	650	3900	368.31	98.63%	0.15	81.94%	0.22	84.63%
4	700	3500	373.31	99.96%	0.16	77.39%	0.21	81.54%
2	625	3800	361.25	96.74%	0.16	75.65%	0.21	80.36%
3	700	3500	373.30	99.96%	0.16	75.20%	0.21	80.17%
4	625	3500	290.71	77.85%	0.16	74.53%	0.21	79.73%
4	675	3700	253.80	67.96%	0.16	74.36%	0.21	79.64%
3	625	3500	337.97	90.50%	0.16	73.76%	0.20	79.24%
2	700	3900	370.82	99.30%	0.17	69.64%	0.20	76.58%
2	675	3900	370.70	99.27%	0.17	68.66%	0.20	76.15%
4	675	3600	248.76	66.61%	0.17	68.79%	0.20	76.08%
2	650	4000	369.74	99.01%	0.17	67.65%	0.19	75.54%
4	675	3500	364.04	97.48%	0.17	66.74%	0.19	75.10%
3	675	3500	373.44	100.00%	0.17	65.10%	0.19	74.17%
3	750	3500	372.94	99.87%	0.18	64.56%	0.19	73.86%
<b>Optimum</b>			<b>373.44</b>		<b>0.13</b>		<b>0.26</b>	

The best option shifts from the p-tuple 2-500-3800 to the p-tuple 4-575-3600. Even four of the previously first ten acceptable alternatives become unacceptable.

Again, the decision changes in terms of the objective function and its subjective parameters. And the choices are going to affect the performance of the generation algorithms of the subset of alternatives to be simulated.

**GENERATION ALGORITHMS**

There exist several algorithms to generate the subset of alternatives that are going to be simulated. The main first distinction has to be made among the analysis of the WHOLE SET (like the example presented in the previous section) or just part of it, SUBSET, which depends on the size of the set,  $I$ , and therefore, on the nature and size of the input factors.

If all the factors are discrete, the size of the available set has a combinatorial status. Its size value is the multiplication of the number of the settings for each parameter ( $f_p$ ):

$$I = \prod_{p=1}^P f_p$$

If any factor  $p$  is continuous, then  $I$  grows to infinity since  $f_p$  is also infinity. The only possibility is then to make the possible factor settings discrete so  $f_p$  is finite.

**Whole Set**

If the total number of runs is manageable, meaning that NTOT can be performed, simulating each and every alternative in the feasible set is possible and should be done. Therefore, the probability of not running and alternative is  $p(R=0)=0$ .

In terms of the characteristics, efficiency is then not a problem, 100% efficacy is fully achieved and confidence in the solution is total.

## Subset

One proposed classification for the SUBSET generation algorithms is the following (Jacobson and Schruben 1989):

- Path search methods: in which a direction of improvement is determined and followed in repeated steps. The main examples are Response Surface Methodology (RSM) (Rees et al. 1985), Stochastic Approximation (SApprox) (Kiefer and Wolfowitz 1951), Perturbation Analysis (PA) (Ho 1984) and the Likelihood Ratio Method (LRM) (Glynn 1987).
- Pattern search methods: in which a pattern in the behavior of the observations is obtained. The main examples are Hooke-and-Jeeves method (HJ) (Hooke and Jeeves 1961) and the Simplex Method (SM) (Nelder and Mead 1965).
- Random methods: in which a randomly selected subset of alternatives is analyzed.
- Integral methods: in which the analysis is done by space- or region-covering and it is specially designed for global optimization.

Several new methods have arisen in the last decades. They have been named as combinatorial methods, or metaheuristics (April et al. 2003). Three different subgroups might be mentioned:

- Simulated annealing (Eglese 1990), in which a new alternative is selected in the neighborhood of the last simulated alternative.
- Evolutionary algorithms like tabu search (TS) (Karaboga and Kalinli 1997), and genetic algorithms (GA) (Goldberg 1989). They search by building and evolving a subset of alternatives.
- Metamodels, like neural networks (Hopfield and Tank 1985). They are used to algebraically represent the simulation model, facilitating then the optimization procedure.

All these methods have been reported to be very useful when  $I$  is very large compared to the other methods mentioned, since they have good convergence properties. The solution of the scheduling problem in flexible manufacturing systems (for example, Fazlollahi and Vahidov 2001) is one of the examples most commonly mentioned.

The main thrust of this group of methodologies is to evaluate the smallest possible subset, that is, the smallest percentage of alternatives, so a satisficing alternative is found in the required time. That means that  $p(R=0) > 0$ , increasing efficiency, but then 100% efficacy and credibility is not guaranteed. A good compromise between efficacy and efficiency is sought, even knowing that the relationship efficacy-efficiency

relies heavily on the initial solution (Shelokar et al. 2004). For that reason, several improvements on the original methodologies are being proposed. For example, in order to generate a good set of initial solutions for a genetic algorithm, a linear program is used (Yokohama 2002).

## QUALITATIVE COMPARISON OF REPORTED SOMs

An extensive literature review has been performed to see the combinations of objective functions and generation algorithms that have been reported. Using electronic versions of databases, the keywords "simulation and optimization" provide many references. For example, Science Direct includes 3854 articles, Springer 214 and Kluwer 649.

The vast majority of references include SUBSET SOMs with an aggregate average cost function that is being optimized using simulated annealing, an evolutionary algorithm or a neural network or a combination of them. Obviously, the research focus is on trying to improve the convergence of the algorithms to the optimum value in problems with many alternatives.

Among the reported combinations, genetic algorithms have been used to find a good local area, which is studied using hill-climb algorithms (Hart et al. 1998). Simulated annealing finds a subset of good-quality alternatives which are compared using comparison and selection (Ahmed and Alkhamis 2002). Neural networks are used as a filtering step prior to simulating alternatives generated using Genetic Algorithms (Glover et al. 1996, Johnson and Rogers 2001, Laguna and Martí 2002, Yu and Liang 2001). Scatter search (Glover 1977) follows the same principle but with tabu search and neural networks.

In terms of the WHOLE SET, the number of references is not high, and again an average cost function is what is reported in conjunction with the traditional comparison and selection (CS) method. The major drawback is that the possible number of alternatives to evaluate cannot be too high (usually less than 20) for the selection errors not to be too large (Law and Kelton 1991).

The only combination reported in which the objective function is the capability index is GESAS (Otamendi 1999), with its upcoming natural continuation GESAS II (Otamendi 2004). This SOM has been successfully applied in studies with many alternatives: one with 2646 (Otamendi 1999), other with 288 (Rivera 1998).

What follows is a comparison of SOMs by studying in detail the characteristics of the two opposite leading methodologies: GESAS, which tries to evaluate the WHOLE SET of alternatives (increasing the speed

while maintaining 100% credibility), and pure EA, which tries to evaluate a small SUBSET in which the optimum is included (improving credibility while maintaining efficiency). A combination of the two methods is also presented.

### Whole Set - GESAS II

It is the first attempt to use a multicriteria process capability index (MPCI) in a WHOLE SET simulation study. The focus is on trying to reduce  $NTOT = I * R_{mean}$  via reducing  $R_{mean}$ , making it possible to evaluate more alternatives I, while keeping efficacy and credibility at 100%.

The use of a MPCI instead of an average cost function allows for the definition of a rejection algorithm (Otamendi 1999) that discriminates easily between good and subpar alternatives. The number of repetitions R is not fixed anymore but dependent on the fitness of the alternative. The satisfaction limits (USL and LSL) are also updated as better alternatives are found (Otamendi 2004). For feasible alternatives,  $R_{max}$  runs are performed, but for subpar alternatives a lower number of executions of the model is needed.

Hence, the distribution of the number of runs  $\Omega$  in the whole range of possible value of R ( $R=0, \dots, R_{max}$ ) is:

$$P_{GESASII}(R) = \begin{cases} p(0) = 0 \\ p(1) = p(1) \\ p(R) = 0 \quad \forall R = 2, \dots, R_{max} - 1 \\ p(R_{max}) = 1 - p(1) \end{cases}$$

Since all the alternatives are evaluated,  $p(R=0)$  is 0. And the performance of the algorithm is such that usually a large percentage of the alternatives are simulated only once, and the rest  $R_{max}$ . To facilitate the upcoming comparison, the small percentage that will be run between 2 and  $R_{max}-1$  is not considered.

The resulting quantification of the efficiency is:

$$\begin{aligned} NTOT &= I \sum_{R=0}^{R_{max}} R p_R = I * R_{mean} \\ &= I * \langle 1p(1) + R_{max} [1 - p(1)] \rangle \end{aligned}$$

The value of  $R_{mean}$  and NTOT for GESASII then depends on  $R_{max}$  and  $p(1)$ . Figure 2 shows the possible values for  $R_{mean}$  if  $p(1)$  is varied between 0 and 0.999 and  $R_{max}$  between 10 and 100.

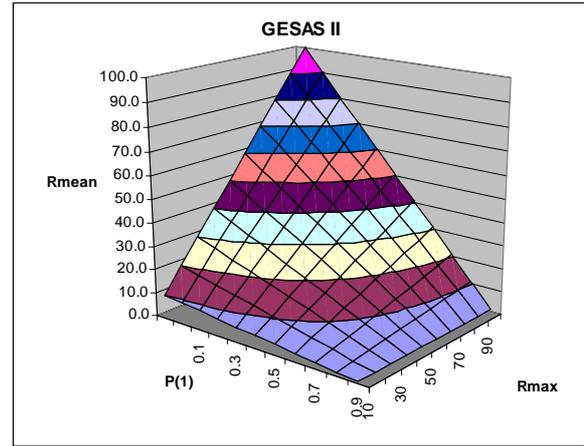


Figure 2.  $R_{mean}$  for GESAS II

For common values of  $R_{max} = 50$  and  $p(1) = 0.80$ ,  $R_{mean} = 10.8$ .

### Subset - Evolutionary algorithms

It is clear that the main research is being done in the area of evolutionary algorithms. The reason is simple: there are many problems in which only a subset of alternatives is to be simulated, with the corresponding possible loss in efficacy and in psychological credibility.

The focus in this group is then primarily in increasing credibility, via convergence to the optimum solution, and efficiency, via reduction of the simulated subset.

The objective function is an *Average* function, and the number of runs is constant within the simulated set. Hence, the distribution of runs in this case is:

$$P_{SA}(R) = \begin{cases} p(0) = p(0) \\ p(1) = 0 \\ p(R) = 0 \quad \forall R = 2, \dots, R_{max} - 1 \\ p(R_{max}) = 1 - p(0) \end{cases}$$

Many alternatives are not simulated and  $R_{max}$  runs are made of the rest. The total number of runs is:

$$\begin{aligned} NTOT &= I \sum_{R=0}^{R_{max}} R p_R = I * R_{mean} \\ &= I * \langle 0p(0) + R_{max} [1 - p(0)] \rangle \\ &= I * \langle R_{max} [1 - p(0)] \rangle \end{aligned}$$

The value of  $R_{mean}$  and NTOT for EA then depends on  $R_{max}$  and  $p(0)$ . Figure 3 shows the possible values for  $R_{mean}$  if  $p(0)$  is varied between 0 and 0.999 and  $R_{max}$  between 10 and 100.

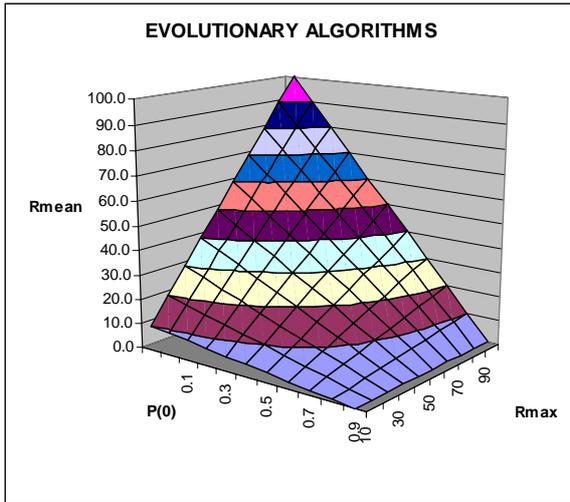


Figure 3.  $R_{mean}$  for Evolutionary Algorithms

The graph looks very similar to the one for GESAS II. Using the same values of  $R_{max} = 50$  and  $p(0)=0.80$ ,  $R_{mean} = 10.0$ .

### THE NEW METHODOLOGY – GESAS II+EA

As a summary of what has been presented so far is that WHOLE SET methodologies favor credibility and must improve efficiency and SUBSET methodologies favor efficiency and must improve credibility.

In terms of the desirable characteristics of a methodology, credibility is increased as  $p(0)$  is reduced and  $R_{max}$  is increased, but then efficiency is reduced.

The MPCII and the rejection algorithm of GESASII are powerful tools that must be used to reject subpar alternatives without performing  $R_{max}$  runs, so more alternatives are analyzed in the same amount of time (bigger SUBSET). If still the WHOLE SET cannot be analyzed, evolutionary algorithms look like a good choice, but need to improve the convergence.

An improvement on EA might be achieved by the inclusion of  $C_{pm}F$  as the objective function instead of the aggregated average cost. Again, the performance of the search algorithm might be dependent on the objective function as well as on the stopping conditions and the initial alternative.

If EA and GESAS are combined into a new SOM, GESAS II + EA, an improvement looks possible. More alternatives will be evaluated per unit time, improving the efficacy for the same number of repetitions.

The distribution of  $\Omega$  for GESAS II + EA is:

$$P_{GESASII+EA}(R) = \begin{cases} p(0) = p(0)_{EA} \\ p(1) = (1 - p(0)_{EA}) p(1)_{GESAS} \\ p(R) = 0 \quad \forall R = 2, \dots, R_{max} - 1 \\ p(R_{max}) = (1 - p(0)_{EA}) (1 - p(1)_{GESAS}) \end{cases}$$

so the quantification of the efficiency is:

$$\begin{aligned} NTOT &= I \sum_{R=0}^{R_{max}} R p_R = I * R_{mean} \\ &= I * (1 - p(0)_{EA}) (p(1)_{GESAS} + R_{max} [1 - p(1)_{GESAS}]) \end{aligned}$$

Figures 4, 5 and 6 show the possible values for  $R_{mean}$  if  $p(0)_{EA}$  is varied between 0 and 0.999 and  $p(1)_{GESAS}$  is varied between 0 and 0.999 for different values of  $R_{max}$ .

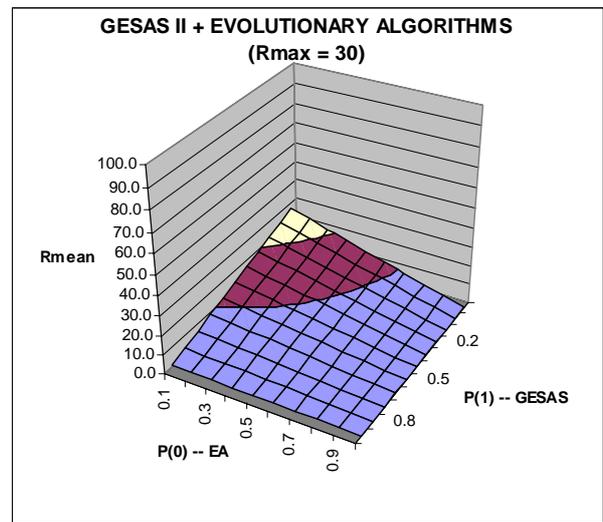


Figure 4.  $R_{mean}$  for Combination ( $R_{max}=30$ )

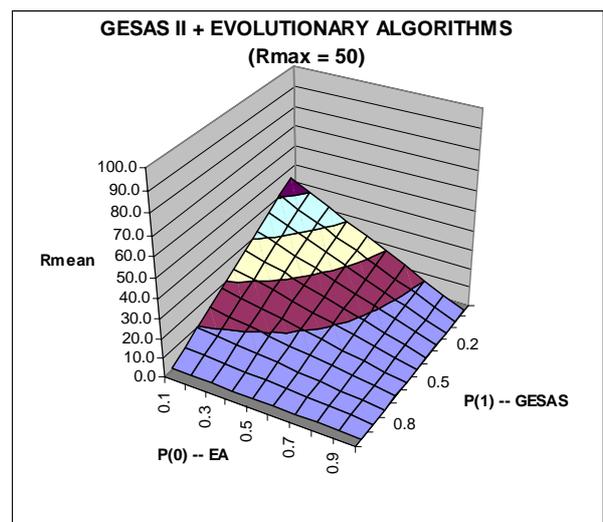


Figure 5.  $R_{mean}$  for Combination ( $R_{max}=50$ )

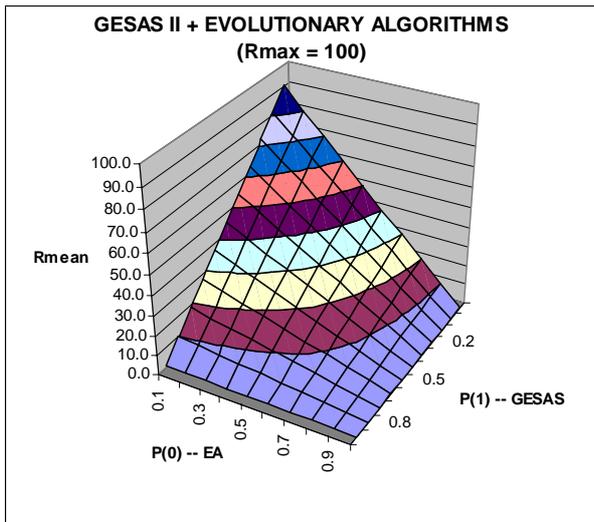


Figure 6.  $R_{mean}$  for Combination ( $R_{max}=100$ )

Using the same values of  $R_{max} = 50$ ,  $p(0)_{EA} = 0.80$  and  $p(1)_{GESAS} = 0.80$ ,  $R_{mean} = 2.2$ , five times lower than in the two separated SOMs.

### QUANTITATIVE COMPARISON OF SOMs

The inventory problem has been used again but with a small modification. The size of the space has been cut in half by reducing the set of available values of the *review interval* to 2, 3 and 4, eliminating the subpar 5, 6 and 7.

The following methodologies are analyzed (Table 8), all with the  $C_{pm}F$  index as the objective function:

- SuperOptimum: infeasible methodology in which one simulation is performed for each alternative except for the selected one, which is run  $R_{max}$  times
- Mesh: each alternative is run  $R_{max}$  times
- GESAS I: mesh with variable number of repetitions
- GESAS II: GESAS I with updating of satisfying limits
- SA
- GA

Table 8. Comparison of SOMs with PCI fitness

METHODOLOGY		VARIABLES				EFFICACY	EFFICIENCY		
Name	Rmax	I	RI	RP	ML	Efficacy	NTOT	Reps/ Alt.	Speed Ratio
SuperOptimum	30	1323	4	575	3600	100.00%	1352	1.02	1.00
MESH	30	1323	4	575	3600	100.00%	39690	30.00	29.36
GESAS I	30	1323	4	575	3600	100.00%	7548	5.71	5.58
GESAS II	30	1323	4	575	3600	100.00%	3205	2.42	2.37
SA	30	1323	2	425	4000	73.23%	199	0.15	0.15
GA	30	1323	3	675	3800	92.39%	636	0.48	0.47

It looks like the efficiency of GESAS II is very good since uses only 2.39 more time that the super optimal alternative. SA, which is included for comparison, is the quicker methodology but it does not achieve the optimum. GA almost attains the optimum, and its convergence rate is about 5 times greater than that of GESAS II alone.

### CONCLUSIONS AND FURTHER RESEARCH

The area of simulation optimization is too complex to develop universal search methodologies. It is very difficult to come up with a multicriteria objective function that includes risk and that might be evaluated with a small number of runs. The selection of the search algorithm is also not easy. So the combination of both objective function and search algorithm is complicated.

It has been shown in this article that the performance of the simulation optimization methodology depends heavily on the definition and evaluation of the objective function, although most of the ongoing research is on the improvement of just the search methodologies over an average response surface.

The analysis has been performed introducing a detailed factorization of the variable NTOT, or total number of repetitions made in the analysis. The separation between runs per alternative, which might be variable, and total number of alternatives, which is usually a just a small subset, helps define and quantify the probability distribution  $\Omega$  of the number of runs per alternative.

The use of evolutionary algorithms, on one hand, helps reduce the number of runs by reducing the size of the subset to run. The use of a capability index, on the other hand, helps by reducing the number of runs per alternative. The combination of both techniques then reduces drastically the total number of runs without reducing credibility in the solution, as demonstrated quantitatively in the inventory example.

So the future looks promising. Improving each technique independently and conjointly is the way to go. Studies are being carried in the quality area to define new aggregate indices, so stronger rejection algorithms might be developed, and larger problems simulated exhaustively using GESAS II.

There is research being done in the evolutionary algorithms area to increase the convergence rate and the credibility of the solution.

And tests must be performed to relate both the objective function, crucial part of any SOM, and the search method.

## REFERENCES

- April, J.; F. Glover; J.P. Kelly; and M. Laguna. 2003. "Practical Introduction to Simulation Optimization." In *Proceedings of the 2003 Winter Simulation Conference* (New Orleans, LA, Dec.7-10). IEEE, Piscataway, NJ, 71-78.
- Ahmed, M.A. and T.M. Alkhamis. 2002. "Simulation-based Optimization Using Simulated Annealing with Ranking and Selection." *Computers and Operations Research* 29, No.4, 387-402.
- Barba-Romero, S. and J.C. Pomerol. 1997. *Decisiones Multicriterio: Fundamentos Teóricos y Utilización Práctica*. Universidad de Alcalá, Madrid.
- Eglese, R.W. 1990. "Simulated Annealing: A tool for Operational Research." *European Journal of Operational Research* 46, No.3, 271-281.
- Fazlollahi, B. and R. Vahidov. 2001. "Extending the Effectiveness of Simulation-Based DSS Through Genetic Algorithms." *Information and Management* 39, No.1, 53-65.
- Gama Caldas L. and L.K. Norford. 2002. "A Design Optimization Tool Based on a Genetic Algorithm." *Automation in Construction* 11, No.2, 173-184.
- Glover, F. 1977. "Heuristics for Integer Programming Using Surrogate Constraints." *Decision Sciences* 8, No.7, 156-166.
- Glover, F.; J.P. Kelly; and M. Laguna. 1996. "New Advances and Applications of Combining Simulation and Optimization." In *Proceedings of the 1996 Winter Simulation Conference* (Coronado, CA, Dec.8-11). IEEE, Piscataway, NJ, 144-152.
- Goldberg, D.E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, New York.
- Glynn P.W. 1987. "Likelihood Ratio Gradient Estimation: An Overview." In *Proceedings of the 1987 Winter Simulation Conference* (Atlanta, GA, Dec.14-16). IEEE, Piscataway, NJ, 366-375.
- Hart, R.P.S.; M.T. Larcombe; R.A. Sherlock; and L.A. Smith. 1998. "Optimisation Techniques for a Computer Simulation of a Pastoral Dairy Farm." *Computers and Electronics in Agriculture* 19, No.2, 129-153.
- Hilgers A. and B.J. Boersma. 2001. "Optimization of Turbulent Jet Mixing." *Fluid Dynamics Research* 29, No.6, 345-368.
- Ho Y.C. 1984. "Likelihood Ratio Gradient Estimation: An Overview." In *Proceedings of the 1984 Winter Simulation Conference* (Dallas, TX, Nov.28-30). IEEE, Piscataway, NJ, 171-173
- Hooke R. and T.A. Jeeves. 1961. "A Direct Search Solution of Numerical and Statistical Problems." *Journal of the Association for Computer Machinery* 8, 21-229.
- Hopfield J. and D. Tank. 1985. "Neural Computation of Decisions in Optimization Problems." *Biological Cybernetics* 52, 141-152.
- Jacobson, S.H. and L.W. Schruben. 1989. "Techniques for Simulation Response Optimization." *Operation Research Letters* 8, No.1 (Feb), 1-9.
- Johnson, V.M. and L.L. Rogers. 2001. "Applying Soft Computing Methods to Improve the Computational Tractability of a Surface Simulation-Optimization Problem." *Journal of Petroleum Science and Engineering* 29, 153-175.
- Karaboga, D. and A. Kalinli. 1997. "A Tabu Search Algorithm for Combinatorial Optimisation Problems." In *Proceedings of the 1997 European Simulation Multiconference* (Istanbul, Turkey, Jun.1-4). SCS.
- Kiefer J. and J. Wolfowitz. 1952. "Stochastic Approximation of the Maximum of a Regression Function." *Annals of the Mathematical Statistics* 23, 462-466.
- Kotz, S. and C.R. Lovelace. 1998. *Process Capability Indices in Theory and Practice*. Arnold, London.
- Laguna, M. and R. Martí. 2002. "Neural Network Prediction in a System for Optimizing Simulations." *IIE Transactions* 34, No.3, 273-282.
- Law, A. and W.D. Kelton. 1991. *Simulation Model and Analysis*. McGraw-Hill, New York.
- Nelder J.A. and R. Mead. 1965. "A Simplex Method for Function Optimization." *Computer Journal* 7, 308-313.
- Otamendi, J. 1999. "G.E.S.A.S. Methodology: A Methodology for the Generation, Evaluation and Selection of Alternatives via Simulation." *Simulation* 73, No. 2, 91-99.
- Otamendi, J. 2001. *Técnicas de Simulación Avanzada para el Tratamiento de Situaciones de Riesgo*. Ph.D. Dissertation. Universidad Politécnica de Cartagena, Murcia, Spain.
- Otamendi, J. 2004. "G.E.S.A.S. II: A Better Relationship Between Efficiency and Efficacy While Experimenting With Simulation Models." In Press, *Simulation*.
- Rees, L.P.; E.R. Clayton; and B.W. Taylor. 1985. "Solving Multiple Response Simulation Models Using Modified Response Surface Methodology Within a Lexicographic Goal Programming Framework." *IIE Transactions* 17, No.1, 47-57.
- Rivera, J., 1998. *Estudio del Proceso de Clasificación de una Empresa de Paquetería vía Simulación*. Creative Component. Universidad Carlos III de Madrid.
- Saaty, T.L. 1980. *The Analytic Hierarchy Process*. McGraw-Hill, New York.
- Shelokar, P.S.; V.K. Jayaraman; and B.D. Khulkarni. 2004. "An ant colony approach for clustering." *Analytica Chimica Acta* 509, No.2, 187-195.
- Taha, H. 1988. *Simulation Modelling and SIMNET*. Prentice Hall, New York.
- Wager, T.D. and T.E. Nichols. 2003. "Optimization of Experimental Design in fMRI: a General Framework Using a Genetic Algorithm." *NeuroImage* 18, No.2, 293-309.
- Yokohama, M. 2002. "Integrated Optimization of Inventory-Distribution Systems by Random Local Search and a Genetic Algorithm." *Computers & Industrial Engineering* 42, No.2-4, 175-188.
- Yu, H. and W. Liang. 2001. "Neural Network and Genetic Algorithm-Based Hybrid Approach to Expanded Job-Shop Scheduling." *Computers & Industrial Engineering* 39, No. 3-4, 337-356.
- Zeleny, M., 1982. *Multiple Criteria Decision Making*, McGraw-Hill, New York.

## AUTHOR BIOGRAPHIES



Javier Otamendi Fernández de la Puebla received the B.S. and M.S. degrees in Industrial Engineering at Oklahoma State University, where he developed his interests in Simulation and Total Quality Management. Back in his home country of Spain, he received a B.S. in Business Administration and a Ph.D. in Industrial Engineering. He is currently a simulation and statistics consultant and professor. His e-mail address is jotamendi\_30@yahoo.com.

# MODELLING OF STEAM TEMPERATURE DYNAMICS OF A SUPERHEATER

Imre Benyó, Jenő Kovács, Jari Mononen and Urpo Kortela  
University of Oulu, Systems Engineering Laboratory  
P.O.Box 4300, FIN-90014 University of Oulu, Finland  
Fax.: +358-8-553-2439, e-mail: [imre@paju.oulu.fi](mailto:imre@paju.oulu.fi)

## Keywords

steam superheater, non-linear models, Wiener system, Hammerstein system

## Abstract

The paper presents a MISO Wiener-Hammerstein cascade model describing the thermal process of a single steam superheater stage. The non-linear static part is based on the energy balance of the medias involved in the process. The linear part involves the time dependence of the process.

The derivative of the model can be easily obtained, which allows applying the Levenberg-Marquadt method for identification. The identification and the validation of the model is presented on the measurement of the second superheater stage of a 180 MW fluidized bed boiler.

The proposed model is simple and transparent, its identification is not demanding. According to the validation result, the model is suitable to test superheater control structures on it.

## Nomenclature

T	temperature,	K
A	surface area,	m <sup>2</sup>
V	volume,	m <sup>3</sup>
m	mass,	kg
$\rho$	density,	kg/m <sup>3</sup>
c	specific heat,	kJ/(kg·K)
h	enthalpy,	kJ/kg
Q	heat flow,	kJ/s
$\alpha$	heat transfer coefficient,	kJ/(s·K)

### indices

st	steam,
fg	flue gas,
m	
fg_rad	representative temperature of the combustion chamber,
m	metal of the pipe,
in	inlet,
out	outlet,
conv	convective,
rad	radiative.

## Introduction

During modelling the steam superheating process can be understood as a combination of two subprocesses: the hydrodynamic and the thermal. Dynamic time constants of these two processes are significantly different, the hydrodynamic changes happen in tenth of seconds meanwhile the magnitude of the time constants of the thermal process are the tens of second. Therefore these processes are usually modelled separately. The model presented in this paper concerns only the thermal process.

Different approaches for modelling the superheater are already available in the literature. The aim of these models is to describe the dynamic behaviour of the temperature of the steam leaving the superheater. The models could be sorted according to the extension of the derived process. Namely whether it covers only the steam side process assuming the transferred heat flow being known; or it also contains the heat transfer phenomena (approximating the heat transfer coefficients) as well, or even the combustion process (heat release).

The models are generally considering the first-principle equations (mass, momentum and energy balances) and the phenomenological correlations (*e.g.* heat transfer correlations). According to the time and spatial distribution of the temperature in the superheater, the process is described by partial differential equations, that solution may be complex.

Zima (2001) presented a model applying the powerful method of the finite difference method for the solution of the partial differential equations. In his model, the heat transfer coefficients were assumed to be known and constant.

The distributed parameter problem was also addressed by the Profos model (Profos 1962). The model based on the one-pipe approximation of the process. This linearized model is derived by performing Laplace transformation once by the time and again by the spatial variable on the partial difference equations. The extended Profos model presented by Czinder (1996) incorporates the dynamical behaviour of the outlet temperature of the flue gas flow.

Oda *et al.* (1995) introduced a simplified model for testing a model reference controller. The model implements the same phenomena as the previous ones, but the distributed parameter problem was solved by applying two of the same concentrated parameter model-block. The fuel flow also appears among the model inputs, and the temperature of the flue gas is estimated. In the model, only radiative heat transfer was assumed. The presented validation data shows good matching between the estimation and measurement of the steam temperature.

Maffezzoni (1997) presented a model to describe the boiler turbine dynamics. The proposed simple linearized model concerns only the steam side phenomena utilising the heat flow assumed to be known.

For the simulation of the superheater process, black box models are also proposed in the literature. For example, Alippi and Piuri (1995) reported a computationally simple, distributed non-linear model; however their model covers not only the superheating process, but the whole power plant. The identification of their neural model was performed on a 320 MW one-through boiler. The inputs of their superheater model were the same as the inputs of the model proposed in this paper.

In this paper, the Wiener-Hammerstein cascade model is applied for the modelling of the superheated steam temperature behaviour. The Wiener and Hammerstein models are widely used for modelling of non-linear process, because of their transparency, the capability to capture well the behaviour of the process and because of the feature to be identified easily. The heat transfer coefficients are not known, thus the model identification must include the approximation of this parameter. For the identification the gradient based Levenberg-Marquadt algorithm was applied.

## Wiener Hammerstein Cascade process model

In many cases, the behaviour of non-linear process can be approximated by linear transfer functions for describing the system dynamics, and a non-linear static function describing the non-linearity. Wiener and Hammerstein structures are typical examples of such structures.

The Wiener and Hammerstein models have several advantages, the most important ones:

- the function can be derivated by the parameters of the static part and by the parameters of the linear dynamic parts as well;
- the linearity of the dynamic part simplifies not only the parameter estimation, but also the (closed loop) system analysis, modelling of disturbance, and controller design;
- the *a priori* knowledge about industrial processes usually concerns the steady state relations. With

this model structure it is easy to incorporate it into the model.

The simple Wiener-Hammerstein cascade model (Haber and Keviczky 2002) consists of linear dynamic parts and one non-linear static term connected in series, as shown in Figure 1.

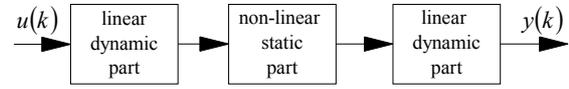


Figure 1. The Wiener-Hammerstein cascade model

## The superheater model

In the case of the superheater, the static part is a multi-input single-output function. Thus the first linear dynamic part in Figure 1 contains several linear dynamics. The more detailed model for the superheater is given in Figure 2

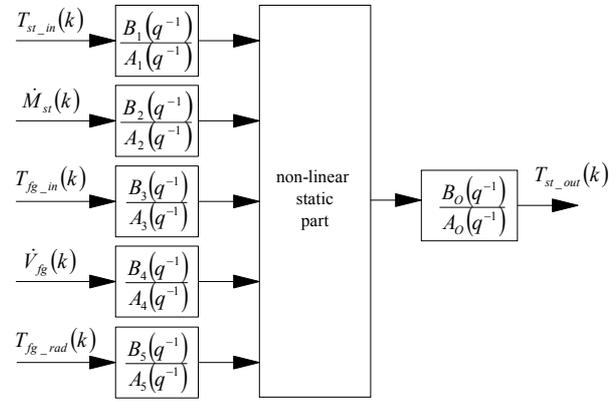


Figure 2. The Wiener-Hammerstein cascade model for the superheater

All the dynamic parts are chosen to be second order transfer function with unit gain. Thus one dynamic model has only three parameters:

$$A_i(q^{-1}) = 1 + a_{i,1}q^{-1} + a_{i,2}q^{-2} \quad (1)$$

$$B_i(q^{-1}) = b_{i,0}q^{-1} + (1 + a_{i,1} + a_{i,2} - b_{i,0})q^{-2} \quad (2)$$

The nonlinear static part is based on a concentrated parameter static model of the superheater process. The output temperature is expressed from the energy balance equations of the steam (3), flue gas (4) and wall (1), and from the convective (5,6), and radiative heat transfer (7) equations.

The energy balance of the wall:

$$\dot{Q}_{st} = \dot{Q}_{fg} + \dot{Q}_r \quad (3)$$

The energy balance of the steam flow:

$$\dot{Q}_{st} = c_{st} \cdot \dot{m}_{st} (T_{st\_out} - T_{st\_in}) \quad (4)$$

The energy balance of the flue gas flow:

$$\dot{Q}_{fg} = c_{fg} \cdot \dot{V}_{fg} \cdot \rho_{fg} \cdot (T_{fg\_in} - T_{fg\_out}) \quad (5)$$

The heat transfer from the wall into the steam

$$\dot{Q}_{st} = \alpha_{st} (T_m - T_{st}) \quad (6)$$

The convective heat transfer from the flue gas flow to the wall:

$$\dot{Q}_{fg} = \alpha_{fg} (T_{fg} - T_m) \quad (7)$$

The radiative heat transfer from the combustion chamber to the wall:

$$\dot{Q}_r = \alpha_{rad} (T_{fg\_rad} - T_m) \quad (8)$$

The average steam and flue gas temperature.

$$T_{st} = \frac{T_{st\_in} + T_{st\_out}}{2} \quad (9)$$

$$T_{fg} = \frac{T_{fg\_in} + T_{fg\_out}}{2} \quad (10)$$

To facilitate the expression of the output steam temperature the following approximations were applied:

- the radiative heat transfer is approximated to be linear to the temperature difference (7);
- the representative temperatures for the convective heat transfer calculations are the linear average of the inlet and outlet steam and flue gas temperatures (8,9), and not the logarithmic means, as it is suggested in the literature.

In this function the convective heat coefficients are approximated as linear functions of the fluid flows around the surface, thus

$$\alpha_{st}(k) = a_{st} \cdot \dot{m}_{st}(k) + b_{st} \quad (11)$$

$$\alpha_{fg}(k) = a_{fg} \cdot \dot{V}_{fg}(k) + b_{fg} \quad (12)$$

After a series of substitutions and arrangements the output steam temperature can be expressed:

$$T_{st\_out} = f(T_{st\_in}, \dot{m}_{st}, T_{fg\_in}, \dot{V}_{fg}, T_{rad}) \quad (13)$$

## Identification and Validation

The identification of the model was performed on the measurement data of a 185 MW Bubbling Fluidized Bed Boiler.

The aim of the identification is to determine the model parameters: the  $b_{i,1}$ ,  $a_{i,1}$  and  $a_{i,2}$  coefficients of the dynamic parts and the  $a_{st}$ ,  $b_{st}$ ,  $a_{fg}$ ,  $b_{fg}$  and  $\alpha_{rad}$  parameters of the static part. The parameters to be identified are put into the  $\theta$  vector.

Most of the input variables (steam mass flow, steam inlet temperature, flue gas inlet temperature and representative temperature of the combustion chamber) were taken directly from the measurement. The flue gas volume flow was calculated by an Adaptive Neuro-Fuzzy Interference System (ANFIS) that describes the combustion process. The applied ANFIS model is presented by Himer (2003) in details.

The minimization was performed by a gradient based second order method, the Levenberg-Marquadt algorithm as it is given in Ikonen (2001). The cost function is:

$$J(\theta) = \frac{1}{2} R(\theta)^T R(\theta) \quad (14)$$

where the components of the  $R$  vector are

$$r_k = T_{st\_out}(k) - T_{st\_out\_meas}(k) \quad (15)$$

where  $k=1,2,\dots,N$ , and  $N$  is the number of data records.

The Levenberg-Marquadt iteration is given:

$$\theta(l+1) = \theta(l) - [G(\theta)^T G(\theta) + \mu(l)I]^{-1} G(\theta)^T R(\theta) \quad (16)$$

where the elements of the  $G$  matrix are:

$$g_{k,p} = \frac{\partial r_k}{\partial \theta_p} \quad (17)$$

and the  $\mu(l)$  is increased whenever the step would result to an increased value of the cost function, otherwise reduced.

Since the optimization algorithm is a gradient based method, the derivatives of the cost function according to the parameters to be optimized must be calculated. (This is the main reason why the steam outlet temperature must have been expressed explicitly, and (7-9) approximations were needed.)

During the iteration, the estimated  $A_i$  polynomials can happen to become unstable. To avoid the unlikely result of applying the unstable transfer function, in every iteration round the new parameters are checked. If instability was encountered, the parameters were projected towards the stable region.

The estimated and the measured steam outlet temperature on the identification data range are shown in Figure 3. The inputs (steam inlet temperature, mass flow, etc) are illustrated in Figure 4.

The validation of the identified model was performed on another measurement series from the same boiler. The model performance is presented in Figure 5; the inputs of the model are given in Figure 6.

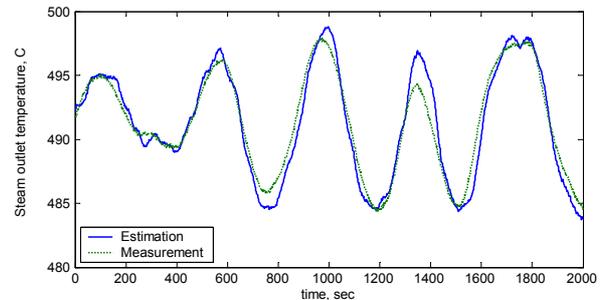


Figure 3. The measured and estimated outlet steam temperature in the identification

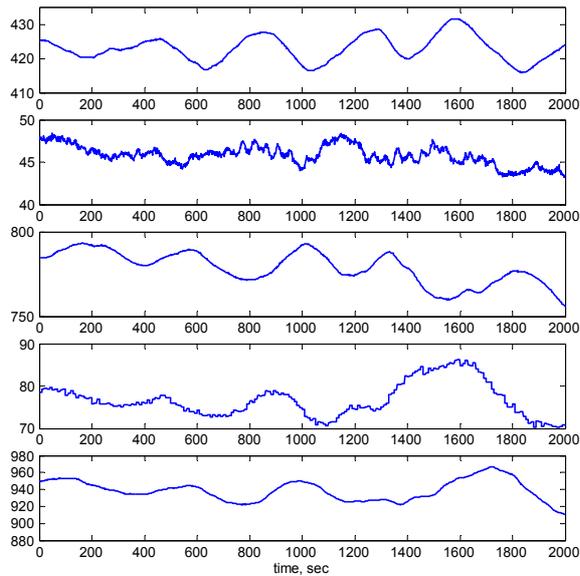


Figure 4. The model input variables (steam inlet temperature, steam mass flow, flue gas inlet temperature, flue gas volume flow, flue-gas representative temperature in the combustion chamber respectively) on the time range applied in the identification

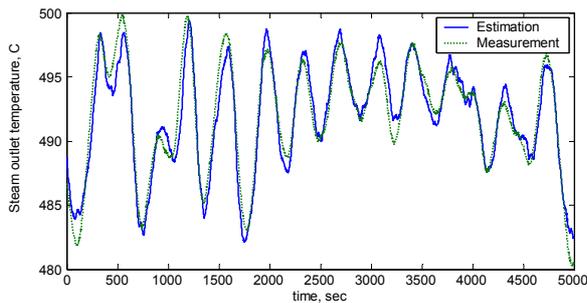


Figure 5. The measured and estimated outlet steam temperature in the validation

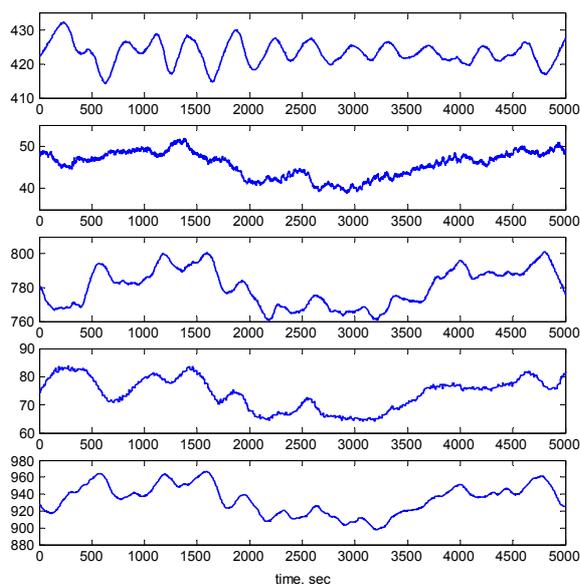


Figure 6. The model input variables in the validation (same as in Figure 4)

## Conclusions

This paper described a Wiener–Hammerstein cascade non-linear model for superheater steam temperature. The model steady state characteristic is based on the phenomenological equations; meanwhile the dynamic behaviour is merely identified.

The validation data shows good results. The applied structure seems satisfactory for this problem and undemanding from the computational burden point of view. The model is suitable to test superheater control structures on it.

## Acknowledgement

This research was supported by the Academy of Finland, project no. 73281 and 48545. The research work was also supported by the Graduate School of Electronics, Telecommunication and Automation.

## References

- Alippi C. and Piuri V. 1995. Identification of Non-linear Dynamic Systems in Power Plants., Proceedings of the 38th Midwest Symposium on Circuits and Systems, Pages:493 - 496 vol.1
- Barin I, 1989. Thermochemical data of pure substances, Part I and II, VCH, Weinheim, 1989
- Czinder J, 1996. Linear modelling of heating surface, the extension of Profos-model, Hungarian Energy, 1996/6
- Haber R, Keviczky L, 2002. Nonlinear System Identification – Input-Output Modelling Approach, Kluwer Academic Publisher, Boston
- Himer Z, Wetz V, Kovács J, Kortela U, 2004. Neuro-Fuzzy model of flue-gas oxygen content, Proceedings of 23<sup>rd</sup> International Conference on Modelling, Identification and Control, Grindelwald, Switzerland
- Ikonen I, Najim K, 2001. Advanced Process Identification and Control, Marcell & Dekker.
- Maffezzoni C. 1997. Boiler-turbine Dynamics in Power-Plant Control, Control Engineering Practice, Vol. 5, No. 3, 1997.
- Oda K, Toyoda Y, Nakamura H, 1995, Model Reference Control for Steam Temperature of Power Plant at Start-up Stage, Proceedings of SICE 1995.
- Profos S. 1962. Die Regelung von Dampfanlagen. Springer Verlag.
- Zima W, 2001. Numerical modelling of dynamics of steam superheaters, Energy vol. 26, page 1175-1184.

## Author Biographies



**IMRE BENYÓ** was born in Budapest, Hungary in 1975. He was graduated at the Technical University of Budapest as mechanical engineer. He is researching at the System Engineering Laboratory, University of Oulu, Finland. His research area covers the predictive control, system identification problems, and its applications in the power plant control.



**JENŐ KOVÁCS** (M.Sc. 1991 Budapest, Hungary, Ph.D. 1998 Oulu, Finland) is a senior assistant at the Systems Engineering Laboratory, University of Oulu, Finland. His research interests include adaptive control, constrained control, advanced modelling and their application to energy systems and power plant control problems.



**JARI MONONEN** (M.Sc. 1995 University of Oulu, Finland) is a researcher at the Systems Engineering Laboratory, University of Oulu, Finland. He has a long experience in the modelling and control of power plants. His main research area is the development of combustion control and in emission optimisation in full-scale applications. Currently, he is completing his Ph.D. studies concerning identification and control of non-linear systems.



**URPO KORTELA**, born in Finland, 1945, is the head professor of the Systems Engineering Laboratory, University of Oulu, Finland. He graduated as M.Sc. in Technical Physics in 1970 at the University of Oulu, Finland. He received the Licentiate of Technology in 1973 at the University of Oulu and the Doctor of Technology in 1981 at the University of Helsinki, Finland. His interest lies in the research in control engineering and system theory: state and parameter estimation and advanced control methods. The application field consists of power plant modelling and control, control and fault diagnosis of pulp and paper processes, and field bus technology.

# PLANNING OF ORDER PICKING PROCESSES USING SIMULATION AND A GENETIC ALGORITHM IN MULTI-CRITERIA SCHEDULING OPTIMIZATION

Balázs Molnár  
Budapest University of Technology and Economics  
Department of Transportation Technology  
Bertalan L. u. 2, 1111  
Budapest, Hungary  
E-mail: bmolnar@kku.bme.hu

## KEYWORDS

Simulation, Genetic Algorithms, Scheduling, Order Picking

## ABSTRACT

In profit-oriented environments, such as warehousing, the minimization of labor costs, and thus the flexibility of labor force is an increasingly important issue.

Such an operating policy requires effective capacity planning methods to determine the number of personnel and equipment per activity and scheduling procedures to define the sequence of tasks by considering their strict deadline.

In the following, the use of a discrete event simulation model for multi-criteria scheduling optimization of order picking activities in a warehouse with genetic algorithm (GA) is presented.

The operative planning system consists of a database, a discrete event simulation model, an application for capacity estimation and a scheduling algorithm. The system was designed to support operative warehouse management personnel in order picking process scheduling and planning.

## INTRODUCTION

Flexibility of labor in a warehouse means that available personnel are redeployed during shifts to activities (storage, order picking, replenishment, etc.) where extra capacity is required. In case the available labor capacity is not sufficient, temporary staff can be hired from specialized agencies.

Order picking – retrieval of products from storage to meet customers' demand – is often the most labor intense activity in a warehouse. The human hand as a "handling equipment" is hard to replace and economical automation of retrieval of products is seldom possible. Therefore, the costs of order picking may amount to about half of the operational costs in a warehouse, so

any improvement in this field may result in significant cost reduction (Van den Berg 1999, Roodbergen 2001).

## THE ORDER PICKING PROCESS

### Planning and Disposition

Based on the customers' orders, the operative planning of the order picking process in a *Warehouse Management System* (WMS) is completed in the following steps (Ten Hompel and Schmidt 2002):

1. Download of customers' orders from the Host system;
2. Separation of orders into one-unit and multi-unit orders;
3. Appointment of the pick position of each product;
4. Drawing up internal orders (pick lists) based on the external (customers') orders;
5. Calculation of the number of empty pallets needed per pick list;
6. Determination of the retrieval (lead) time of each pick list;
7. Disposition of resources to each pick list by considering different constraints.

### Retrieval of Products

A so called *picker-to-part* order picking method is implemented in the presented planning system. The order pickers receive information about their next task at a designated location (depot) in the warehouse. First they drive their pallet truck to pick an empty pick device (pallet). Then the order pickers ride along the pick positions in the aisles of a multi-block warehouse. The order pickers pick the proper quantity manually from the lowest level of the racks to the pallet on the truck. Following each pick, they confirm the action and then they read information about the next pick location. When the pick device is full or the picking list is completed, the pickers ride to the area of the warehouse where the shipments are controlled and prepared before loading them on the trucks. After dropping off the pick device the pickers return to the depot to receive the next task (Gudehus 1999, Tarnai 2000).

## THE MODEL

The object of the developed model is to determine on the one hand the number of order pickers, on the other hand the sequence of the retrieval of the pick lists so that the total cost of order picking is minimal. The objective function describing the optimization problem consists of the following three terms (Kljajić 2002):

- Minimization of the labor costs;
- Minimization of earliness/tardiness costs;
- Maximization of resource utilization.

The labor costs are determined by the number of order pickers and the specific labor costs which may vary per each shift.

The pick lists must be ready for shipping by the internal deadlines calculated by the tour-planning module of the WMS based on delivery dates of the customers' orders. Deviation from these cut off times induces incidental expenses. If the orders are prepared earlier, then these must be stored temporarily, so storage costs occur and in addition, useful space is occupied from other warehousing activities. However, if the trucks must wait because the orders are not picked on time, then transports may arrive late to customers, service level of delivery declines and extra labor and other costs may occur.

In order to convert the objective function into a minimization problem, the maximization of resource utilization is formulated as the minimization of idle times. Idle times refer to the times when order pickers do not work during the shift and they are not taking a rest but have no task to perform.

The model determines the optimal sequence of the completion of the pick lists in three phases:

- Experiment;
- Estimation;
- Optimization.

In the first phase the time needed to pick each list is evaluated. Based on the mean times, the number of required order pickers is estimated for each shift. In the third phase, the optimal schedule can be produced (Figure 1). In the following, the three phases of the model are described.

## EXPERIMENT

The model supports the last two steps of the operative planning of the order picking process described previously, so it is assumed that the pick lists are already produced. The lists must contain the ID number, the pick location and the pick quantity of each stock keeping unit (SKU) on the list, and the deadline of completion. The lists are processed in tab delimited file (.txt) format in order to make the system independent from the type of the database in which the lists are stored.

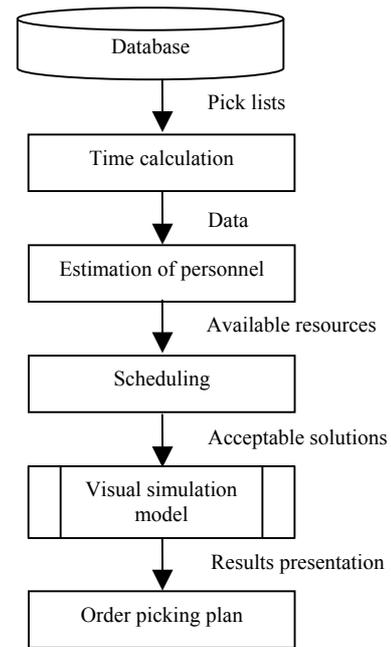


Figure 1: Scheduling and decision support process for order picking planning

To simulate order picking processes, the times needed for:

- Preliminary activities,
- Picking up and dropping off loads,
- Acquisition of information, recording of data,
- Concluding activities (fastening of the load etc.),

and the physical properties, the speed and acceleration values of the applied equipment must be analyzed. For validation purposes and to handle the changes in the performance of the order pickers, the stochastic of the activities, the structure of the shifts, and to complete planned experiments the model was implemented in Enterprise Dynamics 5.1, a visual interactive modeling and animation simulation package. The model is depicted in Figure 3.

In the *Experiment* phase, the retrieval time of each and every list is measured separately. The number of measurements is a variable set by the user. The results – the single time values, the average and the standard deviation of times measured – of an Experiment of 10 runs with 5 pick lists are shown in Figure 2.

Results of Experiment in ED						
Number of runs in ED: 10						
	7	8	9	10	Mean/Avg.	Std.Deviation
PickList1	1019	1015	1034	1017	1017	7
PickList2	420	416	421	422	420	2
PickList3	974	959	986	967	972	9
PickList4	964	975	971	962	965	5
PickList5	570	573	575	577	574	3

Figure 2: Results of the *Experiment* phase

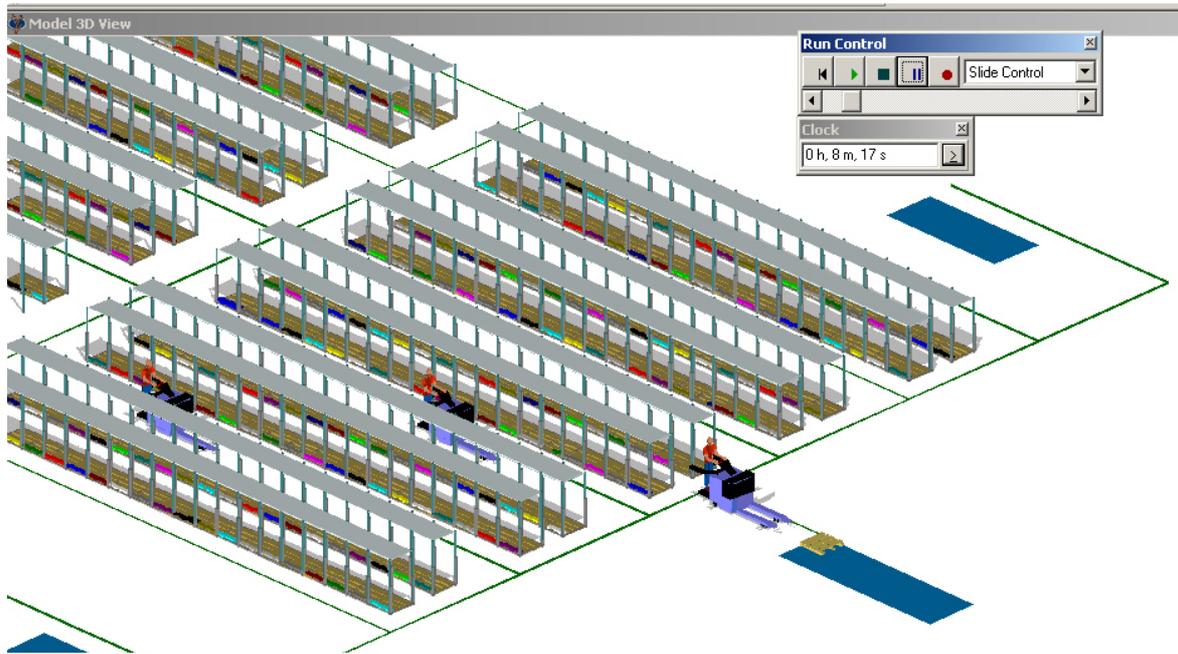


Figure 3: The Simulation Model for the Order Picking Process

## ESTIMATION

A good order-picking schedule means good resource utilization while respecting cut off times and other constraints in the warehouse (pauses during the shifts etc.).

Based on the deadline of completion, the lists are sorted into shifts. In this phase the deadlines inside the shifts are not taken into consideration. The application – developed in the Delphi programming environment – estimates the number of order pickers needed per shift so as to complete all lists during the given shift by the most balanced load on the pickers. The input of the application is the result table of the *Experiment* phase with the average retrieval times of the pick lists. The output of the *Estimation* phase is shown on Gantt charts. Figure 4 represents 34 pick lists in a shift picked by five order pickers.

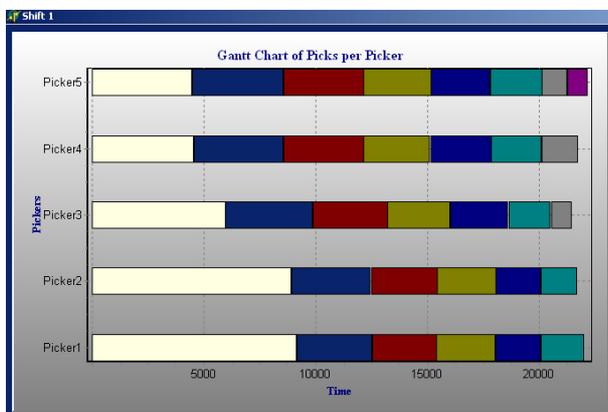


Figure 4: Gantt chart – Estimation

## OPTIMIZATION

The change of the management interest from productivity improvement to inventory reduction, the introduction of information technologies that enable it, and the emergence of new management philosophies like Just-in-Time (JIT) production, demand warehouses to deliver lower volumes but more frequently with shorter response times from a significantly larger assortment of products.

As a result of these trends, the number of customer orders has increased and the punctuality of deliveries has become essential. Since the complexity of the scheduling problem grows factorially with the number of tasks to be carried out, traditional methods are not able or consume too much time to find a suitable solution.

Constraint programming (CP) is a technique to solve non-linear problems, mostly in planning and scheduling. The problems are solved by imposing constraints and choosing an appropriate search strategy. Genetic algorithms are proven search methods with good quality/speed ratio (Goldberg 1997).

### Genetic Algorithm

Genetic algorithms represent the solutions of a problem by a set of parameters. These parameters are regarded as the *genes* of a *chromosome*, and a set of chromosomes is named *population*. In case of permutation problems, the best representation technique is to use the indexes of the tasks in the permutation, and so is one chromosome a sequence of integer numbers (Derhán 2002).

The chromosomes of the GA in the Optimization phase represent the order in which the pick lists are released to the order pickers to be retrieved from storage. Figure 5 shows the logic of the optimization with GA.

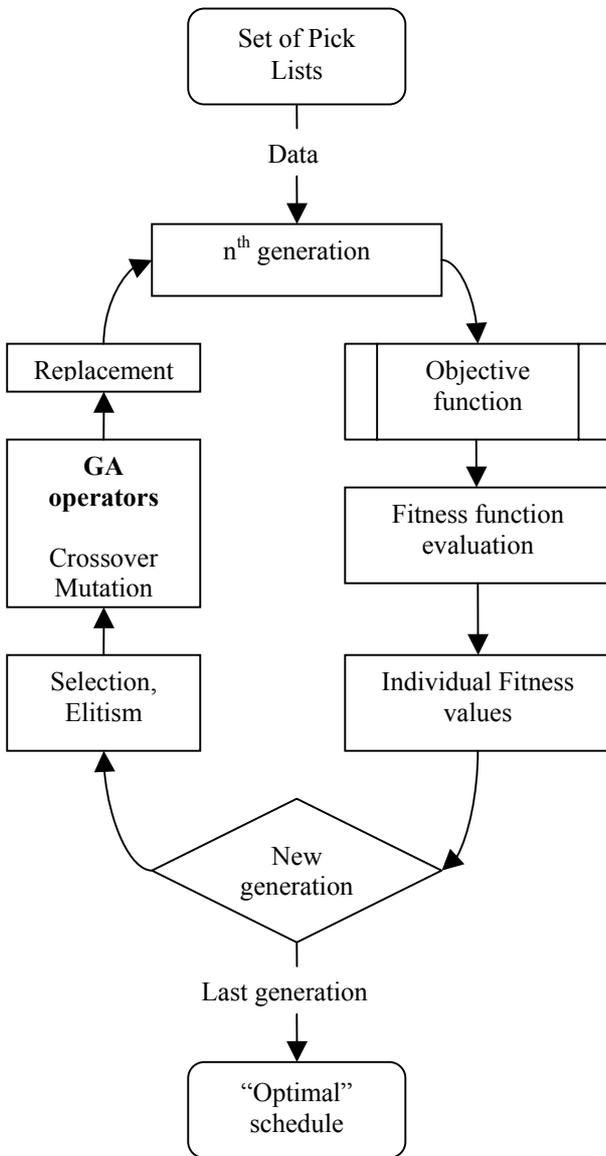


Figure 5: Optimization with GA

The GA prepares an initial population (1<sup>st</sup>) of chromosomes based on the data extracted from the database. The number of genes in a chromosome is equal to the number of pick lists in the scheduling problem. The number of the chromosomes in each population is fixed and may be set by the user at the beginning of the scheduling. With the help of the simulation model, the objective functions of each chromosome in a generation is evaluated. Based on the objective function values, the fitness of each member of the given generation is calculated. According to their fitness, parent chromosomes are selected and they form an offspring with GA operators, like crossover and mutation. The offspring is placed in the new (n<sup>th</sup>) generation.

Offspring are created until a new generation is not complete. The new generated population is used in the further run of the algorithm. The creation of new populations is repeated until the end condition is not satisfied. From the last generation the best solution is returned.

The program for scheduling optimization was developed in the Delphi programming environment, and for the evaluation of the objective functions, the model in Enterprise Dynamics 5.1 – described in the Experiment phase – was used.

The objective function ( $O$ ) can be described in the following discrete form:

$$O = \Sigma(P_i * C_{Li}) + E * C_e + T * C_t + I * C_i \quad (1)$$

where

- $O$ : the total cost of order picking in the planning time horizon
- $i$ : number of shifts
- $P_i$ : number of order pickers in shift  $i$
- $E$ : total time of earliness
- $T$ : total time of tardiness
- $I$ : total idle time of pickers
- $C_{Li}$ : labour cost of a picker in shift  $i$
- $C_e$ : the unit time earliness penalty (cost)
- $C_t$ : the unit time tardiness penalty (cost), in practice, generally  $C_t > C_e$
- $C_i$ : the unit idle time penalty (cost)

The operative warehouse management personnel set the  $C_{Li}$ ,  $C_e$ ,  $C_t$  and  $C_i$  values. The simulation model evaluates each chromosome in a population separately, calculates the time difference between the cut off time and the actual finish time of every pick list and stores the idle time of the pickers.

The objective value  $O$  is mapped into a fitness value  $F$ , by the Power Law Scaling method, where the actual fitness value is taken as a specific power ( $k$ ) of the objective value (Man 1999):

$$F_i = (O_{max} - O_i)^k + O_{min} \quad (2)$$

where

- $O_i$ : objective value of chromosome  $i$
- $O_{max}$ : the largest objective value in the population
- $O_{min}$ : the smallest objective value in the population
- $F_i$ : fitness value of chromosome  $i$
- $k$ : constant, set by the user

The selection technique selects two parents at a time and employs the Roulette Wheel Mechanism (Man 1999). After selecting two parents, the offspring is formed by applying the Order Crossover (OX) technique, described in Figure 6. In case of Order Crossover one randomly selected segment (8,7,3) from Parent 1 is copied to the same place in the Offspring. The remaining empty positions in the Offspring are filled from Parent 2, by keeping the sequence but skipping the already used points.

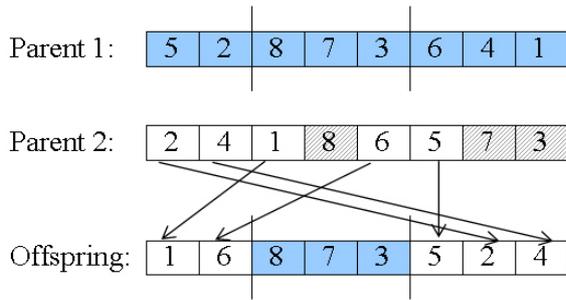


Figure 6: Order Crossover

The Offspring is mutated with a mutation probability – which also can be set at initialization – using the Inversion mutation technique (Figure 7). This method inverts the genes between two randomly selected segments of the chromosome.

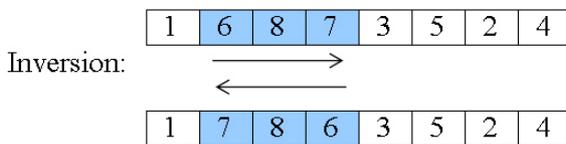


Figure 7: Inversion Mutation

The chromosomes created with the GA operators after selection are added to the new population. Elitist selection is the other method to add chromosomes to the new population. Elitism means that the fittest chromosomes survive. A user set parameter determines the number of fittest chromosomes which are selected and placed in the new population without any manipulation. When the new population is complete, the previous population is replaced and the search for the solution continues.

When the end condition is fulfilled, the Algorithm stops the search process. The result is the sequence of the pick lists in which they should be released to the floor to achieve the best solution by considering all of the predefined constrains.

## Order picking process scheduling and planning

As mentioned in the *Estimation* phase, the number of pickers is evaluated per shift, based on the total number of pick lists but only partially considering their final deadline. It is also stated that good scheduling means good resource utilization while respecting constraints of the order picking process. To achieve good resource utilization, the number of pickers evaluated in the Estimation phase must be tested and – if necessary – refined.

The model's feature is that the number of pickers per shift is not a fixed number but can be varied in an interval. The middle of the interval is the number of pickers defined in the Estimation phase, and the radius can be set by the user. The default value of the radius is 1, which means, in case of an estimated number of 5 pickers in a shift, the model will also evaluate the scenarios, when 4 and 6 pickers are working in the given shift. If there are three shifts in a warehouse, that means 27 possible scenarios.

The model evaluates all the possible scenarios separately and searches for the best sequence of the pick lists. When the best solution for each scenario has been determined the results are presented to the user. As the output of the model, the operative warehouse management personnel receive the following data per scenario:

- Number of order pickers in each shift
- Best sequence of the pick lists returned by the Genetic Algorithm
- Total labor cost
- Total cost of earliness/tardiness and idle times

Based on these data, the operative warehouse management personnel can decide the number of order pickers to be deployed per shift to retrieve orders from the warehouse. For every scenario, the best sequence of fulfilment of the pick lists is presented. By analysing the costs, the management can decide if they undertake the risk of preparing some orders later than the deadline determined by the tour-planning module of the WMS or employ more order pickers with higher labor costs, if necessary.

## CONCLUSION

Optimization and simulation are both tools that support decision making. Optimization uses fixed input data, avoids uncertainty and details. Optimization models simplify the complexity of the real system and some factors are even not considered. Simulation is not creative like optimization, but can cover uncertainty and complexity of dynamic systems in detail.

The combination of optimization and simulation (simulation optimization) can be defined as the process of finding the best set of input variables without evaluating each possibility. The objective of simulation optimization is to minimize the resources spent (i.e. time) while maximizing the quality of information gained in the experiment.

The model represented in this paper also uses the benefits of simulation optimization. The designed system supports operative warehouse management personnel in order picking process scheduling and planning. By evaluating a number of scenarios, the number of the order pickers per shift, and the best sequence of releasing the pick lists to be retrieved from storage are determined.

It is the management's responsibility to monitor and control the order picking activities in the warehouse continuously and force the adherence to the schedule. If all order picking activities are realized according to the schedule, then the planning of the replenishment of the order picking places is also possible. The goal of the author of this paper is to further develop the above described planning system and include the scheduling of these activities, too.

The connection to the database of the WMS with the simulation model already exists and so it is possible to determine when the last products will be picked from each picking place and when replenishment is necessary. By applying advanced search methods – like Genetic Algorithms – the optimal schedule for the replenishment of the picking places can be evaluated. The objective function must reflect the goal of planning the replenishment process so that the order picking processes can be executed continuously and undisturbed – products are available at the picking place and the congestion in the aisles is avoided.

## REFERENCES

- Derhán D. 2002. *Optimalizáló eljárás készítése genetikus algoritmus alkalmazásával*. M.S. thesis, Budapest University of Technology and Economics.
- Goldberg, D.E. 1997. *Genetic algorithms in search, optimization & machine learning*. Addison Wesley.
- Gudehus, T. 1999. *Logistik 2*. Springer Verlag, Berlin.
- Kljajić, M.; Bernik, I.; Breskvar, U. 2002. „Production Planning using Simulation Model and Genetic Algorithms“. In *Proceedings of the IASTED International Conference*, (Marina del Rey, CA, USA, May 13-15.). ACTA Press, Anaheim, 54-58.
- Man, K.F.; Tang, K.S.; Kwong, S. 1999. *Genetic Algorithms, Concepts and Design*. Springer Verlag, Berlin.
- Roodbergen, K.J. 2001. *Layout and Routing Methods for Warehouses*. Ph.D. thesis, Erasmus University, Rotterdam.
- Tarnai J. 2000. „A raktári kommissiózó rendszerek fejlesztése.“ *Anyagmozgás - Csomagolás* 45, No.6, 8-11.
- Ten Hompel, M.; Schmidt, T. 2002. *Warehouse Management*. Springer Verlag, Berlin.
- Van den Berg, J.P. 1999. „A literature survey on planning and control of warehousing systems.“ *IIE Transactions* 31, 751-762.

## ACKNOWLEDGEMENT

The author would like to thank Prof. György Lipovszki at the Budapest University of Technology and Economics for his encouragement and valuable suggestions.

## AUTHOR BIOGRAPHY

**BALÁZS MOLNÁR** was born in Debrecen, Hungary and went to the Budapest University of Technology and Economics, where he studied transportation technology and logistics and graduated in 2002. He is now a PhD candidate at the Department of Transportation Technology and his research field is organization and planning of order picking processes.

# SIMULATION SUPPORTED OPTIMISATION OF INVENTORY CONTROL PROCESSES BY APPLICATION OF GENETIC ALGORITHMS

Krisztián Bóna

Budapest University of Technology and Economics  
Faculty of Transportation Engineering – Department of Transportation Technology  
1111 Budapest, Bertalan Lajos utca 2.  
Hungary  
E-mail: kbona@kku.bme.hu

## ABSTRACT

The inventory control systems are responsible for the optimal operation of the inventory processes of a company. Generally, the optimisation of the inventory control system manifests itself in a target conflict representing the implementation of the optimal operation in economic and reliability terms. For the process optimisation, the control parameters of the regulation system should be defined. Their actual settings determine the time of placing orders and the required quantities for the optimal operation of the processes defined above. This article presents a particular method of exploitation of the opportunities provided by the computer aided simulation and the genetic algorithms for the optimisation of inventory control systems applying classical inventory mechanisms.

## THE PROBLEM

A stock is generally composed of several stock keeping units (SKU). We supposed in our examinations that the optimisation of the inventory processes for each SKU provides the optimum of the entire inventory system as well. This is the principle of the so called SKU based inventory optimisation (Chikán 1983). Hereafter, this paper analyses the problems of the SKU based inventory optimisation.

The SKU based inventory optimisation is a complex, multi-criteria optimisation problem. The main points are the following:

The basic problem is the expansion of the reliability of the system and the reduction of the operation related costs are conflicting requirements in terms of inventory planning. The continuous operation of the process requires the expansion of the stock levels, while the economic efficiency demands their reduction.

In addition, the parameters of the processes triggering the inventory system change in time, i. e. the dynamic features of the processes can not be disregarded. The varying character of the customers' demand in the inventory control system influences the operation of all

other processes in the system shows a simple example for this statement. Consequently, the actual values of the parameters controlling the system should also be set dynamically, i. e. an adaptive inventory control system should be established being able to determine the actual optimum values of the control parameters considering the changes taking place in the inventory system. This is an essential criterion, as - due to the temporal changes in the system - the actual optimum parameter setting may not be optimal in the future.

Accordingly, the optimisation of the inventory processes implies the obtainment of minimum total costs concerning the inventory process for a given period and/or maximum probability of the satisfaction of all demands entering the system within a specified period. Theoretically, the achievement of these objectives is possible by controlling the inventory processes by setting of the various control parameters so that the above mentioned costs and reliability indicators tend to the right direction.

When establishing the inventory control system - among others - two basic questions should be answered (Chikán 1983):

- (1.) What control parameters are required for the optimisation of the process according to the criteria indicated above?
- (2.) How to determine dynamically the values of the control parameters to achieve the objectives?

The various inventory mechanisms and the inventory models modelling their operation can be applied to give exact answers to these questions. The control parameters are exactly determined by the inventory strategies applied, while for the calculation of the actual parameter values exact mathematical models are available. Up to now, the operations research specialists developed approximately four hundred various inventory models to model the various inventory strategies and describe them mathematically. However, experiences show that there are very few practical applications (as compared to the number of models).

The main reasons of this situation are:

- The application of the models is frequently tied to constraints that can not be met in the real stochastic processes.
- The application of an exact mathematical formalism to define the target function required to the optimisation is usually very difficult or even impossible.
- The target function of the models is able to manage only the cost or only the reliability parameters.
- If the target function is available, the next problem is the exact solution of the extreme value searching problem.

Consequently, the goal is to develop such an inventory modelling method which eliminates the above described problems when calculating the actual values of the control parameters.

### ADAPTIVE DYNAMIC INVENTORY MANAGEMENT

Surveys about practical inventory management show that there is demand for the development of an inventory management system comprising both the above mentioned inventory mechanisms and the methods modelling their operation (Ten Hompel and Schmidt 2002). A system like this could be used to the backing of the decision preparation process required to the inventory planning and the automation of the adaptive, dynamic inventory management. The main units and unit parts of the system are as follows:

#### Input unit:

- Query database (to store the data created during the operation of the real inventory system);
- Data conversion system (to create the data groups required to the simulation of the processes taking place in the inventory system);
- Input database of simulation (to store the data groups created by the data conversion system);

#### Core unit:

- The simulator of the inventory mechanisms (to simulate the operation of the inventory system considering the specified inventory mechanism);
- Output database of simulation (to store the data groups created during the simulation);
- Comparative system (to compare the data created by the simulation system model and the real system, i. e. verification of the inventory simulator);

#### Output unit:

- Optimizer (to specify the actual values of the control parameters of the applied inventory mechanism);
- Database of optimum solutions (to store the output data groups of the optimisation);

- Optimizer adjuster (to set the optimum values of the optimisation parameters).

The above described control system is able to use adaptively and dynamically the data created during the operation of the stockpiling system for the creation of a data structure modelling the real system. A further advantage is that, simulation techniques and special optimisation procedures can be integrated in the system by which the actual value settings of the parameters controlling the system can be calculated.

Nowadays, various enterprise resource planning systems (ERP) are yet able to log continuously the transactions in the inventory system. The data set needed to the estimations are mostly available or they can easily be created from the stored data (query database). A more problematic issue is to evaluate the possible fields of application and the most rewarding ways of application, i. e. how to extract the data groups important for the inventory planning from these logged data (input and output databases of the simulation) and how to determine the actual values of the control parameters applied in the inventory management system in the simplest, most proper and most dynamic way (database of optimum solutions). The simplified process of the operation of an adaptive, dynamic inventory management system is shown in Figure 1.

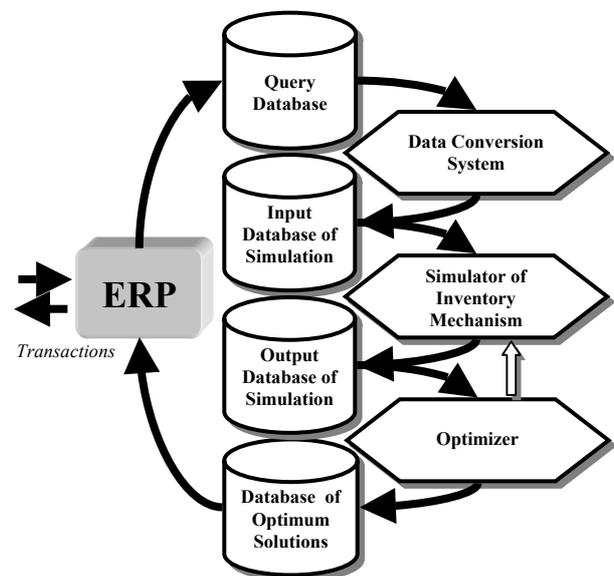


Figure 1. Process of the dynamic inventory management

The simulator of the inventory mechanisms and optimizer are in tight connection with each other, as one part of the data required to the optimisation is provided by the simulator of the inventory mechanisms through the output database of the simulation.

Hereafter, the presentation of the operation of the simulator of the inventory mechanisms will be explained and a simulator of the inventory mechanisms

and genetic optimizer developed by the author using MS Excel and Visual Basic will be presented.

## SIMULATION OF INVENTORY PROCESSES

The job of the simulator of the inventory mechanisms is the simulation of the operation of the inventory system by the specified inventory mechanism using the data contained in the input database of simulation. In case of the application of classical inventory mechanisms this can involve the following strategies (Chikán 1983):

- [t;q] – placing fixed orders (q) in fixed intervals (t);
- [t;S] – placing orders in fixed intervals (t) and ordering such a quantity, which - when received - completes the stock level to a previously specified maximum level (S);
- [s;q] – the order should be placed, when the stock level falls below a specified minimum (s) and the quantity to be ordered is fixed (q);
- [s;S] – the order should be placed, when the stock level falls below a specified minimum (s) and the order specifies a volume, which - when received - completes the stock level to a previously specified maximum level (S).

It is obvious for every classical mechanism, the optimum values of two control parameters (hereinafter A and B) should be found. In fact, the simulator of the inventory mechanisms is controlled by the optimizer. In every iteration step the optimizer runs the simulator of the inventory mechanisms by setting the actual control parameter. After having run the simulation, the output results should be stored in the output database of the simulation. The optimizer can reach all the required data from this database any time.

The *input database of simulation* should contain the following important data by SKU:

- the statistical parameters describes the elementary processes taking place in the real system (e. g. type of distributions and their parameters describing the demand and supply processes);
- specific cost parameters concerning the operation of the real system (e. g. specific warehouse unit cost, ordering costs by commodities);
- reliability parameters concerning the operation of the real system (required probability levels of the satisfaction of demands);
- parameters characterising the analysed period (e. g. time sections, length of the time sections);
- other auxiliary parameters (e. g. opening stock level, simulation constants).

When initiating the simulation running, the discrete event simulator coded in the simulator of the inventory mechanisms starts to function. The simulator generates the vectors indicated in Table 1. considering the control parameters of the mechanism applied and the

parameters characterising the analysed SKUs for the examination period (T):

Table 1. The simulation vectors

Process element	Interval vector	Quantity vector
Customers' orders	$\underline{t}_{VR}$	$\underline{q}_{VR}$
Deliveries	$\underline{t}_{ki}$	$\underline{q}_{ki}$
Warehouse orders	$\underline{t}_{RR}$	$\underline{q}_{RR}$
Supply processes	$\underline{t}_p$	–
Intakes	$\underline{t}_{be}$	$\underline{q}_{be}$

Using the elements of the vectors indicated in Table 1. the following output parameters for every time slot of the examination period (i) can be determined:

- opening stock level –  $Q_{ny}^{(i)}$ ;
- quantities demanded by the customers –  $q_{VR}^{(i)}$ ;
- quantity delivered –  $q_{ki}^{(i)}$ ;
- unsatisfied demand –  $q_{st}^{(i)}$ ;
- quantity ordered by the warehouse –  $q_{RR}^{(i)}$ ;
- received quantity –  $q_{be}^{(i)}$ ;
- stock in transit –  $Q_{ut}^{(i)}$ ;
- closing stock level –  $Q_z^{(i)}$ ;
- operation related total cost –  $K^{(i)}$ .

These output parameters should be stored in the *output database of simulation*. The above indicated parameters enable the calculation of all important parameters (e. g. costs, reliability) concerning the simulated operation of the inventory system for the examination period.

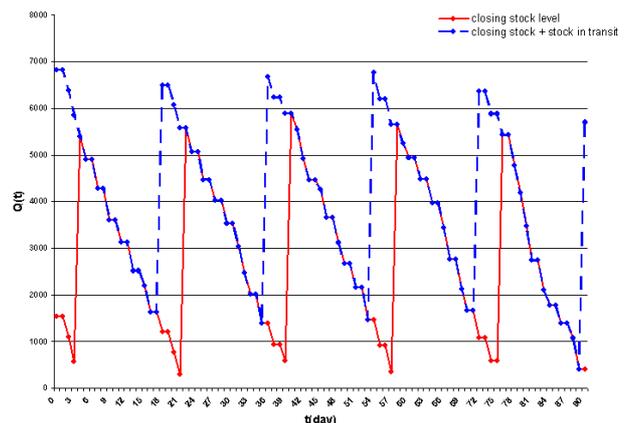


Figure 2. Operation of the [t;q] mechanism (stock levels)

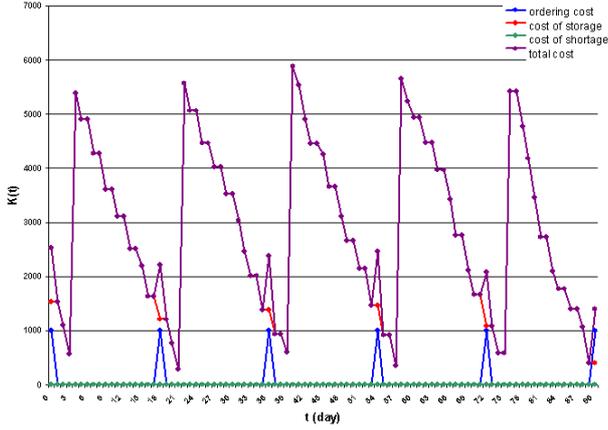


Figure 3. Operation of the [t;q] mechanism (costs)

As show on Figures 2. and 3. the simulated operation of a classical [t;q] mechanism. The graphs indicate clearly the quantitative processes taking place in the inventory system and the time function of the characteristic costs. The most important relations between the above parameters are shown below:

$$Q_{ny}^{(i)} = Q_z^{(i-1)}, \quad (1)$$

$$Q_z^{(i)} = Q_{ny}^{(i)} + q_{be}^{(i)} - q_{ki}^{(i)} - q_{st}^{(i)}, \text{ where} \quad (2)$$

$$q_{be}^{(i)} = q_{RR}^{(i-t_p)}. \quad (3)$$

$$\text{If } Q_{ny}^{(i)} + q_{be}^{(i)} \geq q_{VR}^{(i)}, \quad (4)$$

$$\text{then } q_{ki}^{(i)} = q_{VR}^{(i)} \text{ and } q_{st}^{(i)} = 0.$$

$$\text{If } Q_{ny}^{(i)} + q_{be}^{(i)} < q_{VR}^{(i)} \text{ and } Q_{ny}^{(i)} + q_{be}^{(i)} > 0 \quad (5)$$

$$\text{then } q_{ki}^{(i)} = Q_{ny}^{(i)} + q_{be}^{(i)} \text{ and } q_{st}^{(i)} = q_{VR}^{(i)} - q_{ki}^{(i)}.$$

$$\text{If } Q_{ny}^{(i)} + q_{be}^{(i)} < q_{VR}^{(i)} \text{ and } Q_{ny}^{(i)} + q_{be}^{(i)} < 0 \quad (6)$$

$$\text{then } q_{ki}^{(i)} = 0 \text{ and } q_{st}^{(i)} = q_{VR}^{(i)}.$$

The size of the travelling stock ( $Q_{ut}^{(i)}$ ) depends always on the date(s) of orders placed before the  $i$ th time element, the quantity (quantities) ordered and the expected date of intake(s).

## OPTIMISATION OF INVENTORY PROCESSES

In case of SKU based inventory process optimisation the conditions of the optimisation specified earlier should be met, namely:

- the total cost relating to the inventory process for the given examination period should be minimized and/or
- the probability of the satisfaction of the demands entering the system should be maximised.

Of course, the efficiency of the inventory system depends greatly on the mechanism chosen and on the

actual values of the control parameters. The basic condition of the process optimisation is the existence of a *target function* by which the above set of (contradicting) criteria can be managed and the optimal settings of the control parameters of the mechanism chosen can be found.

The determination of the inventory related costs is not problematic, as with the application of the specific costs of the process related total cost can be calculated. The two basic specific cost parameters of the model presented are as follows:

- ordering cost ( $k_f = \text{HUF/order}$ ), and
- specific warehousing unit cost ( $k_r = \text{HUF/pcs*day}$ ).

One possible solution for the evaluation of the reliability is a target function being a cost function in which even the reliability of the system is expressed in the form of cost. By means of such a target function both optimisation aspects could be handled on cost basis. A solution for this purpose is the specific deficit cost ( $k_f = \text{HUF/pcs*day}$ ), indicating the losses arising when the system is unable to satisfy the customers' demands. The analyses confirmed unambiguously the points summarised in Table 2. which can be explained by the learning ability of the system.

Table 2. Correlation between the specific deficit cost and the system's reliability

$k_f$ and $k_r$ relation	P(deficit)	Reliability
$k_f \gg k_r$	Small	Big
$k_f \approx k_r$	Medium	Medium
$k_f \ll k_r$	Big	Small

where P(deficit) = the probability of the deficit

The effective reliability of the system can be calculated after the simulation runs and compared with the required reliability. The exact value of the specific costs depends always on the inventory system under review and the commodity. The *target function* (7) can be written in the form:

$$\sum_{i=1}^T K^{(i)} = \sum_{i=1}^T K_t^{(i)} + \sum_{i=1}^T K_r^{(i)} + \sum_{i=1}^T K_f^{(i)} \Rightarrow \text{MIN!}, \quad (7)$$

$$\text{where } \sum_{i=1}^T K_t^{(i)} = \sum_{i=1}^T q_{RR}^{(i)} \cdot k_t, \quad \text{furthermore}$$

$$\sum_{i=1}^T K_r^{(i)} = \sum_{i=1}^T Q_z^{(i)} \cdot \frac{\text{sgn}(Q_z^{(i)}) + 1}{2} \cdot k_r, \text{ and}$$

$$\sum_{i=1}^T K_f^{(i)} = \sum_{i=1}^T Q_z^{(i)} \cdot \frac{1 - \text{sgn}(Q_z^{(i)})}{2} \cdot k_f.$$

## OPTIMISATION WITH A GENETIC ALGORITHMS

The optimum values of the control parameters of the inventory mechanism chosen will be determined by a binary genetic optimizer. The two parameters (A and B) will be coded by the algorithm in binary form, handled in binary form during the running of the genetic algorithm, then the optimised parameters ( $A_o$ ,  $B_o$ ) will be decoded. The operation of the genetic algorithm is shown in Figure 4. (Goldberg 1997).

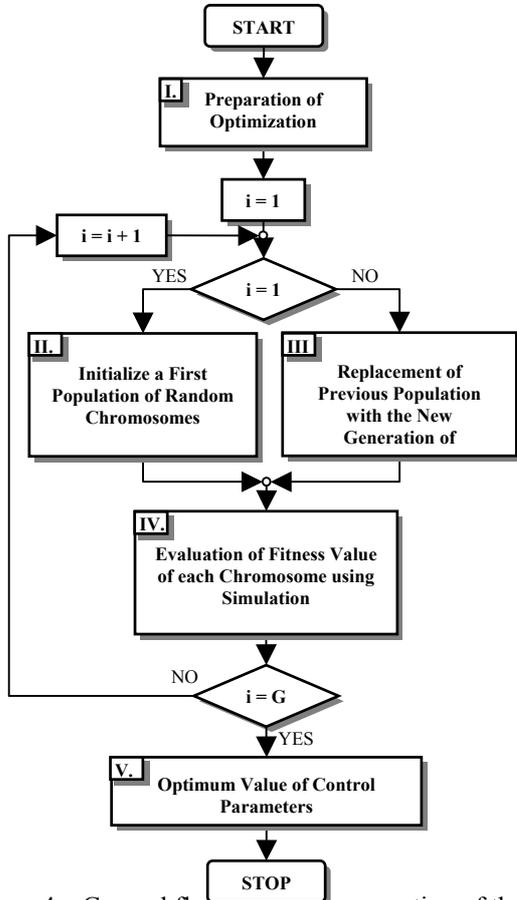


Figure 4. General flowchart of the operation of the binary genetic optimizer

The main phases of the preparation (I.) of the optimisation:

- selecting the inventory mechanism to be optimised;
- setting the upper and lower limits of the parameters to be optimised ( $A_{min}$ ;  $A_{max}$ ), ( $B_{min}$ ;  $B_{max}$ );
- setting the number of the simulation running cycles (N);
- setting the number of the generations (G);
- setting the number of individuals in the generation ( $E_G$ );
- setting the number of the offspring in the generation ( $U_G$ );
- setting the crossover probability ( $p_k$ );
- setting the mutation probability ( $p_m$ );
- setting the power of the selection pressure (k).

In case of random initialisation of the chromosomes (II.), a specified ( $m^{th}$ ) individual of the generation will be created in such a manner that the control parameter "A" is coded in binary form at the upper  $A^{bit}$  bites,

while the control parameter "B" at the lower  $B^{bit}$  bits. The coding will be carried out by the following scaling formulae (8) and (9) (Man and Tang and Kwong 1999):

$$E_m = [110...101][001...110] \text{ considering that}$$

$$A_{min} \rightarrow A_{min}^{BIN} = [000...000] \quad A_{max} \rightarrow A_{max}^{BIN} = [111...111]$$

$$B_{min} \rightarrow B_{min}^{BIN} = [000...000] \quad B_{max} \rightarrow B_{max}^{BIN} = [111...111]$$

$$\frac{A_m - A_{min}}{A_{max}^{BIN} - A_{min}^{BIN}} = \frac{A_{max} - A_m}{A_{max}^{BIN} - A_{min}^{BIN}} \quad (8)$$

$$\frac{B_m - B_{min}}{B_{max}^{BIN} - B_{min}^{BIN}} = \frac{B_{max} - B_m}{B_{max}^{BIN} - B_{min}^{BIN}} \quad (9)$$

The connection between the simulator of the inventory mechanisms and the genetic optimizer will be established at the determination (IV.) of the fitness values belonging to the new individuals of the generation. The fitness values belonging to the actual individual will be calculated by the genetic algorithm pursuant to the mathematical expectation ( $M(\sum K)$ ) of the total cost determined by the simulator of the inventory mechanisms by "N" simulation runs. Before the simulation runs, the control parameters belonging to the examined individual should be decoded, as the simulator of the inventory mechanisms manages the control parameters in decimal system. After the calculation of the mathematical expectation of the total cost belonging to the  $m^{th}$  individual, the genetic optimizer computes the fitness values by the following formula (10) (Man and Tang and Kwong 1999):

$$F_m = M\left(\sum K\right)_{min} + \left(M\left(\sum K\right)_{max} - M\left(\sum K\right)_m\right)^k \quad (10)$$

The continuous refreshment (III.) of the chromosomes of the generation takes place by the roulette-wheel selection method following the sequencing of the individuals by fitness value, four-point crossover and mutation. Two parents will be selected randomly, and two offsprings are formed, if the relation  $\text{Random}(0;1) < p_k$  is met. The mutation of the offspring created by the crossover will take place, if  $\text{Random}(0;1) < p_m$ . This process ensures that the iteration results better and better solutions (combinations) and, simultaneously, the diversity of the combinations is maintained (to avoid that the algorithm sticks at a local optimum). After the creation of the appropriate number of generations (G) (V.) - i. e. the execution of the specified number of iteration steps - the algorithm selects the optimum setting values of the control parameters from the last generation by the target function and decodes them by the following formulae (11) and (12) :

$$A_o = A_{\min} + A_m^{\text{BIN}} \cdot \frac{(A_{\max} - A_{\min})}{2^{A^{\text{bit}}} - 1}, \text{ where} \quad (11)$$

$A_m^{\text{BIN}}$  is the binary conversion of parameter A.

$$B_o = B_{\min} + B_m^{\text{BIN}} \cdot \frac{(B_{\max} - B_{\min})}{2^{B^{\text{bit}}} - 1}, \text{ where} \quad (12)$$

$B_m^{\text{BIN}}$  is the binary conversion of parameter B.

## RESULTS

To test the [t;q] mechanism, 15 experiments were carried out. The binary genetic optimum searching algorithm was run with varying parameter settings (e.g. number of generations, number of entities or offsprings), but identical input data 100 times. According to the experiences, the search for the optimum parameter setting of the search algorithm is a quite time consuming process, it requires several testing operations and the statistical evaluation of the results. Based on the results of the evaluations, the optimum parameter settings for a given job can be approached by successive approximation. The optimum of the 15 experiments will be shown below.

*Parameter settings of the genetic algorithm:*

$G=100$ ;  $E_G=50$ ;  $U_G=10$ ;  $N=20$ ;  $p_k=85\%$ ;  $p_m=5\%$ ;

*Results (at a confidence level of 95%):*

12.37 days <  $t_o = 13.65$  days < 14.92 days  
 3935.28 pieces <  $q_o = 4313.12$  pieces < 4690.95 pieces  
 HUF 428340.6 <  $K_o =$  HUF 455023.6 < HUF 481706.5

The second important parameter – the runtime of the optimisation – is depending largely on the capabilities of the computer used. A test carried out on a computer of 1.33 GHz and 128 MB SDRAM, resulted in 69.24 sec <  $T_{\text{opt}} = 69.99$  sec < 70.73 sec. The distribution of the results and the runtime is shown in Figures 5., 6. and 7.

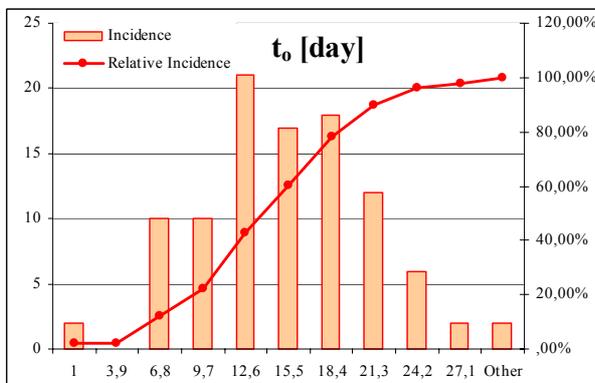


Figure 5. Distribution of the ordering interval

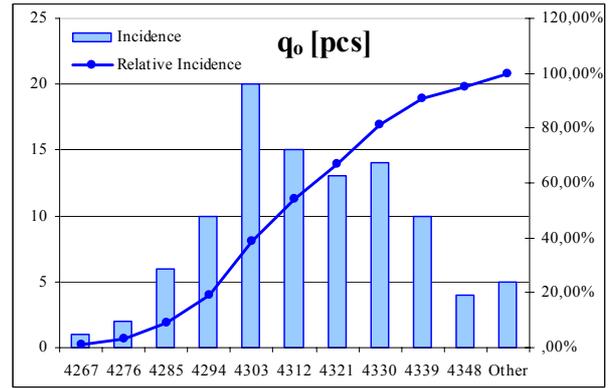


Figure 6. Distribution of the ordered volume

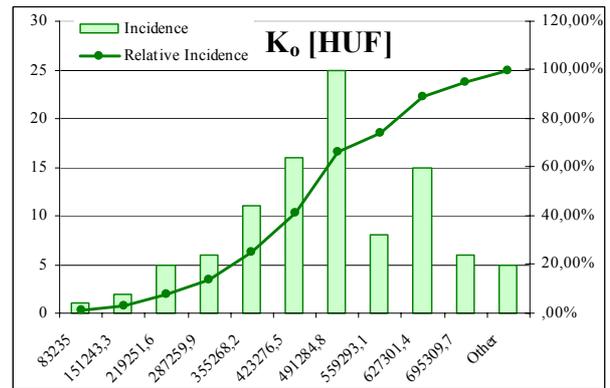


Figure 7. Distribution of the costs

## SUMMARY

The simulation inventory model presented in this paper and the binary genetic optimising algorithm determining the control parameters showed beneficial properties in managing of stochastic inventory processes. For the establishment of proper applications, it is worthwhile to examine also the services rendered by the genetic algorithms operating with real number representation, as it is possible that this type of algorithm is able to provide the same results in a faster, more accurate way. Experiences show that the inventory processes in the future may constitute a special application field of the simulation supported optimisation with genetic algorithms.

## LITERATURE

- Álmos A.; Györi S.; Horváth G.; Várkonyiné K. A.: *Genetikus algoritmusok*. Typotex Kiadó. Budapest, 2002.
- Chikán A.: *Készletezési modellek*. Közgazdasági és Jogi Könyvkiadó. Budapest, 1983.
- Goldberg, D.E. 1997. *Genetic algorithms in search, optimization & machine learning*. Addison Wesley.
- Man, K.F.; Tang, K.S.; Kwong, S. 1999. *Genetic Algorithms, Concepts and Design*. Springer Verlag, Berlin.
- Ten Hompel, M.; Schmidt, Th.: *Warehouse management*. Springer Verlag. Berlin, 2003.

# NUMERICAL SIMULATION OF AMPEROMETRIC BIOSENSORS PERFORMANCES

A. BENYAHIA, and S. BACHA  
Laboratoire de génie de l'environnement  
Université Abderrahmane Mira de Béjaia.  
Route de Targa Ouzemmour 06000 Béjaia, Algeria.  
E-mail: hal\_ben@yahoo.fr

## KEYWORDS:

amperometric biosensor, homogeneous enzyme kinetic, enzyme membrane, diffusion, modelling, simulation.

## ABSTRACT:

A mathematical model is used to analyse the transient response of amperometric biosensors has been developed. The model is based on non stationary diffusion equations containing a non linear term related to Michaelis-Menten kinetics of the enzymatic reaction. The numerical simulation is performed with finite volume method and results are compared with known analytic solutions obtained for some restricted parameters value's at steady state.

Numerical results have shown that best performances are obtained for biosensor when the biosensor operate under internal diffusion control and this may be obtained either by a high loading of the membrane with enzyme and/or by ensuring a good stirring of the solution.

## INTRODUCTION

Biosensors are defined as analytical devices incorporating a physicochemical transducer or transducing system and biologically active material integrated with it. Biosensors yield a signal witch is proportional to the concentration of a measured analyte or group of analytes.

Amperometric biosensor is one of the most types of biosensors witch typically rely on an enzyme system that catalytically converts electrochemically non-active analyte into products that can be oxidized or reduced at a working electrode. This is electrode is maintained at an appropriate potential with respect to a reference electrode.

Modelling of biosensors is of a crucial importance to understand their behaviour. Biosensor modelling have began with the work of (Gough et al. 1985) which have simulated the performance of a cylindrical biosensor

for glucose monitoring at steady state. An other mathematical model have been used for the description of steady state and transient behaviour of a multi-membrane multi-enzyme amperometric biosensor (Jobst et al. 1996), a finite difference scheme was used for the discretization of the model equation. (Caras et al.1985a,b and c) have developed a model to simulate the behaviour of a potentiometric biosensor where only a membrane enzyme layer zone is considered, a method of line was used to solve numerically the model. (Bacha et al. 1995 and 1996) have developed a model that take into account a variety of configuration designs to describe the behaviour of amperometric biosensor for glucose monitoring. A finite volume method is used to solve the model.

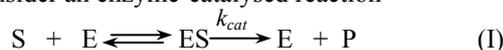
Recently (Baronas et al 2003a) have developed a mathematical model to examine the dynamic response of amperometric biosensors as well in stirred solutions as in non-stirred ones. The authors in another paper examine the influence of the thickness of the membrane on the response of amperometric biosensors (Baronas et al 2003b). The authors use a finite difference scheme to solve the model equations. The same authors (Baronas et al 2002) have already used slightly the same model to study the response of biosensors to a mixture of compounds.

In the present work, we have developed a mathematical model in order to describe and evaluate the performance of amperometric biosensors. The chosen configuration is the most used in the design of nowadays enzymatic biosensor realizations such as the use of polymeric matrices as an enzyme support and the mass production of biosensors by the screen printing technique.

## DESCRIPTION OF THE MODEL

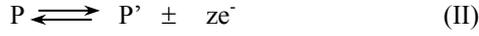
The sensors considered in this study consist of a metallic electrode surface, a membrane hold against the surface electrode. According to operating conditions, a diffusion layer adjacent to the outer surface of the membrane is considered.

Consider an enzyme-catalysed reaction



In this scheme the substrate (S) binds to the enzyme (E) and converts to the product (P).

P is electroactive; it can be oxidized or reduced - whether it is considered as a reductor or an oxidant respectively- electrochemically at a specified potential.



Let us assume the symmetrical geometry of the electrode and homogeneous distribution of immobilised enzyme in the membrane. Coupling the enzyme catalysed reaction in enzyme layer with the one-dimensional-in-space diffusion, described by Fick's second law, leads to the following equation:

$$\frac{\partial[C]}{\partial t} = D_c \frac{\partial^2[C]}{\partial x^2} + R_c \quad (1)$$

Where [C] is the concentration of any species involved in both enzymatic and electrochemical process,  $D_c$  its diffusion constant and  $R_c$  is a term related to its production or consumption by the enzyme kinetics.

$$R_c = \pm \frac{v_{\max}^v [C]}{k_s + [C]} \quad (2)$$

$v_{\max}^v$  is the maximal rate of enzymatic reaction which is obtained with a given amount of enzyme [Et]:

$$v_{\max} = k_{cat} [Et] \quad (3)$$

The current density is calculated by estimation the gradient of the electroactive species at the surface of electrode:

$$i(t) = zFD_p \left. \frac{\partial[P]}{\partial x} \right|_{x=0} \quad (4)$$

$z$  is the number of electrons involved in the electrochemical reaction,  $D_p$  is the diffusion constant of the electroactive species and  $F$  is the Faraday number ( $F=96485$  C/mole).

The biosensor configuration is given in figure 1 and In order to work with a minimal amount of parameters, mass balance equations of involved species are transformed to dimensionless ones where parameters of reference are the Michaelis-Menten constant  $k_s$ , the membrane thickness  $l_m$  and the substrate diffusion constant  $D_{sm}$ .

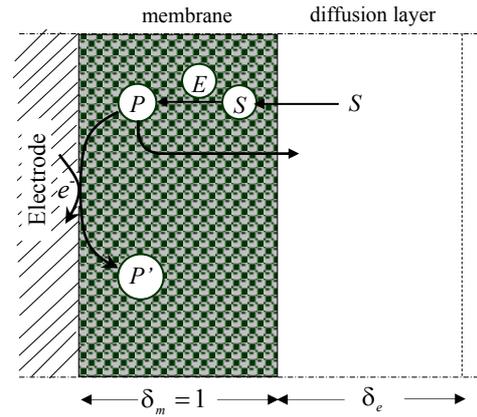


Figure 1: Scheme of simulated biosensor and build-up

The following table summarize the whole parameters used in this work.

Table 1: Dimension and Dimensionless Parameters

Parameter	Dimensional	Dimensionless
Time	$t$ (s)	$T = \frac{t D_{sm}}{l_m^2}$
Membrane thickness	$l_m$ (m)	$\delta_m = \frac{l_m}{l_m} = 1$
Diffusion layer thickness	$l_e$ (m)	$\delta_e = \frac{l_e}{l_m}$
Distance from electrode	$x$ (m)	$X = \frac{x}{l_m}$
Substrate concentration	$[S]$ (mol.m <sup>-3</sup> )	$S = \frac{[S]}{k_s}$
Product concentration	$[P]$ (mol.m <sup>-3</sup> )	$P = \frac{[P]}{k_s}$
Michaelis-Menten constant	$k_s$ (mol.m <sup>-3</sup> )	$K_S = \frac{k_s}{k_s} = 1$
Substrate diffusion coefficient		$\bar{D}_{sm} = 1$
1- Membrane 2- Diffusion layer	$D_{sm}$ (m <sup>2</sup> .s <sup>-1</sup> ) $D_{se}$ (m <sup>2</sup> .s <sup>-1</sup> )	$\bar{D}_{se} = \frac{D_{se}}{D_{sm}}$
Product diffusion coefficient		$\bar{D}_{pm} = \frac{D_{pm}}{D_{sm}}$
1- Membrane 2- Diffusion layer	$D_{pm}$ (m <sup>2</sup> .s <sup>-1</sup> ) $D_{pe}$ (m <sup>2</sup> .s <sup>-1</sup> )	$\bar{D}_{pe} = \frac{D_{pe}}{D_{sm}}$
Current density	$i$ (A.m <sup>-2</sup> )	$\Psi = \frac{i l_m}{zF D_{sm} k_s}$

By converting equation (1) to dimensionless form, a dimensionless quantity arises in the source term which is similar to the Thiele Modulus. This parameter describes the relative importance of diffusion and reaction in the enzyme layer:

$$\Phi = l_m \sqrt{\frac{v_{\max}}{D_{sm} k_s}} \quad (5)$$

When  $\Phi$  is small, the kinetics are the predominant. In contrast, when the Thiele modulus is large, diffusion limitations are the principal determining factor.

Another dimensionless number, namely the Biot number is needed to express the ratio of the internal mass transfer resistance to the external one is given by:

$$Bi = \frac{l_m / D_{sm}}{l_e / D_{se}} = \frac{D_{se} l_m}{D_{sm} l_e} \quad (6)$$

### Transient Mass Balance Equations:

The governing equations of the model are written in the two regions which represent the physical domain of both biosensors, namely, the membrane and the diffusion layer in which mass transfer is occurring only by diffusion. In the bulk solution the transport is carried by convection and the concentration of all species is assumed to remain constant thanks to the stirring of the solution.

In the membrane, the model is described by the two following equations in which, three terms are present, the accumulation term, the diffusion term and the kinetic one.

$$\frac{\partial S}{\partial T} = \frac{\partial^2 S}{\partial X^2} - \Phi^2 \frac{S}{1+S} \quad (7)$$

$$\frac{\partial P}{\partial T} = \bar{D}_{pm} \frac{\partial^2 P}{\partial X^2} + \Phi^2 \frac{S}{1+S} \quad (8)$$

In the diffusion layer, equations are the same as in the membrane except there's no kinetic term within the diffusion layer:

$$\frac{\partial S}{\partial T} = \bar{D}_{se} \frac{\partial^2 S}{\partial X^2} \quad (9)$$

$$\frac{\partial P}{\partial T} = \bar{D}_{pe} \frac{\partial^2 P}{\partial X^2} \quad (10)$$

### Boundary Conditions

The substrate is not electroactive, its mass flux density is thus nil at the surface of electrode:

$$\left. \frac{\partial S}{\partial X} \right|_{X=0} = 0. \quad (11)$$

The electrochemical reaction is assumed to be fast enough to ensure a nil concentration at the electrode

$$P|_{X=0} = 0. \quad (12)$$

At the membrane-diffusion layer interface the mass flux density of both  $S$  and  $P$  must be continuous, this is provided by:

$$\left. \frac{\partial S}{\partial X} \right|_{X=1^-} = \bar{D}_{se} \left. \frac{\partial S}{\partial X} \right|_{X=1^+} \quad (13)$$

$$\bar{D}_{pm} \left. \frac{\partial P}{\partial X} \right|_{X=1^-} = \bar{D}_{pe} \left. \frac{\partial P}{\partial X} \right|_{X=1^+} \quad (14)$$

At the diffusion layer-bulk interface, during biosensor operation all concentrations are supposed to be constant.

$$S|_{X=1+\delta e} = S_0 \quad (15)$$

$$P|_{X=1+\delta e} = 0. \quad (16)$$

The dimensionless current density due to the mass flux of  $P$  at the electrode is provided by:

$$\Psi(T) = \bar{D}_{pm} \left. \frac{\partial P}{\partial X} \right|_{X=0} \quad (17)$$

### Description of the Numerical Method:

The previous equations governing the dynamic of the biosensors considered as well as boundary conditions are solved by the finite volume method (FVM) which is a simple variant of the well known finite element method (FEM). The FVM use the same weighting function in the entire domain.

$$W(X) = \begin{cases} 1 & X \in [X_w, X_E] \\ 0 & \text{otherwise} \end{cases}$$

The entire physical domain  $\Omega$  (membrane + diffusion layer) which is divided into subdomains  $\Omega_i$  ( $\Delta X_i$ ) in which mass balance equations are integrated assuming a piecewise profile of the dependent variable  $S$  and  $P$ .

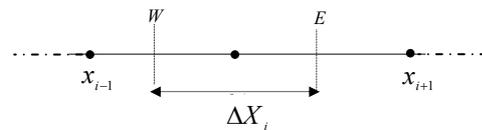


Figure 2: One dimension finite volume discretization scheme.

A great advantage of the finite volume integration is that there's no problem if the space grid is enlarged or reduced in any part of the integrated domain. For example, applying this basis to the governing equation relative to the substrate, this leads to:

$$\int_W^E \frac{\partial S}{\partial T} dX = \int_W^E \frac{\partial^2 S}{\partial X^2} dX - \int_W^E \Phi^2 \frac{S}{1+S} dX$$

After linearization of the non-linear term related to the enzyme kinetic and integration of each term we obtain a system of algebraic equations that has the following form:

$$a_i^n S_{i-1}^{n+1} + b_i^n S_i^{n+1} + c_i^n S_{i+1}^{n+1} = d_i^n$$

Where  $a_i^n, b_i^n, c_i^n$  and  $d_i^n$  are constant that depend on previous calculated values of  $S$ .

$S_{i-1}^{n+1}, S_i^{n+1}$  and  $S_{i+1}^{n+1}$  are the current value of  $S$  at points  $\sum_{k=1}^{i-1} k\Delta X_k, \sum_{k=1}^i k\Delta X_k$  and  $\sum_{k=1}^{i+1} k\Delta X_k$  at time  $(n+1)\Delta t$ .

Applying the same treatment to the product mass balance equation we obtain finally two systems of algebraic equations to solve simultaneously:

$$\begin{aligned} A_s^{(n)} * S^{n+1} &= D_s^{(n)} \\ A_p^{(n)} * P^{n+1} &= D_p^{(n)} \end{aligned}$$

$A_s^{(n)}$  and  $A_p^{(n)}$  are tridiagonal matrixes of  $a_i^n, b_i^n$  and  $c_i^n$  elements respectively for  $S$  and  $P$ .

$D_s^{(n)}$  and  $D_p^{(n)}$  are second member vectors respectively for  $S$  and  $P$ .

The two systems of algebraic equations are solved numerically by Thomas algorithm. The algorithm is written in Visual Fortran and run on a Pentium III PC (733 kHz) processor.

### General Considerations:

One of the most problems encountered in biosensor modelling is the lack of information about numerical value of each parameter and the scope of their variation. Many parameters are not accessible or not known with sufficient accuracy. Explicitly; independent experiments must be carried out before simulation. Nevertheless, this can be avoided if one studies the influence of a group of parameters on biosensor's output in stead of examining the influence of each parameter taken lonely. It is still possible when introducing dimensionless number such as Thiele modulus and Biot number.

### RESULTS AND DISCUSSION:

The model presented below is applied to simulate the response of the biosensors versus many parameters such as the amount of the immobilized enzyme, the membrane thickness and the thickness of the diffusion layer witch is estimated by the Levich relation for the rotating electrode:

$$l_e = 1.61 D^{1/3} \nu^{1/6} \omega^{-1/2}$$

$D$ : diffusion coefficient (m<sup>2</sup>/s).

$\nu$ : viscosity of the solution equal to  $1.02 \cdot 10^{-6}$  m<sup>2</sup>/s.

$\omega$ : Speed of rotation of the electrode (rad/s).

$l_e$  is inversely proportional to the stirring strength.

### Influence of the Internal Resistance on the Signal Magnitude:

In Order to study the effect of the internal resistance, a diffusion layer thickness is taken deliberately to equal to 0. This leads to a value of  $Bi = \infty$ . The following parameters are taken for all the simulation experiment (Bacha et al. 1996) and (Baronas et al 2003a):

$$\begin{aligned} l_m &= 50 \mu\text{m}, D_{sm} = D_{pm} = 3.0 \times 10^{-10} \text{ m}^2/\text{s}. \\ D_{se} &= D_{pe} = 2 \times D_{sm}. \\ ks &= 0.1 \text{ mol/m}^3, [S_0] = 0.1 \text{ mol/m}^3. \end{aligned}$$

Value of  $\nu_{\max}$  are taken in such a manner to have a decades of  $\Phi$  ranging from 0.1 to 7.

Equations (7)-(10) and the corresponding boundary conditions (11)-(16) are solved to simulate the behaviour of the biosensor. Initial conditions are taken according to following assay protocol:

$$\begin{aligned} \text{At } t=0 \quad S &= 0 \text{ for } 0 \leq X < 1 \\ S &= S_0 \quad X \geq 1 \\ P &= 0 \quad \forall X \end{aligned}$$

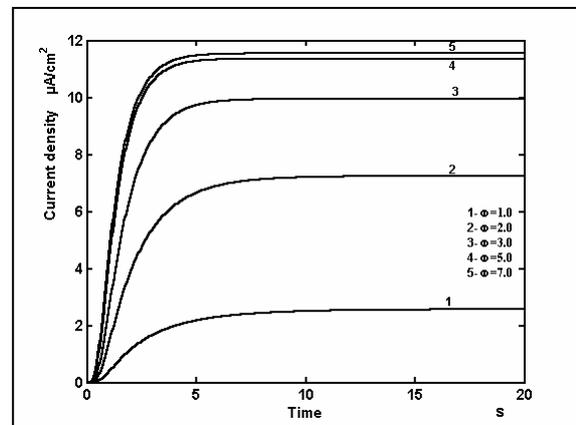


Figure 3a : The dynamics of the biosensor current  $i$  at va Thiele modulus values ranging from 1 to 7.0  
Other parameters:  $l_m=50\mu\text{m}, [S_0]=ks=0.1 \text{ mol/m}^3$  and  $Bi=\infty$

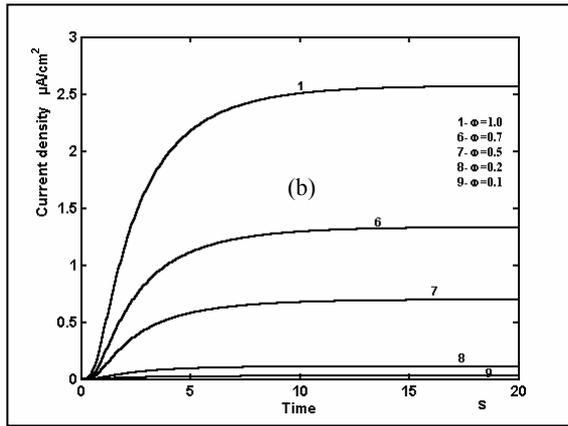


Figure 3b: The dynamics of the biosensor current  $i$  at various Thiele modulus values ranging from 0.1 to 1. Other parameters:  $l_m=50\mu\text{m}$ ,  $[S_0]=k_s=0.1\text{ mol/m}^3$  and  $Bi=\infty$

Figures 3a and 3b show the response to the biosensor for a wide range of  $\Phi$  value. One can see that the maximal magnitude of the biosensor which corresponds to the steady state current increases with the increase of the amount of enzyme since  $v_{\text{max}}$  is directly related to the enzyme concentration. All current's curve have an S shape and the response reaches 99,99 % (which is considered the response time) of the steady state value as faster as the value of  $\Phi$  is important.

Response time is less than 5 seconds for a  $\Phi$  value of 5.0 or beyond (i.e. for a response controlled by diffusion).

Another interesting feature is that the maximum current density does not exceed a limit of  $11.56\ \mu\text{A}/\text{cm}^2$  regardless the amount of immobilized enzyme. One can take advantage of that remembering the deactivation of the enzyme during the storing of the biosensor and this may keep the same magnitude as long as the response is controlled by diffusion. Figure 4 illustrate this well, as we can see the maximum

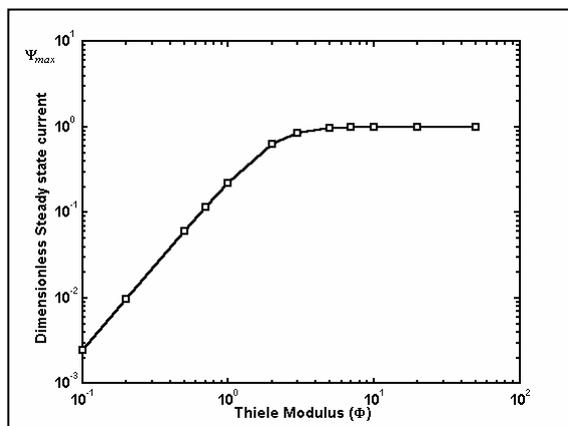


Figure 4: Dimensionless maximal current density versus Thiele modulus.  $l_m=50\mu\text{m}$ ,  $[S_0]=k_s=0.1\text{ mol/m}^3$  and  $Bi=\infty$

dimensionless current density  $\Psi_{\text{max}}$  versus Thiele modulus is linear up to a value of  $\Phi=2$ . No further increase of  $\Psi_{\text{max}}$  is obtained beyond this value whatever is the Thiele's modulus value.

### Effect of External Resistance to Mass Transfer

To study the effect of external resistance to mass transfer, values of  $\Phi=7$  and  $\Phi=1$  are taken to perform the simulation. The diffusion layer thickness  $\delta_e$  is chosen to ensure value of  $Bi$  from 0.5 to 50 (i.e. for diffusion layer thickness  $\delta_e$  varying from 200 to  $2\mu\text{m}$ ). Figure 5a show the biosensor response obtained with a value of  $\Phi=7$  which correspond to diffusion control. The magnitude of the response increase with the decrease of the diffusion layer thickness. One can see response 7 and 8 in figure 5a obtained respectively for  $Bi=1$  and  $Bi=0.5$  show a peak of  $6.6\ \mu\text{A}/\text{cm}^2$  approximately at  $t=4\text{s}$ , this may be attributed to the excessive thickness of the diffusion layer ( $100\mu\text{m}$  for  $Bi=1$  and  $200\mu\text{m}$  for  $Bi=0.5$ ). This layer was

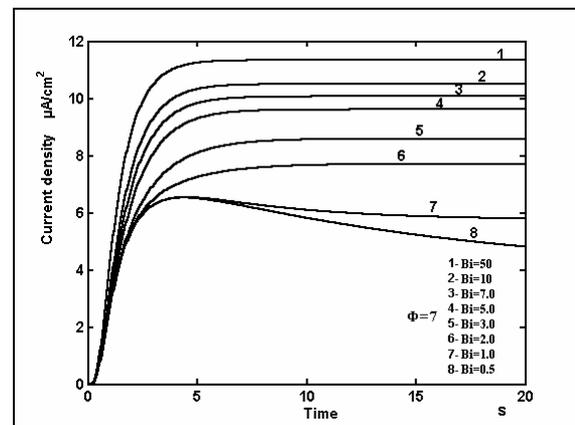


Figure 5a: The dynamics of the biosensor current versus Biot number at  $\Phi=7$ .

Other parameters:  $l_m=50\mu\text{m}$ ,  $[S_0]=k_s=0.1\text{ mol/m}^3$

initially loaded with the substrate, which induced fast reaction in the membrane, when most of substrate had been consumed, diffusion from the bulk of the solution was too slow to maintain the high initial reaction rate. The maximum current density at  $Bi$  value of 50 is (well stirred solutions) 1.8 times its value at  $Bi$  value of 1 (weak stirred ones).

Responses obtained when reaction is the limiting factor (i.e. for  $\Phi=1$ ) in figure 5b does not show any peak. This is well understood because the diffusion rate of the substrate in the diffusion layer has at some extent the same order of magnitude as its consumption in the membrane.

The obtained current densities are ranging from 2.5 to  $3\ \mu\text{A}/\text{cm}^2$  and do not exceed the later value despite the wide scope of  $Bi$  value. In such situation stirring would decrease the maximal current density (curve n°4 Figure 5b). Biosensor's responses obtained are slower than those

obtained previously with  $\Phi=7.0$  and the strength of stirring seems not to have commonly a drastic influence on response magnitudes.

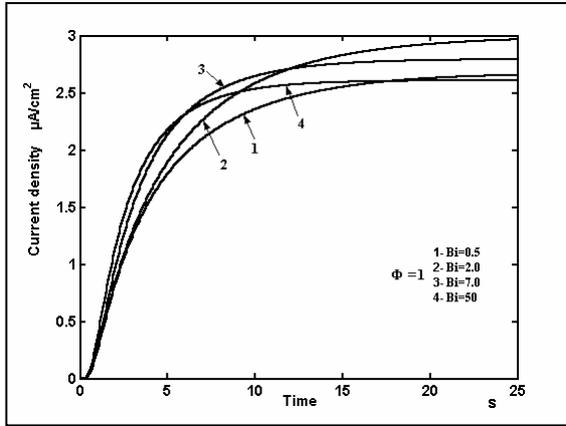


Figure 5b: The dynamics of the biosensor current versus Biot number at  $\Phi=1$ .  
Other parameters:  $l_m=50\mu\text{m}$ ,  $[S_0]=ks=0.1\text{ mol/m}^3$

### Effect of Internal Resistance on the Linearity of the Biosensor Response

As we have seen in the previous section, best and fast responses are obtained for biosensor operating under diffusion control, investigations are made to check the linearity of the biosensor response versus the amount of enzyme loaded in the membrane. Figure 6 show that the widest interval of concentration in which the relationship with the steady state current density is linear.

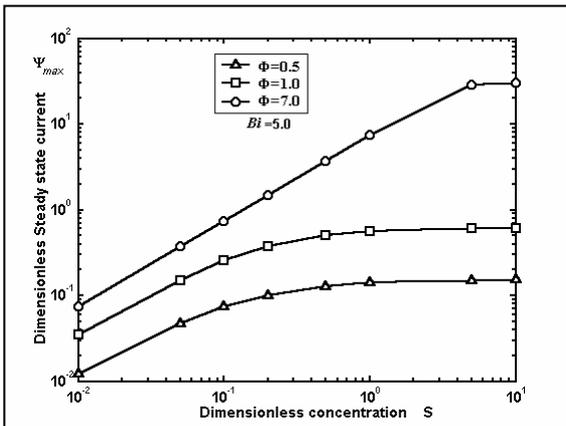


Figure 6: Steady state current density versus dimensionless concentration.

### MODEL VALIDATION

The model is validated by comparing numerical solutions obtained at steady state to analytical ones obtained for some limited cases. For example the substrate masse balance equation (7) may be approximated to the following equations:

$$0 = \frac{d^2 S}{dX^2} - \Phi^2 S \quad (a)$$

Or

$$0 = \frac{d^2 S}{dX^2} - \Phi^2 \quad (b)$$

According to whether we assume  $S \gg 1$  (equation a) or  $S \ll 1$  (equation b). Equations a and b are linear ordinary second order differential equations which can be integrated analytically without any problem and provide as well the steady state current density as the steady state substrate or product profiles.

Figure 7 show the comparison between steady state current density obtained by both finite volume method and finite difference method (FD) in one hand to the analytical solution obtained at steady state (equation b) in the other hand in the case  $S \ll 1$ . Comparison shows that the relative error in percent does not reach 0.1% when the number of space points  $I_{max}$  exceed 51. This demonstrate the power of the FVM for a relatively little effort of calculation.

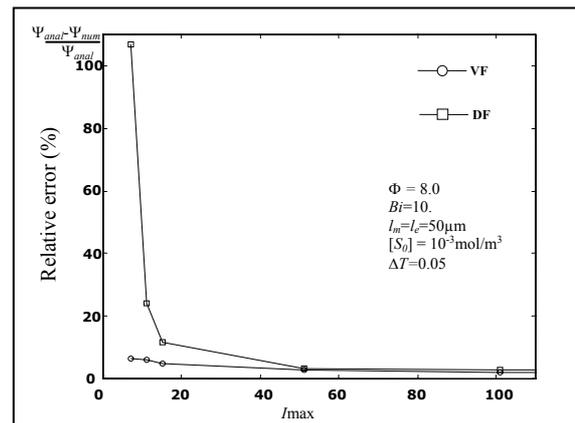


Figure 7: Relative error evolution versus the maximal number of space points.

The comparison of both substrate and product numerical profiles at steady state and analytical ones show that the numerical method (the FVM) is in good agreement all above in the region located near the electrode (see product profile).

### CONCLUSION

The described model has shown interesting features in investigating amperometric biosensor performance. It is useful to explain the sensing mechanism and provide a basis for interpretation of the response in order to incorporate further improvements in amperometric biosensor design.

Numerical results demonstrate that best performances are obtained when the biosensor operate under internal

diffusion control. High loading membrane with enzyme will provide, in addition to a maximal and fast signal for a given substrate concentration, a wide range of linearity and may keep somewhat a regular biosensor signal during the storing of the biosensor as long as the biosensor is under diffusion control.

## REFERENCES

Bacha S., A. Bergel and M. Comtat. 1995. "Transient response of multilayer electroenzymatic biosensors". *Anal. Chem.* 67, pp 1669-1678.

Bacha S., M. Montagné and A. Bergel. October 1996. "Modelling Mass transfer with Enzymatic Reaction in Electrochemical Multilayer Microreactors". *AICHE journal*. Vol. 42, No.10, pp 2967-2976.

Baronas R., E. Ivanauskas and J. Kulys. 2002. "Computer simulation of the response to Mixtures of compounds". *Nonlinear analysis: Modelling and control*, Vol 7, N°2, pp 3-14.

Baronas R., E. Ivanauskas and J. Kulys. 2003. "Computer simulation of the response of amperometric biosensors in stirred and non stirred solution". *Nonlinear analysis: Modelling and control*, Vol 8, N°1, pp 3-18.

Baronas R., F. Ivanauskas and J. Kulys. 2003. "The influence of the enzyme membrane thickness on the responses of amperometric biosensors", *Sensors* 2003, 3, pp 248-262.

Caras S.D., J. Janata, D. Saupe and K. Schmit. 1985. "pH based enzyme potentiometric sensors". Part 1. "Theory". *Anal. Chem.* 57, pp 1917-1920.

Caras, S.D., D. Petelenz and J. Janata. 1985. " pH based enzyme potentiometric sensors". Part 2. "glucose- sensitive field effect biosensor". *Anal. Chem.* 57, pp 1920-1923.

Caras, S.D. and J. Janata. 1985. "pH based enzyme potentiometric sensors". Part 3. "Penicillin-sensitive field effect biosensor". *Anal. Chem.* 57, pp 1924-1925.

Gough, D.A., J.Y. Lucisano, and P. H. S. Tse. 1985. "Two-Dimensional Enzyme Electrode Sensor for Glucose". *American chemical society*, pp 2351-2357.

Jobst G., I. Moser and G. Urban. 1996. "Numerical simulation of multi-layered enzymatic sensors". *Biosensors & Bioelectronics*. Vol. 11. No 1/2, pp 111-117.

# THE EFFECT OF DIFFERENTIATION ON PRISON POPULATION: A SIMULATION STUDY OF THE SWEDISH PRISON SYSTEM

Fredrik Persson, Anna Palmerius, and Joakim Barkman  
Department of Production Economics  
Linköping Institute of Technology  
S-581 83, Linköping, Sweden  
E-mail: fredrik.persson@ipe.liu.se

## KEYWORDS

Prison, Discrete Event Simulation, Queuing systems.

## ABSTRACT

This paper describes the case of a simulation study of the Swedish Prison and Probation system regarded as a queuing system. The situation in Swedish prisons today is not acceptable with a utilisation over 100 per cent. The short-term solution of the problem has been to overcrowd the prisons. The purpose of this work is to describe the relations between waiting time in remand prison, the official number of cells in prisons and the degree of differentiation regarding to the expected torrent of criminals. Differentiation is the attempt to keep different categories of clients apart during their prison time. The degree of differentiation is defined as the number of defined categories used for client placement. The conclusions from this study is that there are not enough number of prison cell available. The situation for male inmates/clients is more critical than for female clients. The results from the study suggest that there needs to be approximately 11,000 prison cells in total to meet the unofficial recommendation of a maximum of seven days in remand prison.

## INTRODUCTION

The situation in Swedish prisons is strained. This has resulted in long waiting times for inmates (here called clients) before they can receive a suitable placement to serve their prison time. The queues in remand prison are not acceptable and the degree of occupancy has gone over 100 per cent according to the official statistics. The short-term solution of the problem has been to overcrowd the prisons. In the beginning of the year 2004 there was officially 4,571 number of prison cells to be shared between clients of both sexes. The target is to build 1,150 new places prior to 2007 which will result in a total capacity of 5,721 places. It is not clear that this expansion will result in an elimination of waiting time in remand prison, or even a reduction to the regulated maximum of seven days after that sentence have been made official.

The degree of differentiation is one of the regulations that makes this planning problem suitable for a more in-depth analysis. Differentiation is the attempt to keep different categories of clients apart during their prison time. The degree of differentiation is defined as the

number of defined categories used for client placement. The degree of differentiation have a direct influence on the waiting time in remand prison. With a high degree of differentiation (many different client categories) the chance of receiving a suitable placement decreases as the number of options for placements are few.

Tarling (1986) reviews the use of statistical analysis in criminology and describes a simulation model depicting the interrelation between the police, court, prison, and probation systems. With the proposed model it is possible to e.g. evaluate different strategies for trial priorities in order to reduce waiting time before the trial. Most statistical applications reported in Tarling (1986) covers the prediction of prison population. Barnett (1987) follows in the same field research with a projection model for future prison population based on historical data. Lattimore and Baker (1997) studies the effect of limited prison capacity on the average time a client serves. They use a input/output process model with feedback.

Other, similar applications of simulation has been reported in the healthcare sector. In Ridge et al. (1998) a simulation model (coded in Pascal) is used to capture the capacity need in an intensive care unit. The capacity planning in this environment is similar to that of the remand prison queuing system. The waiting is to be kept to a minimum by planning the capacity in the later parts of the flow.

The purpose of this work is to describe the relations between waiting time in remand prison, the official number of prison cells and the degree of differentiation regarding to the expected torrent of criminals. This analysis is done with use of discrete event simulation where the process of waiting to be placed, the placement and the serving time is considered as a queuing system. Each client will be provided with his/her own treatment program during the prison time. This treatment program is called a treatment chain since the client is routed through different instances of treatment in a specific order.

The methodology, using discrete event simulation, is a novel application for the National Prison and Probation Administration of Sweden. On hindsight, the most useful results came from the fact that a conceptual model was created that eventually all concerned personnel could agree upon. This conceptual model depicted a

system that none of the personnel had the complete understanding about. The methodology and results are of interest for other practitioners in the same field, but also for fellow researchers that are interested in the queuing system of a prison. There are also some modelling aspects of these kinds of systems that would be of interest for both practitioners and researchers.

## DIFFERENTIATION

All government activities are regulated by law or more loosely, by recommendations. This work is no exception. Statutory instruments must be followed and if it is possible, the visions and goals of the Swedish Prison and Probation Service. One of their goals is the degree of differentiation which influences the prison system behaviour and sets the boundaries for possible solutions.

There are three main grounds for differentiation; (i) sex; male and female clients are separated, (ii) security; open prisons, closed prisons, or closed prisons with extra high security, and (iii) age; younger, first time clients, are separated from older clients with a high return rate to crime and prison. Besides these three differentiation grounds, clients are also separated based on the nature of their crimes or on their need for treatment for drug or alcohol abuse.

Clients that have committed sex-related crimes are separated from other clients because the sex-offenders rank lowest among criminals. There is a risk for reprimands from those who rank higher in the unofficial ranking system. Another group that is separated is the clients with drug abuse. These clients are often in different treatment programs and need special care and competent supervision.

Today, there exists twenty-six different categories for differentiation, see table 1 for a limited selection of categories. As defined earlier, the degree of differentiation is equal to the number of available categories. The question is what categories should be utilised and how many places (cells) should each category have?

Table 1: Categories of Differentiation

Male	Female
Normal (open)	Normal
Normal (closed)	Treatment for drug abuse
Drug free	...
Youth	
Youth with motivation	
High security	
Treatment for drug abuse	
Treatment for alcoholics	
Sexual offenders	
...	

Laws and recommendations given by the Swedish government clearly states that a client must be placed in a suitable category or remain in remand prison. Temporarily, the normal category is utilised as placement while the client waits for a suitable categorised place to become available. There is also a desire not to move clients between different categories or placements and in some cases between different prisons. Each time a client is moved, he or her must adapt to a new environment and new prison personnel.

## SIMULATION MODEL

The used simulation methodology follows the steps described in Persson (2003) and does not differ from the methodologies described in *e.g.* Law and Kelton (1991), and Banks (1998). The first step (i) is the project planning or problem formulation where the outline of the study is determined. The next step (ii) is the conceptual modeling. The conceptual model describes the system under investigation. The conceptual model is validated as the next step (iii). The computer-based model is created as step (iv). This model must be verified (v) and validated (vi). Model verification aims at estimating if the simulation model is a valid representation of the conceptual model while model validation aims at estimating if the model is a valid representation of the system. The experimentation step (vii) consists of experimental runs with the simulation model. The results of these runs are then analysed (viii) and the result of that analysis is the base for the recommended decision or implementation (ix).

### Conceptual Model

The conceptual model is described in different levels. The highest level describes the model as a whole and contain few details. The sub levels are more detailed and shows the exact flow of each of the treatment chains. A treatment chain consists of the different instances that the client is routed through during the prison time in the system (the prison organisation). Note that the treatment chains are the ideal sequence of client activities during the prison time.

An example of a treatment chain is the treatment for a client that has problems with drug abuse; he/she gets an initial placement at a motivation wing. After the motivation (if the client is motivated), the next step is the treatment wing. After a successful treatment or when the prison stay has ended and the client is ready to leave, he/she has the possibility to continue the care outside the prison on a contract basis. Figure 1 shows the treatment chain for drug addicted males. Note that if no cell is available for in the treatment wing or in the motivational wing, the clients stays in remand prison until prison capacity is released.

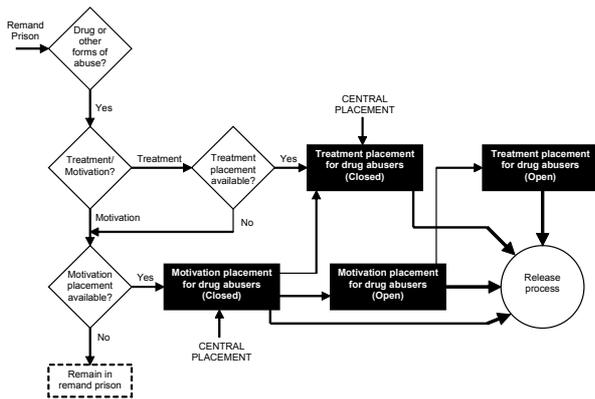


Figure 1: Treatment chain for male drug abusers

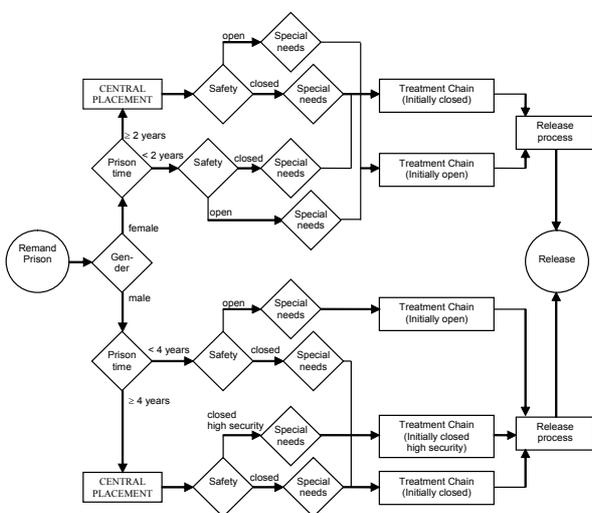


Figure 2: Conceptual Model – Highest Level

The high level conceptual model is depicted in figure 2. The model includes all initial placements to prison and placements in-between prisons. Individual prisons are not modelled, just the cells with different treatment programs. The remand prison has infinite capacity since the objective is to study the queues or waiting time for all differentiation categories.

The conceptual model is represented as a computer-based model in the simulation software Arena (version 7.0). A graphical user interface is provided in MS Excel. The user is able to adjust parameters that affects the output without any interaction with Arena. Parameters to use in experimentation is; torrent of criminals per year, number of places in different categories and the percentage share in different categories, both from remand prison and for persons liable to be detained but not yet deprived of liberty.

### Computer Model

The computer based model (the discrete event simulation model) is constructed in ARENA (version 7.00).

This section contains information about some special features included in the simulation model.

Clients with drug abuse have the opportunity to get the last part of the sentence in care outside the prison. As these clients get care outside the prison, capacity is made available at an earlier stage than anticipated. To implement this into the model all clients that were suitable for care outside the prison were given a reduction of their sentence. If the client had a sentence shorter than 30 days this treatment is not an option.

To measure the time that a client had been waiting for the requested treatment, time-attributes had to be added before entering and leaving the queue (remand prison). The difference between the two attributes is then measured as the time in queue. The number of days in queue are stored in a counter for statistics, which later is exported to MS Excel.

During the modeling effort, one of the larger problems was the warm up period. As the model runs for ten years the simulation should have a relative long warm up period. This did not work due to that the model showed an instable behavior due to the lack of capacity. Therefore, a shorter warm up period were chosen than was intended.

### Model assumptions

The following assumptions are made when constructing the computer based model.

- Different prisons are not separated in the model, all prison cells are kept together as one whole unit.
- The number of cells in remand prison is unlimited. This will not affect the final result.
- The client will stay in remand prison as long as there are no places available. With this assumption it is possible that the client serve the whole sentence in remand prison.
- When changing the degree of differentiation, clients previously belonging to categories that are removed are placed in the normal wings (open or closed).

### Model Validation

Basically, a simulation model can be divided into two distinct parts; i) the model logic, and ii) the statistical data for time between events and the stochastic occurrences of events. In this case, the model logic is easy to validate. Although there are several different opinions about how the prison system works, consensus about the model was reached by the system experts. The statistical data are more difficult to validate. Much data were collected from the official statistics of the National Prison and Probation Administration and must therefore be associated with a high degree of credibility. Other data that were needed for the simulation model proved difficult to obtain and were in the end estimated by the

proper personnel. These data might have less credibility but are still the best estimation possible.

When dealing with people's opinions about whether or not a model is valid, there is always a risk of including personal beliefs in the model validation process. The validity of a model that has been validated with subjective methods like a model walkthrough can be criticised due to the subjectivity in the technique itself. Nevertheless, subjective validation methods are, due to the simplicity in application, much used in practise. In this case, the model was validated using a walkthrough where experts found the model valid.

### **Graphical User Interface**

One of the objectives stipulated by the National Prison and Probation Administration is an easy-to-use interface for the simulation model. The model will be used as a planning tool at the National Prison and Probation Administration by personnel who is unfamiliar with simulation methodology. A graphical user interface is therefore used to define each scenario and start the simulation. The user interface (built in MS Excel) is connected to ARENA and the user needs only a basic familiarity with MS Excel. After a simulation run, experiment data are collected in the same interface and communicated by graphs and numbers.

To be able to get a high acceptance of the user interface, a hand drawn model of the MS Excel sheet was initially presented to the intended users. This model was changed many times before it was finalised and coded in MS Excel. In all, five intended users were used to test the interface before all involved were satisfied. The connection to ARENA was created with Visual Basic.

## **EXPERIMENTS AND RESULTS**

The experiments provide data to answer the following three questions:

- Q 1 What are the utilisation of available places of different categories as a function of the chosen degree of differentiation?
- Q 2 What are the minimum number of places in each category and the composition of categories to be able to reach the objective of maximum seven days in remand prison?
- Q 3 Given the forecasts of number of new clients, what is the number of places needed in ten years time?

The first experiment is based on the idea that a large number of different categories will show a low overall cell utilisation since some categories can be empty and not used by other categories. The second experiment follows the same reasoning since a small number of categories with a large number of cells will keep the

waiting time to a minimum. In the last experiment, the proposed expansion of an additional 1,150 prison cells, is tested and evaluated. If the number of clients continue to increase, this expansion will be insufficient in a couple of years.

The experiments are controlled by the MS Excel user interface. Each experiment starts with a warm-up period of one year and continue with a run length of ten years.

### **Experiment 1**

To find the relationship between the degree of differentiation and utilisation, five scenarios are created with an increasing degree of differentiation. The first scenario contains the lowest degree (as stated by law) and the fifth scenario contains the maximum degree of differentiation that is possible to obtain.

The results of experiment 1 shows that with today's number of clients, the utilisation soon reaches 100 % in each category. Even with the expansion with 1,150 new prison cells, the utilisation still reaches maximum. The situation is better with female clients than for male clients. Without a heavy expansion in prison capacity for male clients, it is impossible to find a relationship between utilization and degree of differentiation.

### **Experiment 2**

The objective concerning a maximum of seven days of waiting time in remand prison is evaluated by the best scenario in experiment 1. The percentage of clients that are placed in the correct category within seven days after that the sentence has been finalised is varying between 0 % (non of the clients) and 100 % (all clients). On average, 19.8 % of the clients in remand prison are given a correct placement within the stipulated seven days.

### **Experiment 3**

The third experiment examines the influence of a forecasted increase in clients during the next ten years. Also in this case it is clear that more capacity is needed to be able to estimate the effect of an increasing number of clients. The poor results in experiments 1, 2, and 3 resulted in a fourth experiment to capture the need of extra capacity.

### **Experiment 4**

The result of this experiment shows that the total prison system needs to be expanded with 6,700 new cells for male clients with the lowest degree of differentiation. This massive expansion is totally out of the scope of the planned expansion of 1,150 prison cells.

## **Results**

The main result from this study is that the system in its present form very quickly becomes full and the waiting

time in remand prison steadily increases. For female clients, the situation is within the desired specifications. A scenario with shorter waiting time in remand prison then seven days can be found for a high degree of differentiation. In the case of the male clients, an additional 6,700 prison cells are needed. This number is outside the scope for this project since the expansion is planned for a modest 1,150 prison cells.

## CONCLUSION

It is clear that the situation in Swedish prisons is critical. The planned expansion of 1,150 prison cells is far from enough. The situation for male clients is more critical than for female clients, who exhibit a far better situation. The female client population is also very small compared to the number of male clients (approximately 5 % of the prison cells are dedicated to female clients).

According to the results from the simulations the prison system should be expanded to include 11,000 prison cells in a few years time. It will otherwise be impossible to meet the recommendation of a maximum of seven days waiting time in remand prison.

These results are valid for a constant torrent of clients of 10,200 every year and a minimum degree of differentiation. There may be some uncertainty connected to the statistics due to the fact that some input data are estimates done by prison personnel. Nevertheless, this is the still the best estimate available.

The risk of heavily overcrowded prisons will force the National Prison and Probation Administration to increase the rate of expansion from the planned 1,150 new prison cells to the suggested 11,000 cells. Other activities that could decrease the need for traditional prison cells are the use of electronic supervision such as foot cuffs.

## ACKNOWLEDGEMENT

This work has been carried out with support from the National Prison and Probation Administration of Sweden during the autumn of 2003 and the beginning of 2004.

## REFERENCES

- Banks, J. (1998) *Handbook of Simulation*, John Wiley & Sons, New York.
- Barnett, A. (1987) Prison Populations: A Projection Model, *Operations Research*, Vol. 35, No. 1, pp. 18-34.
- Lattimore, P. K., and Baker, J. R. (1997) Demand estimation with failure and capacity constraints: An application to prisons, *European Journal of Operational Research*, Vol 102, Issue 3, pp.418-431.
- Law, A. M. and Kelton, W. D. (1991) *Simulation Modelling & Analysis*, 2nd Ed., McGraw-Hill, New York.
- Persson, J.F. (2003) *Discrete Event Simulation of Supply Chains – Modelling, Validation and Analysis*, Doctoral Thesis, Department of Production Economics, Linköping Institute of Technology, Linköping.
- Ridge, J.C., Jones, S.K., Nielsen, M.S., and Shahani, A.K. (1998) Capacity planning for intensive care units, *European Journal of Operational Research*, Vol 105, pp.346-355.
- Tarling, R. (1986) Statistical Applications in Criminology, *The Statistician*, Vol. 35 No. 3, pp. 369-388.

## AUTHOR BIOGRAPHIES

**FREDRIK PERSSON** is an Assistant Professor the department of Production Economics at Linköping Institute of Technology, Sweden. His research interests include modelling and simulation of manufacturing systems and supply chains. Of special interest are simulation methodology and validation methods.

**JOAKIM BARKMAN** received his Master of Science in Communication and Transport Engineering at Linköping Institute of Technology, Sweden, in 2004.

**ANNA PALMERIUS** received her Master of Science in Communication and Transport Engineering at Linköping Institute of Technology, Sweden, in 2004.

# GAME THEORY BASED TASK PLANNING IN MULTI ROBOT SYSTEMS

Krzysztof Skrzypczyk  
Department of Automatic Control  
Silesian University of technology  
Akademicka 16, 44-100, Gliwice, Poland  
e-mail:kskrzypczyk@ia.polsl.gliwice.pl

## KEYWORDS

Game theory, motion planning, multi-agent systems

## ABSTRACT

In the paper we discuss a problem of planning and coordination in a multi robot system. We consider a team of robots that performs a global task in a human-made workspace of complex structure. A hybrid architecture of the team motion control system is considered. The system can be split into two layers: the planner module and the behavior based collision free motion controller, that is designed to perform several elementary navigation tasks. The role of the planner is to plan and coordinate execution of elementary tasks by individual agents to obtain performance of global task. We presents the method of elementary tasks planning based on  $N$ -person game. The Nash equilibrium concept of the game solution is applied. An algorithm of multi robot workspace exploration is presented as an example of application of the proposed method. Simulation of the algorithm was carried out, and its result is presented and discussed in the paper.

## INTRODUCTION

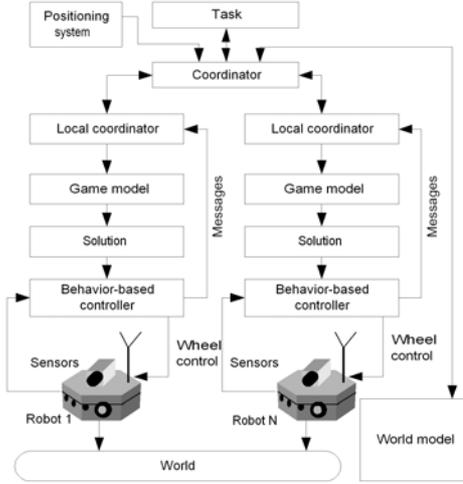
One of the fundamental requirements for mobile robot control system is its autonomy which is defined as an ability of operation without control of a human operator in an environment model of which is unknown and react a dynamical changes of this environment. Meeting these requirements implies the system has to cope with and solve many complex problems like task planning, collision free trajectory generation, operating on the basis of imprecise and uncertain information, environmental model building. Mobile robot control systems can be generally divided into three groups: deliberative, reactive (behavior-based) (Althaus and Christensen 2003; Arkin 1998; Skrzypczyk 2002) and hybrid systems. The most effective approach seems to be the third one that brings together advantages of reactive and deliberative systems (Fox et al. 1998; Shim et al. 2000). Operating by a robot in a real-world environment of a complex structure such as an office for instance, without any a priori information about the environment usually leads to inefficient task execution. Therefore, if robot is designed to work inside definite workspace it is reasonable to introduce partial knowledge of the

workspace to the control system (Althaus and Christensen 2003). When we consider a work of a team of robots, that are intended to perform some complex task (workspace exploration for instance) the additional problem of coordinating actions (tasks) of individual robots needs to be taken into account. Wrong coordination may lead to ineffective task execution or even to inability of completion the task. Therefore a lot of attention has been paid to this problem (Burgard et al. 2000; Gerkey and Mataric 2002; Golfarelli and Meuleu 2000; Lawton et al. 2003; Sequiera and Ribeiro 2001; Schneider-Fontan and Mataric 1998). The problem of coordination of multiple robots can be stated as a conflict situation between individual robotic-agents and can be modeled as a decision making problem. The game theory is convenient tool for modelling problems of conflict nature. Therefore applications of game theory in the context of multi-agent coordination have been widely reported in the literature (Golfarelli and Meuleu 2000; LaValle 2000; Li and Payandeh 2001). Unfortunately most of the works lack the treatment of application aspect of game theoretical approach. They only consider the problem as a theoretical one, without taking into account limitations of a control system framework. In this paper we discuss an approach to coordination of multiple robots operating in an environment of a complex structure, performing complex task which example is the environment exploration one. We model the problem of coordination of elementary tasks as a  $N$ -person, noncooperative team decision problem. We present a hybrid architecture of a system that is designed to control the work of agents that perform the stated task. In the end of this work we present and discuss the simulation results of the proposed system.

## SYSTEM OVERVIEW

A general structure of the control system is presented on the block diagram below (fig.1). The system can be split into two layers. The first one that is intended to be implemented on mobile platform and the second, that provides information of robots location and communication between robots. The first one that is considered in this work has typical hybrid structure. It consists in behavior-based motion controller that is responsible for executing elementary navigational tasks (modeled further by operators) and non-cooperative game based planner. The role of the planner is to choose from admissible actions the one that provide execution

of a part of primary mission (in our case the mission is the workspace exploration). Of course the planner has to



Figures 1: A Diagram of the Hybrid, Multi-Robot Control System

take into account all possible actions of other agents and provide their proper coordination.

### The world model

Robots are intended to operate inside a well structured, complex human made workspace. In order to simplify the navigation problem a partial knowledge about the environment is introduced to the system. In the fig. 2a an exemplary office environment plan is presented. An overall workspace is divided into regions named *sectors*. Each sector represents an area occupied by a room, corridor or a part of a corridor. Moreover passages between rooms are distinguished and introduces to the model as *door-objects* called further *door* for simplicity. The workspace model is stored onto two layers: topological and geometrical. The first one is given by weighted graph:

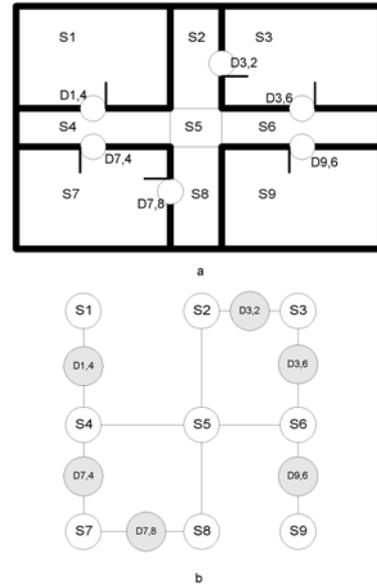
$$M = (V, W) \quad V = \{v_1, v_2, \dots, v_M\}, \quad W \subset V \times V \quad (1)$$

which nodes represent objects of the environment: sectors  $V_S$  and doors  $V_D$  where  $V = V_S \cup V_D$ . On the topological level each object is described by a real number that is an "cluttering coefficient"  $c_i$  in case the object is a sector and by a probability of being opened  $o_j$  when the object is a door. The first coefficient reflects the number of objects placed inside of the given sector. On a geometrical level  $i$ -th sector is represented by coordinates of its top left corner  $(x_i^t, y_i^t)$ , bottom right corner  $(x_i^b, y_i^b)$ , and a center point  $(x_i^p, y_i^p)$ . Similarly the  $j$ -th door-object is described by a circle of radius  $r_j$  and a center of the circle  $(x_j^p, y_j^p)$ . The edges of the graph define relations of neighborhood between environmental objects related to vertices of the graph. Weighting factors fixed to the edges of the graph are

related to some costs of moving robot from the  $i$ -th to the  $j$ -th vertex (object). In this work we define the costs using some heuristic formulae:

$$w_{ij} = \begin{cases} L_{i,j} (1 + c_i) (1 + e^{-a o_j}) & \text{if } v_i \in V_S \cap v_j \in V_D \\ L_{i,j} \frac{1}{o_i} (1 + c_j) & \text{if } v_i \in V_D \cap v_j \in V_S \\ L_{i,j} [1 + 0.5(c_i + c_j)] & \text{if } v_i \in V_S \cap v_j \in V_S \end{cases} \quad (2)$$

It is worth of noticing that the model describes only invariable features of the workspace. The layout of objects (furniture, equipment) placed inside sectors is not known. Moreover it can undergo dynamic changes.



Figures 2: An Exemplary Workspace Layout (a) and its Topological Model (b)

### The behavior-based controller

The role of this module is to execute elementary navigational tasks. It is designed using the behavior-based idea of control (Arkin 1998). It is composed of six behaviors that process the state and sensory information into proper set-points for motion controller which are values of linear and angular velocity. The coordination of behaviors activities is made by fixed priority arbiter. For more detailed description of this module refer to (Skrzypczyk 2002). For the purpose of this article that is enough to consider the module as the one which is able to execute four different elementary tasks, represented by following operators:

- **FindDoor(D)** - the task of moving the robot inside the area of door-object D;
- **TraverseDoor(D,S)** - the task of going through the door-object D to the sector S;
- **Wait()** - the simple command that stop the robot;

• **GoTo(S1,S2)** - the task of moving robot from the sector  $S1$  to  $S2$ ;

All of the operators are related to a task of collision free moving robot from a given initial location to a desired one which are specified by arguments of the operator. The difference between individual tasks lies in the set of parameters associated with a given operator that is sent to the controller. The parameters are priorities of individual behaviors as well as velocity limits.

### The planner

From the perspective of this work the planner is the core of the system. It consists of three modules: local coordinator, decision process modeling module and the solution computation one. The work of the planner can be briefly described in a following way. The local coordinator receives information of location of all of the robots. Moreover it is provided with a world model and a primary task data. Depending on a type of the task, location of all of teammates and a state of the task execution, a model of a decision process is built. The model is in fact the cost function that depends on actions made by individual agents and a state of task completion. Next the problem is solved and the solution computed. The solution of the problem determines an elementary action which is optimal for a given agent from the point of view of primary task execution. Detailed description of the process of building the model (taking an exploration task as an example) and the methods of solution shall be presented in further sections.

### AN EXPLORATION PROBLEM FORMULATION

As we mentioned before, we want to present the method of coordination of multiple robots that provides completion of the task of exploration of the workspace the topological model of which is known to the system. This task can be generally stated as a problem of visiting a given part of the workspace by teammates with a cost as low as possible. Exemplary interpretation of the exploration task is delivering parts in a large plant by multiple robots. In terms of this work the exploration task is defined as visiting a part  $M_V \subset V_S$  of the workspace  $M$  in a number of steps as small as possible. Here in this work, we model the problem as a sequence of one stage, non zero sum games in a normal form.

### Modelling the Problem of Exploration

Let us first introduce a notation that will be used hereafter. The state of a team of robots is denoted by a set:

$$X = \{x_i\} \quad i = 1, 2, \dots, N, \quad x_i \in V \quad (3)$$

what is equivalent to the fact that there is the  $i$ -th robot inside the area described by the vertex  $x_i$ . The set of all

possible robot actions described by operators is given by the set:

$$A = \{a_1, a_2, \dots, a_M\} \quad (4)$$

where  $M$  is a number of all operators (in this work  $M=4$ ). A set of possible actions of the  $i$ -th robot in the state  $x_i$  is defined by :

$$A_i(x_i) = \{a_1, a_2, \dots, a_K\} \quad (5)$$

and it is determined by precondition lists of individual operators. In our case they are as follows:

#### FindDoor(D)

$$\text{preconditions} = \{x_i \in V_S, w_{x_i, D} \neq \infty\}$$

#### TraverseDoor(D,S)

$$\text{preconditions} = \{x_i = D \in V_D, w_{D, x_i} \neq \infty\}$$

#### Wait()

$$\text{preconditions} = \phi$$

#### GoTo(S1,S2)

$$\text{preconditions} = \{x_i = S1 \in V_S, w_{S1, S2} \neq \infty\}$$

In the terms of the decision making process model of an action  $a_k \subset A_i$  is a mapping:

$$a_k : x_i^n \rightarrow x_i^{n+1} \quad x_i^n \subset X \quad x_i^{n+1} \subset V \quad a_i \subset A_i \quad (6)$$

where  $x_k^n$  is the current state of the  $i$ -th robot, and  $x_k^{n+1}$  is a state the robot will be in as a result of the action  $a_k$ . The primary task of the team of robots is to visit all objects defined by a set  $M_G \subset V_S$ . We introduce an auxiliary set defining objects that have already been visited and we denote it by  $M_V \subset M_G$ . Using this notation we can precisely formulate a goal of the team as satisfying the equality:  $M_V = M_G$ . The task of the planning algorithm is to choose for each robot one of the possible action, that applied to the robot will result in performing a part of the primary task. The problem of selection of proper action is in this work perceived as a game between individual robotic-players. The result of the game related to the defined task depends on decisions made by individual game participants. Moreover the task of exploration has a specific nature that can be classified as a team-work problem where all of the players (robots) want to optimize one performance index. Although the environment is in principle dynamic we model the problem of action planning as a sequence of static  $N$ -person games in a normal form. Therefore we need to define for each stage of the planning process the single cost function value of which depends on actions made by all of teammates and on the task completion state. We propose to define the cost function as a sum of three components:

$$I(a_1, \dots, a_i, \dots, a_N) = -I_R + I_E + I_D \quad (7)$$

The first one is related to a some value of "reward" that is given to robots for exploring unvisited part of the workspace and it is given by:

$$I_R(a_1, \dots, a_N) = \sum_{i=1}^N R_i(a_i) \quad (8)$$

$$R_i = \begin{cases} \frac{1}{k} R > 0 & \text{if } x_i^{n+1} \in M_G \cap x_i^{n+1} \notin M_V \\ 0 & \text{otherwise} \end{cases}$$

where  $R$  is a positive number that denotes the reward value. The  $k$  is a number of robots visiting the same object as a result of their actions. The value of the second component  $I_E$  is dependent on an amount of energy necessary to make an action  $a_i$  which is proportional to a cost of transition of robots between environmental objects defined by the model  $M$ :

$$I_E(a_1, \dots, a_N) = \sum_{i=1}^N w(x_i^n, x_i^{n+1}) \quad (9)$$

Third component denotes cost of moving the robot to the nearest (in the sense of costs defined by  $W$ ) unexplored object. Let us first denote a path of minimal cost between an object  $n$  and  $m$  as:

$$p_{\min}(n, m) = \{v_n, \dots, v_k, \dots, v_m\} \subset V \quad (10)$$

and let the set of unexplored objects is given as  $M_U = \{u_i\} = M_G \setminus M_V$ . Then the cost  $I_D$  is given by:

$$I_D(a_1, \dots, a_N) = \sum_{i=1}^N D_{\min,i} \quad (11)$$

$$D_{\min,i} = \min_l [p_{\min}(x_i^{n+1}, u_l)]$$

where  $D_{\min,i}$  is the cost of moving the  $i$ -th robot from the state  $x_i^{n+1}$  to the "nearest" unexplored object.

### Solution

In the previous section we derived a model for a single stage of the exploration process. Now, we look for a solution of the problem defined above. The solution will be a set of actions of individual robots  $S^* = \{a_{10}, a_{20}, \dots, a_{N0}\}$  that if performed lead to execution of a part of primary task. The problem of exploration is in principle cooperative one but taking into account a fact that robots can not communicate each other during an action execution implies that the problem has features of noncooperative one. Thus we try to find the solution for a single decision-making stage, considering the problem as a noncooperative one. One of the best known concept of solution of such problems is the Nash equilibrium one (Basar and Oldster 1982). Applying it

to our problem, the solution is the equilibrium point if following inequalities are satisfied by the set of the decisions  $\{a_{10}, \dots, a_{N0}\}$ :

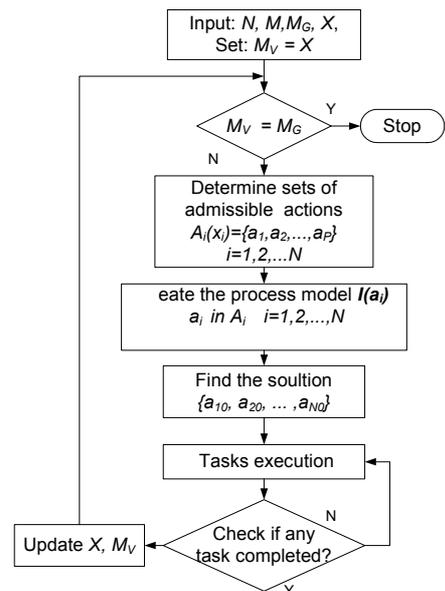
$$\begin{aligned} I_1(a_{10}, \dots, a_{N0}) &\leq I_1(a_1, a_{20}, \dots, a_{N0}) \\ &\vdots \\ I_N(a_{10}, \dots, a_{N0}) &\leq I_N(d_{10}, d_{20}, \dots, d_{N0}) \end{aligned} \quad (12)$$

In case of a team-problem, where  $I_i = I$ ,  $i = 1, 2, \dots, N$  the solution reduces to minimization of the cost function  $I$ . Thus we have:

$$S^* = \{a_{10}, a_{20}, \dots, a_{N0}\} = \min_{a_1, \dots, a_N} I \quad (13)$$

### THE ALGORITHM OF EXPLORATION

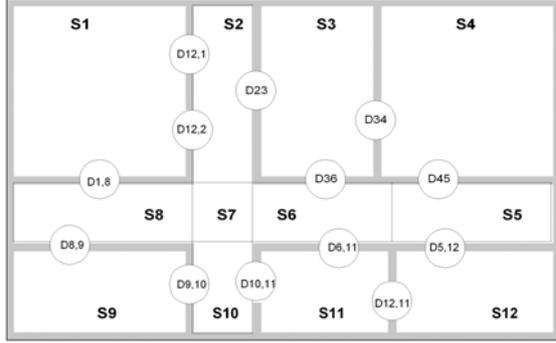
In this section we present an overall algorithm of workspace exploration for a team of robots. The block diagram of the algorithm is presented in the fig.3. First, the global task is stated by the set  $M_G$  and the state of the robot team  $X$  is determined using information of robots location. In the next step of the algorithm admissible actions of individual robots are determined taking into account the current state  $X$  of robots and environmental model  $M$ . Then, the model of the process is calculated according to equations (7)-(11). The solution of the problem is computed using formulas (12),(13) and it determines for each robot an elementary task (action) to be performed. The action code is sent to the motion-controller. A robot that executes an action as a first sends a message to the coordinator. That causes the state  $X$  of a team as well as the state of the primary task completion are updated. The process described above is repeated until the task is completed  $M_G = M_V$ .



Figures 3: The algorithm of exploration

## SIMULATION EXAMPLE

In order to show how the approach discussed in the paper works we present a result of an exemplary simulation. We implemented the method using a simulation environment *M.A.S.S.* (Multi Agent Systems Simulator) which was created and has been developed by the author. This application allows to create a model of a well-structured workspace and simulate the work of differentially-driven mobile robots inside of the modeled workspace. The layout of a workspace we modeled is shown in the fig.4.



Figures 4: The Layout of a Workspace Used for a Simulation

It is an example of a typical office environment. It consists of twelve sectors  $V_s = \{v_1, v_2, \dots, v_{12}\}$  that represent rooms and parts of corridors, and thirteen passages (doors) between rooms  $V_d = \{v_{13}, \dots, v_{25}\}$ . The area of the workspace is of a size  $15 \times 10$  [m] (width, height). Various pieces of furniture and equipment are placed inside of individual rooms (fig. 5) and it is modelled by different cluttering coefficients  $c_j$  fixed to each sector. In the presented simulation they are equal  $C = \{0.7, 0, 0.1, 0.6, 0, 0, 0, 0.2, 0, 0.3, 0.2\}$ . All of probabilities  $o_j$  that doors are opened are equal one except the door  $D_{3,6}$  which is open with probability  $o_5 = 0.2$ . That corresponds to the fact that this door most time is closed. The robot model used for the simulation is typical differentially driven one, with 8 range sensors placed around the disc-shaped platform of diameter 0.55[m]. We consider the following exploration task. A team of three robots ( $N=3$ ) is intended to explore the workspace what is equivalent to visiting all of the sectors. Thus the goal of exploration is defined as  $M_G = V_s$ . Initially, robots are located inside of sectors  $X_{init} = \{v_4, v_9, v_2\}$  so  $M_V = X_{init}$ . In the fig.5 successive stages of simulation of exploration process are presented. A sequence of operators that was used to perform the task is presented in the table 1. The symbols *FD*, *TD*, *GT*, *W* denote operators: *FindDoor*, *TraverseDoor*, *GoTo*, *Wait*. We can see that algorithm works well, providing completion of the task stated above. One can notice that robot 1 perform only a small part of the task, exploring only one room. But it has to be taken into account, that the door  $D_{3,6}$  is modeled as

the one that is almost for sure closed. That is the reason of this "strange" task execution. Yet another aspect is worth of commenting. The algorithm presented in paper make only "one step ahead" planning. It causes that the task execution may be not optimal. Using other planning algorithms we would obtain better or even optimal solution. But such an approach would be valid if the environment was static as well as we assumed perfect result of each action.

Table 1: A Sequence of Operators that Provides Completion of the Exploration Task

n	Robot 1	Robot 2	Robot 3
1	FD('D3,4')	FD('D9,10')	FD('D1,2 1')
2	TD(D3,4,S3)	TD(D9,10,S10)	TD(D1,2 1,S1)
3	FD(D2,3)	FD(D10,11)	FD(D1,8)
4	TD(D2,3,S2)	TD(D10,11,S11)	TD(D1,8,S8)
5	W()	FD(D11,12)	GT(S7)
6	W()	TD(D10,11,S12)	GT(S6)
7	W()	W()	GT(S5)

## SUMMARY AND CONCLUSION

In this paper we discussed a problem of planning and coordination of tasks in a multi robot system. We considered a team of robots that was intended to perform a task of exploration of a human-made workspace of complex structure. We proposed both the hybrid architecture of a control system and method of coordination of multiple robotic agents. In the paper we also presented an algorithm of exploration of workspace. The core of the algorithm is the model of the process that is stated as a noncooperative game in a normal form. We applied the Nash equilibrium concept to generate a solution of the problem. Although the result of only one simulation was presented, we had made a number of simulation experiments using both various parameters and workspace configurations. In all cases we obtained correct task execution. On the basis of simulations we carried out we can conclude that this algorithm works well and provides effective exploration of even very complex-structured environments. However, the algorithm can not guarantee optimal task performance. It is caused the algorithm uses only one-step-ahead planning method. But this approach on the other hand has other advantage - it allows to track dynamical changes of the environment and it does not need the assumption that a given action is always executed in a perfect way. The simulation environment allowed us to verify this approach. Taking into account the results of simulations we can state that the approach presented in the paper seems to be promising. Therefore our future researches shall be focused on applying it to a real multi robot system.

## Acknowledgments

This work has been supported by KBN Grant No. 3 T11A 012 26 in the year 2004.

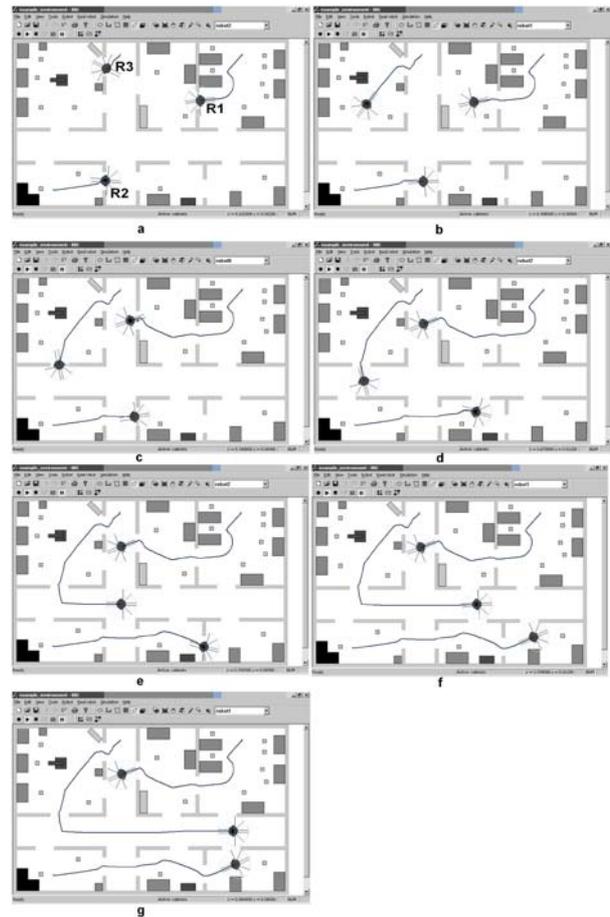
## REFERENCES

- Althaus, P. and H.I. Christensen. 2003. "Behaviour coordination in structured environments." *Advanced Robotics*, 17(7), 657–674.
- Arkin, R.C. 1998. *Behavior-Based Robotics*. MIT Press, Cambridge, MA.
- Basar, M. and G.J. Olsder. 1982. *Dynamic Noncooperative Game Theory*. Mathematics in Science and Engineering, Academic Press Inc. Ltd, London.
- Burgard, W. et al. 2000. "Collaborative Multi-Robot Exploration." *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, San Francisco CA.
- Fox, D.; W. Burgard; and S. Thrun. 1998. "A Hybrid Collision Avoidance Method For Mobile Robots." *Proc. of the IEEE International Conference on Robotics and Automation*, Leuven, Belgium.
- Gerkey, B. and M.J. Mataric. 2002. "Pusher-watcher: An Approach to Fault-Tolerant Tightly-Coupled Robot Coordination." *Proc. of the IEEE International Conference on Robotics and Automation (ICRA2002)*, Washington DC.
- Golfarelli, M. and N. Meuleu. 2000. "Multi Agent Coordination in Partially Observable environments." *Proceedings of the 2000 IEEE Int. Conf. of Robot and Automation*, San Francisco, 2777- 2782.
- LaValle, S. 2000. "Robot motion planning: A game-theoretic foundation." *Algorithmica*. 26(3), 430–65.
- Lawton, J.R. et al. 2003. "A Decentralized Approach to Formation Maneuvers." *IEEE Transactions on Robotics and Automation*, Vol 19, No 6, 933–941.
- Li, Q. and S. Payandeh. 2001. "On Coordination of Robotics Agents Based on Game Theory", *ICRA 2001*, Seoul, Korea, (private communication).
- Sequeira, J. and M. I. Ribeiro. 2001. "A Negotiation Model for Cooperation Among Robots", *Proc. of European Control Conference*, Porto, Portugal.
- Shim, H.S. et al. 2000. "A hybrid control structure for vision based soccer robot system." *Int. J. Intelligent Automation and Soft Computing*, 6: (1), 89–101.
- Schneider-Fontan, M. and M.J. Mataric. 1998. "Territorial Multi-Robot Task Division." *IEEE Transactions on Robotics and Automation*, Vol 15, N.5, 815-822.
- Skrzypczyk, K. 2002. "Behaviour-based approach to a synthesis of a mobile vehicle control system." *IV-th Domestical PhD Workshop OWD*, Istebna-Zaolzie, Poland.



### CHRISTOPHER T. SKRZYPCZYK

was born in Tarnowskie Gory, Poland and went to the Silesian University of Technology in Gliwice, where he studied automatics and robotics and obtained his degree in 2000. In the same year he started his PhD studies at the Department of Automatic Control at the same university. His interests have been focused on mobile robots collision free motion planning and coordination methods in Multi Agent Systems based on the game theory. His e-mail is: [kskrzypczyk@ia.polsl.gliwice.pl](mailto:kskrzypczyk@ia.polsl.gliwice.pl).



Figures 5: The result of the exemplary simulation

# MINORITY GAME WITH COMMUNICATION: AN AGENT BASED MODEL

Marco Remondino  
Department of Computer Science  
University of Turin  
C.so Svizzera 185  
10149 Turin, Italy  
E-mail: remond@di.unito.it

Alessandro Cappellini  
PhD Student in Simulation  
University of Turin,  
C.so Unione Sovietica 218 bis  
10134 Turin, Italy  
E-mail: cappellini@econ.unito.it

## KEYWORDS

Minority Game, El Farol Bar problem, social network, multi agent simulation, communication

## ABSTRACT

The Minority Game (MG) is a simple, generalized framework, belonging to the Game Theory field, which represents the collective behaviour of agents in an idealized situation where they have to compete through adaptation for some finite resource. It generalizes the study of how many individuals may reach a collective solution to a problem under adaptation of each one's expectations about the future. It is assumed that an odd number of players take a decision at each step of the simulation; the agents that take the minority decision win, while the others loose. The Minority Game in its original formulation state that there is no communication among the agents involved in the simulation; the idea in this paper is to introduce in the model a sort of a social network, in order to see how the links among certain agents can change the results of the simulation. A software model is built, in which the user can define the number of the agents involved and the number of links among them; some examples are studied and analyzed in order to find some general rule. Besides, two communication protocols are implemented in the model: the asynchronous one, in which the agents act sequentially. So the first agents which act take their decision, and from then on they reply to the other agents with the new decision taken. The synchronous protocol states that the agents always communicate to the others their original opinion: they broadcast their opinion to all the agents which are linked to them. Finally, after having collected all the opinions of their friends, they reconsider their choice. We examine the difference among the two protocols using the same starting parameters in the simulation.

## INTRODUCTION

Game Theory is a distinct and interdisciplinary approach to the study of strategic behaviour. The disciplines most involved in game theory are mathematics, economics

and the other social and behavioural sciences. Game theory (like computational theory and so many other contributions) was founded by the great mathematician John von Neumann. The first important book was The Theory of Games and Economic Behaviour, which von Neumann wrote in collaboration with the great mathematical economist, Oskar Morgenstern. Certainly Morgenstern brought ideas from neoclassical economics into the partnership, but von Neumann, too, was well aware of them and had made other contributions to neoclassical economics.

The key link between neoclassical economics and game theory was and is rationality. Neoclassical economics is based on the assumption that human beings are absolutely rational in their economic choices. Specifically, the assumption is that each person maximizes her or his rewards - profits, incomes, or subjective benefits - in the circumstances that she or he faces. This hypothesis serves a double purpose in the study of the allocation of resources. First, it narrows the range of possibilities somewhat. Absolutely rational behaviour is more predictable than irrational behaviour. Second, it provides a criterion for evaluation of the efficiency of an economic system. If the system leads to a reduction in the rewards coming to some people, without producing more than compensating rewards to others (costs greater than benefits, broadly) then something is wrong. Pollution, the overexploitation of fisheries, and inadequate resources committed to research can all be examples of this.

In neoclassical economics, the rational individual faces a specific system of institutions, including property rights, money, and highly competitive markets. These are among the "circumstances" that the person takes into account in maximizing rewards. The implications of property rights, a money economy and ideally competitive markets is that the individual needs not consider her or his interactions with other individuals. She or he needs consider only his or her own situation and the "conditions of the market." But this leads to two problems. First, it limits the range of the theory. Wherever competition is restricted (but there is no monopoly), or property rights are not fully defined, consensus neoclassical economic theory is inapplicable, and neoclassical economics has never produced a generally accepted extension of the theory to cover these cases. Decisions taken outside the money economy were also problematic for neoclassical economics.

Game theory was intended to confront just this problem: to provide a theory of economic and strategic behaviour when people interact directly, rather than through the market. In game theory, "games" have always been a metaphor for more serious interactions in human society. Game theory may be about poker and baseball, but it is not about chess, and it is about such serious interactions as market competition, arms races and environmental pollution. But game theory addresses the serious interactions using the metaphor of a game: in these serious interactions, as in games, the individual's choice is essentially a choice of a strategy, and the outcome of the interaction depends on the strategies chosen by each of the participants. On this interpretation, a study of games may indeed tell us something about serious interactions.

In neoclassical economic theory, to choose rationally is to maximize one's rewards. From one point of view, this is a problem in mathematics: choose the activity that maximizes rewards in given circumstances. Thus we may think of rational economic choices as the "solution" to a problem of mathematics. In game theory, the case is more complex, since the outcome depends not only on my own strategies and the "market conditions," but also directly on the strategies chosen by others, but we may still think of the rational choice of strategies as a mathematical problem - maximize the rewards of a group of interacting decision makers - and so we again speak of the rational outcome as the "solution" to the game.

## THE MINORITY GAME

The Minority Game (MG) is a simple, generalized framework, belonging to the Game Theory field, which represents the collective behaviour of agents in an idealized situation where they have to compete through adaptation for some finite resource.

While the MG is born as the mathematical formulation of "El Farol Bar" problem considered by (Arthur, 1994), it goes way beyond this one, since it generalizes the study of how many individuals may reach a collective solution to a problem under adaptation of each one's expectations about the future. In (Arthur, 1994) the "El Farol Bar" problem was posed as an example of inductive reasoning in scenarios of bounded rationality. The kind of rationality which is usually assumed in economics - perfect, logical, deductive rationality - is extremely useful in generating solutions to theoretical problems, but it fails to account for situations in which our rationality is bounded (because agents can not cope with the complexity of the situation) or when ignorance about other agents ability and willingness to apply perfect rationally lead to subjective beliefs about the situation. Even in those situations, agents are not completely irrational: they adjust their behaviour based on what they think other agents are going to do, and these expectations are generated endogenously by information about what other agents have done in the past. On the basis of these expectations, the agent takes an action, which in turn becomes a precedent that

influences the behaviour of future agents. This creates a feedback loop: expectations arise from precedents and then create the actions which, in turn, constitute the precedents for the next step.

The original formulation of "El Farol Bar" problem is as follows:  $N$  people, at every step, take an individual decision among two possibilities. Number one is to stay at home; number two is to go to a bar. Since the space in the bar is limited (finite resource), the time there is enjoyable if and only if the number of the people there is less than a fixed threshold ( $aN$ , where  $a < 1$ ). Every agent has his own expectation on the number of people in the bar, and according to his forecast decides whether to go or not. The only information available to the agents is the number of people attending the bar in the recent past; this means that there is no deductively rational solution to this problem, but there can be plenty of models trying to infer the future number according to the past ones.

The other very interesting aspect of the problem is that if most agents think that the number of people going to the bar is  $> aN$  then they won't go, thus invalidating their own prevision. Computer simulations of this model shows that the attendance fluctuates around  $aN$  in a  $(aN, (1 - a)N)$  structure of people attending/not attending. The "El Farol Bar" problem has been applied to some proto-market models: at each time step agents can buy (go to the bar) or sell an asset and after each time step, the price of the asset is determined by a simple supply-demand rule.

The MG has been first described in (Challet and Zhang, 1997) as a mathematical formalization and generalization of "El Farol Bar" problem. It is assumed that an odd number of players take a decision at each step of the simulation; the agents that take the minority decision win, while the others loose. Stepping back to "El Farol Bar" problem, we can see it as a minority game with two possible actions:  $a_1 = 1$  (to go to the bar) and  $a_2 = -1$  (not to go to the bar). After each round, the cumulative action value  $A(t)$  is calculated as the sum of each value given to the single actions. The minority rule sets the comfort level at  $A(t) = 0$ , so that agent is given a payoff  $-a_i(t)g[A(t)]$  at each time step with  $g$  an odd function of  $A(t)$ .

## INTRODUCING COMMUNICATION AMONG AGENTS

The "El Farol Bar" problem, as well as the Minority game in its original formulation state that there is no communication among the agents involved in the simulation; the idea in this paper is to introduce in the model a sort of a social network, in order to see how the links among certain agents can change the results of the simulation. A social network is defined as "a set of nodes - e.g. persons, organizations - linked by a set of social relationship - e.g. friendship, transfer of funds, overlapping membership - of a specific type" (Laumann, et al., 1978).

In our case the minority rule will be very easy: a set of  $N$  agents will have to choose between (-1) and (1). Who is in the minority (denoted with  $n < N$ ) wins and gets a payoff equal to  $N/n$ : the fewer agents stay in the minority, the higher the payoff. Also the social network involved will be quite simple, just linking an agent to others with a relation limited to the possibility of asking a question: “will you choose (-1) or (1)?”. Not all the agents will be connected, though, so that some of them will have to make a prevision just considering the past few results, exactly like in the original MG. The described situation is depicted in figure 1 (in which we have twelve agents and eight links).

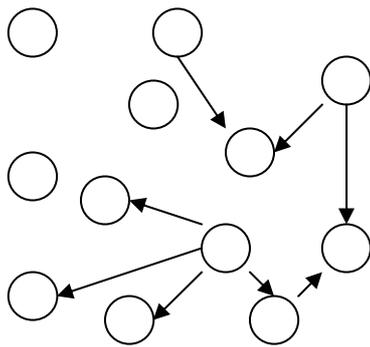


Figure 1: A Simple Social Network

It’s important to notice that the versus of the arrow is crucial; it means that agent A can ask agent B, but not necessarily agent B can ask agent A. An example of the possible relations is given in Table 1.

	1	2	3	4	5
1		x	x	x	
2			x		
3	x				x
4	x				x
5					

Table 1: definition of relations among agents

In the example we have five agents involved in the simulation: agent 1 can ask agents 2, 3 and 4, while agent 2 can ask agent 3 and number 3 can ask number 1 and number 5; agent 4 can then ask number 1 and number 5, while number 5 is a lonely agent (he can’t ask anyone, even if two other agents can ask him what he will do).

## AGENT BASED SIMULATION

In (Ostrom 1988), agent based simulation is described as a third way to represent social models, being a powerful alternative to other two symbol systems: the verbal argumentation and the mathematical one. The former, which uses natural language, is a non computable way of modelling though a highly descriptive one; in the latter, while everything can be done with equations, the complexity of differential systems rises exponentially as the complexity of behaviour grows, so that describing complex individual behaviour with equations often becomes an intractable task. Simulation has some advantages over the other two: it can easily be run on a computer, through a program or a particular tool; besides it has a highly descriptive power, since it is usually built using a high level computer language, and, with few efforts, can even represent non-linear relationships, which are tough problems for the mathematical approach. According to (Gilbert, Terna 2000):

*“The logic of developing models using computer simulation is not very different from the logic used for the more familiar statistical models. In either case, there is some phenomenon that the researchers want to understand better, that is the target, and so a model is built, through a theoretically motivated process of abstraction. The model can be a set of mathematical equations, a statistical equation, such as a regression equation, or a computer program. The behaviour of the model is then observed, and compared with observations of the real world; this is used as evidence in favour of the validity of the model or its rejection”*

In Remondino (2003) we read that computer programs can be used to model either quantitative theories or qualitative ones; simulation has been successfully applied to many fields, and in particular to social sciences, where it allows to verify theories and create virtual societies. In order to simulate the described problem, multi-agent technique is used. Agent Based Modelling is the most interesting and advanced approach for simulating a complex system: in a social context, the single parts and the whole are often very hard to describe in detail. Besides, there are agent based formalisms which allow to study the emergency of social behaviour with the creation and study of models, known as artificial societies. Thanks to the ever increasing computational power, it’s been possible to use such models to create software, based on intelligent agents, which aggregate behaviour is complex and difficult to predict, and can be used in open and distributed systems. The concept of Multi Agent System for social simulations is thus introduced: the single agents have a very simple structure. Only few details and actions are described for the entities: the behaviour of the whole system is a consequence of those of the single agents, but it’s not necessarily the sum of them. This can bring to unpredictable results, when the simulated system is studied.

There are many toolkits and frameworks that can be used to build agent based simulations; for this work JAS was selected (<http://jaslibrary.sourceforge.net>) since it includes graph support for Social Network Analysis. In the basic model we present in this paper we only examine how many agents change their own opinion, when increasing the number of direct relations among them; further work will address some other issues, such as the correctness of the agents' choice, and so on.

## THE SIMULATION FRAMEWORK

At the beginning of the simulation, during the setup, we create a simple world populated by  $N$  agents. These agents can be considered as the vertexes of a social network and the links among them (relations) as the edges. The network is directed and every arc is composed by two edges with opposite directions. Every agent has a list of  $F$  (friends) other agents (called friendsList) to whom he can ask. This list is composed by the neighbors, i.e. the vertexes linked to the examined vertex (the agent).

Here follows a brief description of the simulation process:

- At the beginning of each simulation step, every agent has its own forecast. The forecast is absolutely random between two choices  $-1$  and  $+1$ .
- The decision taken by each agent (before communicating with others) is denoted with a "certainty index" equal to 1 (100%).
- Now an agent is randomly chosen. He starts asking to the first in the list; if this one has the same prevision, then the certainty index is increased by a value of  $1/F$ , while if the prevision is different, than the certainty index is lowered by  $1/F$
- After having asked to all the friends in his list, the agent takes the final decision: if the certainty index is equal or greater than 1, then the decision will be the original one. If it's lower than 1, then the decision will be the other possible one
- Another agent is then randomly chosen, and so on (the same agent can't be chosen twice during the same turn). Note that an agent that's been asked can still change his mind, basing on the agents he will in turn ask

Before starting the simulation, we can change two core parameters: the number of the agents involved and the number of the links among the agents. Here we examine three runs of the simulation, one with 1000 agents and 500 total links (an average of one link every two agents); the other one with 100 agents and 500 links (an average of five links for every agent) and the last one with 100 agents and 5000 links (fifty links for every agent). In every run we iterate the minority game for 1000 times.

The model could be considered as some groups of friends that must choose between two alternatives: pub and disco. They communicate the selected choice to

their friends, elaborate them and then take a final decision.

In the output graph we can read the time on x-axis (1000 iterations of the game), and we plot two lines: the red one (the lower one in the graphs) depicts the decisions changed while the blue one (the upper one) is for unchanged decisions.

In y-axis we read the number of decisions (changed or not) the scale ( $10^1$ ,  $10^2$ ,  $10^3$ ) depends from agents number.

We choose as standard example a world of 100 agents and 500 relations (figure 2), in which an average of 65 out 100 preserve their original decisions.

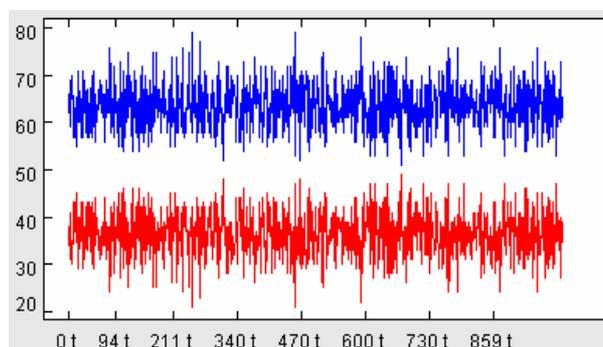


Figure 2: 100 agents and 500 relations

In a second run we imagine a different situation, in which the agents have many more relations among them: an average of fifty for every inhabitant (figure 3).

A simple common sense rule states that the more relations, the higher is the probability to change opinion.

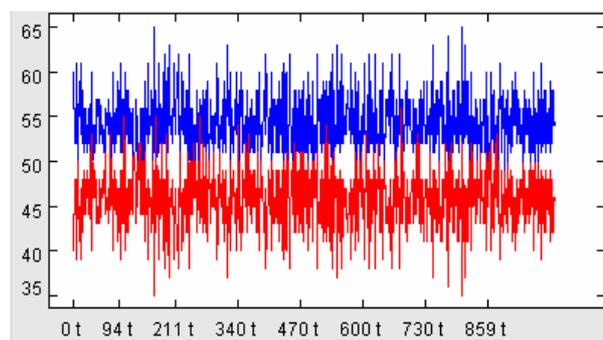


Figure 3: 100 agents and 5000 relations

This example proves the rule to be right and our model to be consistent with real world results; we can now try a counter example, i.e. a poor relations world, as the one in figure 4; one thousand inhabitants with a total of just five hundred relations.

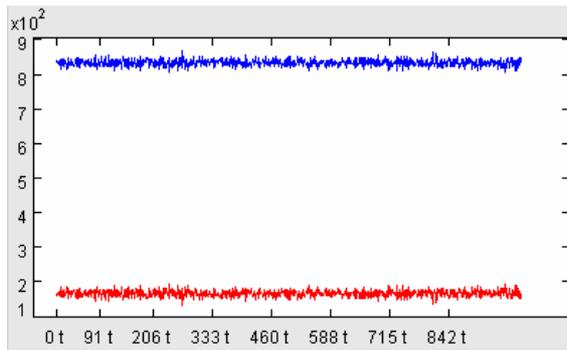


Figure 4: 1000 agents and 500 relations

Here we can observe that less than 20% of the agents changed their opinion. In order to test the extreme situation, we also imagined a world with no relations among the agents (like in the original MG).

Obviously in a world with one thousand unlinked agents we have no changing of opinion (figure 5).

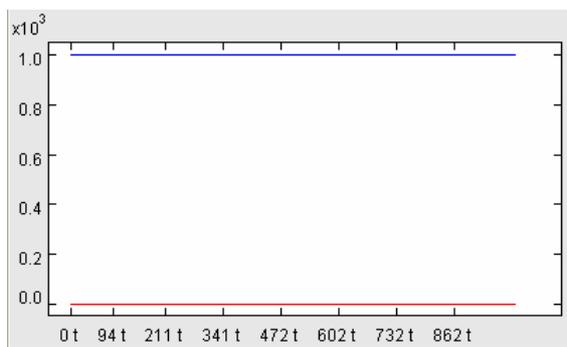


Figure 5: 1000 Agents and Zero Relations

## SYNCHRONOUS COMMUNICATION

A step further is the implementation of a different communication protocol among agents.

The first we used is an asynchronous one: the agents act sequentially. So the first agents to act take a decision, and from then on they reply to the other agents with the new decision taken. We wonder if this method can be realistic, so we decided to explore also a synchronous communication process, which seems more similar to the one we would have in a real world.

Now the agents always communicate to the others their original opinion: they broadcast their opinion to all the agents which are linked to them. Finally, after they collect all the opinions of their friends, they evaluate the certainty index and reconsider their choice.

We executed the simulation with the new rule and the same parameters as before.

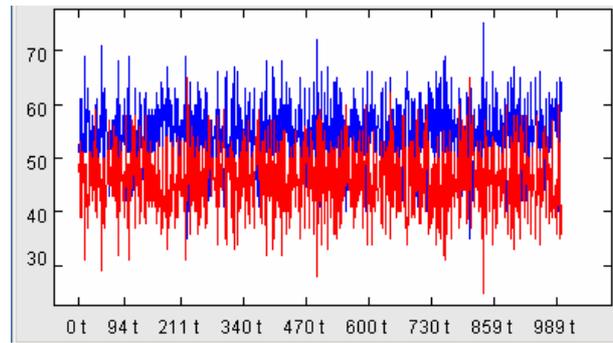


Figure 6: 100 agents and 500 relations

In the first example (figure 6) we have a ten percent more changed opinions, than we had in the sequential model.

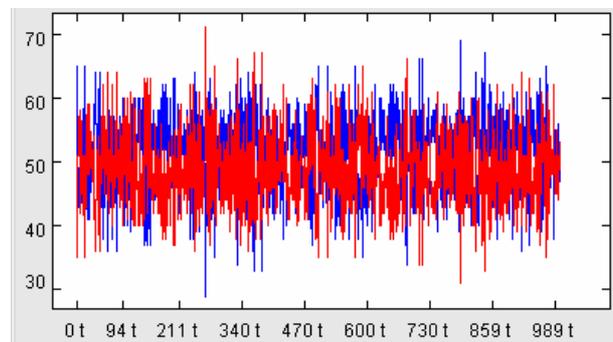


Figure 7: 100 agents and 5000 relations

The best result is in the second run (figure 7): the world rich of relations. The two lines are quite overlapped (even if there is a high variance in data).

We can now express a second simple rule coming from this analysis: a synchronous communication among the agents increases their attitude to change opinion, which is at least ten percent higher.

The proof is the third run, in which again we have an higher result when compared to the asynchronous case.

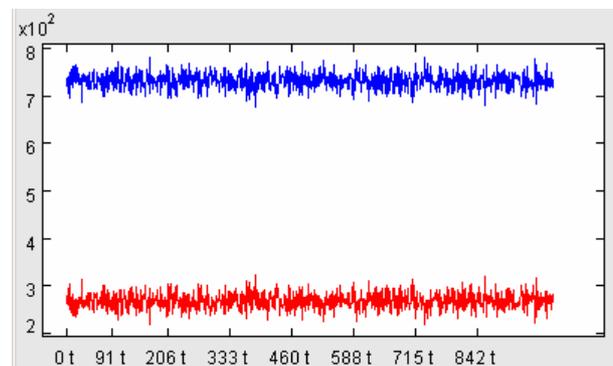


Figure 8: 1000 agents and 500 relations

## CONCLUSIONS

While the original Minority Games states that the agents involved must take a decision based on the historical data, their own experience and the forecasts about what the others will choose, in this paper we introduced communication among them, in order to see how the decision process would change. The stress here is not on the decision taken, be it the best or the worst, but on how the agents can change their decision when they are linked in a social network; in particular, we tried to find the empiric proof to a common sense rule: with a fixed number of agents, the more the links, the higher is the probability to change opinion. We built an agent based simulation, tested some real world parameters and analyzed the results we obtained.

We examined two different communication protocols among the agents: the asynchronous one and the more realistic synchronous one, in order to see how this could affect the way the agents changed their opinions. Using the synchronous communication, the one in which an agent communicates with all the ones linked with him at the same time, we saw that the attitude to change opinion is at least 10% higher than in the asynchronous case, in which the agents act sequentially.

## REFERENCES

- Arthur W.B. 1994, "Inductive Reasoning, Bounded Rationality and the Bar Problem", *Am. Econ. Assoc. Papers and Proc.* 84, 406
- Challet D., Zhang Y.C. 1997, "Emergence of cooperation and organization in an evolutionary game", *Physica A*246, 407
- Gilbert, N. and Terna, P. 2000. "How to build and use agent-based models in social science", *Mind & Society* 1, 57-72
- Laumann E.O. et al. 1978, "Community Structure of Interorganizational Linkages", *Annual Review of Sociology*, 4, pp 455-484
- Morgenstern O., von Neumann J. 1944, "Theory of Games and Economic Behavior", Prince-ton University Press
- Ostrom T. 1988, "Computer simulation: the third symbol system", *Journal of Experimental Social Psychology*, vol. 24, 1998, pp.381-392.
- Remondino M. 2003, "Emergence of Self organization and Search for Optimal Enterprise Structure: AI Evolutionary Methods Applied to ABPS", *ESS03 proceedings*, SCS Europ. Publish. House
- Sonnessa M. 2004, "JAS 1.0: New features", presented at the *SwarmFest2004 conference*, USA, May 9-11, 2004

## AUTHORS BIOGRAPHY



**MARCO REMONDINO** was born in Asti, Italy, and studied Economics at the University of Turin, where he obtained his Master Degree in March, 2001 with 110/110 cum Laude et Menzione and a Thesis in Economical Dynamics. In the same year, he started attending a PhD at the Computer Science Department at the University of Turin, which will last till the end of 2004. His main research interests are Computer Simulation applied to Social Sciences, Enterprise Modeling, Agent Based Simulation, Multi Agent Systems and BDI agents. He has been part of the European team which defined a Unified Language for Enterprise Modeling (UEML). He is also participating to a University project for creating a cluster of computers, to be used for Social Simulation.



**ALESSANDRO CAPPELLINI** was born in Turin, Italy. He studied Economics at the University of Turin, where he obtained his Master Degree in July 2003, with a Thesis in Mathematical Economics concerning stocks market simulation with artificial and natural agents. He started attending a PhD (ending in 2006) on simulation at the University of Turin. His main research interests are Computer Simulation and Experiments applied to Finance, Economics and Social Sciences. He is also interested in behavioural finance, and is a founder of the Italian behavioural finance association ("A.I.FIN.C"). Nowadays he works in Analysis and Strategic Planning office of Sanpaolo IMI Group.

# MASIM: A METHODOLOGY FOR THE DEVELOPMENT OF AGENT-BASED SIMULATIONS

André M. C. Campos

Anne M. P. Canuto

Jorge H. C. Fernandes

Eliane C. M. de Moura

Informatics Applied Mathematics Department  
Federal University of Rio Grande do Norte, Brazil  
{andre,anne,jorge,eliane}@dimap.ufrn.br

## KEYWORDS

Multi-Agent Simulation, Development Methodology.

## ABSTRACT

This paper presents the general aspects that motivated the construction of the MASim methodology, aimed for development of agent-based simulations. MASim employs features common to the development of agent-based software as well as to the development of simulation models. MASim is described in terms of agent-based concepts. It also borrows concepts used in mainstream software engineering process frameworks, defining workflows where users, simulation modelers, software developers, testers and experts of the simulation domain collaborate with the purpose of streamlining the development and reuse of simulations and agent components.

## INTRODUCTION

The development of agent-based software has been largely increased in the last years (Deravi et al. 2003; Canuto et al. 2003). Examples of successful application of agent-based software can be found in electronic commerce, industry, web, etc. As a consequence, several methodologies for developing agent-based software have been proposed to help workers in developing efficient agent-based software. Most of the existing methodologies, such as CommonKADS (Schreiber et al. 2000), GAIA (Wooldridge et al. 2000), TROPOS (Bresciani et al. 2003), MASE (Wood 2000) etc., are mainly focused on the specification of concurrent software components (agents), describing its roles, goals, functions, communication etc. Those methodologies are supposed to be suitable for most agent-based applications.

On the other hand, simulation systems provide the capability of deriving statistically meaningful conclusions for a computer generated synthetic world. Simulations are crucial in providing advice to natural resource managers, for training and management purposes. The use of the

multi-agent paradigm for building simulation leads to a more powerful and user-friendly computer environments often based on parallel processing, being also able to model the spatial data (positions in the environment) of a system, hardly represented in some mathematical modeling approaches. Multi-agent simulations are becoming increasingly relevant in the simulation field (Campos and Hill, 1998), and can be applied to various areas, to simulate social systems (Gilbert and Troitzsch, 1999).

However, the development of simulations is not as usual as information systems development, given that in the design of a simulation it is necessary not only to model the software components themselves, but also the simulation model behind the application. The process of specification, development and calibration of simulation models may require several iterations, demanding stronger approaches to the tasks of verification and validation. Thus, a methodology for simulation should adapt standard software development methods including V&V processes.

As a solution for the aforementioned problem, this paper proposes a methodology to develop agent-based simulations, named MASim – Multi-Agent Simulation Methodology. MASim employs aspects common to the development of agent-based software as well as to the development of simulation models. MASim is described in terms of agent-based simulation concepts (agents, roles, resources, dependencies, interactions, etc). It also borrows concepts used in mainstream software engineering process frameworks, defining workflows where users, simulation modelers, software developers, testers and experts of the simulation domain collaborate with the purpose of streamlining the development and reuse of simulations and agent components. The use of workflow-based processes emerged from the need to better organize the development process of simulation environments. MASim has been applied to the design and implementation of a simulation environment dealing with human organizations, referred to as SimOrg.

The remainder of this paper is divided as follows: Section 2 describes the state of the art in software development methodologies, focusing on the main agent-based development methodologies. It also presents the main differences between methodologies for agent-based development and methodologies for developing simulation models. Section 3 presents the general idea of the proposed methodology. Afterward, roles and phases of the methodology are exposed in Sections 4 and 5 respectively. The last section is dedicated to final remarks on this work and presents further work.

## BACKGROUND

Simulation and reality are at the extremes of a spectrum of systems, and there is a myriad of intermediate situations between them. In general, the goal of a simulation is to model a real system whose nature is sometimes marked by concrete (vs. informational or abstract), physical (vs. symbolic), analog (vs. digital), or continuous (vs. discrete) aspects. In order to obtain this simulated representation, the system must be mapped to the discrete computational domain. This differs from the goal of usual information systems, which is to enhance the manipulation of information already formalized in human-centric organizations. Given this need for translation from a natural to a synthetic domain, simulation models are developed through several iterations, each one producing an enhanced model of a system, that is defined, implemented, verified and validated. This cyclical process goes on until the model satisfies the objectives of the model user (as Minsky says, *“To an observer B, an object A\* is a model of an object A to the extent that B can use A\* to answer questions that interest him about A”* (Minsky 1965)).

Thus, the approach of an iterative simulation development differs from the approach of most information systems engineering processes. Indeed, a simulation is also a software system. The point is that, while in the simulation development process, iterations aim to evolve from a single initial model until a useful (consistent) one, the cyclical iterations in the information systems engineering process aim to achieve new and interrelated system and software functions. This is achieved by an incremental and modular construction of pre-planned chunks, usually driven by user scenarios, like use cases or user stories. In other words, in the development of information systems, the various functions to be developed are generally well known and have a representation close to the information that must be used in the user organization. This way, the information systems usage tends to provoke changes over its user organization. Conversely, when developing simulation applications, the simulation model outcomes are meant to give indications on how close the model is approaching the real system that is of interest to the user. However, the usage of the simulation does not induce or precludes a change to the real system.

The success of a simulation is thus measured on how close it is from the real system it may mimic, while. On the other hand, the success criterion of an information system is how close it is to provoke enhancement on the manipulation of the information model in the user organization.

## Methodologies for the development of Agent-oriented software

In this subsection, some of the existing agent-based development methodologies are described. The main advantages and disadvantages of each methodology are presented, considering its application in development of multi-agent simulation software.

Most agent-based methodologies are proposed for the development of information systems software. Due to this intended application, sometimes, these methodologies lack some important specific aspects of other specific types of applications. In analyzing the advantages and disadvantages of each methodology, this paper aims to understand the degree of suitability of these methodologies for the specific domain of multi-agent simulation.

### *Gaia*

Gaia is a methodology for agent-oriented software analysis and design. The Gaia methodology is both general, in that it is applicable to a wide range of multi-agent systems, and comprehensive, in that it deals with both the macro-level (societal) and the micro-level (agent) aspects of systems (Wooldridge 2000). Gaia is intended to allow an analyst to go systematically from a statement of requirements to a design that is sufficiently detailed that it can be implemented directly. Analysis and design can be thought of as a process of developing increasingly detailed models of the system to be constructed. During the analysis phase, role and interaction models are created and agent, services and acquaintance models are created during the design phase.

The possibility that different agents may be implemented using different programming languages, architectures, and techniques as well as the fact that it deals with both the macro-level (societal) and the micro-level (agent) aspects of systems are very important aspects of this methodology. However, Gaia does not explicitly attempt to deal with systems in which agents may not share common goals (self-interested agents) and in conflict situation. Furthermore, the environment modeled by Gaia is closed and static. For a multi-agent simulation, these aspects are very important and this is a drawback of this methodology.

Nevertheless, even if a multi-agent simulation requires some features not provided by the Gaia methodology, it

could be observed that it is essential only in the macro-level. The idea of roles and its junction into agent types can be very useful for the design of the agents which compose the simulation model.

#### *MASe*

The MASe (Multiagent Systems Engineering) methodology has focused on the development of practical agents (Wood 2000). It defines multi-agent systems in terms of agent classes and their organization. MASe defines its organization in terms of which agents can communicate using conversations. There are basically two phases in MASe: analysis and design. The first phase, Analysis, includes three steps: capturing goals, applying use cases, and refining roles. In the Design phase, it transforms the analysis models into constructs useful for actually implementing the multi-agent system. The Design phase has four steps: creating agent classes, constructing conversations, assembling agent classes, and system design.

The main advantage of this methodology is its simplicity since it defines clearly how a designer should act at any moment of the process of the development of a software. However, MASe lacks some abstractions which are important in the modeling of agent-based software. In this case, it is very similar to an object-oriented methodology and this is an important drawback of this methodology.

#### *TROPOS*

Tropos is a software development methodology allowing a designer to exploit all the flexibility provided by agent-oriented programming (Bresciani et al 2003). TROPOS is intended to support all analysis and design activities in the software development process, from application domain down to system implementation. There are five main development phases of the Tropos methodology: Early requirements, late requirements, architectural design, detailed design and implementation.

One of the main features of Tropos is the crucial role played by the early requirements, which has been added to the phases of an agent-based development process. It helps in the development of agent-based software, including multi-agent simulation. In addition, the way in which the phases are distributed allow the designer to have a great understanding of the system as a whole. Finally, the fact that the environment and system models are being built in an incremental way, being refined and extended at each step, is very attractive for the development of multi-agent simulation. The main drawback of this methodology is the complexity behind the models to be created, which becomes even worse in complex applications.

Due to the aforementioned facts, it can be concluded that Tropos is a very attractive methodology for developing also multi-agent simulations, mainly in a macro-level perspective.

### **Useful characteristics from software engineering methodologies**

The most known generic framework for software engineering is the Unified Process (Jacobson 1999). The unified process is marked by the adoption of the following principles: Iterative; Risk-driven; Requirements based; Architecture based; Visual modeling-based; Continuous quality assurance; and Change management

Furthermore, the Unified process states that there are nine specialization areas, called disciplines, that involve the work of diverse experts. The areas are: Business modeling; Requirements Analysis and Design; Implementation; Test; Deployment; Environment; Project Management; and Configuration and Change Management

The activity inside all disciplines is marked by a structured work around the concepts of roles, tasks and workflows. Some of these disciplines are closely related to technical aspects of simulations and agents, namely: Business modeling, Requirements, Analysis and Design, Implementation and Test.

The point is that in order to streamline the development of simulations, it is important to provide a clear set of paths for development and validation of systems. Given that the methodology is aimed to guide the work of simulation developers, working inside an organization, it is suitable that the methodology has to lead towards an application on the same concepts employed in the development of the simulation.

### **MASIM: A METHODOLOGY FOR DEVELOPING MULTI-AGENT SIMULATIONS**

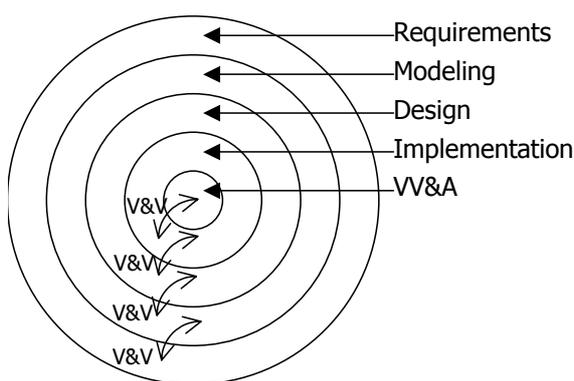
MASim is a process framework specifically focused on the development of multi-agent simulations. Before presenting the methodology it is worth commenting on its purposes and the type of applications it tackles. As multi-agent simulations are well-suited for studying complex-adaptive systems (Gilbert and Troitzsch 1999), the proposed methodology is supposed to be appropriated to develop large-scale applications for simulating such type of systems. It is also intended to provide, through the simulation, foundations for modifying the system itself.

Agent concepts are used in the methodology mainly to model the process roles, the interactions, the dependencies, etc. It considers the individuals running the methodology as components of a multi-agent system, which compose indeed a real organization.

MASim consists of five phases, presented below:

- **The requirements phase**, which consist in identifying the scope of the simulation model as well as the needs of the application for handling such a model;
- **The modeling phase**, which aims to construct an abstract model of the system, presenting its elements and the dependences between themselves and the system as a whole;
- **The architectural and design phase**, which translate the model into a set of concrete specifications able to be easily implemented. This phase also intends to identify patterns to be reused in other simulations of the same domain;
- **The implementation phase**, where the specifications are coded in programming language;
- **The verification, validation and accreditation (VV&A) phase**, which confronts the overall simulation results with the real system and determines if the simulation application is suitable for the initially required purposes.

The whole process is cyclical, as most of the modeling methodologies. However, it also goes forward and backward through the phases, verifying the codification of a specification in a higher level of abstraction and validating the proposed specification, as shown in the Figure 1. This idea of several V&V phases follows the pragmatics observations of Swartout and Balzer, who says that any specification  $S$  might be seen as an implementation of another specification  $S'$  of a higher level of abstraction (Swartout and Balzer 1982). Nonetheless, the last phase is explicitly concerned with V&V in order to confront the produced simulation software with the requirements initially exposed. In other words, it analyzes the implementation of lowest level of abstraction in the perspective of the highest-level specification.



**Figure 1 - Phases of the methodology and their internal verification and validation**

MASim encourages reusability in the architectural and design phase by the architect role. The architect is responsible for defining structures patterns for the whole application and for specifying how those patterns (e.g

organizational structures and interaction protocols) must be implemented.

## MASIM ROLES

In large-scale systems development, it is impractical that the same individual performs all tasks in different phases of the system development. Tasks requiring different abilities must be executed by individuals fulfilling their specific requirements. In order to optimize this process according to the abilities required in its activities, MASim defines seven roles that different individuals might play.

Those roles are also introduced in order to better characterize the objectives and responsibilities of each individual in the process. Nevertheless, for small-scale simulation applications, a unique individual might play different roles. The same happens when developing non agent-based systems. A software process, like RUP, identifies several roles, but this does not mean that there is a person for each role. A typical scenario for the development of simulation systems consists at the minimum a domain expert and a developer. However, if one consider the development of large-scale simulation, it is necessary to involve several individuals, each one playing a specific role in the software development process. An example is some simulation games like SimCity, where several individuals are involved.

MASim preview they following roles:

- **End-user**, an individual (or organization) for whom the simulation application is under development. The end-users are responsible for setting up the requirements for the application. They should target the objective of the simulation and, defining what they expect from the simulation as results;
- **Domain expert**, an individual (or several ones) who has deep knowledge about the domain being simulated;
- **Modeler**, who is responsible for, along with the end-user and the domain expert, to construct the simulation model;
- **Software architect**, who is responsible for defining software patterns and/or simulation model components that might be reused in other simulations as well as defining the whole simulation framework;
- **Designer**, who is responsible for transforming the simulation model constructed by the modeler into a software design, able to be more easily implemented;
- **Developer**, who is responsible for implementing the model designed by the designer;

- **Tester**, who is responsible for verifying and validating the application according to the pre-established scenarios.

Figure 2 shows the whole process of MASim, involving all the roles previously described. The diagram, an adapted version of the workflow diagram of the Unified Process (Jacobson et al. 1999), shows the process in the following steps:

- 1) The process begins when the end-user and the domain expert describe the organization and what is expected from the simulation through a collection of scenarios;
- 2) The scenarios are then used by the modeler and the domain expert to construct the model of the organization (macro level), its individual elements (micro level) as well as their interrelationships.
- 3) The architect identifies patterns in the previous model and defines standards for designing the simulation.

They aim to achieve reusable components of software and simulation model.

- 4) The computational model is constructed by the designer from the organizational model and the standards previously mentioned. The main different between this model and the organizational one is the approach (concept vs. coding-oriented) and their level of detail.
- 5) It is then used by the developer to implement the simulation software, which it will be used by a tester in a process of verification and validation and by the end-user for accreditation.
- 6) The process starts a new iteration, taking into account the results from the current developed simulation.

The next sections present the phases of the methodology and how the previously mentioned activities are involved in each one of them.

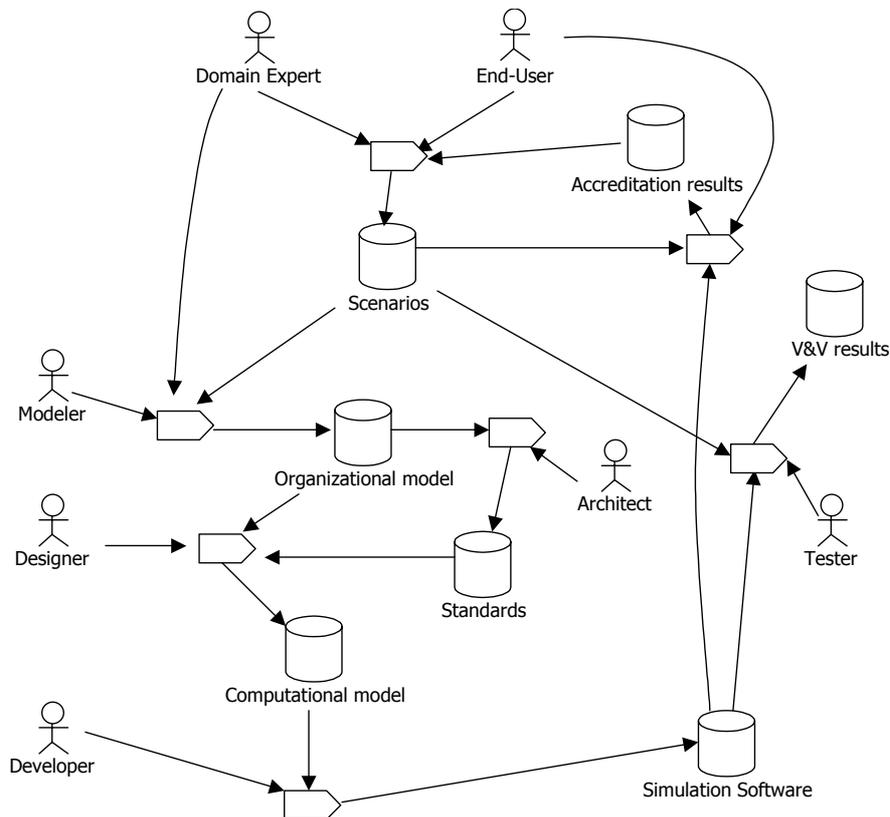


Figure 2 – Workflow showing the activities performed by individuals playing different roles

## MASIM DEVELOPMENT PHASES

### Requirements phase

All requirements in MASim are expressed as scenarios. As in information system requirement analysis, those scenarios

also describe what is expected from the application being developed. However, instead of being based on which changes the application must introduce in the real system, the requirements in MASim are based on what is expected to reproduce from the real system. The observations provided in the scenarios are used to define the scope of the simulation model and application.

As shown in Figure 2, the scenarios might come from two sources: the end-user, who knows what the application must produce, and the domain expert, who knows how to produce. The scenarios are described in an informal way through a schema containing the involved elements, a description of the initial conditions of the elements, the events identified in the scenario case and a description of the conditions of the elements after the events. It is also possible to make some notation about the mechanisms responsible for the condition changes. The Figure 3 shows an example of scenario schema.

Scenario schema: <i>scenario name</i>	Version: <i>num</i>
Involved elements:	
<i>Subsystems involved.</i>	
Pre-conditions:	
<i>Description of the initial situation of the elements</i>	
Events:	
<i>Description of events happening to the elements</i>	
Pos-conditions:	
<i>Description of the final situation, after the events</i>	
Possible mechanisms:	
<i>Description of the mechanism or the hypothesis about the element changes</i>	

**Figure 3 – Schema of a scenario form**

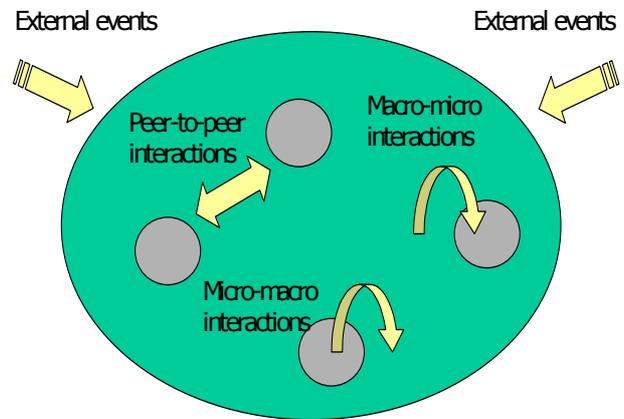
The objective of such schema is not to extensively document all possible scenarios of the organization in order to provide all the information (as the unique input) for the modeling phase. The scenarios are mainly used to define the scope of the simulation model as well as the application handling such a model. It specifies what the activities in the modeling phase must focus on.

In the proposed methodology, requirements are also driven by scenarios describing what it was intended for the application. However, instead of presenting requirements based on changes (early and late requirements), the requirements are based on what it is observed from the real system. The observations compose several scenarios that are used to define the scope of the model.

The scope of the model is defined through four types of scenario. The first one specifies the boundaries of the system being modeled. It describes external events that influence the system as a whole and how the system reacts to them (exogenous events). The other types of scenarios, illustrated in Figure 4, are related to internal interactions in the system (endogenous events), describing:

- **Macro-micro interactions:** this type of scenario reflects how the behavior of the system as a whole changes the behavior of its subsystems.
- **Micro-macro interactions:** on the opposite way, this type of scenario aims to present how individual subsystem activities alter the whole system.

- **Peer-to-peer interactions:** this scenario identifies the dependencies between two or more subsystems and how they influence each other.



**Figure 4 – Scenarios used for defining the scope of the model**

### Modeling phase

The modeling phase consists in identifying, from the scenarios previously specified and face-to-face interactions between the modeler and the domain expert, entities, groups, roles, tasks, activities and dependencies between those elements. Its main aim is to construct a model of the organization in a macro (society) and micro-level (individual).

#### Modeling activities

During the modeling phase, several *views* of the system are constructed, each one provided by a different modeling activity. They are:

- **Resource modeling:** It consists in identifying and modeling the existing roles and actors playing those roles within the system. It is important to emphasize that the roles defined in this activity are not the same as the ones presented in the previous section. While the former represents roles in the methodology, the latter represents roles in the system model. In MASim, roles are abstract concepts representing general activities performed by individual elements. Consider, for instance, that a human organization, like a business company, is being modeled. The company has roles such as CEO, directors, and so on. Those roles are positions with well-established objectives, activities and permissions (or restrictions). They will typically correspond to: individuals or groups (for instance, a department of a company). Roles are then defined in terms of: objectives, activities and permissions. The agent modeling will identify the agents which play one or more roles in the system. Actors can be viewed as concrete instances of the existing roles. There is no actor in the system without playing a specific role

within it. Furthermore, there is no one-to-one mapping between roles and actors. An agent might play several roles as well as a role might be played by several agents. However, agents playing the same role might behave in a different way to achieve the objective of such a role. It depends on the agent beliefs and behavior patterns (mental models).

- **Dependency modeling:** It consists in specifying the dependencies between roles as well as the activities performed by individuals playing different roles. Roles and activities dependencies provide a basis for modeling the organization structure. It defines, for instance, how hierarchical the organization is (role dependencies), how groups are constituted (role dependencies), and how centralized the decision-making processes are (activity dependencies).
- **Interaction modeling:** It consists in defining the set of protocols of interactions used by the actors for accomplishing the objective of their roles. While the dependency modeling activity models what is dependent for a specific activity, the interaction model details how a specific activity is performed. It points out the mechanisms used to transform the resources and how those transformations flow during the execution of the different activities.

### Architectural and Design phase

In this phase, the conceptual model must be detailed according to the programming approach to be used in the next phase.

This phase is mainly concerned by the system architect and the designer. The former is responsible for defining the system global architecture according to previously developed simulation in the same domain. It must identify multi-agent architectural patterns for the software as well as for the simulation model, in terms of component reusability. For instance, consider that a simulation has been developed for the department of marketing of a company in which resources, activities, individuals and roles are similar to a simulation of another company which a simulation has already been developed (notice that the role of the end-users differs from the role of the developers, and so they – end-users – might be clients of a simulation development company – the developers). In this case, several agent-based components must be reused in order to facilitate the simulation development. Reusability, as a key concept of software engineering, is one of the strongest points of the MASim methodology.

The designer takes charge of the specification of the internal design of each component, using modeling languages like AUML (Bauer et al. 2001), where objects, agents, and agent communication protocols are well represented.

### Implementation phase

MASim is intended to be independent of a programming paradigm. However, the approach presented here concerns the object-oriented paradigm due to the lack of enough mature agent-oriented programming languages. As consequence, most of the multi-agent applications currently developed is implemented in an object-oriented approach. Although the existence of several tools has been proposed to implement multi-agent simulations, like MadKit (Gutknecht and Ferber 2000), they fall down in the object-oriented paradigm as they are implemented in object-oriented languages (Java and Smalltalk respectively).

Other approaches might be suitable in case they provide a robust way for transforming the agent-oriented design into a computational language. Among the existing agent-oriented languages, Brahms (Sierhuis et al. 2000) seems to fulfill the requirements for developing a multi-agent simulation where the system is viewed as an organization.

### Verification, Validation and Accreditation phase

Before a simulation model can be used with confidence by the end-user, it must be verified, validated and accredited. Those activities are performed in this phase of the methodology. As mentioned before, verification and validation activities are performed during the various phases, checking errors generated when implementing specifications of a lower level of abstraction (verification) and checking if the specification corresponds to the model of a higher level of abstraction (validation). However, this phase was introduced to link the first specification (requirements) to the last implementation (software).

Verification and Validation can be a complex matter. Indeed, depending on the characteristics of the model, they might become a hard task to perform. Several methods are proposed in the literature, including grounding, calibration and statistical comparisons. For such reasons, those activities are performed by an individual playing a specific role, the tester.

While the V&V process is performed by the tester, the accreditation activity is performed by the end-user. In fact, accreditation refers to evaluate how useful is the simulation application (and model) for the specific purposes it was targeted to. The only person able to execute this task is the end-user.

### MASIM USE CASE: SIMORG

The proposed methodology has been used for the construction of applications and tools for simulation human organizations, through a project named SimOrg - *Simulation of human Organization*. SimOrg aims to define, implement to validate an agent-based computational model

able to represent complex organizational behavior arising from people interactions.

As SimOrg deals with the study of organizational behavior, various experts in organizational psychology are jointly working in the project as domain experts, tackling the following aspects:

- The process of planning, elaboration and evaluation of working flows in order to describe and systematizes the efficacy required in organizational positions and functions;
- The process of recruiting new collaborators, including the usage of evaluation methods and techniques, whose objective is to assist in the identification of most adequate candidates for well-defined functions within the organization;
- The process of moving people inside an organization taking into account the current context of the organization, the individual antecedents and perspectives as well as their psychological and motivational aspects.

The simulations provided by SimOrg will provide strong basis for the validation of theories dealing with the mentioned aspects, where profiles, processes and group dynamics must be analyzed. The final objective is to estimate how they can be defined in an organization in order to provide a well-functioning organizational structure with a better productivity in a long-term perspective.

The SimOrg project is starting its second year of execution. MASim is being used in SimOrg as a general methodological framework for producing the simulations required by such a project.

## FINAL REMARKS AND FURTHER WORK

In this paper, we have described MASim, a methodology for developing multi-agent simulations. MASim proposes a system development framework employing common aspects of agent-based software development and simulation modeling field. For this, it borrows concepts used in mainstream software engineering process frameworks, defining workflows where users, simulation modelers, software developers, testers and domain experts.

MASim is based on the authors' background in developing multi-agent simulations (Hill et al. 2000) and is currently being used in the development of a simulation environment dealing with human organizations, referred to as SimOrg. The first year of the project execution gave us valuable feedback about the usefulness of the methodology. Nevertheless, the methodology has not been fully validated and several issues in the agent-based modeling practices are not supported yet. One of the main missing points in its current version is that it does not provide enough directives nor formal specifications for the organizational model. In

next versions of the methodology, it is intended to fulfill this gap through the support to represent all the agent-based concepts described in Section 5.2. More specifically, it is intended to provide tools to formally describe roles, objectives, permissions, activities, agents, beliefs, behavior patterns, dependencies and interactions.

## ACKNOWLEDGEMENTS

This work has the financial support of CNPq (Brazilian Research Council), under process number 552431/2002-8.

## REFERENCES

- Bauer, B., Müller, J., Odell, J. "Agent UML: A Formalism for Specifying Multiagent Interaction". In *Agent-Oriented Software Engineering*, pp.91-103. Springer-Verlag, Berlin, 2001.
- Campos, A. and Hill, D., "An Agent-Based Framework for Visual-Interactive Ecosystem Simulations". In *Transaction of SCS*, 15(4):139-152, December, 1998.
- Campos, A., "Perceptive agents: modeling non-intentional interactions". *Proceedings of the 5<sup>th</sup> World Multiconference on Systemic, Cybernetics, and Informatics*. Orlando, FL, 2001.
- Canuto, A., Gottgroy, M., Lucena, M., Bezerra, V., Medeiros Jr., J., Oliveira, A., "RetDat: A multi-agent architecture for the retrieval of information in heterogeneous databases". *Proceedings of the 7<sup>th</sup> IASTED International Conference on Artificial Intelligence and Soft Computing*, pp.321-326, 2003.
- Deravi, F., Fairhurst, M., Guest, R., Canuto, A., Mavity, N., "Intelligent Agents for the Management of Complexity in Multimodal Biometrics". In *Journal of Universal Access in the Information Society*. Springer-Verlag, 2(4):293-304, 2003.
- Gilbert, N. and Troitzsch, K.G., *Simulation for the Social Scientist*. Open University Press. 1999.
- Gutknecht, O., Ferber, J., "MadKit: a generic multi-agent platform". *Proceedings of the fourth international conference on Autonomous Agents*, pp.78-79, 2000.
- Hill D., Mechoud S., Campos A., Coquillard P., Gueugnot J., Orth D., Michelin Y., Christophe P., L'Homme G., Carrère P., Lafarge M., Loiseau P., Micol D., Brun J.-P. Decuq F., Dumont B., Petit M., Teuma M., "Modélisation de l'entretien du paysage par des herbivores en moyenne montagne: une approche multi-agents". In *Ingénieries – ETA*, 21:63-75. March, 2000.
- Jacobson, I., Booch, G., Rumbaugh, J. *The Unified Software Development Process*. Addison-Wesley, 1999.

- Minsky, M. "Matter, Mind and Models". Proceeding of International Federation of Information Processing Congress, vol. 1, pp.45-49, 1965.
- Bresciani P., Giorgini P., Giunchiglia F., Mylopoulos J. and Perini A.. "TROPOS: An Agent-Oriented Software Development Methodology". In *Journal of Autonomous Agents and Multi-Agent Systems*. Kluwer Academic Publishers, 2003.
- Schreiber, A., et al. *Engineering of Knowledge and Management: The COMMONKADS Methodology*, MIT Press. 2000.
- Sierhuis, M., Clancey, W., Hoof, R., Hoog, R., "Modeling and Simulating Human Activity". Proceedings of Autonomous Agents 2000 workshop on Intelligent Agents for Computer Supported Cooperative Work: Technology and Risks (Ed, Petsch, M.) Barcelona, Spain. 2000.
- Swartout W., Balzer R., "On the Inevitable Intertwining of Specification and Implementation". In *Communications of ACM*, 25(7):438-440. July 1982.
- Wood, M. F. *Multiagent Systems Engineering: A Methodology for Analysis and Design of Multiagent Systems*, Air Force Institute of Technology – AFIT. Master thesis, 2000.
- Wooldridge, M., Jennigns, N.R. e Kinny, D. "The Gaia methodology for agent-oriented analysis and design". In *Journal of Autonomous Agents and Multi-Agent Systems*. 3(3):285-312. 2000.



# **MICROSIMULATION**



# CALCULATION OF THE RISK PREMIUM IN A SELF-SUSTAINING, INCOME-CONTINGENT STUDENT LOAN SYSTEM

Edina Berlinger PhD.  
Tamás Makara PhD.  
Department of Finance  
Budapest University of Economic Sciences and Public Affairs  
H-1093 Budapest, Fővám tér 8., Hungary  
E-mail: [edina.berlinger@bkae.hu](mailto:edina.berlinger@bkae.hu)  
[t.makara@alarmix.net](mailto:t.makara@alarmix.net)

## KEYWORDS

Income contingent student loan, Risk premium, Microsimulation, Income paths, „Top-down“ model,

## ABSTRACT

In order to maintain the financial stability of the zero-profit student loan system continuous control and periodical intervention is inevitable. In this article we focus on the problem of the calculation of the risk premium, which assure the self-sustaining operation, when the repayment rate is given. We introduce a so called “top-down” simulation technique to create individual income paths which is quite simple to use and fits well the available cross-sectional database. We conclude that in a society where individuals income relative to others can easily change in time the risk premium of the student loans can be much lower.

## SELF-SUSTAINING, INCOME-CONTINGENT STUDENT LOAN SYSTEM

In 2001 a new institution was implemented in Hungary whose aim is to provide loans to the students of higher education. The repayment is income contingent (ICL-scheme), conditions (e.g. eligibility, interest rate, repayment rate, maximum allowance per month) are the same for everybody and it is declared that the system must operate in a self-sustaining (zero-profit) way, which means that the default risks and operating costs should be financed by the risk cohort of the debtors so in principle it must work without any direct state subsidy. Almost every student borrows up to the maximum possible amount of the loan, which is administered on individual accounts and the actual interest rate starts to accumulate. The repayment period begins just after graduation. There is always one interest rate and repayment rate valid for everybody but these parameters can be changed year to year. The income-contingent repayment lasts until full repayment or retirement (in case of early death or disability debt is cancelled).

Until now the Student loan Centre Plc. made over 170 000 contracts and granted loans of over 70 billion HUF. Estimations show that the mature system (in 20-

25 years) can be compared to the biggest retail banks in Hungary considering number of clients and aggregate loan stock.

The origin of this Hungarian model goes back to Milton Friedman who was the first to introduce the idea of a self-sustaining, income contingent student loan system (Friedman 1962). Several authors joined to this thought (Cohn and Geske 1996] but it was tested in practice only in the 70's in the United States when some universities (Duke, Yale, Harvard etc.) set up income-contingent plans. The most famous example is the TPO-plan at Yale University designed by James Tobin. In this special model one had to pay a fixed percent of her/his income until the whole debt of the cohort was repaid. Later, differences in income led to such a cross-financing effect that wasn't acceptable any more, so in 1999 debts were cancelled and the system was abolished. The failure of Tobin's model contributes to the belief that an income-contingent student loan system cannot work in a self-sustaining manner.

The first national ICL system was created by the Australian government in 1989, New Zealand and Great-Britain followed suit in 1992 and 1996 (Chapman 2002; Barr 2001). These student loan systems are considered fundamentally successful, but the plans are considerably subsidised (interest rates are often lower than corresponding treasury bond yields), not surprisingly they go along with huge government-expenditure. Another difference relative to the Tobin's model is that debt is registered on individual, rather than cohort level, so cross financing cannot be so high.

If the Hungarian model proves to be sustainable it can serve as an example for other developing countries or countries in transition whose governments cannot increase significantly the budget deficit.

## MICROSIMULATION OF INDIVIDUAL INCOME PATHS – A TOP-DOWN MODEL

The field of dynamic micro simulation originates from a paper by Orcutt (Orcutt 1957). He suggested the development of simulation models using micro-agents for policy use. In recent decades with the development

of information technology micro simulation models were elaborated in a lot of countries (USA, Canada, UK, France, Australia) for several purposes (i.e. analysis of redistribution effect of the tax system, pension and health care systems, financing of education etc.) (Harding 1996).

In these models hundreds of stochastic equations and deterministic algorithms are used to represent complex life events. Events are mainly represented by transition matrices or multi-nominal logit relations which are supposed to be dependent on the actual status of the individual or in the simpler models are constant over time. Events influencing job history and income path can be: carrier effect, job change, unemployment, disability, geographical mobility etc. Wage equations can include age, race, sex, education, experience, marital status as explanatory variables with the residual randomized to create some earnings mobility. These residuals represent nonsystematic effects and are independent from one individual to another.

The main advantage of this kind of modeling is that it allows to capture the full distributional impact of some policies, whose full effects take a considerable amount of time to filter through. However model building requires enormous computing, data and manpower resources. Another problem is that greater complexity increases the risk that the model functions as a 'black box' and the validation of the model also requires special considerations.

In the case of student loaning it is quite straightforward that we should use micro simulation techniques. *First*, if everybody would be exactly like a representative agent, there would be no problem at all. Every time we use averages and not the whole distribution of income we under- or overestimate the risk-premium needed to self-sustaining operation. *Second*, the task is relatively simple: we have to simulate the basis of the repayment: the official yearly gross income. We do not have to take into consideration black revenues or the effects of taxes or social insurance. Third, according to the government decree, the Hungarian Student Loan Centre Plc. is annually obliged to calculate the zero-profit risk premium, so it's worth to make an effort to construct a complex model. And *finally*, after some decades of operation the student loan company will have all the representative panel data on the income of graduates.

Actually we face some fundamental problems:

- We do not have enough data to evaluate regression equations yet.
- In our economy in transition we do not know how the present relations and relevant macroeconomic factors can evolve in a medium or a long time.

We were searching for a micro simulation model which doesn't need a huge and detailed database but captures the relevant characteristics of the income paths of

individuals and can be calibrated to some available cross-sectional information about current income-distribution.

The main characteristics of income paths in reality are:

- High diversity.
- Positive trend (nominal decreasing is less probable than increasing).
- Positive autocorrelation along the income path.
- Special carrier pattern (income increase more in the first years than later).
- Lognormal-nature of cross-sectional income distribution.

The available database, we had:

Survey on graduates' income in 2001. From OMMK (National Labour Centre).

The steps of the simulation:

### Simulating Income Paths (using Mathematica)

1. We simulate only one generation, because we do not want to allow for intergenerational redistribution. It means that in theory every generation should operate in a self-sustaining manner.
2. The simulation operates in discrete time. One period is one year.
3. The  $t=0$  point represents the minute just after graduation. At the end of the first year every debtor is 23 years old. The model covers  $N=39$  years until the retirement age (62 years).
4. At the beginning of the repayment period the generation consists of  $Q_0=10\ 000$  individuals, who have the same amount of debt:  $H_0=2.4$  millions of HUF.
5. Every year,  $d$  percent of the debtors disappears definitely (died, disabled, emigrated etc.). Their total debt is cancelled. While probability of death and disability can be known precisely, other types of disappearance are rather difficult to estimate and they have much larger effect. Taking  $d$  constant the number of the debtors in the cohort in the  $t$ -th year ( $Q_t$ ) equals:
 
$$Q_t = Q_0 \cdot (1 - d)^t$$
 We used  $d=1\%$ .
6. Those who have not disappeared can be inactive, unemployed or working. It is only  $w=80\%$  of the remaining population who work.
7. Inactive and unemployed debtors always have to pay  $\alpha=6\%$  of the official minimum wage

( $M_t$ , which is currently  $M_0=53 \cdot 12=636$  thousands of HUF per year). The minimum wage is assumed to increase by  $a=1,071$  year by year just like the average nominal income of all employees. (Which comes from 5% inflation and 2% real income growth.)

$$M_t = M_0 \cdot a^t$$

8. We suppose that the aggregate cross-sectional distributions of income above the actual minimum wage ( $M_t$ ) follow lognormal distribution with mean  $\mu_t$  and standard deviation  $\sigma_t$ .
9. We estimated average carrier growth factors of the  $t$ -th year ( $c_t$ ) emanating exclusively from the advancement of the carrier (promotion or job change) using the OMMK cross-sectional database. The standard deviations were estimated the same way. We supposed a steady-state world in the sense that we assumed that the carrier path remains the same.
10. The essential point of our model is that the  $t$ -th year's gross income of the  $i$ -th individual is determined by his income rank,  $k_{i,t}$ , which is a natural number between 1 and  $Q_t$ . The individual with income rank 1 has the lowest income of the population and the individual with income rank  $Q_t$  has the highest income. The 20 percent of the population with the lowest income ranks are inactive or unemployed and they earn the minimum wage,  $M_t$ . The  $t$ -th year's gross income of the individuals who work is determined by the following function ( $l$ ):

$$B_{i,t} = \begin{cases} M_t, & k_{i,t} \leq (1-w)Q_t \\ M_t + l(k_{i,t}, q_t, \mu_t - M_t, \sigma_t), & k_{i,t} > (1-w)Q_t \end{cases}$$

$i=1,2,\dots,q_t$

where  $l$  represents the discretisation of the lognormal distribution with mean  $\mu_t - M_t$  and standard deviation  $\sigma_t$  and  $q_t = wQ_t$ , the number of people who work.

11. The mean income of the working population in the  $t$ -th year ( $\mu_t$ ) is determined as follows:

$$\mu_t = \mu_0 \cdot \prod_{j=1}^t a \cdot c_j$$

and  $\mu_0$  (the mean of the average income above the minimum wage at time  $t=0$ ) is  $110 \cdot 53 \cdot 12 = 57 \cdot 12 = 684$  thousands of HUF.

12. In the  $t$ -th year, the  $k_i$  rankings come from a special stochastic "reshuffling" process. It is quite probable that the ranking of a given individual changes a little bit every year, but there is little chance to have a big jump. We use two kinds of normally distributed random numbers: one for the noise around the latest ranking, and another for the big jumps that represent significant changes in the individuals carrier or social status. We use a Bernoulli random number for signaling if there is jump or not. Adjusting the parameters of the random numbers' distributions we can achieve high or low variation of rankings and so we can analyse the effects of the income-variability.

This model can easily be generalized by using parameters depending on time, age, sex, industry etc. One can replace the lognormal distribution with another – empirical or hypothetic – distribution. It seems however that already in this simplistic form, this model is able to catch all the main characteristics of income paths mentioned above. In a few years we will have enough data to further refine the parameters and to bring the model to the real world phenomena. Figure 1. and Figure 2. shows some income paths given by the model in four special cases.

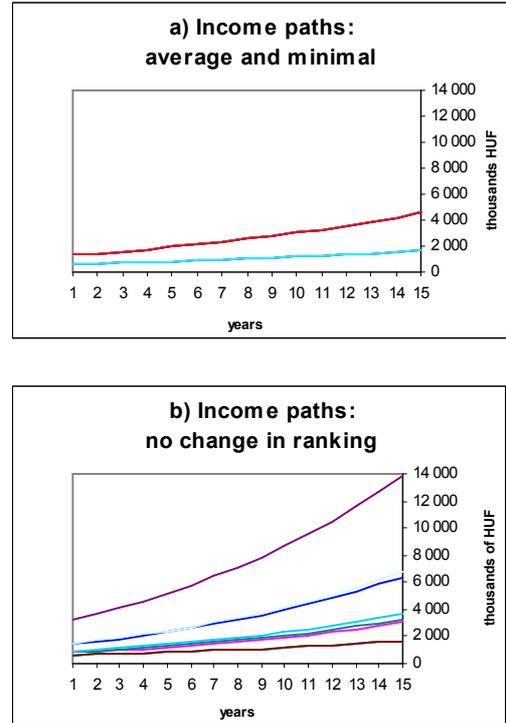


Figure 1: Income paths without reshuffling

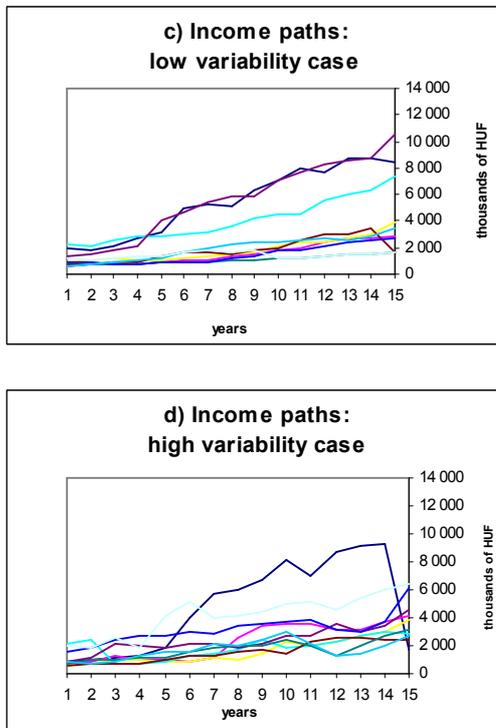


Figure 2: Income paths with reshuffling

- Case a): The official risk premium calculation takes the assumption, that there is only two status of every debtor still in the cohort: average wage and minimal wage.
- Case b): We introduce the lognormal distribution of incomes, but still disregard the “reshuffling” effect: everybody preserves her/his initial ranking.
- Case c): Lognormal distribution with light “reshuffling” effect.
- Case d): Lognormal distribution with strong “reshuffling” effect.

The logic of this and the traditional income simulation are quite different. We accepted the available cross-sectional income distribution relevant and stable but we did not examine what forces had created it. This is why

we call this model “top-down”: we start with the aggregate relationship and force individuals to match it. In traditional microsimulations the distribution of incomes is only the result of the complicated correlations estimated from long panel data (“bottom-up”). In traditional microsimulations individuals have independent noises in their incomes, here ranking is noisy as well, but a kind of interdependence still exist.

### The Calculation of the Risk Premium

We had 10 000 income paths in each of the four cases. Keeping these income paths fixed we examined the effect of possible risk premiums on the net profit/loss of the Student Loan Company. Profits or losses are expressed in HUF and in present value. (The present value of the whole debt is  $Q_0 \cdot H_0 = 24$  billions of HUF.)

1. Calculating individual debts year by year ( $H_{i,t}$ ) along every income path.

$$H_{i,t} = \max(r \cdot H_{i,t-1} - \alpha \cdot B_{i,t}; 0)$$

where  $r$  is the interest factor of the student loan, which consists of two elements: factor of the cost of financing ( $f$ ) and risk premium ( $p$ ):  $r = f + p$ . We assumed  $f=1,071$ .

2. Calculating individual repayment cash-flows ( $C_{i,t}$ ) along every income paths.

$$C_{i,t} = \min(r \cdot H_{i,t-1}; \alpha \cdot B_{i,t})$$

3. Calculating aggregate profit/loss of the lender ( $\pi$ ):

$$\pi = \sum_{i=1}^{Q_0} \sum_{t=1}^N C_{i,t} \cdot f^{-t} - H_0 \cdot Q_0$$

Results are summarized in the Figure 3.

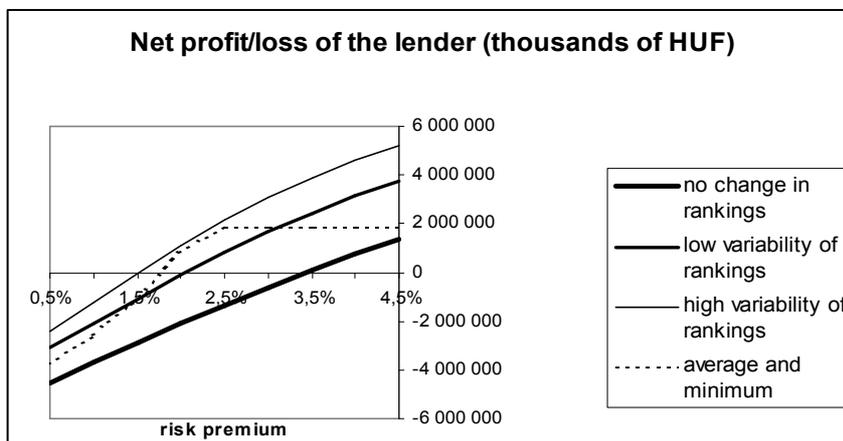


Figure 3: Net profit/loss of the lender

The lender wants to operate on zero-profit level, thus we can numerically determine the risk premium where profit is just zero. Table 1. shows zero-profit risk premia in cases a)-b)-c)-d).

Table 1: Zero-profit risk premia

average and minimal	1,79%
no change in ranking	3,44%
low variability case	2,07%
high variability case	1,51%

It is interesting that taking income distribution without “reshuffling” effect causes as high risk premium as 3,44%. Inserting the “reshuffling” effect will reduce the risk premium to 2,07% and 1,51% depending on the variability. We can summarize that from the point of view of the lender the most favorable situation is when the standard deviation of the cross-sectional income distribution is low, but the reshuffling of rankings is high. We can draw another conclusion as well: as current practice disregards both the cross-sectional standard deviation and the reshuffling of rankings it can happen that the two effects roughly compensate for each other and the official estimate of the risk premium is quite accurate.

## REFERENCES

- Barr, N. 2001. *The Welfare State as Piggy Bank*. Oxford University Press.
- Chapman, B. and Ryan, C. 2002. “Income Contingent Financing of Students Charges for Higher Education: Assessing the Australian Innovation.” Australian National University, Centre for Economic Policy Research, Discussion Paper No. 449.
- Cohn, E. and Geske, T.G. 1990. *The Economics of Education*. 3rd edition, Pergamon Press, 1990.
- Friedman, M. 1962. *Capitalism and Freedom*. University of Chicago Press.
- Harding, A. 1995. “Financing higher education: an assessment of income-contingent loan options and repayment patterns over the life cycle.” *Education Economics*, 3, 173-203.
- Harding, A. 1996. *Microsimulation and Public Policy, Contributions to Economic Analysis*. Vol 221, Amsterdam: North Holland
- Jain, S. K. and Wagner, H.M. 1975. “Comparative Analysis of Income Contingent Plans.” Northwestern University CMS-EMS Discussion Paper N.134.
- Orcutt, G. H. 1957. “A New type of Socio-Economic Systems.” *Review of Economics and Statistics*, 58. 773-797.
- Simonovits, A. 1992. “Indexált kölcsönök és várakozások matematikai elemzése.” *Közgazdasági Szemle*, 39. 262-278.
- Woodhall, M. 1992. “Student loans in developing countries: feasibility, experience and prospects for reform.” *Higher Education*, 23, 4.

# DEVELOPING A MICROSIMULATION SERVICE SYSTEM

Joseph Csicsman and Cecilia Fenyés  
Department of Information and Knowledge Management  
Budapest University of Technology and Economics  
H-1111, Budapest, Hungary  
*E-mail:* [csicsman@calculus.hu](mailto:csicsman@calculus.hu)  
[fenyesc@itm.bme.hu](mailto:fenyesc@itm.bme.hu)

## KEYWORDS

Microsimulation, simulation, household-statistics.

## ABSTRACT

### The term of microsimulation

The microsimulation procedure examines social and economic changes by assessing the effect of each provision with small units and the description of the overall effects is derived from these assessments. Relevance of the results relating to the society as a whole is ensured by the database which is a national representative sample of the units or households of such size that guarantees the required statistical reliability. Naturally, the range of social and economic changes that can be modelled in this way is defined by the information available on the microsimulation units in the database.

### Objectives of the project

In the past few years the Hungarian Central Statistical Office (KSH), the Ministry of Economic Affairs (GM), the Ministry of Finance (PM) and other departments of the government have had no opportunity to analyse large data sets to establish policy proposals.

The available analytical systems – usually based on EXCEL – are, in methodological respect, questionable,

because it is hard to accept the reports based on small data sets both in mathematical and in economical sense.

A research group was founded for the project jointly by the Research Centre for Financial Economics of the Budapest University of Technology and Economics and Új Calculus Bt. The aim of the project is to develop and maintain a model system which allows analysing data collections and databases available in the administration for economists. Naturally, this methodology can also be used for processing other types of data too.

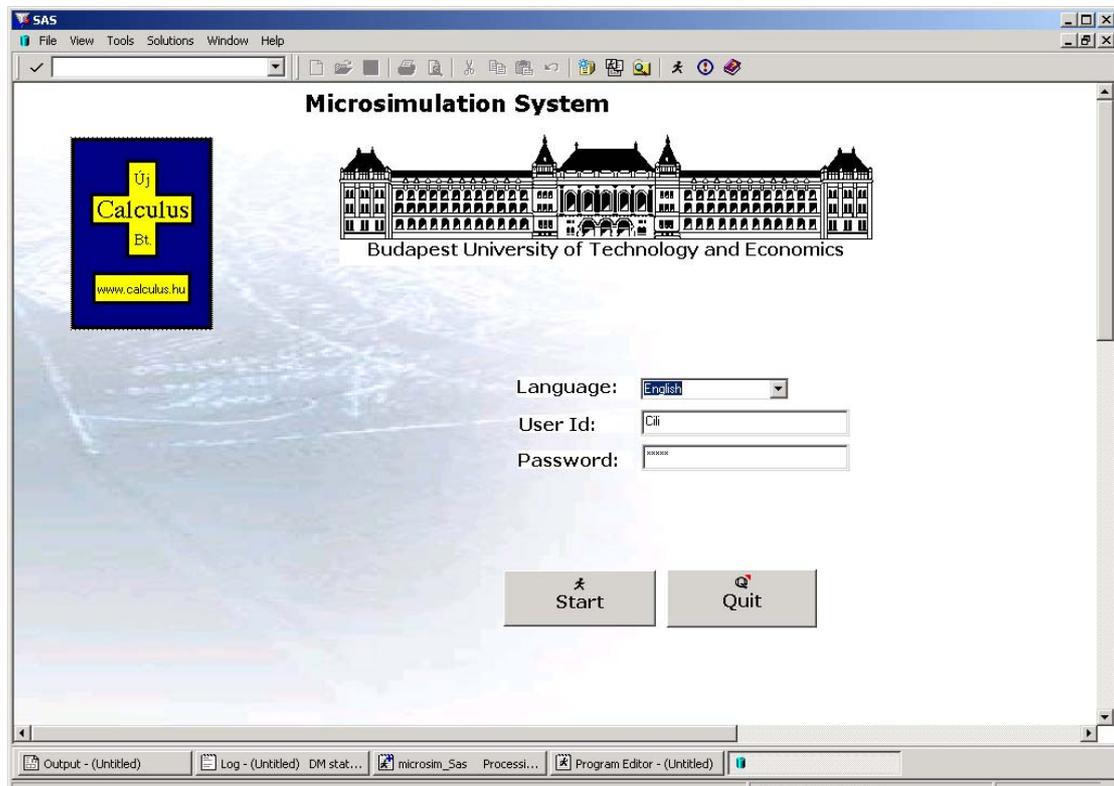
In accordance with the international applications the research group creates technical and methodological conditions for analysing large data sets and developing a SAS Software based microsimulation modelling system.

Till now a methodology to assess the impact of government programs has not been available. By utilizing research experience of the university more exact analyses could be prepared to qualify and quantify policy options.

As an outcome of this project a Microsimulation Modelling System will be developed which will be suitable for modelling the decisions of economic and social policy, and – complying with the national and international requirements – well-founded analyses can support the policy proposals of the government.

The following description will present the recent results of the project in a framework in which great number of university theses have been developed.

# THE MICROSIMULATION SERVICE SYSTEM

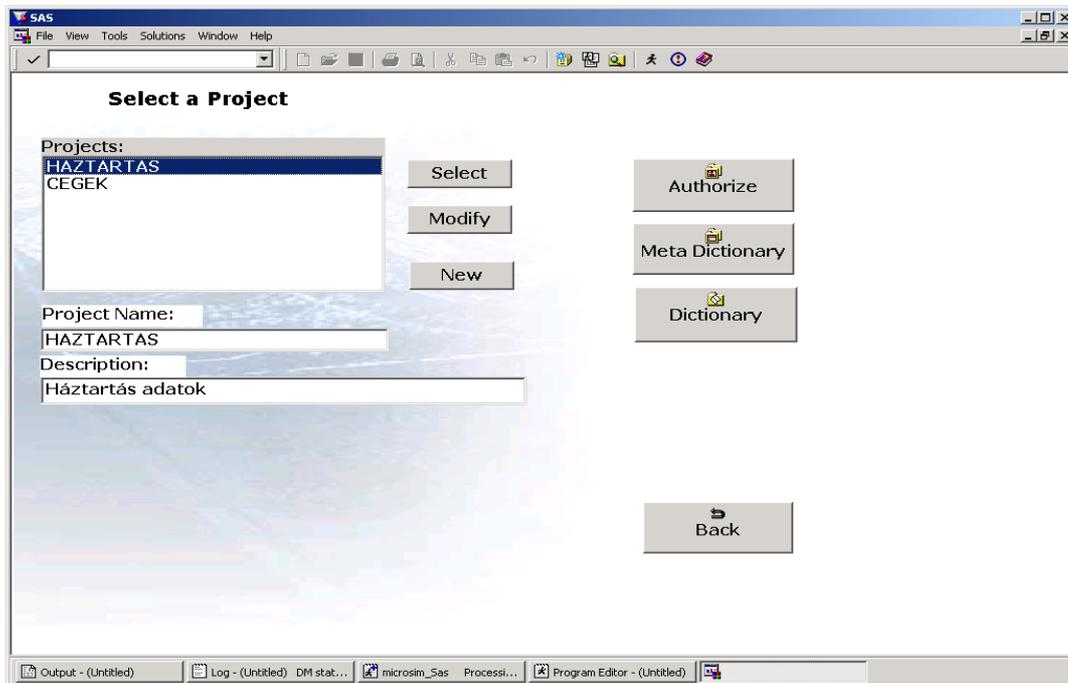


As it can be seen in the Main menu, the language of the program is adjustable. At the moment English and Hungarian versions are available, but because the system works from a dictionary and the compilation doesn't need any special knowledge (only the

command of the language), it is very easy to translate it into any other language.

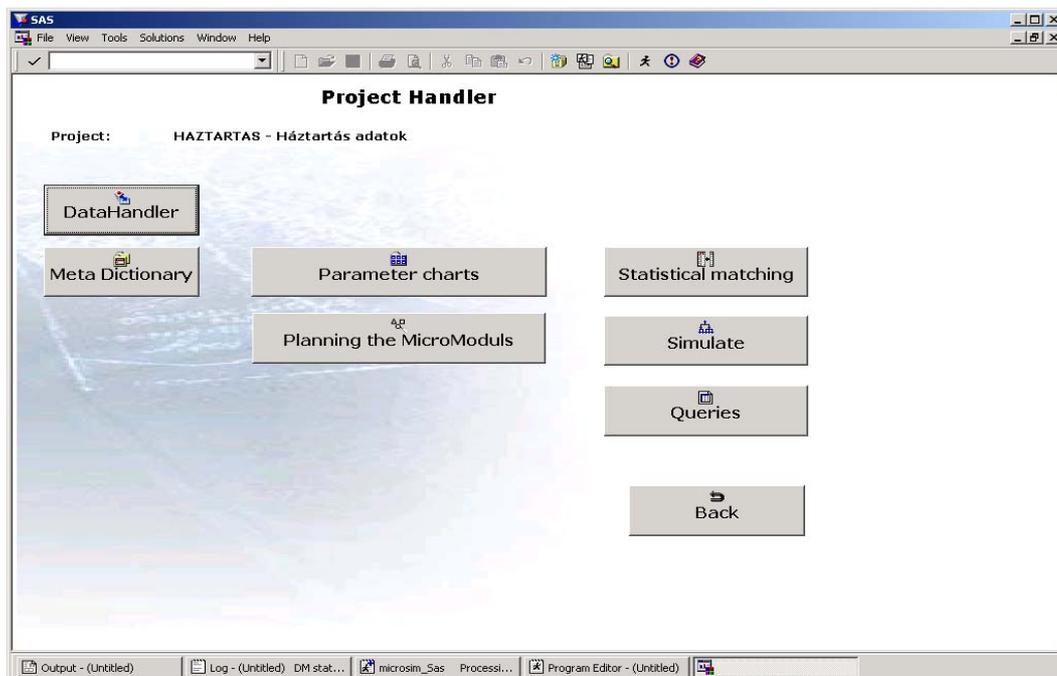
## Project selection

Prior to running a particular simulation a Project is to be selected (or created, if it does not exist), because the next steps are related to the data we want to use.



Without selecting a project 3 tasks can be chosen. By selecting 'Authorize' rights of the users can be set. By pushing the button 'Meta dictionary' the data's of meta can be read or set, which are independent in any project

(nomenclatures and numeric values). By pressing Dictionary the dataset of languages can be found. After selecting or creating the Project or selecting a task, the Project Handling frame will appear:



## Datahandling

### Manipulating the input data

The system can handle text form and SAS files. Expanding the types of importable external files is one of the tasks of the next developing period.

### Data protection

Due to the public environment use of a password entry and different levels of users is unavoidable.

## Meta Dictionary

The Meta Dictionary contains all the information about the data and datasets: identifier, type, length and name of nomenclatures and pointers, structures of input datasets. Also the file catalog is the part of the meta. On the following screenshot the list of nomenclatures can be seen with a user friendly screen to view or

Project: HAZTARTAS - Háztartás adatok  
Structure: SZEMELY

Name of chart: EARNINGROW - Modeling the growth of earningsd

Dimensions	Dimension's items	Vector datas	
MEGYE - Code of county	BUDAPEST	Distribution	Value
AGECAT - Categories by ages	26 - 45		
		0.6	0.04
		0.2	0.06
		0.1	0.1
		0.1	0.2

Buttons: Save, Save as ..., View chart, Back, Add, Delete, OK

modify it. Numeric values, structures of datasets, and the file catalog will be shown in similar frames.

## Estimation algorithms

The parameter charts of estimation algorithms can be filled with the help of a graphical user interface, so economists can determine the internal algorithms without any SAS programming knowledge.

One of our most important goals is to complete these mathematical algorithms to provide opportunity for making appropriate analysis.

One of the frames, which helps to fill in the 'Value assignment upon distribution' parameter chart can be seen below:

## Micromodules

Micromodules are Base Sas codes. During the simulation the selected micromodules will run on every record of the input dataset. This means that micromodules set the changes, which will occur. Here

we prepared the platform for making micromodules without the knowledge of Base Sas. The code will be generated from the rows of modul steps (down, left) made by the user using almost only the mouse.

Project: HAZTARTAS - Háztartás adatok  
 Structure: SZEMELY

## Micromodul modification

Micromodul:  Description:   Use technical file

Called step

Stepnumb:   
 Identifier:   
 Description:

Expression:   
 True:

**Steptype**

1 - Simple condition  
 2 - Complex condition  
 3 - Set value

**Operators**

AND  
OR  
NOT  
<  
>  
=  
+  
-  
\*  
\*\*  
/  
(  
)  
||  
: SUM

**Input variables**

MEGYE -- Code of county  
 RND -- Véletlen szám  
 RUN --  
 SBKER1 -- Főállású bruttó  
 SCSAP2 -- Family status  
 SFEDR1 -- FEOR number  
 SNEME2 -- Sex  
 SNYUG1 -- Nyugdíj, járulé  
 SNYUJ1 -- Le: nyugdíjjárul

**Set of values**

Called stepid:

Value:

Text:

**Estimating tables**

TANULMPENZ  
 FIZEMELES  
 KER\_96\_03

**Global variables**

DEPENDENTS -- Number  
 AGE -- The age of the p

**Local variables**

**Technical variables**

Stepnumber	Step ID	Steptype	Expression
20	b	1 - Egyszerű feltétel	SBKER1 >650000 AND SBKER1 <1350000
30	c	1 - Egyszerű feltétel	SBKER1 >1350000
40	d	1 - Egyszerű feltétel	SBKER1 >0 AND SBKER1 <2368850
50	e	1 - Egyszerű feltétel	SBKER1 >=2368850

### Running the simulation

This frame is to set all the parameters of the simulation. Input file(s) can be chosen, name of the output file must

be set, and here the micromodules can be selected to run from the list of the premade micromodules.

**SAS**

File View Tools Solutions Window Help

**Versions: Set parameters and Run**

Project Name:

Versions:   Version:    Works with a Structured File Actual Year:

**Files**

Layer:   
 Input file:   
 Output file:   
 Technical file:   
 Structures:

**MicroModuls**

Layers:    Run Before Next  Run After Next

MicroModuls:  Selected:

Output - (Untitled) Log - (Untitled) DM stat... microsim\_Sas Processi... Program Editor - (Untitled) C:\WINNT\system32

## Analysing

After running the simulation it is very important to have opportunity to analyse the input and output data. The

## Statistical Matching

Statistical Matching is to pair the records of two data sets without having any key variable. The records of the secondary data sets are separated into groups by their selected attributes (they can be defined in the section of parameter charts, by selecting 'statistical matching - parameter chart of teams' from the list of type of table). The statistical matching goes through on every records

Analyse function of the system can help analysts, who are not experienced in SAS programming. For Analysis the procedures of SAS can be used very well.

of the primary data set. By the attributes of the record the member of the appropriate group (which is a record of the secondary data set) will be paired with this record.

If there is only one data set and the records of the same dataset are intended to be paired (like the simulation of marriage), it is also solved in the microsimulation system.

Projekt: HAZTARTAS - Háztartás adatok **Statistical Matching**

Select datasets and parameter charts

Base dataset (D1) személy

Secondary dataset (D2) haztartas

Output dataset test

Name of output structure

- New test  - Select

Parameter chart of teams

Distribution chart

Repeated  Yes \ No

Import existing structure

Select columns from input datasets

Columns of base ds	Columns of secondary ds	Operators	Number
SBKER1	D4V	+	
SNYUG1	HCSP01	-	
SQSSZ1	D5V	*	
SZJAD1	HLANM1	/	
SNYU11	HSZ0B1	(	
SQSSZ2	HSZ0B2	)	
RUN	HFLUR2	SUM	
RND		{	

Identifier(8) MEGYE = MEGYE\_D1

Meta datas

Megnevezés (30)	Megyekód
Megjegyzés (200)	
Leírás (30)	Megyekód
Érvényes től (date)	
Érvényes ig (date)	
Hossz	8

Columns of output ds

## PRACTICE

Real application of the Microsimulation System started in the Hungarian Central Statistical Office last year.

The known errors of Household Statistics 2002 were corrected using standard simulation technics.

Income Distribution was modified with statistical matching, based on administrative income registration system.

# MICROSIMULATION MODEL DEVELOPMENT ENVIRONMENTS

Istvan Molnar

School of Business, Department of Computer and Information Systems

Bloomsburg University of Pennsylvania,  
Bloomsburg, Pennsylvania, 17815, U.S.A.

E-Mail: imolnar@bloomu.edu

**KEYWORDS:** Microsimulation, Socioeconomic applications, Microsimulation software, Web-services,

## ABSTRACT

This paper describes some of the results of a search for a microsimulation software and analyzes new software technologies (among others Web-based application development), which can be applied to develop microsimulation models.

First, the application field and some of its major characteristics and requirements are introduced. After a short introduction and classification of microsimulation models, basic technological approaches for microsimulation model implementations are presented. Next, two major technologies, the database-oriented and the web-oriented approaches are discussed in detail. The two major solutions for realizing Web-services are J2EE and .NET. These two technologies are compared and their impact on microsimulation and related software technology shortly discussed. Final remarks conclude the paper.

## INTRODUCTION

Microsimulation is a method able to handle complex socioeconomic systems by creating and studying a model that makes intensive use of the statistical data of the observed objects. These objects are the so-called *micro units* of the socioeconomic system; the *person*, the *family* or the *household*. The microsimulation models use simulation techniques in order to study the behavior of micro level units in time (see also Orcutt et al. 1961).

Microsimulation is generally accepted by decision-makers and widely used in Australia, Canada, Europe and the USA to prepare political decisions (see O'Donoghue 2001). Not just highly developed economies, but economies in transition also face many problems especially in demography, pension systems, health care, and taxation. Microsimulation can be a very useful tool to a model-based study of related problems and possible solutions.

Generally, two different microsimulation model classes were developed in order to build realistic

models: *data-driven models* and *agent-based models*. Despite the different modeling approaches, both model classes handle model data and methods in a similar way; in both cases, significant amount of data must be analyzed and processed.

One of the most important technical problems of microsimulation model implementations is the integration and usage of different data sources available for microsimulation models. Historically, three different approaches have been developed:

- File processing approach (e.g., Heike et al. 1994)
- Database-oriented approach (e.g., Sauerbier 2002)
- Agent-oriented approach (e.g., Pryor et al. 1996)

These approaches use mainframe or PC technology and as such, are not portable and architecture neutral. In the 90-ies, new network-oriented technologies were developed in order to support applications (like model-based analysis) using heterogeneous hardware and/or software platforms. Nowadays, the development of networked multi-platform microsimulation applications is not just necessary but also technically possible; beyond networked data access distributed computing can also be realized.

In Hungary, after joining the European Union, more and more signs indicate an increasing demand for instruments of macroeconomic analysis and prediction, coupled with a tendency of more willingness to budgetary spending for microsimulation. There is an urgent need to find and/or develop advanced software tools, which support the achievement of the following goals:

- Network oriented data and model access.
- Distributed model execution.
- Multi-platform hardware and software solutions.
- Open standard based software solutions.
- Data and network security.
- User friendliness
- Efficiency

## MICROSIMULATION

*Microsimulation models* have different data elements: *initial model data*, *intermediate* and/or *final simulation data*; all of these data are stored for further analysis. *Model behavior* in micro-simulation models is described using algorithms, which reflect the behavior

processes of the micro units and represent their environment. Special care is taken to do the *data analysis* and the estimation of simulation *model parameters*. The microsimulation model is working in an *experimental framework* in order to study the effects of policy changes on the microsimulation model behavior. See Figure 1.

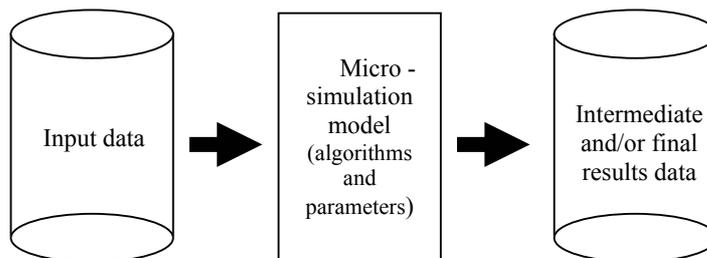


Figure 1: Microsimulation model

Input data, intermediate and/or final results' data are carefully analyzed and special techniques have been developed to improve data quality (e.g., imputing, merging, synthetic data). Model verification and validation also uses different methods and techniques (see e.g., Rubin 2004; Little and Rubin 2002; Schofield and Polette 1998)

## TECHNOLOGICAL ALTERNATIVES

Considering the available network-oriented technologies, Web-enabled microsimulation models can be developed using different technologies. Following the IT industry main trends, two basic approaches can be distinguished':

- Database-oriented approach
- Web-service approach

### Database-oriented Approach

The database-oriented approach is based on RDBMS technology — models are implemented and used in a RDB environment using different analytical tools and technological standards (incl. also Web technologies, mathematical and or statistical program packages and advanced user control and interface).

The network-oriented RDBMS provides a possibility to develop advanced microsimulation applications using architecture as depicted in Figure 2.

This architecture emphasizes data management and provides a multi-platform accessibility for microsimulation data over the network. Applications

processing these data are not considered as “special,” but rather as “usual” DB applications.

Several commercial software products are available that can take over the main burden of networked database management (e.g., Oracle Application Server 2003). Microsimulation algorithms can be developed using Java; data analysis and parameter estimation can be prepared by special mathematical and statistical software tools (e.g., SAS, SPSS).

The elements of this technology are widely available and well proven, industry-standard, vendor-neutral and platform-independent solutions are extensively used.

### Web service Approach

Web service is a transition to service-oriented, component-based, distributed applications. Web services are applications implemented as Web-enable components with well defined interfaces, which offer certain functionality to clients via the Internet. Once deployed, Web services can be discovered, used/reused by consumers (clients, other services or applications) as building blocks via open industry-standard protocols. Web service architecture is built on open standard, vendor-neutral specifications. Services can be implemented in any programming language, deployed and then executed on any operating system or software platform.

The software architecture of Web services is presented in Figure 3.

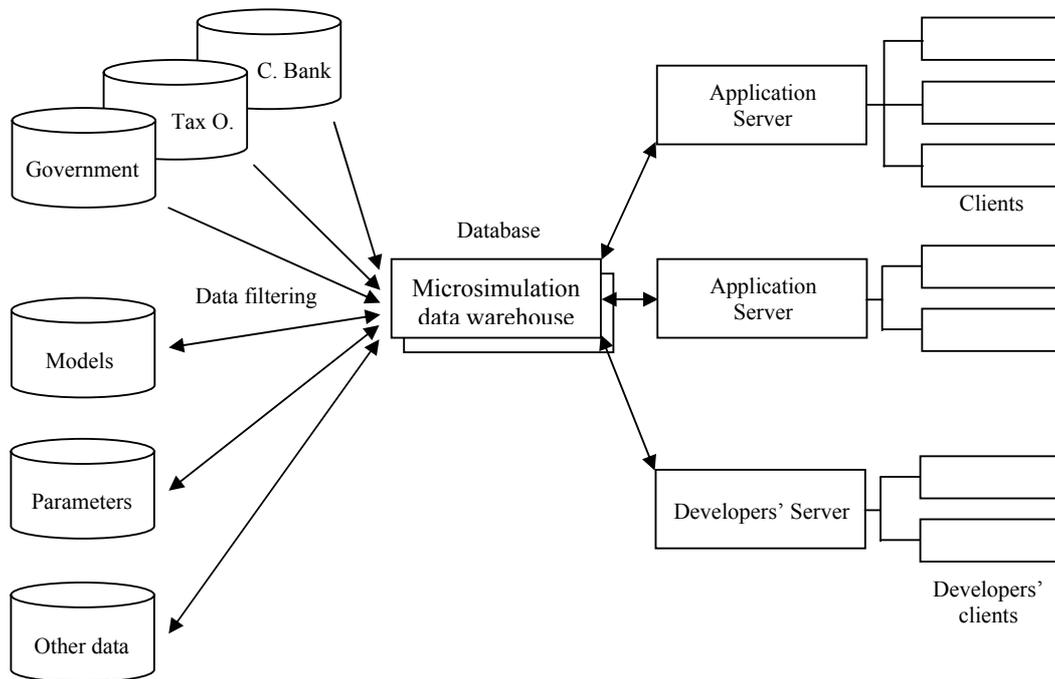


Figure 2: Database-oriented architecture

The model behind Web services is a loosely coupled architecture, consisting of different software components working together. Before invoking a service, consumers must first locate the application offering the service needed, discover the interface and then configure their software in a way that it is able to collaborate with the Web service.

Consuming Web services is based on open standards managed by broad consortia. The Universal Description Discovery and Integration (UDDI) is responsible for publishing, locating and binding Web

services to consumer software. The service, requested by the consumer is determined by a contract between the service provider and the client, who will consume the service. The contract can be formulated using the Web Services Description Language (WSDL). Using the Simple Object Access Protocol (SOAP) and Hypertext Transfer Protocol (HTTP), the parties involved will agree upon a common message and protocol. The data interchange format, used during all communications, is also standardized; the Extensible Markup Language (XML) is used.

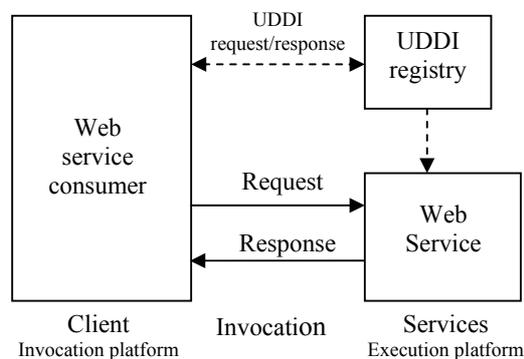


Figure 3: Web service architecture

Currently, two main platforms are used to develop Web services: Microsoft .NET and J2EE.

These technological developments will fundamentally question and/or extend previous network-oriented simulation technologies (e.g., Miller et al. 1998 and Miller et al. 2001). HLA-based solutions (e.g., Lantzsch et al. 1999) can further extend the possibilities of Web services and allow the user to develop large, multi-platform, network-oriented microsimulation models.

Unfortunately, sufficient experience is not available yet. The obvious advantages will certainly further attract simulation software developers in the future; consequently Web services will play a viable alternative for microsimulation software development as well.

#### *Microsoft .NET Technology*

Microsoft .NET is a new Windows platform for developing and deploying web services. The platform is a combination of new and old tools and approaches; .NET is optimized for XML and designed around XML-based Web Services. .NET extends the Visual Basic paradigm to all programming languages on the platform; any language can be used and they can also inter-operate. The most important .NET languages are Visual C# .NET and VB .NET. The platform incorporates a run-time environment called Common Language Runtime (CLR), which is similar to Java VM and referred to as "managed execution environment", which allows compiled programs to run on any platform. The framework also consists of other Web service related components like ASP.NET and Microsoft Internet Information Services (IIS)

#### *Java Technology*

Java technology includes the Java programming language, the runtime environment, the platform editions, and the application programming interfaces (APIs). The current platform editions are: J2ME, J2SE, and J2EE (Micro, Standard and Enterprise Editions), each of them with the Java language and its portable bytecode at the core. The major J2EE application server vendors are IBM, SUN, Borland, Oracle, and Macromedia. J2EE runs on several platforms, but only supports the Java language.

In a J2EE-based Web-service environment, different software elements, provided by different vendors, must work together, which can make the program development more difficult. Therefore development requires more professional attention, because process and code generation is not a highly automatic process, like in the case of .NET. To increase programming efficiency, it is strongly suggested to use an IDE (Integrated Development Environment), like IBM WebSphere or SUN One.

One of the most critical elements of the development is the data communication between client and service provider (see Figure 3); both client and service provider must have the same elementary data type or have the map the given data types between Java and SOAP.

#### *Comparison of .NET and Java Technologies*

Comparing and evaluating application software, decision-makers usually examine the following criteria: TCO, software performance, development and other abilities (CGI Group Inc. 2002).

From the software developer point of view, the arguments and opinions are also very extensively discussed (see Benchmark comparison 2001 and The Middleware Company, 2003).

.NET was designed as "the platform for XML Web Services", while Web service technologies are not yet standardized in J2EE. Despite these facts, .NET and J2EE can, must and will coexist. Businesses or governmental institutions rejecting single vendor lock-in or preferring high level reliability, security or stability, might avoid .NET, but will miss out advantages, offered by the .NET platform.

## **CONCLUSIONS**

The increased interest in microsimulation in Hungary creates an obligation to review the existing practices and get acquainted with the new technologies, which can be applied to new projects.

Converting traditional microsimulation models into Web-enabled ones provides for their effective and efficient use in modern integrated information systems. New technologies like Web service provide user-friendly and powerful tools for new model development.

Web-enabled microsimulation models can be used to develop Web-enabled model bases consisting of different type of microsimulation models and creating Web-enabled decision-support systems. The presented software-environments are generally available in contemporary IT environments.

## **REFERENCES**

- Benchmark comparison 2001. "Building XML-based Web Services in Microsoft .NET vs. IBM WebSphere 4.0". Revision 2.0, (Dec.).
- CGI Group Inc. 2002. "Microsoft .NET or Java 2 Enterprise Edition: Is it just a question of platforms and languages?", White Paper, (Sept.)
- Heike H.-D.; Beckmann K.; Kaufmann A.; Sauerbier T. 1994. „Der Darmstädter Mikro-Makro-Simulator-Modellierung, Software Architektur und Optimierung“. In Proceedings: Faulbaum, F. (Ed.).

- SoftStar'93–Advances in Statistical Software 4. Fischer. Stuttgart/New York, 161–169
- Lantzsch, G.; Straßburger, S.; Urban, C. 1999. "HLA-basierte Kopplung der Simulations-systeme Simplex3 und SLX". In Proceedings der Tagung Simulation und Visualisierung '99 der Otto-von-Guericke-Universität Magdeburg (Hrsg.: Deussen, O., Hinz, V., Lorenz, P.), Institut für Simulation und Graphik, (4.-5. März), Magdeburg.
- Little, R. J. A. and Rubin, D. B. 2002. "Statistical analysis with missing data". Wiley-Interscience. New York.
- Miller, J.; Fishwick, P.; Taylor S.; Benjamin, P.; Szymanski B. 2001 "Research and Commercial Opportunities in Web-Based Simulation". Simulation Practice and Theory, Special Issue on Web-Based Simulation, Elsevier Science.
- Miller, J.; Ge, Y.; Tao J. 1998. "Component-Based Simulation Environments: JSIM as a Case Study Using Java Beans," In Proceedings of the 1998 Winter Simulation Conference, Washington DC. (Dec.) IEEE, 373-381
- O'Donoghue, C. 2001. "Dynamic micro-simulation: A methodological survey". Brazilian Electronic Journal of Economics, vol.4, no. 2.
- Orcutt, G.; Greenberger, M.; Korbel, J.; and Rivlin, A. 1961. „Microanalysis of socioeconomic systems: a simulation study". Harper & Brothers. New York
- Oracle Application Server 2003. "Oracle Application Server 10g J2EE and Web Services". White Paper, ORACLE, (Sept.)
- Oracle Application Server Portal 2003. "Oracle Application Server Portal 10g, Technical Overview". White Paper, ORACLE, (Sept.)
- Pryor, R., Basu, N., Quint T. 1996. "Development of Aspen: A microanalytic simulation model of the US economy". Sandia Report #SAND96-0434, Sandia National Laboratories, Albuquerque, NM, (Febr.)
- Rubin, D. B. 2004. "Multiple Imputation for Nonresponse in Surveys". John Wiley & Sons. New York.
- Sauerbier, Th. 2002. "UMDBS - A New Tool for Dynamic Microsimulation". Journal of Artificial Societies and Social Simulation, vol. 5, no. 2
- Schofield, D. – Polette, J. 1998. "A comparison of data merging methodologies for extending a microsimulation model". NATSEM STINMOD Technical Paper No. 11. National Centre for Social and Economic Modelling, University of Canberra.
- The Middleware Company, 2003. "J2EE and .NET (RELOADED) Yet Another Performance Case Study". The Middleware Company, (June)

#### **AUTHOR BIOGRAPHIES**

Dr. Molnar was born in Budapest and educated at the Budapest University of Economic Sciences, where he received his MSc. and PhD. He has completed his postgraduate studies in Darmstadt, Germany and took part in different research projects in Germany as guest scientist in the 80-ies and 90-ies. In 1996 he has received his CSs. degree from the Hungarian Academy of Sciences. Currently, he is an Associate Professor at the Bloomsburg University of Pennsylvania. His main fields of interest are currently microsimulation, simulation optimization, simulation software technology, and simulation education. Dr. Molnar is a senior member of SCS International, member of the Editorial Board of SCS-European Publishing House and a former member of the European Council Board.

**SIMULATION IN  
BUSINESS, ECONOMY,  
FINANCE AND  
COMMERCE**



# BUSINESS PROCESS MODELLING USING SIMUL8

Vlatka Hlupic  
Brunel University  
Department of Information  
Systems and Computing  
Uxbridge,  
Middlesex UB8 3PH, UK  
E-mail:  
[vlatka.hlupic@brunel.ac.uk](mailto:vlatka.hlupic@brunel.ac.uk)

Vesna Bosilj-Vuksic  
University of Zagreb  
Faculty of Economics,  
Department of Business  
Computing  
Trg J.F.Kennedya 6,  
10000 Zagreb, CROATIA  
E-mail: [vbosilj@efzg.hr](mailto:vbosilj@efzg.hr)

**KEYWORDS:** Business process modelling, Business process renovation, simulation modelling, Simul8, case study

**ABSTRACT:** It is apparent that developing dynamic models of business processes prior to their change could increase the success of business renovation (BR) projects. Simulation has an important role in modelling and analysing the activities in introducing BR since it enables quantitative estimations of influence of the redesigned process on system performances. An example is presented to investigate some of the potential benefits and outcomes of introducing new or redesigning existing processes that could be assessed in advance by using simulation modelling.

## 1. INTRODUCTION

In a period of commercial metamorphosis, organisations, large and small, are finding it increasingly difficult to deal with, and adjust to, the demands of the current business environment. Process renovation is a re-engineering strategy that critically examines current business policies, practices and procedures, rethinks them and then redesigns the mission-critical products, processes, and services (Prassad, 1999).

Many leading organizations have conducted business renovation (BR) in order to improve productivity and gain competitive advantage. However, regardless of the number of companies involved in re-engineering, the rate of success of re-engineering projects is less than 50% (Hammer and Champy, 1993). Some of the frequently mentioned problems related to BR projects include the inability to accurately predict the outcome of radical change, the difficulty in capturing existing processes in a structured way, the lack of creativity in process redesign, the level of costs incurred in implementing the new process, and the inability to recognize the dynamic nature of the processes. The methods of BR, which combine business process modelling and simulation modelling, enabling quantitative estimations of alternative renovated business processes (Harmon, 2003), are one of the possible approaches to address the above-mentioned problem of the evaluation of alternative solutions.

The main objective of this paper is to develop a simulation model of the IT Support function of a multinational construction firm using simulation software tool Simul8. A brief overview of simulation and business process modelling methods is presented in Section 2. A problem definition and model design using Simul8 is provided in Section 3. The evaluation of "AS-IS" model results and "TO-BE" model development are presented in Section 4. Finally, Section 5 outlines the main findings of this research and provides concluding remarks.

## 2. SIMULATION AND BUSINESS PROCESS MODELLING

Many different methods and techniques can be used for modelling business processes in order to give an understanding of possible scenarios for improvement. IDEF0, IDEF3, Petri Nets, System Dynamics, Knowledge-based Techniques and Discrete-Event Simulation are only some examples of widely used business process modelling techniques (Eatock, et.al, 2000, Seila, 2003). As noted by Hommes and Van Reijswound (2000) the increasing popularity of business process modelling results in a rapidly growing number of modelling techniques and tools. The list of the available business process modelling tools supporting simulation includes over 50 names (Hommes, 2001). This makes the selection of the proper tool very difficult. In (Kettinger et al, 1997), an empirical review was made of existing methodologies, tools, and techniques for business process change. The authors also developed a reference framework to assist the positioning of tools and techniques that improve re-engineering strategy, people, management, structure, and the technology dimensions of business processes.

Simulation modelling is being widely used in manufacturing, but also in areas such as health care, the service industry, network communications, traffic modelling and the military. The simulation of business processes is suggested for use in BR projects as it allows the essence of business systems to be understood, the processes for change to be identified, process visions to be developed, new processes to be designed and prototyped and the impact of proposed changes on key

performance indicators to be evaluated (Greasley and Barlow, 1998). The reasons for the introduction of simulation modelling into process modelling can be summarized as follows: simulation allows for the modelling of process dynamics, the influence of random variables on process development can be investigated, re-engineering effects can be anticipated in a quantitative way, process visualization and animation are provided, and simulation models facilitate communication between clients and an analyst. The final reason for using simulation modelling is the fact that it can be increasingly used by those who have little or no simulation background or experience (Irani et al, 2000).

Despite the numerous advantages of simulation software, it is apparent that some user requirements are still not adequately met. The survey on the use of simulation software conducted by Hlupic (2000) revealed that there are two different groups of users: academics and industrial experts. Over three-quarters of academic users and over half of industrial users use simulators. Both groups stated that the main positive features are ease of model development and visual facilities, while the main problems for industrial users were the lack of flexibility (in comparison to simulation and general purpose programming languages), the lack of links with other packages (software compatibility) and the lack of interfaces for data input. It is obvious that no single simulation package could incorporate all desirable features and its selection depends on the application area and the problem complexity.

### 3. PROBLEM DEFINITION AND MODEL DESIGN

The chosen case study is based around the IT Support function of a multinational construction firm, more specifically, the support of approximately 3,000 users within the London offices. Support is provided through the provision of two helpdesks. After an outsourcing agreement, the network infrastructure, hardware and standard office automation application would be managed by the external contractor (Helpdesk 1). This left the company to provide support for the engineering applications and CAD design hardware (Helpdesk 2).

*Helpdesk 1:* Receiving approximately 139 calls per working day, this helpdesk employs 21 full time technicians, of which 5 are employed to receive and log

telephone calls. Upon receipt of a call, the operator logs the callers details and a brief description of the problem. The operator then attempts to resolve the problem over the telephone, a tactic that resolves around 20% of all received calls. However, if the telephone operator is unable to resolve a problem over the telephone, the call is assigned to the team of technicians located in the users building. The technicians check the system for calls, and visit the user at the earliest opportunity. Average waiting time for users is estimated at 2 hours and 30 minutes dependant on workload and staff availability.

*Helpdesk 2:* This helpdesk is staffed by one operator, who allocates calls to any of the eleven technicians, dependent upon the nature of the problem. Dealing with approximately 30 calls per working day, each call is logged and prioritised according to the urgency of the problem. The helpdesk application then acts upon the call, sending an e-mail to the nominated technician, informing them of the call, and a description of the problem. The technician, upon receipt of the e-mail, acknowledges the call, and takes action to resolve the problem, which must be completed within a predefined time.

*Problem areas:* There are several problems related to the chosen case study. Primarily, each helpdesk refers a substantial number of calls to the other because users are not sure of (a) who to call and (b) whether the problem is with the software application, or the underlying hardware or network. This problem is further exacerbated by the two helpdesks using incompatible software applications, resulting in greater delays for users.

Inefficiencies are also evident in the operations of Helpdesk 2, as technicians spend a fair amount of time travelling between the numerous company buildings.

Translating the analysis documentation that had been prepared previously into initial model outlines was quite a simple. Process maps of "AS-IS" model were based on flowcharts (Figure 1) as a very useful, simple and well-known graphical modelling technique (Giaglis, 2001). The next step was to translate the graphical representation of "AS-IS" model into the simulation model using SIMUL8 software model-building tool (Hauge and Paige, 2001).

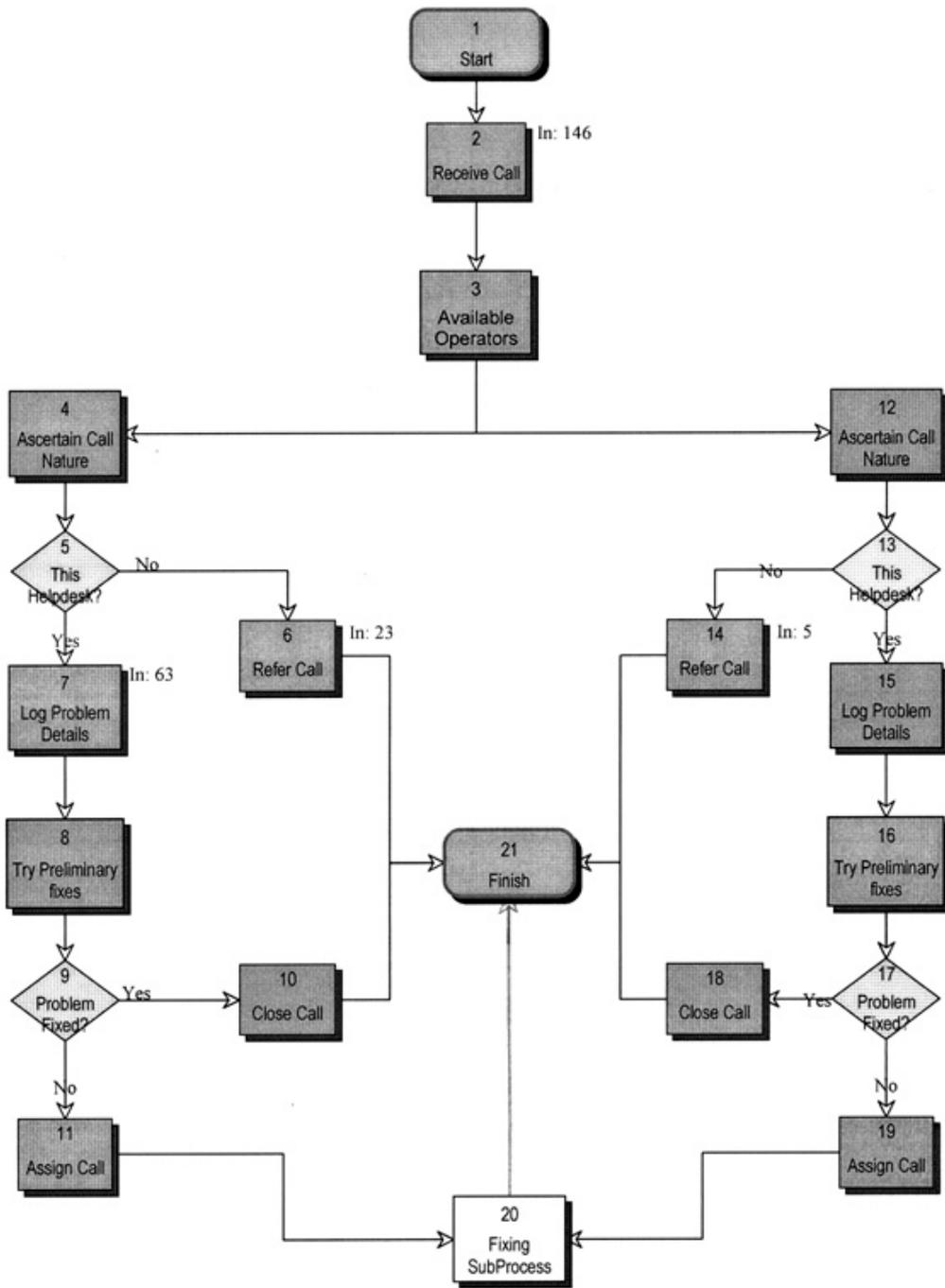


Figure. 1: Diagrammatic representation of “AS-IS” model

ID	Name	Priority	Type	Calendar	Time Option	Path Routing	Path ID	Fixed Time	Probability	Required Resource	Qty	Basis	Usage
1	Start	5	Starter	Total Open Hours	Suspend	Probability	To Receive Call	Exp. (4.633m)	100%				
2	Receive Call		Queue			On Demand	To Available Operators		-				
3	Available Operators	5	Normal	Total Open Hours	Suspend	Parallel	To Ascertain Call Nature		-	Helpdesk Operators (Early T.)	1	Activ.	Inspect
										Helpdesk Operators (Late T.)	1	Activ.	Inspect
4	Ascertain Call Nautre	5	Normal	Team 1 Shift	Finish Task	Probability	To This Helpdesk?	Normal (2m,0.25m)	100%	Helpdesk Operators (Early T.)	1	Activ.	Concurrent
										Caller	1	Activ.	Concurrent
5	This Helpdesk ?	4	Normal	Team 1 Shift	Suspend	Probability	To Refer Call		18%				
							To Log Problem Details		82%				
6	Refer Call	4	Normal	Team 1 Shift	Suspend	Probability	To Finish	1m	100%	Helpdesk Operators (Early T.)	1	Activ.	Concurrent
										Caller	1	Activ.	Concurrent

Table 1: Model definition table

Having sequentially established the processes and decisions of each submodel, the behavioural characteristics were defined (Table 1): the activities were labelled, a priority was assigned, a calendar was assigned (defining what hours it would be operational between), the path ID for the activity was defined (specifying the next activity in the model, and the

statistical distribution data defining the time this process takes to complete). Having specified the behavioural details of the submodels, the resources for each helpdesk were defined. Information about the resources cost, hours and usage were defined. Finally, all the submodels and subprocesses were linked together and presented using SIMUL8 objects and parameters.

Name	Type	Distrib.	Av.	Std. Dev.	Repl.	Routing In	Routing Out	Resources	Actions
Calls Received	Work Entry	Exp.	29.17						
Received Calls Queue	Queue								
Ascertain Call Nature	Work Centre	Normal	2.7	0.5	1	Received Calls Queue	20% Refer Call 80% Log.Assign. Prioritise	Helpdesk Operator Caller	Callers Wage x 25 Helpdesk Operator Wage x 8
Refer Call	Work Centre	Fixed	1		1	Ascertain Call Nature	100% Completed Calls	Helpdesk Operator Caller	Callers Wage x 25 Helpdesk Operator Wage x 8
Log.Assign. Prioritise	Work Centre	Normal	7	1.8	1	Ascertain Call Nature	100% Queue for Technicians	Helpdesk Operator Caller	Callers Wage x 25 Helpdesk Operator Wage x 8
Queue for Technicians	Queue								
Fixing Process	Work Centre	Fixed	0		11	Queue for Technicians	100% For Call Priority Routing		
Close Call	Work Centre	Fixed	3		1	Problem Fixing (FC) Problem Fixing (FN) Problem Fixing (FS) Problem Fixing (RH) Problem Fixing (RN) Problem Fixing (RS)	100% Completed Calls	Helpdesk Technician	Helpdesk Technicians Wage x 20

Table 2: Model definition table in Simul8

Using the symbol set available within Simul8, the work entry points, work centres, queues and work exit points that constituted the top level of the submodels were defined (Table 2). All the objects were connected together using the simple arrow facility, and the finer details of the models were entered using the properties tab (Figure 2).

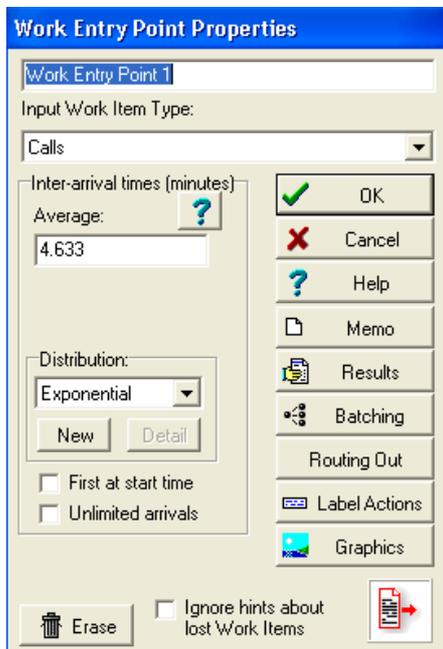


Figure. 2: Simul8 Activity Properties Tab

As mentioned before, to handle the complexity of some of the models, several layers were necessitated. Each submodel was validated and tested both, under its normal working conditions, as well as under increased volumes of calls and reduced levels of staffing.

#### 4. EVALUATION OF MODEL RESULTS AND “TO-BE” MODEL DEVELOPMENT

From the experiment conducted, it appears that operators are operational for 23% of the business day, with the technicians not far ahead on only 26% and most alarmingly the supervisors being involved with helpdesk for 2% of their business day. Calculating the mean resource utilisation, indicates that the personnel resources for Helpdesk 2 are only occupied for 18% each working day. An examination of the resource utilisation in Helpdesk 1 illustrates that they are occupied for 75% each working day.

Prior to initiating the actual reengineering activity, a selection of "what-if" queries has been run on the implemented model. One of the questions examined was: What if both helpdesks were merged together? The aim is to ascertain whether, by pooling the resources of both Helpdesk1 and Helpdesk2, there would be a noticeable drop in the waiting times and quantity of calls

being left until the next working day. The results show (Table 1) that, without fundamental redevelopment of the model, there is a marginal improvement in the time taken to get a response from an operator, but a profound improvement in the responsiveness of the technicians, with the maximum waiting time nearly third of its present value.

As the “what-if” analysis proves (Table 3), the first problem that needed to be addressed was the matter of how to improve the response times to callers, both for the operators and technicians. It was proven that the overall time taken to deal with user calls was only marginally improved when the number of operators was merged between the two helpdesks. To overcome this problem, the reengineered process would comprise a pre-screening operator, responsible for picking up calls as soon as they arrive and quickly ascertaining what their problem is, from where the call is routed to the correct people. Behind this, helpdesk 1 would have reorganised their shift system as follows: 2 operators on the early shift, 1 operator on the normal business day shift, 2 operators on the late shift. This would ensure that the busiest times of the day were covered and hopefully lead to an improvement in the response times. With regard to improving the response times of the helpdesk technicians, the reengineered model would have had the 11 technicians from helpdesk 2 located in teams in the different buildings, so as to get breadth of knowledge but to benefit from the decreased travelling times. Furthermore, the adoption of the information system that is presently in use in helpdesk 2, is proposed. This would mean that technicians would no longer have to log into the system to find their calls, they could simply read their e-mails.

Results	Old Times	New Times
Wait for an operator (av.)	24s	14s
Wait for an operator (max.)	7m 22s	6m 57s
Wait for a technician (av.)	63m 59s	2m 24s
Wait for a technician (max.)	165m 13s	60m
Work items in	131	157
Work items out	75	97
Work in progress	56	60

Table 3: The results of the “what-if” analysis

#### 5. CONCLUSION

It is evident from the material presented within this research that simulation modelling is the "cost-effective" method of exploring "what-if" scenarios quickly, and finding a solution to or providing a better understanding of the problem, as this method is supported by a number of software tools (similar to Simul8) that provide a graphical representation of the business processes through executable models.

By engaging dynamic modelling techniques, an examination was made of a chosen case study. Based on the data presented from the modelling already undertaken, a reengineered business process was proposed and refined. Additionally, the effects of reengineered model were created by performing "what-if" analysis. In this phase of the research a "prototype" of the "TO-BE" model was developed. The improvements made in the process were evaluated presenting the simulation results to the managers and end-users. The model was well accepted by both of them and management was impressed enough to plan to make simulation modelling an integral part of its business renovation plans. The authors plan to explore the benefits of the developed model through further research and the model implementation

## REFERENCES

- Eatock, J., Giaglis, G.M., Paul, R.J., and Serrano, A. 2000. "The Implications of Information Technology Infrastructure Capabilities for Business Process Change Success". In: Henderson, P. (Ed.), *Systems Engineering for Business Process Change*. Springer-Verlag, London, 127-137.
- Giaglis, G.M. (2001). "A taxonomy of business process modeling and information systems modeling techniques", *International Journal of Flexible Manufacturing Systems*, Vol. 13, No. 2, 209-228.
- Greasley, A. and Barlow, S. 1998. "Using simulation modelling for BPR: resource allocation in a police custody process", *International Journal of Operations & Production Management*, Vol. 18, No. 9/10, 978-988.
- Hammer, M. and Champy, J. 1993. "Reengineering the Corporation: A Manifesto for Business Revolution", London, Nicholas Brealey Publishing.
- Harmon, P. 2003. "Business Process Change: A Manager's Guide to Improving, Redesigning and Automating Processes", Morgan Kaufmann Publishers, San Francisco.
- Hauge, Jaret W., and Kerrie N. Paige. 2001. *Learning SIMUL8: The Complete Guide*. Bellingham, Washington: PlainVu Publishers.
- Hommel, B. 2001. "Overview of Business Process Modelling Tools" [URL:<http://is.twi.tudelft.nl/~hommel/scr3tool.html>]
- Hommel, B. and Van Reijswoud, V. 2000. "Assessing the Quality of Business Process Modeling Techniques". 33rd Hawaii International Conference on System Sciences, Vol. 1, January 4-7, Maui, Hawaii.
- Irani, Z., Hlupic, V., Baldwin, L.P. and Love, P.E.D. 2000. "Re-engineering manufacturing processes through simulation modeling", *Logistics Information Management*, Vol. 13, No. 1, 7-13.
- Hlupic, V. 2000. "Simulation software: An operational research society survey of academic and industrial users". In Joines, J.A., Barton, R.R., Kang, K. and Fishwick, P.A. (Eds.), *Proceedings of the 2000 Winter Simulation Conference*, 1676-1683.
- Kettinger, W.J., Teng, J.T.C., and Guha, S. 1997. "Business process change: a study of methodologies, techniques, and tools", *MIS Quarterly*, 21: (1), 55-80.
- Prasad, B. 1999. "Hybrid re-engineering strategies for process improvement", *Business Process Management Journal*, 5: (2), 178-197.
- Seila, A.F. and Ceric, V. 2003. "Applied Simulation Modeling. Thomson Learning, Southbank, Australia.

## BIOGRAPHY

VLATKA HLUPIC is a Senior Lecturer in the Department of Information Systems and Computing at Brunel University and a Visiting Research Professor at Delft University of Technology, Department of Systems Engineering. She received a Dipl.Econ. and an M.Sc in Information Systems from the University of Zagreb, and a Ph.D. in Information Systems at the London School of Economics, England. She has published over 120 papers in journals, books and conference proceedings mainly in the area of simulation modelling, business process re-engineering and knowledge management. She acts as a consultant for a variety manufacturing and service companies, as well as having held a variety of lecturing posts in England and Croatia. Dr Hlupic is a Chartered Engineer, European Engineer and a member of several professional organisations including the British Computer Society, and the director of the Brunel Centre for Knowledge and Business Processes Management at Brunel University.

VESNA BOSILJ VUKSIC received a Dipl.Econ., M.Sc and Ph.D. in Information Systems from the University of Zagreb. She is an associated professor of Simulation Modelling and Business Computing at the Faculty of Economics, University of Zagreb, at the Department of Information Sciences. Her current research interests focus on graphical methods in simulation modelling, business process reengineering, information systems development and knowledge management. She is a former president of the Croatian Society for Simulation Modelling (CROSSIM).

# SUITABILITY OF PROCESS MAPS FOR BUSINESS PROCESS SIMULATION IN BUSINESS PROCESS RENOVATION PROJECTS

Mojca Indihar Stemberger

Jurij Jaklic

Ales Popovic

Department of Information and Management Science

University of Ljubljana, Faculty of Economics

Kardeljeva ploscad 17, Ljubljana, Slovenia

E-mail: {mojca.stemberger, jurij.jaklic, ales.popovic}@ef.uni-lj.si

## KEYWORDS

Business Process Renovation, Modelling Technique, Modelling Tool, Process Maps, Process Modelling, Process Simulation

## ABSTRACT

Many different methods and techniques can be used for modelling business processes within business process renovation (BPR) projects. There are several techniques and tools that attempt to effectively represent all modelling perspectives and fulfil all goals and objectives but as such generate complex models that are hard to understand and reduce their ease of use. One of widely used techniques for process modelling is process maps. They are based on flowcharts and one of their most important advantages, that is extremely important in early phases of BPR projects, is that models are easily understandable to all members of a project group. It is believed that this technique can provide only basic facilities in representing processes and is inappropriate for simulation. The main objective of this paper is to show that enhanced process maps have all elements required for simulation, they can serve as a foundation for IS modelling, and they are very suitable for business renovation. However, a single technique cannot cope with a variety of different aspects related to modelling without becoming too complex and thus less useful in phases where communication of models is very important.

## 1. INTRODUCTION

Business process modelling has emerged as an important research and application area within organizational and information system design. Business process models can be used to serve a wide number of applications, for example to drive a strategic organizational analysis, to renovate existing processes as a part of business renovation, to derive requirements and specifications for information systems design, or to support (semi)automated execution of processes or so called work flows (Paul et al. 1999). Curtis et al. (1992) had identified several modelling goals and objectives: facilitate human understanding and

communication, support process improvement, support process management, automated guidance in performing process, and automated execution support.

The focus of this paper is on process modelling with the purpose of business process renovation. This is a re-engineering strategy that critically examines current business policies, practices and procedures, rethinks them and then redesigns the mission-critical products, processes, and services (Prasad 1999). The objectives of modelling during business renovation projects are usually to fulfil the first two of the above-mentioned goals.

Models of business processes play an important role in different phases of business process (re)design regardless of the methodology used (Desel and Erwin 2000). Several definitions of business processes can be found in literature but, as observed by (Giaglis et al. 1999), all of them have something in common. Most authors agree that processes have internal or external customers and have to produce an output for them. Business processes are decomposed into a number of more elementary steps (activities) that are being executed according to certain rules. During their execution, activities have to be coordinated (Desel and Erwin 2000). Resources have to be provided where needed for the execution of activities. A process has to be described in a way specifying which activities have to be executed in what order and what resources are needed for the execution of these activities.

The need to deal with business processes has caused an increased need for suitable techniques and tools for their identification, modelling and analysis. The increasing popularity of business process modelling results in a rapidly growing number of modelling techniques and tools. Kettinger et al. (1997) report about at least 72 techniques and 102 tools, while Hommes' (2001) survey revealed about approximately 350 business process modelling tools. No single technique or approach can capture the whole spectrum of requirements posed by different people and applications. The choice of a modelling technique for a particular project should be based on matching the

virtues and limitations of various techniques with the objectives of the project (Paul et al. 1999). Giaglis (2001) proposed an evaluation framework and a taxonomy of modelling techniques.

There are several techniques and tools that attempt to effectively represent all modelling perspectives and fulfil all goals and objectives. As already observed in (Curtis et al. 1992) such modelling techniques generate complex models and reduce the ease of use for any single particular application. On the basis of our involvement in some business renovation projects we came to a conclusion that complex models are a big obstacle especially in early phases of such projects when the focus is on human understanding and communication.

On the other hand, processes need to be analysed and different scenarios have to be evaluated to support their improvement. The methods of business renovation, which combine business process modelling with analysis of process performance are one of the possible approaches that can be used for the analysis of the existing processes and for the evaluation of redesigned processes. Simulation can provide a valuable mechanism for addressing the problem of quantitative and qualitative evaluation of business processes. It can facilitate experimentation with and study of multiple perspectives of organizations, thus contributing toward increasing the quality of change decisions.

Flowcharting is one of the first graphical modelling techniques. Nowadays flowcharts are very useful as a simple, graphic means of communication, intended to support understandable descriptions of processes (Giaglis 2001).

The paper is structured as follows: In the next section simulation modelling of business processes is discussed. Section 3 describes process maps modelling techniques and illustrates a case. Process maps suitability for simulation modelling in BPR projects is analysed in section 4. Conclusion remarks and some further research directions are the content of the last section.

## **2. BUSINESS PROCESS SIMULATION MODELLING**

Business process modelling and the evaluation of different alternative scenarios for improvement are usually the driving factors of the business renovation process (Bosilj-Vuksic et al. 2003). Techniques that enable modelling business processes, evaluation of their performance, experimenting with alternative configurations and process layouts, and comparing between diverse proposals for change, are highly suitable for organizational design. Computer based simulation models of business processes can help overcome the inherent complexities of studying and

analyzing organizations and therefore contribute to a higher level of understanding and designing organizational structures (Giaglis et al. 1999). Simulation of business processes creates added value in understanding, analysing, and designing processes by introducing dynamic aspects. It enables the migration from a static towards a dynamic process model (Aguilar and Rautert 1999).

Simulation is generally defined as a set of numerical and programming techniques for representing stochastic models and conducting sampling experiments on those models using a digital computer (Seila et al. 2003). It is important to note that simulation is a set of techniques belonging to analysis methodology. It is not a specific type of model, as would be implied by the often-used term simulation model. Instead, simulation involves methodology for extracting information from a model by "observing" the behaviour of the model with the use of a digital computer. The term simulation model actually means a model that has been adapted to be analyzed with the use of simulation.

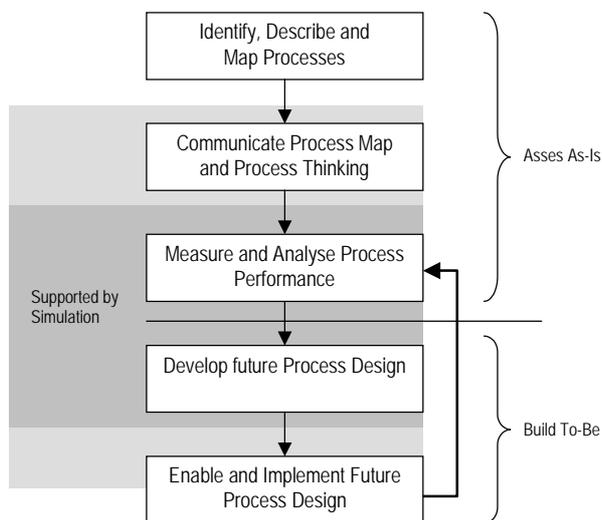
There are some modelling requirements specific to simulation-assisted business renovation modelling (Giaglis and Paul 1996):

- Processes need to be formally modelled and documented.
- Modelling should take stochastic nature of business processes into account, especially the way in which they are triggered by external factors.
- There is a need to quantitatively evaluate the value of proposed alternatives.
- The evaluation is highly dependent on the objectives of the particular study.
- Modelling tools should be easy to use to allow users of the processes to be involved in the modelling process.

Simulation can serve as a tool for deriving new knowledge on current business processes, such as additional in-depth understanding of how the process is executed and the identification of the sources of the problems observed during the process execution (Bosilj-Vuksic et al. 2003). A first phase of business renovation project usually consists of identifying, describing and mapping (modelling) the processes of a company. The results have to be communicated carefully so that everybody in the company understands the concept of process orientation and the mapping results. By introducing dynamic parameters of the process, like times, volumes, capacities and costs simulation fundamentally enhances process performance analysis. It provides a much better picture

of bottlenecks, hand-over times and dynamic performance than a static analysis. In order to detect weak points and opportunities for improvement process performance is evaluated and benchmarked (Aguilar and Rautert 1999).

The main impact of simulation is directed towards performance analysis and design of future processes as illustrated in Figure 1 (Aguilar and Rautert 1999). With a help of process simulation tools we normally assign values to activities and then run a number of cases to see how the business process will respond (Harmon 2003). Thus, simulation has also an important role in analyzing the activities before changes are introduced, since it enables quantitative estimations to be made on the influence of the redesigned process on system performances (Bhaskar et al. 1994). Any envisaged change in process design can be anticipated and evaluated by simulation. The experimentation results can significantly contribute to the decisions about future process design.



**Figure 1: The main impact of simulation in BPR (Aguilar and Rautert 1999)**

Simulation further supports the communication and implementation steps illustrated by the light grey areas in Figure 1. Modelling and simulation of entire process helps all participants adapt a process perspective, understand their contribution to the process result and reflect about the interactions with others in the process. Therefore simulation facilitates communication and redirects people to the most important objective: improving process performance (Aguilar and Rautert 1999).

Thus, simulation is a technique that uses a model to make predictions about a system or process (Harmon 2003). There are different types of simulation, some more informal and some more formal. The technique that is the most suitable for simulation of business

processes and is also implemented in the majority of simulation software is the discrete-event simulation - DES (Seila et al. 2003). Discrete simulations allow system quantities to change only at discrete points in time that are called events (for example arrival of a new customer). Computer-based discrete-event simulation relates to a symbolic representation of processes in ways that can be made persistent, replayed, dynamically analyzed, and reconfigured into alternative scenarios (Paul et al. 1999).

Simulation offers a wide range of possibilities for analysing time/cost/resources aspect of a business process. In the context of other process improvement methodologies, there are two general areas to which simulation modelling may contribute uniquely. These areas dynamically measure activity utilization and system workload. There is a need for integrating component simulation results obtained from alternative process design considerations (Paul et al. 1999). Simulation can aid business decision makers in prioritizing improvement actions and resource allocation decisions. Simulation in BPR can be used not only for modelling the As-Is and To-Be processes but also for marketing, communication, educational, and benchmarking purposes (Bhaskar et al. 1994). Mature process companies will maintain simulations and routinely use them for process improvement projects (Harmon 2003).

There are many techniques and tools that support them for simulation modelling of business processes. Since our focus is on modelling to support human understanding and communication and to support process improvement, it is very important that a model is understandable, because communication between members of a project team is extremely important (Kawalek and Kueng 1997). Visual interactive simulation (VIS) meets this request. The basic features of VIS can be summarised as the ability to build and modify simulation models on-screen, execute graphic simulation models, animate models as they execute, present simulation output graphically, and interact with the model during execution (Seila et al. 2003). It can contribute to communication of process thinking and to the acceptance of simulation results.

Simulation of business processes is very useful in many areas and many tools for its implementation are available. Besides simulation usefulness and tools availability there are still many open issues around to which practitioners and researchers have been devoting their attention. Among these issues there are data collection issues (many data have to be collected for running simulation, which is sometimes very time demanding), hierarchical decomposition modelling issues and granularity issues (the level of details has to be balanced with project goals). Some of these issues are discussed in next sections and can be overcome by

a suitable process modelling technique and tool selection.

### 3. PROCESS MAPS

Processes can be modelled with different techniques – most of them are graphical. One of the most popular is process maps. Process maps are a proven analytical, communication and management tool intended to help process participants understand real business processes, make improvements to them or to implement a new process-driven structure in order to renovate business processes (Hunt 1998). They were initially developed and implemented by General Electric as part of their integrated strategy to significantly improve their bottom-line business performance (Boehringer 2003).

Process maps are based on flowcharts. A flowchart is a graphic representation of all major steps in a process. It is used to provide understanding of processes, help in identifying critical stages of processes, locate problem areas and show relationships between different steps in a process. By reviewing articles in the fields of business process renovation and business process management a set of standard flowchart symbols most commonly used to model business processes can be identified (see symbols 1 – 3 in table 1). Many authors address flowcharts and process maps as synonyms. Some advanced flowcharts might show some of the inputs, but rarely take into account all of the process information. Enhanced process maps (EPM) on the other hand, provide additional process information opposite to simple process maps (van Ackere et al. 1993). EPM consider information as time, resources (personnel, material and equipment), environment (functions or departments), outputs, etc. For each step in the flowchart we add the EPM information and thus have a real understanding of the process. Each graphic symbol (see table 1) can be additionally described in a structured text format and for the purpose of simulation tools usually enable associating these information to graphic symbols. Association of information to graphic symbols is usually not seen on the model graphic.

The authors of this paper understand EPM as a technique for graphic representation of logical steps in a process by considering activities (including duration, resources, constraints and costs), resources (types, number and costs), process delays, hierarchical decomposition and organizational structure (e.g. departments). Modelling elements (symbols) are connected with links that describe the process flow. EPM are described by activities placed in one or more departments (e.g. organizational units performing these activities). A process can be broken down into sub-processes to get a more detailed view (the level of detail is defined by the goal of the model). Delays are clearly noted in order to ease fast spotting of potential

“bottleneck” areas in the process. At the stage of simulation model execution VIS can be carried out.

Figure 2 shows an example of using the EPM technique, an EPM model of a renovated process that was previously presented in (Jaklic et al. 2003). The figure also shows sample details collected for an activity.

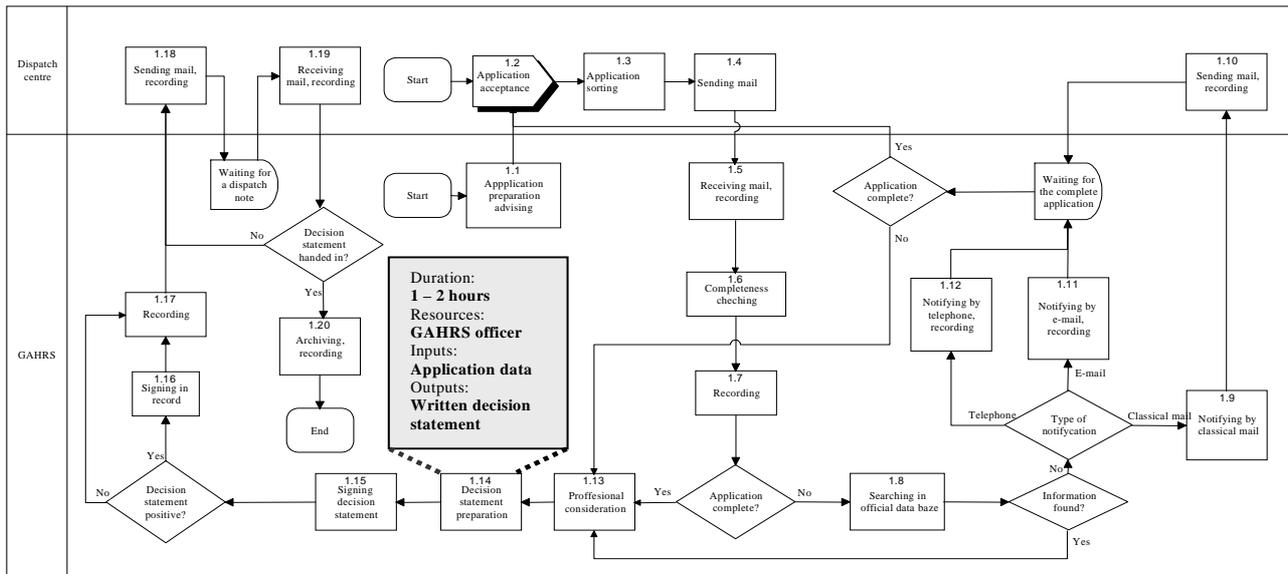
**Table 1: EPM symbols**

	Symbol	Indicates	Examples
1		Start / finish	Receive sales report Customer arrives
2		Activity	Check merchandise Prepare customer invoice
3		Decision point	Approve / Disapprove Accept / Reject
4		Delay	Waiting for customer's response
5		Sub process	Ship merchandise
6		Organisational unit	Sales department Marketing
7		Process flow	

Over the past years, several new software tools have been developed specifically for modelling business processes and their simulation. Most of these tools define business processes using graphical symbols. Special characteristics of each process or activity may then be attached as attributes of the process. Many of these tools also allow for some type of activity-based costing or simulation analysis depending on the sophistication of the underlying modelling technique. Boehringer (2003) divides process mapping tools into three general categories: flow diagramming tools (e.g. ABC Flowcharter, EasyFlow, Flowcharting), CASE tools (e.g. Design/IDEF and Workflow Analyzer, Action Workflow) and simulation tools (e.g. PROMODEL, SimProcess, Simul and iGrafx/Optima!).

### 4. ANALYSIS OF THE PROCESS MAPS SUITABILITY

As discussed before, business process modelling has several modelling goals and objectives which result in different requirements for modelling techniques. Curtis et al. (1992) proposed five different goals and objectives presented in table 2.



**Figure 2: Renovated process Promotion of employees in education to a higher educational title**

In BPR projects process modelling mostly addresses the first two goals, i.e. support human understanding and communication and support process improvement. Overlooking communication is identified as one of four most damaging practices in reengineering work (Bhaskar et al. 1994). Therefore the first requirement or criteria for selecting a technique in a BPR project would be clarity for the intended users.

**Table 2: Goals and objectives of business process modelling (Curtis et al. 1992; Giaglis 2001)**

Modelling Goals and Objectives	Requirements for Modelling Techniques
Support human understanding and communication	Comprehensibility, communicability
Support process improvement	Model process components, reusability, measurability, comparability, support technology selection and incorporation, support process evolution
Support process management	Support reasoning, forecasting, measurement, monitoring, management, and coordination
Support process development	Integrate with development environments, support for process documentation, reusability
Support process execution	Automate process tasks, support co-operative work, automate performance measurement, check process integrity

In Guidelines of Modelling Becker et al. (2000) stressed six principles that are important for business

process modelling: correctness, relevance, economic efficiency, clarity, comparability, and systematic design. Guideline of clarity is extremely subjective and postulates that the model is understood by the model user. Clarity of models is especially important when the objective of business process modelling is to facilitate human understanding and communication or to support process improvement. Although several different graphic notations are used to present process maps, it has to be observed that in the analysis and design phase it is important to use a notation which is easily understandable to the process participants. The actual notation used is of secondary importance; it is more important that the process team feels comfortable (Kawalek and Kueng 1997).

As suggested by Harmon (2003), many different groups are involved in business process modelling. Predictably, different groups use different types of diagrams. The key thing to think about in selecting any notation is who is going to use it. As the main target audience of the process models in BPR projects are people who perform processes and other members of BPR team it is clear that the elements of the process models that are only used to describe software conventions should be omitted. Without a readable, understandable, useful model all other efforts become obsolete.

To accommodate the objectives and goals of BPR, a model must be capable of providing various elements to its users. Such elements include, for example, what activities constitute the process, who performs these activities, when and where the activities are performed, how and why they are executed, and what data elements they manipulate (Giaglis 2001). For the business process analysis in BPR projects the data about who performs activities, other required resources, duration etc. are important regardless of the

fact if simulation is going to be used as an analysis technique or not.

The EPM technique enables modelling of business processes in a way that is as easy to understand as possible, while it is still possible to describe (Harmon 2003) all of the basics that need to be described for the intended purpose i.e. process analysis for BPR and communication. Process maps can be very effective communication tools that facilitate, across internal and external organizational boundaries, in a very simple visual format, transmission of ideas concerning what is actually happening in business and ways to improve the business.

As follows from the work of Curtis et al. (1992), the second criteria for selecting a modelling technique for BPR projects is the possibility to model process components, measurability, compare different scenarios etc. The static model description models the components that make up the system, but it does not tell how they interact (Seila et al. 2003). A complete model description must also include the system dynamics. The dynamic model description provides a set of rules telling how the components interact as time advances.

The process view represents system dynamics by joining a sequence of events and activities to follow the progress of a temporary entity through the system. A process consists of a routine that describes the sequence of related events (where the system state changes) and activities (the interaction between different entities over a specific length of time) for the entity - permanent or temporary (Seila et al. 2003). Giaglis et al. (1999) define process from the simulation point of view as a time-ordered sequence of interrelated events (activities) which describes the entire experience of an entity as it flows through a system. From these definitions one can see that a simulation is just a way in which one can analyze a business process. In other words, simulation modelling techniques are by nature process-oriented (Giaglis et al. 1999).

The main required elements for a DES simulation model are therefore: events, activities, time, permanent (resources) and temporary (tokens) entities, and sequence (flow). Harrel and Field (1996) argue that much of the process definition used in a simulation model is contained in a process map, yet insufficient data are provided in a process map for running simulation. Therefore additional information has to be manually added on the simulation side. Giaglis et al. (1999) state that several attempts of integrating process mapping and simulation have been made, although with not very successful results. In their opinion this is due largely to incompatibilities in both purpose and paradigm. On the other hand Srinivasan and Jayaraman (1997) stated that activity (function) and entity

(information) models contain all the information needed for developing DES simulation models of an enterprise. Van Ackere et al. (1993) argue that *simple* process maps do not typically provide sufficient understanding of the process to know what to change. Van der Aalst (1992) suggests that the intended analysis dictates the type of modelling that is done. The goals of a reengineering effort are most often related to business improvement measures. The process maps helps to understand the problem, but to help in knowing what to change, the process map must be supplemented with quantitative analysis.

Our experience on BPR projects shows that for a detailed analysis of existing business processes, required by process renovation, more elements of the process model than just activities and process flow are required. A complete viewpoint may be difficult to establish using only fixed-process map descriptions. They address aspects of processes for which static activity and data modelling are inadequately suited, because they cannot cope with the impact of resource flow. Resources (permanent entities) and their costs, organizational structure, duration of activities and waiting times, events and their frequency are the necessary elements of a carefully prepared process model for a process (organizational) analysis. Process maps provide detailed information about observed processes, usually presented on separate sheets that are filled in as the process map is developed. This brings to the conclusion that a properly designed process map model contain all the necessary information for DES analysis of a business process.

The only element of simulation that is not fully supported by process maps is an explicit definition (presentation) of events. In a certain other more structured modelling techniques, such as eEPC (extended Event-Driven process Chain) each activity is triggered by an event and the result of each activity is an event. With process maps this is not a case, however we do not consider this as a weak point of the process maps technique considering the purpose of modelling in our case, i.e. process analysis and renovation which requires a great deal of communication and therefore comprehensibility. Also, from the formal point of view an activity consists of a starting event followed by an ending event that is scheduled by the starting even (Seila et al. 2003), which implies that it is not necessary to model events separately. We found more formal technique regarding modelling of events as less understandable.

Another important issue in process modelling and simulation is the support for hierarchical decomposition and design modularity (MacArthur et al. 1994), which is also enabled by the EPM technique. Today many tools for process maps modelling support process decomposition, and appropriate simulation decomposition.

Advances in software technology support integration of technologies such as process mapping and simulation that previously functioned only as stand-alone applications (Giaglis and Paul 1996). The fact is that today many of the more powerful business process tools offer simulation. While a certain amount of expertise is required to build models with most simulation languages, process modelling solutions usually offer a possibility for an easy to use appropriate simulation performance measurement in the assessment of alternative designs for BPR.

Bhaskar et al. (1994) proposed a set of requirements that should be met by tools used for modelling and simulation of business processes. These requirements can be divided into five groups: process documentation, process redesign, performance measurement, communication, and institutional learning. Based on our analysis and our experience from BPR case studies we believe that EPM based tools meet these requirements.

Hommel and van Reijswoud (2000) have developed a framework for the evaluation of business process modelling techniques. They propose eight evaluation criteria which can be divided into two groups: one related to the conceptual modelling in general and another group related to the business process modelling in particular. They refer to the quality of the way of modelling and the way of working of a modelling technique respectively. These criteria are:

- Expressiveness - the degree to which a given modelling technique is capable of denoting the models of any number and kinds of application domains;
- Arbitrariness - the degree of freedom one has when modelling one and the same domain;
- Suitability - the degree to which a given modelling technique is specifically tailored for a specific kind of application domain.
- Comprehensibility - the ease with which the way of working and way of modelling are understood by the participants;
- Coherence - the degree to which the individual submodels of a way of modelling constitute a whole;
- Completeness - the degree to which all necessary concepts of the application domain are represented in the way of modelling;
- Efficiency - the degree to which the modelling process utilises resources such as time and people;

- Effectiveness - the degree to which the modelling process achieves its goal.

It has to be observed that the properties are not orthogonal. For example, van der Aalst (1993) suggests that the complexity and detail are essential to sound analysis, however excessive complexity and detail can impede human understanding of the process.

Using theoretical findings from the reviewed literature and our own experience from several BPR projects we have evaluated the EPM technique based on these criteria. Additionally, we have to evaluate for each criteria/property the importance for the purpose of modelling which is in our case business process renovation.

As seen from the table 3 our estimation is that the EPM technique performs well for the criteria that are of high importance for the BPR projects.

**Table 3: Evaluation of the EPM technique**

Criteria	Importance	Score
Expressiveness	Low	Poor
Arbitrariness	Medium	Good
Suitability	High	Good
Comprehensibility	High	Good
Coherence	Low	Good
Completeness	Medium	Limited
Efficiency	Medium	Good
Effectiveness	High	Good

Levas et al. (1995) discuss some of business process simulation issues (such as problem definition, data collection, socio-political issues, hierarchical and modular modelling, granularity, integration, and multi-perspective issues) needing attention in BPR projects. We found the EPM technique very suitable for solving socio-political issues (by enhancing communication and understanding in a project group), hierarchical issues (by providing hierarchical (de)composition features), and granularity issues (by easily allowing modelling at different levels of details).

## 5. CONCLUSION

Simulation modelling has many benefits for BPR projects for analysing the existing processes and evaluation for alternative scenarios of their improvement. In the paper we have analysed suitability of the process maps technique for simulation based process analysis.

From the literature review and our own research work we can conclude that flowcharts and simple process maps do not have all the required elements for simulation modelling. However, enhanced process maps as defined in the paper have all modelling elements formally required for simulation. With

additional benefits, such as clarity of models, tools based on this technique can be very efficient for the first two modelling goals that are related to BPR projects.

Process models built by using EPM technique can serve as a base for identifying information requirements and planning of information system development projects. They are also very suitable for the introduction of workflow management system. Therefore, enhanced process maps can serve as a foundation for IS modelling. However, a single technique cannot capture all different aspects of modelling without becoming too complex and as a consequence less useful for early phases of BPR where communication of models is very important. Research on bridging the gap between business process and IS modelling is left for our future work.

## 6. REFERENCES

- Aguilar M.; T. Rautert; and P. Alexander. 1999. "Business Process Simulation: A Fundamental Step Supporting Process Centered Management". In *Proceedings of 1999 Winter Simulation Conference* (Phoenix, Arizona, December 5-8), 1383-1392.
- Becker J.; M. Rosemann; and C. von Uthmann. 2000. "Guidelines of Business Process Modeling". In *Business Process Management*, van der Aalst, W. et al. (Eds.), Springer Verlag, Berlin, 30-49.
- Bhaskar R.; H.S. Lee; A. Levas; R. Petrakian; F. Tsai; and B. Tulsjie. 1994. "Analyzing and Re-engineering Business Processes Using Simulation". In *Proceedings of the 1994 Winter Simulation Conference* (Lake Buena Vista, Florida, December 11-14), 1206-1213.
- Boehringer B. 2003. "Process Mapping: How to Streamline and Reengineer Business Processes". Outline of seminar on Process Management, Charlottesville, VA.
- Bosilj-Vuksic V.; M. Indihar Stemberger; J. Jaklic; and A. Kovacic. 2003. "Assessment of E-Business Transformation Using Simulation Modeling". *Simulation*, Vol. 78, No. 12, 731-744.
- Curtis B.; M.I. Kellner; and J. Over. 1992. "Process Modeling." *Communications of the ACM*, Vol. 35, No. 10, 75-90.
- Desel J. and T. Ervin. 2000. "Modeling, Simulation and Analysis of Business Processes". In *Business Process Management*, van der Aalst, W. et al. (Eds.), Springer Verlag, Berlin, 129-141.
- Giaglis G.M. and R.J. Paul. 1996. "It's Time to Engineer Re-engineering: Investigating the Potential of Simulation Modelling in Business Process Redesign". In *Business Process Modelling*, Scholz-Reiter, B. and Stickel E. (Eds.), Springer-Verlag, Berlin, 313-332.
- Giaglis G.M.; R.J. Paul; and V. Hlupic. 1999. "Integrating simulation in organizational design studies". *International Journal of Information Management*, 19, 219-236.
- Giaglis G.M. 2001. "A taxonomy of business process modeling and information systems modeling techniques". *International Journal of Flexible Manufacturing Systems*, Vol. 13, No. 2, 209-228.
- Harmon P. 2003. *Business Process Change: A Manager's Guide to Improving, Redesigning and Automating Processes*. Morgan Kaufmann Publishers, San Francisco.
- Harrel C.R. and K.C. Field. 1996. "Integrating Process Mapping and Simulation". In *Proceedings of the 28th conference on Winter simulation* (Coronado, California, December 8-11), 1292-1296.
- Hommel B. and V. van Reijswoud. 2000. "Assessing the Quality of Business Process Modeling Techniques". In *Proceedings of the 33rd Hawaii International Conference on System Sciences*, Vol. 1 (Maui, Hawaii, January 4-7), 1-10.
- Hommel, B. 2001. "Overview of Business Process Modelling Tools", <http://is.twi.tudelft.nl/~hommel/scr3tool.html>
- Hunt D.V. 1998. *Process Mapping – How to Reengineer Your Business Processes*, John Wiley & Sons, New York.
- Jaklic J.; A. Groznik; and A. Kovacic. 2003. "Towards e-government – the role of simulation modelling". In *Proceedings of 15th European Simulation Symposium* (Delft, The Netherlands, October 26-29), 257-262.
- Kawalek P. and P. Kueng. 1997. "The Usefulness of Process Models: A Lifecycle Description of how Process Models are used in Modern Organisations". In *Proceedings of The Second CAiSE97/IFIP8.1 International Workshop on Evaluation of Modeling Methods in Systems Analysis and Design* (Barcelona, Spain, June 16-17).
- Kettinger W.J.; J.T.C. Teng; and S. Guha. 1997. "Business process change: a study of methodologies, techniques, and tools". *MIS Quarterly* (March), 55-80.
- Levas A.; P. Jain; S. Boyd; and W. Tulsjie. 1995. "Panel Discussion on the Role of Modeling and Simulation in Business Process Reengineering". In *Proceedings of the 1995 Winter Simulation Conference* (Arlington, Virginia, Dec 3-6), 1341-1346.
- MacArthur P.J.; R.L. Crosslin; and J.R. Warren. 1994. "A Strategy for Evaluating Alternative Information System Designs for Business Process Reengineering". *International Journal of Information Management*, 14(4), 237-251.
- Paul R.J.; G.M. Giaglis; and V. Hlupic. 1999. "Simulation of Business Processes". *The American Behavioral Scientist*, Vol. 42, No. 10, 1551-1576.
- Prasad B. 1999. "Hybrid re-engineering strategies for process improvement". *Business Process Management Journal*, Vol. 5, No. 2, 178-197.
- Seila A.F.; V. Ceric; and P. Tadikamalla. 2003. *Applied Simulation Modeling*. Thomson Learning, Southbank, Australia.
- Srinivasan K. and S. Jayaraman. 1997. "Integration of simulation with enterprise models". In *Proceedings of 29th conference on Winter Simulation* (Atlanta, Georgia, December 7-10), Vol. 2, 1352-1356.
- Tarumi H.; T. Matsuyama; and Y. Kambayashi. 2000. "Evolution of business processes and a process simulation tool". In *Proceedings of the Asia-Pacific Software Engineering Conference* (Takamatsu, Japan, December 7-10), 180-187.
- Van Ackere A.; E.R. Larsen; and J.D.W. Morecroft. 1993. "Systems Thinking and Business Process Redesign". *The European Management Journal*, 11 (4), 412-423.
- Van der Aalst W.M.P. 1992. "Modelling and analysis of complex logistic systems". In *IFIP Transactions on Integration in Production Management Systems*, Eindhoven University of Technology, Netherlands, 277-292.

## **AUTHOR BIOGRAPHIES**

**MOJCA INDIHAR STEMBERGER** received her Master in Computer and Information Science degree in 1996, and her Ph.D. in Information Science in 2000 from the University of Ljubljana, Slovenia. Currently she is an assistant professor at the Faculty of Economics, University of Ljubljana. Her research interests include business process renovation, e-business and decision support systems. She is a president of the Slovenian Informatics conference.

**JURIJ JAKLIC** received his Master Degree in Computer Science in 1992 from the University of Houston and his PhD in 1997 from the University of Ljubljana, Slovenia. Currently he is an assistant professor at the Faculty of Economics, University of Ljubljana. His main research interests are business process reengineering, business renovation, e-business, decision support systems, and data and business modelling. He is a member of the program committee at the Slovenian Informatics conference.

**ALES POPOVIC**, B. Sc., is an assistant lecturer for Information Management at the Faculty of Economics, University of Ljubljana, Slovenia. His pedagogical work in the past years was oriented mainly to preparing and performing computer workshops (lab lectures) for undergraduate business students. His main research interest is concentrated on business renovation and business process modelling.

# QUICK GRID TO COPE WITH COUNTERPARTY RISKS

Radek Skoda, Gergely Szalka  
European Bank for Reconstruction and Development  
One Exchange Square  
London EC2A 2JN, United Kingdom

## *Abstract*

This paper describes, analyzes and evaluates algorithms and implementations for calculating potential future credit exposures borne by financial institutions. As potential future exposure belongs among the most important measures of counterparty risk, it needs to be calculated precisely and on a daily basis. This creates a very high demand on computer power where single computers can not cope with the high number of calculations within the given time. This paper focuses on a grid-type computer solution, the optimal number of computers connected to the grid, the speed gain of connecting an individual computer to the grid and the order in which the subtasks are assigned to the grid. The results indicate that using a grid for simulating potential future exposure profiles is a feasible solution for credit risk management practitioners.

**Key words:** potential future exposure, parallelisation, grid, Monte - Carlo simulation

## **1. Introduction**

Let us imagine a model situation: We lend money to a friend for 5 days and then we start to ponder how much we might lose in the unlikely event of the friend dying 3 days from now. If we are engaged too deeply in such pessimistic thoughts, we would rather need a good psychologist, not a complex grid of computers to solve the problem. For a bank where the number of institutional debtors tends to be large, portfolios are more complex and usual investment horizons may be measured in decades it might be not too crazy to try to determine a potential future loss in the case where one of the debtors “dies” or, rather, defaults. Then the complex approach is worthwhile and may end up with the need of heavy calculations. It is especially likely when the bank wants to know every day the potential future loss for every debtor and for the entire maturity term of the portfolio. The remainder of this paper will show one of the ways how this common problem may be effectively tackled in reality.

After a brief introduction of credit markets, a short discussion on advanced measures of risk, namely potential future exposure, is provided in section 2. The main part of the paper describes the calculation the exposure profiles, and how to come to terms with the huge amount of computation power that is needed to estimate a percentile of portfolio prices at a number of points in the future. In section 3 the simulation process is briefly described, followed by section 4 where the management of maintaining an acceptable precision within run-time constraints. In section 5 the findings and results are summarised.

## **2. Credit market developments and the need to know future exposures**

In recent years credit and derivatives markets (Hull [1993], Hull [2002]) have increased in volume, importance and complexity and hence credit risk measurement has grown proportionally. Structured credit instruments have increasingly liquid markets; products where

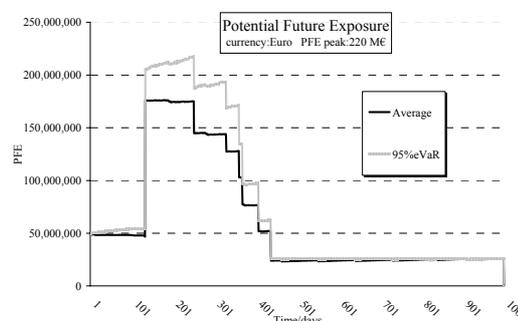
the credit risk can be transferred from one dealer to the other and price risk can be separated from the holder of the underlying instrument. Using credit default swaps, credit risk can be separated from the bond or other obligation one keeps in the balance sheet. Looking at developments in the credit and derivatives markets from another angle shows that handling counterparty transactions at portfolio level has important benefits: e.g. the effect of legal rights like close out and collateralisation can be measured. These considerations can dramatically change the amount of credit risk exposed to a given counterparty, especially when the amount one bank is owing to the trading partner is collateralised. The portfolio context is preferable also as it takes into account correlation between transactions. For effective measurement of all associated risks the portfolio approach requires advanced techniques to be properly deployed.

### Potential future exposure

The starting point of any risk calculation is to calculate the *exposure*, the maximum amount the bank loses in case the debtor or counterparty defaults. *Exposure* today is the current positive value of the portfolio. Potential exposure, however, in a given future point is an estimated percentile of the distribution of the future value(s) of the portfolio – the value of the portfolio depends on the market prices and rates, so it can be thought of as a random variable. Say, that we estimate the potential exposure in such a way that it is not exceeded by the loss with 95% probability. This means that an estimate of the 95% percentile of the distribution of the future value(s) of the portfolio is needed. For a number of reasons the *expected shortfall*<sup>1</sup> is used instead, this being the expected value of the range that exceeds the percentile. In the following example *potential future exposure* will be

<sup>1</sup> The same quantity is also called eVaR (excess value-at-risk) and CVaR (conditional value-at-risk) in the literature.

the expected shortfall of the future value of the portfolio with 95% confidence. Exposure profile will be the potential future exposure numbers plotted against time.



These profiles are good measures of potential exposure by themselves, but are crucial from control aspects allowing risk takers to ensure that they stay within an institution appetite for credit risk on a particular name (that is, comply credit limits).

### 3. Calculating exposure profile

Estimates of the percentile of a potential portfolio value distribution at a given date are given here. First one has to simulate the market drivers that – through the pricing function – will determine the value of the portfolio (Duffie [2001]). A one factor model for yield curves and foreign exchange rates is used here:

$$dr = f(t,r)dt + g(t,r)dz, \quad (1)$$

where  $f(t,r)$  and  $g(t,r)$  functions are described in the model (Rebonato [1996], Jarrow [1998]),  $dz$  denotes the standard Wiener process, and  $t$  stands for time. In fact not one driver but a handful of factors are modelled since each currency has its own process (exchange rate, interest rate, etc.). So, rather, one should write:

$$\Delta R = F(t,R) + G(t,R)\Delta Z, \quad (2)$$

where  $R$  is a vector of market factors,  $F$  is a vector function,  $G$  is a matrix containing all the correlations between the drivers, and let  $Z$  denotes a vector of independent Wiener processes. As soon as the market drivers are simulated, the portfolio can be priced. One goes along all transactions and calculates the price of each of them; having the prices so determined one adds the prices together according to the applicable legal rights (if netting agreements are in place, prices are added together otherwise only the positive values are aggregated). The above can be formalized as:

$$p(t) = P(R_t, t, P_{t-1}, P_{t-2}, \dots). \quad (3)$$

Here,  $p$  is the price of the portfolio at any time  $t$ . The function  $P(\cdot, t)$  defines the price of the portfolio using the  $R(t)$  set of market drivers. The  $P$  function has explicit  $t$  dependence because of the portfolio, and is not continuous<sup>2</sup> in  $t$ . If collateral agreements are present,  $P$  is also conditional on its own previous values.

Considering the above one can easily see that even if a Wiener process is applied for market drivers (i.e. no jumps), the portfolio structure itself introduces non-continuity to the problem. This means that there is no easy way to forecast exposure profile points using previously calculated points<sup>3</sup>.

Having completed the pricing of the portfolio once, one simulated value for the portfolio price distribution is obtained. For estimating the percentile of this distribution one has to repeat simulating market drivers and pricing the portfolio again and again.

Having the percentile estimated for one date point, one has to re-calculate the estimate of the percentile for a number of other date points in order to get the potential future exposure profile.

---

<sup>2</sup> Consider the  $P$  function just before and after a payment has been made.

<sup>3</sup> Except, when having additional information on the portfolio, such as there are no payments, etc.

It is obvious that this process can use immense computing power. Monte Carlo methods are not famous for their speed, and on top of that, the measure to be estimated is the expected shortfall and not the expected value; for estimating 95% percentile one uses only 5% of the simulated data.

The details of why as many simulations as possible are needed will not be addressed here, except to recall the adage: to halve the length of confidence interval four times the work is needed.

### *Why time steps are needed?*

As already discussed the portfolio price function, which takes the input of market drivers and returns the price of the portfolio, is non-continuous. This means that there can be – and in fact there are – discontinuities when moving along the time axis. If the granularity of the time axis is generous and there is a value step-up and a step-down shortly after each other, there is a high probability that a peak is “missed”. Missing a peak is a danger as it may underestimate potential credit exposure. Also the fact that peaks appear on some days and not on other ones undermine the statistical robustness and the reliability of estimating the percentiles.

## **4. To cut a long story short**

There are, generally speaking, two possible ways how to decrease run-time. Either one has to decrease the amount of work to be done or increase the capacity of the processing devices. Optimising the code, trying to make the Monte Carlo method more effective will fulfil the former; enabling the task to use more computing power is the solution to the latter.

### *Variance reduction techniques*

There are existing techniques to increase the efficiency of the Monte Carlo (Glasserman et al. [2002]) however, the outlined problem

is quite specific and does not behave like a "typical Monte Carlo":

- Random number generation is not an issue.
- The goal is to avoid – as far as possible – any calculation, not just speeding up the random number generator.
- Typically the re-simulation of market drivers and random number generation is less than 1% of the total workload for a portfolio where there are more than a handful of trades.

Also, for specific portfolios, some of the standard speed-up techniques may be deployed. It includes antithetic paths, control varieties, stratified sampling or importance sampling to mention just the most obvious candidates. However, as the Monte Carlo calculation is to be used to calculate credit exposure profiles for a variety of different portfolios with very different features, there is no simple recipe which ones should be used.

### ***Increasing computing power***

It is always tempting to increase the 'efficiency' by purchasing new processors, however each computer has a limit whereas the computing power that Monte Carlo estimates can use up is unlimited.

A better solution seems to be parallel processing, which cannot only use multiprocessor machines but also can be run on grid type of hardware configuration. Monte Carlo simulation is the perfect candidate for parallel runs. One has to study the task and determine how the simulation can be cut into smaller blocks. A natural selection would be that one profile point is calculated in one block, thus the whole simulation would be divided into blocks so that each block calculates the distribution of the potential portfolio price, does the statistics, and returns one number: the excess percentile of credit exposure. It is easy to implement, and no large data

movements are necessary. However, it already has been pointed out that different date points are not necessarily independent in the case where collateral management is included in the calculation. Thus, it is necessary to divide the simulation space along another dimension: the number of simulations. The problem here is that one can not do the statistics inside the blocks but will need to store the simulated data, and there should be a block that collects all the data and does the final statistics.

To examine what is the condition for being it worth sending the calculation to the grid, let us break our processes into the following sections:

$$T_{total} = T_{init} + T_{sim} + T_{finish}. \quad (4)$$

$T_{total}$  is the run time of the whole calculation,  $T_{init}$  is the time spent on initialising the task,  $T_{sim}$  is the time spent on the actual simulation, and finally  $T_{finish}$  is the time spent finishing the task, preparing the statistics. Going one step further let us break the simulation into  $N$  blocks as one wants to describe the situation on the grid. Rewriting the formula above yields the following:

$$T_{total} = \hat{T}_{init} + N \cdot (\tau_{init} + \tau_{sim}(N) + \tau_{finish}(N)) + \hat{T}_{finish} \quad (5)$$

Here,  $\hat{T}_{init}$  is the time that the whole process spends on initialising the grid computing components, whereas  $\tau_{init}$  is the time that is needed to initialise one single block (reading input data for example).  $\tau_{sim}$  is the time spent on the simulation inside the block and  $\tau_{finish}$  is the time spent on finishing the process (saving the results to the network/database etc.);  $\hat{T}_{finish}$  is spent on retrieving the results from the network or from a database ( $T_{i/o}$ ) plus the time to

finish the post processing ( $T_{post}$ ), like calculating the statistics, preparing the output. Note, that  $\tau_{finish}$  and  $\tau_{sim}$  might depend on  $N$ . The dependence is thought to be inverse: in case of increasing  $N$ , obviously  $\tau_{sim}$  (as  $\tau_{sim} = T_{sim}/N$ ) and hence  $\tau_{finish}$  should decrease. The reason for the second statement being true is that as the work the block is supposed to do is decreased, it is logical to assume that the amount of data processed is smaller as well. Here, the relation may not be that straight forward as for the simulation. However, it can very well be the case that for one third of work, the time spent on finishing the work is one half. The dependence of  $\tau_{init}$  on  $N$  is thought to be minimal; hence the function argument is omitted from now on.

If all the blocks are running on the same machine (memory is shared) the saving of data at the end of each block ( $\tau_{finish}$ ) and the collecting of the data at completion of the whole process ( $T_{i/o}$ ) can be saved. Thus,  $\hat{T}_{finish} = T_{post}$  and  $\tau_{finish} = 0$ . This is the case when one block after the other is to be run on the same machine<sup>4</sup>. On top of all these, for one machine calculation the initialisation of the grid can be also skipped:  $\hat{T}_{init} = 0$ .

Another special case will be running several blocks on a grid of computers/processors where the number of processors,  $M$  is greater than the number of blocks,  $N$ . Also, for the sake of simplicity let us assume that the speed of all processors is the same. Then,

---

<sup>4</sup> Note, that here it does not make much sense breaking up the process into more than one block, as for each block  $\tau_{init}$  time is needed to get the block initialised, and the simulation does not take any shorter breaking it into many blocks as one follows the other one.

$$T_{total} = \hat{T}_{init} + (\tau_{init} + \tau_{sim}(N) + \tau_{finish}(N)) + \hat{T}_{finish} + N \cdot T_{grid} \quad (6)$$

First, the whole task has to be initialised then it has to be cut into  $N$  pieces. Each piece is running on a different machine, hence the waiting time is the completion time of one block, then the results are to be collected and the statistics done.

The last part of the formula  $T_{grid}$  contains all the time that is spent managing the grid<sup>5</sup>, communicating between the processors, etc. This is increasing when the number of blocks is increasing, however, normally still stays relatively small.

### Pros and cons

To be able to compare single block calculation with grid computing let us recall the general run-time equation (5) and the grid run-time formula of equation (6). For the single block calculation one can write:

$$T_{total} = T_{init} + T_{sim} + T_{finish} = \tau_{init} + N \cdot \tau(N)_{sim} + T_{post} \quad (7)$$

Running the whole calculation in one bunch has one advantage that is worth mentioning: skipping the *collecting* bit of  $\hat{T}_{finish}$  that is  $T_{i/o}$ .

$$T_{finish} = T_{post} = \hat{T}_{finish} - T_{i/o} \quad (8)$$

Let us have a closer look at equation (6) and equation (7). This is to determine whether it is worth breaking a task into blocks and running it on the grid, versus running it as

---

<sup>5</sup> The more blocks one has the more workload on the grid is generated; that is why we have  $N \cdot T_{grid}$  in the formula. That work includes the communication between the machines, auditing the processes, querying the progress made on the jobs, etc.

one task. Comparing the two equations the condition of using the grid boils down to the following:

$$\begin{aligned} & \hat{T}_{init} + T_{i/o} + N \cdot T_{grid} + \tau_{fin}(N) \\ & < (N-1) \cdot \tau_{sim}(N) \end{aligned} \quad (9)$$

Just a quick look on the formula yields the following remarks: the left hand side contains costs of the grid whereas the right hand side the benefits. Supposing that the number of machines on the grid  $M$  is large enough not to violate the assumption of  $M > N$ , and  $\tau_{finish}(N) \ll \tau_{sim}(N)$ ; now it is beneficial to increase the number of blocks. However, there is a limit (apart from the number of processors on the grid) as while  $N$  is increasing,  $\tau_{sim}(N)$  is decreasing faster than  $\tau_{finish}(N)$ .

### *Away from ideal*

Let us summarise the assumptions made about the grid so that the above equation can be derived: (i) the number of computers available is large (ii) all computers are the same (iii) each block to be sent to the grid contains equal amount of work. In the following section let us try to come up with formulas for the less ideal case.

### **Relaxing: number of computers is large**

If  $N$  gets larger than  $M$ , then it is not enough to wait for one block to finish, as there are not enough processors to distribute all the blocks. Hence, the equation (6) is to be modified as:

$$\begin{aligned} & \hat{T}_{init} + T_{i/o} + N \cdot T_{grid} + \zeta_N \cdot \tau_{fin}(N) \\ & + (\zeta_N - 1) \cdot \tau_{init} < (N-1) \cdot \tau_{sim}(N) \end{aligned} \quad (10)$$

and

$$\zeta_N = 1 + \text{int}\left(\frac{N-1}{M}\right). \quad (11)$$

Conceptually this means that it is not enough to wait for only one block, but for  $\zeta_N$  blocks. Also, note that not only the finishing bits are present on the cost side of the equation, but also the initialisation ( $\tau_{init}$ ) bits, that is due to the fact, that if two or more blocks are running after each other, they have to be initialised separately. It is easy to see that the optimum for run-time on the grid is when  $N = M$ , supposing that  $\tau_{finish}(N) < \tau_{sim}(N)$  still holds. However, if  $N$  is just one bigger than  $M$ , then the completion time nearly doubles. One needs to wait

$$\tau_{fin}(N+1) + \tau_{init} + \tau_{sim}(N+1) \quad (12)$$

more, as the last block will run on a single machine while the other machines will be idle. This cost in run-time can be substantial unless one is careful with the issue.

If  $N = 2M+1$ , the situation improves, as the idle machines will be waiting for completion of the last task only for shorter time:

$$\tau_{fin}(2N+1) + \tau_{init} + \tau_{sim}(2N+1) \quad (13)$$

as the workload of one task decreases if we increase the number of tasks. The simulation part for example:

$$1/2 \tau_{sim}(N) = \tau_{sim}(2N). \quad (14)$$

Thus, breaking one task into many blocks seems to be a good idea, as the residual tasks that one might have to wait for are shorter, however, increasing the number of blocks costs run-time (managing the blocks on the grid, initialising each block, etc.).

If the number of machines on the grid fluctuates, or it is not known in advance, the tasks can be split into  $N \gg M$  blocks in order to minimise this residual cost (one or a few blocks remaining running at the end). Of course one also has to keep in mind the additional cost of doing so. This would determine the upper boundary of  $N$ .

In the following section we focus our attention on a grid where the computing units are not the same anymore. Here, we have to address the issue as setting  $M = N$  does not solve the problem any more.

### Relaxing: all computers are the same

The result of equation (10) holds, because of the assumption that all computers are the same in terms of computing power. Relaxing this assumption brings two issues to the fore: (i) the need of measuring the performance of each computer on the grid (ii) the ability of the grid to distribute task blocks accordingly.

The question is not only whether it is worth running the task on a grid or not. It is also the question how to schedule the blocks in such a way that the tasks finish as soon as possible. The scheduler should be able to estimate workload on each processor on the grid, and based on this set of information the scheduler should decide how to distribute the blocks of tasks.

First a function that will measure how many blocks are completed in a given time is constructed. Ignoring the time cost of maintaining the grid, the following can be written:

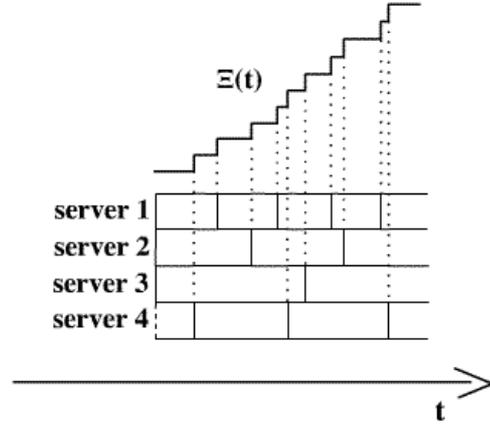
$$\Xi(t) = \sum_{i=1}^M \text{int} \left( v_i \cdot \frac{t}{w(N)} \right), \quad (15)$$

where  $w$  is the time consumption of one block

$$w(N) = \tau_{init} + \tau_{sim}(N) + \tau_{finish}(N) \quad (16)$$

and  $v_i$  is the speed (CPU power) of the processor for the  $i$ -th machine on the grid (for  $i=1..M$ ). Let us illustrate the formula above on the figure below (see fig. 2.). It depicts four servers, each running on different speed, so for each of them calculating the unit block takes different

times (time is denoted by  $t$  on the horizontal axis).

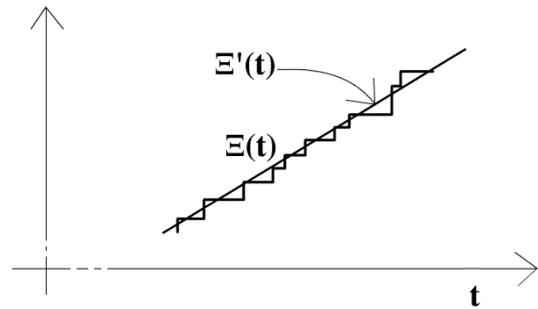


**Figure 2:** Step function describing how many blocks have been finished on the grid by time  $t$ . The blocks plotted next to each server, symbolise one unit of task.

Now, let us write  $\Xi(t)$  in the following form:

$$\begin{aligned} \Xi(t) &= \frac{\langle v \rangle_{i=1..M}}{w(N)} \cdot t + \mathcal{G}(t, M, \dots) \\ &= \Xi'(t) + \mathcal{G}(t, M, \dots) \end{aligned} \quad (17)$$

This would then translate into the following figure:



**Figure 3:** The exact and 'average' step function is plotted against time describing how many tasks have been finished by a given time

The symbol  $\langle v \rangle$  means the average of the machine speed on the grid, and the term  $\mathcal{G}$  incorporates all fluctuation around the linear

term. Thus, on average the time needed to get another block finished is exactly  $w/\langle v \rangle$ .

Going back to the criteria for breaking up the task one finds that as far as the formula below holds

$$T_{grid} + \frac{\tau_{fin}(N) + \tau_{init}}{M} < \frac{\langle v \rangle}{w(N+1)} - \frac{\langle v \rangle}{w(N)}, \quad (18)$$

it is worth increasing the number of blocks, as one gains more on average having shorter blocks than loosing on increasing the cost of doing so on the grid. The context of this formula is probabilistic, as it is not known in advance how the scheduler is going to schedule the blocks, because of the inhomogeneity of the grid. Note, that  $\zeta_N$  is replaced with a smooth term such a way that the expected value of the difference between the original and the new term is zero:

$$\left\langle \zeta_N - \frac{1}{2} - \frac{N}{M} \right\rangle = 0. \quad (19)$$

#### Blocks: not equal workload

The last case left to handle is the case when one cannot guarantee that the workload of the blocks of tasks remains the same. Most of the results from the last section remain however true. There is a trick with which one can decrease the run-time easily: if the workload of each block is known in advance, the large blocks should come first, and the small ones later. This way it can be ensured, that until the very last seconds the full computing power of the grid is utilised, and it is not likely that the completion of the task waits only for one of the largest blocks to finish<sup>6</sup>. Scheduling the jobs this way is similar to packing a suitcase: big shoeboxes come first, filling up the gaps with socks last.

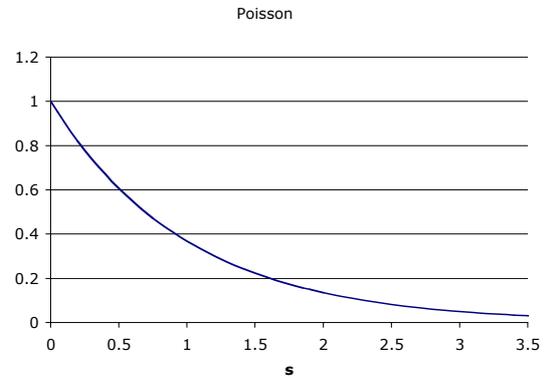
<sup>6</sup> Just by rearranging the order of subtasks, at EBRD a 20% decrease in run-time was realised.

#### How many jobs finishing on the same time?

It is interesting to note, that the fluctuating part of

$$\zeta_N - \frac{1}{2} - \frac{N}{M} \quad (20)$$

is known quite well<sup>7</sup>. If the system is random enough the distribution of the time needed to get another task finished is described by the so-called Poisson statistics<sup>8</sup> developed in the framework of *Random Matrix Theory* [Mehta 1967].



**Figure 4:** Decay of the Poisson spacing statistic

As seen on the density function of the Poisson statistic most of the time intervals are close to zero and for large  $N$  and  $M$  it will have a substantial effect on the network traffic. However, the tail of the distribution is decreasing more slowly than the normal distribution, thus relatively large time lags between two subtasks finishing after each other still occur. If the system is

<sup>7</sup> There are a few assumptions required for this result. The speed of the computers has to be 'heterogeneous enough' meaning that the speed of the computers should be different, also the size of the blocks should vary.

<sup>8</sup> The time between two blocks finishing on the grid has the same statistical properties as e.g. the energy level differences (spacing statistics) of a class of one dimensional quantum graph.

homogenous (lot of computers with the same computation capacity/speed on the grid), the distribution will be even more concentrated around zero. This shows that even in case of a heterogeneous system (different computers, different size of blocks) there will be times when many things will happen at once, like many jobs finishing at the same time (bad news for the grid/network) followed by calmer times when the capacity of the grid will not be fully utilised. The result is warning us that the whole construction of the grid (network, broker, etc.) should be designed such a way to be able to cope this clustered workload peaks. Another way would be to implement such mechanism that would repulse those finishing times that are too close to each other in order to smoothen out the workload on the grid over time.

## 5. Summary

One of the problems when generating potential future exposure profiles for complex portfolios lies in the huge amount of simulations to be performed. The issue can successfully be tackled by deploying grid-type computer architecture. It has been shown that the number of computers connected to the grid contributes to the performance in a step function manner. It was also demonstrated that the speed of each individual computer has a nonlinear impact on the speed of the grid. The last but not least point made confirms that the order in which the subtasks are assigned to the grid has a substantial effect on the total speed. All together the paper demonstrates that using a grid for simulating potential future exposure profiles is an efficient and feasible approach and the associated difficulties are not difficult to surmount.

## 6. References

Hull, J.C. [2002]: *Fundamentals of Futures and Options Markets*, Prentice-Hall International Inc, London.

Hull, J.C. [1993]: *Options, Future and other Derivative Securities*, Prentice-Hall International, Inc.

Duffie, D. [2001]: *Dynamic Asset Pricing Theory*, Princeton University Press, Princeton and Oxford.

Rebonato, R. [1998]: *Interest-rate Option Models*, Wiley, Chicester.

Jarrow, R.A. [1996]: *Modelling Fixed Income Securities and Interest Rate Options*, McGraw – Hill, New York.

Glasserman, P., Heidelberger, P. and Shahabuddin, P. [2000]: *Variance Reduction Techniques for Estimating Value-at-Risk*, *Management Science*.

Mehta, M.L. [1967]: *Random Matrices and the Statistical Theory of Energy Levels*, Academic Press, New York and London.

# TRADING VOLATILITY WITH OPTIONS ON STRADDLE

Levente Zsembery  
Department of Investment, Institute of Finance  
Budapest University of Economics and Public Administration  
H-1093, Fővám tér 8.  
Budapest, Hungary  
e-mail: zsembery@bkae.hu

## KEYWORDS

Option pricing, Stochastic volatility, Volatility trading

## ABSTRACT

The risk of volatility – even if many have not observed this yet – is as old as options. Together with the development of financial markets, newer and newer options, securities and forms of investments with embedded options appeared. As they spread, so grew the number of institutional and private investors running the risk of volatility.

In the eighties and nineties, markets became more and more volatile, thus the frequency of losses originating also grew. In parallel with those losses, demand to reduce the exposure to volatility also emerged. And demand creates its own supply. Banks seemed to be especially active in selling volatility products: they sold swaps and options on volatility. The risk so taken, they tried to hedge by the synthetic creation of the same volatility products. However, a good financial product not only has to be attractive for the market, but also has to be synthetically replicable.

Brenner et al [2002] tried to dissolve this contradiction. They offered investment banks an option on a combination of options (on a straddle) instead of options on volatility. This compound option becomes hedgeable without any major problem.

In this study, this new product was examined in two continuous time models. One of them is the Hull-White model [1987], which has always been considered as a benchmark in literature. In addition, in this case the correlation between the two processes – that of share price and volatility – is also easy to analyse. The other model dealt

with the logarithmic mean-reverting model used by Detemple-Osakwe [2000]. This gives a good description of volatility on the stock market, and Detemple-Osakwe priced volatility options by these processes. Therefore, the results of the paper are comparable with theirs.<sup>1</sup>

This paper shows that although the option on volatility is an adequate asset to hedge the volatility exposure, it can have very special price changes depending on the value of the correlation between the two processes. If the correlation changes, the main characteristics of the option on straddle can change essentially. It means, that banks selling this asset and hedging synthetically will face an other source of risk: the correlation risk.

The results were obtained using the Monte-Carlo simulation method. In the preparation of the programmes and in solving simulation-related problems Gábor Benedek provided valuable help.

## THE OPTIONS ON STRADDLE

Brenner, Ou and Zhang [2002] propose that the underlying of the option should not be volatility itself, but a product sensitive to it, a straddle with an ATM strike price. A great advantage of the latter is that it is traded, investors speculating on volatility have been using it for long, they know it well.

The structure of the contract is the following. The owner of the option can buy an ATM straddle in time  $T_1$  at the maturity of the option, which expires later in time  $T_2$ . The only problem is that it is not yet known what the ATM value

---

<sup>1</sup> (Further below I refer to the Hull-White model as HW, and to the Detemple – Osakwe model as DO. By process I mean the processes driving the share price and volatility.)

will be in time  $T_1$ . Brenner, Ou and Zhang extended their model for both the case of deterministic and stochastic volatility, from which the stochastic one is interesting for us.

Brenner, Ou and Zhang chose a mean-reverting process. Although in their model the stochastic element is a Wiener process, the size of the random element is independent from the current value of volatility.

The process followed by the underlying and that followed by volatility are the following:

$$dS_t = rS_t dt + \sigma_t S_t dZ_t^1 \quad (1.)$$

$$d\sigma_t = \delta(\theta - \sigma_t) dt + \xi dZ_t^2 \quad (2.)$$

where  $\theta$  is the long term mean of volatility,  $\delta$  is the speed of reversion to this mean,  $\xi$  is the volatility of volatility and  $Z^1$  and  $Z^2$  are two independent Wiener processes.

Brenner et al. showed that the value of this product is only a function of volatility and the share price. Delta hedging with the underlying of the primary option results in a volatility product depending only on volatility.

## THE PROCESSES

Literature offers many kinds of procedures to model the evolution of volatility as perfectly as possible. Below we deal with two models. One of them, often treated as a benchmark in literature, is the **Hull-White model**:

$$dS = \mu S dt + \sigma S dw \quad (17)$$

$$dV = \phi V dt + \xi V dz \quad (18)$$

where  $V = \sigma^2$ ,  $dw$  and  $dz$  are correlated Wiener processes, with a  $\rho$  correlation coefficient,  $\mu$  is the expected return of the share,  $\phi$  is the drift (expected growth rate) of the variance,  $\sigma$  is the volatility, and  $\xi$  is the volatility of volatility.

Hull and White assumed that the two processes can be correlated. The application of this model in this paper is also justified by the fact that it makes it possible to examine how correlation between the two processes affects the value of the option on straddle. Brenner, Ou and Zhang did not address this question.

The other model was taken from the article of **Detemple and Osakwe**. Detemple and Osakwe examined several processes, nevertheless they considered the mean-reverting log process as the most important. This was argued by this

model being – as they have shown – a continuous extension of the EGARCH model, hence it gives a good description of the American share prices. Accordingly, the parameters of the equations can be calculated as functions of the parameters used in the EGARCH model.

$$dS = \mu S dt + \sigma S dw$$

$$d \ln(\sigma) = (\alpha - \lambda \ln(\sigma)) dt + \xi dz$$

Where  $\sigma$  is the volatility<sup>2</sup>,  $dw$  and  $dz$  are the two  $\rho$  correlated Wiener processes,  $\mu$  is the expected return of the share,  $\alpha$  is the long-term logarithmic volatility value,  $\lambda$  is the speed of mean-reversion, and  $\xi$  is the volatility of volatility.

Notwithstanding the previously mentioned mean-reverting log model was analysed when pricing another volatility product by this process, that of the option on volatility. The characteristics of a **put** on straddle also were investigated.

For the sake of simplicity, it is assumed all through the analysis that the underlying is a share, which will not pay any dividend during the time examined.

## ABOUT THE NUMERICAL METHODS

### Form of denoting the processes

Similarly to the original articles, the logarithmic form of processes was used. This coincides with the format used by the authors in the original articles, and it ensures that the share price cannot have a negative value. The form of the two processes used in the simulations were the following:

Hull – White model:

$$S_i = S_{i-1} \cdot e^{[(r - V_{i-1}/2)\Delta t + u_i \sqrt{V_{i-1}} \Delta t]}$$

$$V_i = V_{i-1} \cdot e^{[(r - \xi^2/2)\Delta t + \rho u_i \xi \sqrt{\Delta t} + \sqrt{1 - \rho^2} v_i \xi \sqrt{\Delta t}]}$$

Detemple – Osakwe model:

$$S_i = S_{i-1} \cdot e^{[(r - \sigma_i^2/2)\Delta t + \sigma_i u_i \sqrt{\Delta t}]}$$

<sup>2</sup> Note that here the second process is the one driving volatility and not variance as in the Hull-White model!

$$\sigma_i = \sigma_{i-1} \cdot e^{\left[ (\alpha - \lambda \ln(\sigma_{i-1})) \Delta t + \xi \left( \rho u_i \sqrt{\Delta t} + \sqrt{1 - \rho^2} v_i \sqrt{\Delta t} \right) \right]}$$

### The description of the simulation

The length of the period was one year. Both the maturity of the option on straddle and that of the vanilla option was six months. For the sake of clarity, a year contained 360 days, which means 180 days in six months. In the simulations where the results of this paper were compared with that of Detemple and Osakwe, the maturity of the compound option was changed.

The programmes necessary for the simulations were written in Delphi programme language. The Gasdev random number generator was used, often used in economic simulations, to obtain random numbers.<sup>3</sup>

Each realisation was a result of a 1 000x1 000 simulation. To minimise the error in the results, each simulation was carried out ten times, and their arithmetic average was used in the analysis. As the simulation of one phase consisted of 180 periods and two random numbers were generated for one period, each compound option price was built 10x360 360 000 random numbers in order to minimise the noise of the random number generator.

Unfortunately, there is no closed analytical formula for these processes therefore the Black-Scholes setting was used for tests. The programme taking the average of ten realisations converged well to the reference values: as a result of ten realisations the difference (in absolute terms) remained between 0.1% and 0.2%.

Based on the tests, the Detemple and Osakwe model seemed to be more stable, as the results converged faster to reference values.

### Assumptions used for pricing

One of the most important questions when using stochastic volatility models is how to treat the problem of having two risk factors, share price

and volatility, and not having tools to mitigate the risk of the latter. Several ideas came up to solve this problem. One of the simplest is to assume, that there is a product, the underlying of which, is volatility itself; and by using this, volatility, risk can be mitigated. In other words: the market is complete, all risk factors are traded. This assumption is used, for example by Johnson and Shanno [1987].

In this case, the objective is to price such a volatility product. Therefore to assume that such a product already exists would not be too elegant and effective. Hence the other assumption widespread in literature – and also applied among others by Hull and White – was used, according to which volatility risk can be diversified and thus volatility has no systematic risk.

One should notice that from this point on there are two correlations in the models. On the one hand, it is sometimes assumed that processes driving the share price and volatility are correlated (in certain cases they are uncorrelated to some extent), while it is always assumed that correlation between the volatility of the share price and the expected return of the market portfolio is zero. The first assumption will be a subject of the analysis while the second ensures the completeness of the market.

## THE RESULTS

### The values of the options of straddle

In this paper, the effects of the initial volatility, the volatility of volatility, and the correlation of the two processes driving the share price and volatility on the value of options on straddle were analysed. In most cases, the results were similar to that of vanilla options, nevertheless sometimes they were different from Brenner et al's results.

As far as the **initial volatility** is concerned, the option on straddle behaves well: the value of CoST is a positive, while that of PoST is a negative function of initial volatility for both models. But not always a convex function! In the HW model convexity is a function of the correlation coefficient. In case of negative correlation, the value of the call on straddle will be a concave function of the initial volatility.

<sup>3</sup> For more details on the theory and practice of simulations see Knuth [1987] and Benedek [2003].

In the DO model, puts continue to behave regularly, that is their value is a concave function of volatility. In case of the CoST the effect is very interesting. By a correlation coefficient of plus or minus one, the compound option is a concave function of volatility, while in other cases convexity was observed.

These results are not unique. Detemple and Osakwe – using the mean-reverting log process – experienced that the value of options by a high initial volatility became a concave function of volatility.

Interesting was the influence of the **volatility of volatility** as well. The volatility of volatility was examined by Brenner et al., too. In their model, the compound option behaved absolutely “regularly” as a function of the volatility of volatility. But not in these models.

In the HW model, this issue can be addressed only together with that of the correlation coefficient, thus also the next question is answered partly. In the case of zero correlation both CoST and PoST behave regularly, while in case of strong negative correlation the volatility of volatility can have a reducing effect on the value of options in certain phases. The issue is even more important as in practice volatility and share price are typically negatively correlated.

Unique characteristics can also be found in the DO model. The atypical behaviour of option is not only a function of correlation. In case there was zero correlation between the two processes, the CoST did behave as a vanilla option. But not in the case of put options. By higher exercise prices the value of the PoST was a negative function of the volatility of volatility! At the same time by reducing the exercise price the order was restored.

The question is raised whether this relation changes by changing the correlation coefficient. Therefore, the analysis of the value of the PoST should be continued. The value of PoST was analysed by correlation coefficients of minus and plus one.

It can be easily observed that the behaviour observed is independent from the correlation between the two processes, the sensitivity of the option's price to the volatility of volatility will be a function of the exercise price. By a high exercise price, in case of strongly ITM put options, the volatility of volatility reduces the

price of the compound option, while in case of OTM put options, it is going to increase the same.

There was no such problem in case of call on straddle, they behaved as an option on volatility independent from the correlation coefficient.

In the end, it is worth to note, that all throughout the analysis of Brenner et al., the compound option behaved regularly, it was a positive function of the volatility of volatility. Of course, it should be mentioned that they only dealt with call options, in this sense my results do not differ from theirs.

The effects of the **correlation** between share price and volatility processes were also analysed. The issue of correlation was important as it was not addressed by Brenner et al., whereas it is a general observation on the market that there is negative correlation between the volatility and the share price.

In the case of the HW model, the results were interesting. The hint can be put this way: in case of close-to-ATM options, the increase of the correlation coefficient increases the value of both CoST and PoST. As in general, ATM options are the most frequently traded on the market, this rule is going to be the most important one for banks selling these types of combined derivatives. However, it is good to know that when trading with an exercise price different from ATM, the effect might change. Of course in practice, it is also an important question, whether the HW model gives a good estimation of reality, the evolution of the share price and volatility.

In case of the DO model the value of CoST is more sensitive to changes of correlation, than that of PoST.<sup>4</sup> The value of the compound options – both calls and puts – changes parabolically as a function of the correlation between the two processes. While in case of CoST, the price is the highest by zero correlation, in case of PoST is the value of the compound option the lowest.

By CoST, regardless of whether we differ positively or negatively from zero correlation, the value of the option reduces. By PoST, the effect is the other way round. Zero correlation

---

<sup>4</sup> The strength of the effect also depends on the volatility of volatility.

gives the lowest option price. As the absolute value of  $\rho$  grows, PoST becomes more valuable. The importance of the results obtained can again be supported by practice. The value of correlation between share price and volatility can change, therefore the option might be mispriced. When disregarding correlation, call options are systematically overpriced and puts are systematically underpriced.

Changing the exercise price did not cause such a difference as in the HW model. The influence is mostly the same by all exercise prices.

The effects of the correlation between the two processes and of the volatility of volatility are interdependent, the simultaneous increase of the two factors will increase the value of CoST and reduce the value of PoST. In case the latter is zero, the correlation between the two processes ceases to exist. Accordingly the higher the volatility of volatility, the stronger effect will correlation have.

## CONCLUSION

To sum it up, calls on straddle can have special characteristics by other processes than that analysed by Brenner et al.. They not always behave regularly. As it was shown, the correlation of the two processes has a significant effect on the value of the compound options. As a consequence, if we create a volatility neutral

position by using options on straddles, we will have a correlation sensitive position. So we change one source of risk for another one.

On the other hand, the value of puts on straddle was analysed as well. As can be seen, calls and puts can work differently in certain situations. In some cases puts can be more attractive products for hedgers than their call counterparties.

## REFERENCES

- Brenner, M. – Ou, E. Y. – Zhang J. E. [2002]: Hedging Volatility Risk, EFA 2002 Berlin Meetings Discussion Paper,
- Detemple, J. – Osakwe, C. [2000]: The Valuation of Volatility Options, European Finance Review, Vol. IV., 2000. pages 21-50.
- Benedek, G. [2003]: Evolúciós Gazdaságok Szimulációja (*The Simulation of Evolutionary Economies*), PhD Thesis, 2003.
- Hull, J. C. – White, A. [1987]: The Pricing of Options on Assets with Stochastic Volatilities, Journal of Finance, Vol. XLII., 1987 June pages 281-300.
- Johnson, H. – Shanno, D. [1987]: Option Pricing when the Variance Is Changing, Journal of Financial and Quantitative Analysis, Vol. 22. Nr. 2. 1987 June.
- Knuth, D. E. [1987]: The Art of Computer Programming (Hungarian Edition), Műszaki Könyvkiadó, Budapest, 1987.

# SIMULATING MULTI-DIMENSIONAL LATTICES WITH CORRELATION: A CASE STUDY

Javier Otamendi\*

Mark T. Hon

Institute for the Advancement of Business & Technology

Saint Louis University

Madrid Campus, Avenida del Valle 34,

28003 Madrid, Spain

\*Tel: +34 91554 5858 Email: otamendij@madrid.sluiberica.slu.edu

## KEYWORDS

Lattice, Correlation, Monte Carlo Simulation

## ABSTRACT

We employ a new numerical approximation technique for valuing multiple assets that are correlated. For any number of underlying variables, this procedure can simulate a set of correlated assets that follow geometric Brownian processes. In addition to the joint approximation of multiple series based on binomial lattices and Monte Carlo simulations, we introduce a straightforward extension of a structural lead-lag approach to capture the linear relationship between assets in the bid to improve computational accuracy on the proposed algorithm. We show that when applied to a series of correlated assets, the proposed algorithm is consistent, efficient and stable.

## INTRODUCTION

The need for evaluating contingent claims on more than one state variable is not uncommon in most financial investment decisions. Though many numerical methods (most approaches can be classified into three broad classes: Lattice methods of Cox, Ross and Rubinstein, 1979, Monte Carlo simulations of Boyle, 1977, and finite difference methods of Brennan and Schwartz, 1977) have been proposed for pricing options in the financial literature, an outstanding challenge in computational finance is the simulation of portfolios with more than two correlated assets.

Lattices or trees were first introduced by Cox, Ross and Rubinstein (1979) to complement the Black and Scholes (1973) closed-form solution model where asset prices follow stochastic processes with multiplicative sequences of variable drifts (i.e., geometric Brownian motion). To obtain the terminal value for an option, the expected option payoff is weighted by its probability via the lattice and the expectation is discounted back at the risk-free rate. Using discrete binomial distributions to approximate the continuous lognormal distributions in the Black-Scholes model, it

offers an efficient way of generating series of predetermined paths for pricing both American and European options (contingent claims on European options can only be exercised at maturity, while American options may be exercised at intermediated dates, not just at maturity. For a comprehensive exposition on derivatives, see Hull (2000)).

By and large, binomial trees can be configured in different ways according to a set of parameters for probabilities of 'up' and 'down' steps, and step size (see Jarrow and Rudd, 1983 and Leisen and Reimer, 1996 for example). In this paper, we employ the Cox, Ross and Rubinstein (1979) procedure where asset prices are modeled such that the product of up and down price multipliers equals to unity in a risk-neutral valuation environment. Clearly, as the number of replications in the lattice increases, the simulated value converges to that of the Black-Scholes pricing solution.

Existing papers on lattices are usually built upon a single-asset problem with time-dependent volatility, or multiple assets with constant correlation and volatility. Boyle, Evnine and Gibbs (1989) extend the Cox, Ross and Rubinstein approach to multiple assets but admit their technique is vulnerable to negative martingale probabilities when the correlation between assets is too large or if the volatilities are time-dependent. These negative martingale probabilities can lead to unstable pricing structures and if the number of state variables is large, the exponential complexity of the multi-dimensional binomial approach can prove to be a computational burden. Other extensions of the lattice procedure that can be computationally exhaustive include that of Boyle (1988) and Kamrad and Ritchken (1991) where trinomial algorithms are used to estimate multiple state variables.

Monte Carlo simulation on the other hand, entails the generation of asset price paths with a stochastic process. It is known to be ill equipped to handle problems involving early exercises (i.e., American options) but it enjoys computational advantages over pure lattice approximations for multiple-asset scenarios. With a new algorithm that is

computationally efficient and stable, Otamendi and Hon (2004) demonstrate that it is possible to construct a combination of binomial lattice approximation with Monte Carlo simulation where probabilities stay positive even if the multiple state variables are highly correlated geometric Brownian processes.

In this paper, we simulate time paths for a portfolio with multiple correlated assets using a new joint lead-lag binomial lattice-Monte Carlo approximation proposed by Otamendi and Hon on four assets that are arbitrarily chosen from the New York Stock Exchange (NYSE) and the National Association of Securities Dealers (NASDAQ): Boeing Company (NYSE: BA), International Business Machines (NYSE: IBM), Hewlett-Packard (NYSE: HPQ) and Microsoft Corporation (NASDAQ: MSFT).

This article is organized as follows: We first present a numerical exposition of existing two-asset simulation models. The new algorithm for a two-asset case follows. The proposed algorithm is further extended to cover a three-asset example before we present a generalized  $k$  series elucidation of the methodology with a built in lead-lag structure. Finally, we apply the algorithm to a portfolio of four assets to show that the model is consistent, efficient and stable.

## EXISTING TWO-ASSET MODELS

Most existing two-asset models are either based on Monte Carlo or lattice approximations.

### Monte Carlo Simulation

This particular type of simulation is based on the continuous evolution of the series where the movement process follows a Brownian motion, and the variability a Wiener process. If the time period  $D$  is divided by small intervals  $T$  ( $t=0, \dots, T$ ), the recursive equation to estimate  $V_t$  is (see Haug 1998):

$$V_t = V_{t-1} \exp \left[ \left( r - \frac{1}{2} v^2 \right) \frac{D}{T} + v \varepsilon_t \sqrt{\frac{D}{T}} \right]$$

where:

$V_t$  = value of the series at time  $t$

$r$  = annual interest rate

$v$  = volatility of the series

$D$  = time horizon (expressed in years)

$T$  = number of intervals per time period  $D$

$\varepsilon_t \sim N(0, 1)$

Note that the consecutive independent samples follow a normal distribution. To estimate  $V_T$  and  $T$ , independent samples must be generated for each time period  $t$ .

For a correlated two-asset series, the recursive equations (Haug 1998) are:

$$V_{1,t} = V_{1,t-1} \exp \left[ \left( r - \frac{1}{2} v_1^2 \right) \frac{D}{T} + v_1 \alpha_{1t} \sqrt{\frac{D}{T}} \right]$$

$$V_{2,t} = V_{2,t-1} \exp \left[ \left( r - \frac{1}{2} v_2^2 \right) \frac{D}{T} + v_2 \alpha_{2t} \sqrt{\frac{D}{T}} \right]$$

where:

$V_{k,t}$  = value of the asset  $k$  at period  $t$

$\alpha_{1t} = \varepsilon_{1t}$

$\alpha_{2t} = \rho \varepsilon_{1t} + \varepsilon_{2t} \sqrt{1 - \rho^2}$

$\rho$  = correlation between the two assets

## Binomial Tree

With a binomial tree, the path taken by the asset at each time step is either a positive upward ( $u$ ) jump with a probability  $p$ , or a negative downward move  $d$ , with a probability  $(1-p)$ . At time  $t=0$  with an initial value  $V_0$ , the successive values  $V_t$  are obtained with the following recursive equations:

$$V_t = \begin{cases} V_{t-1} u & \forall U[0,1] \leq p \\ V_{t-1} d & \forall U[0,1] > p \end{cases}$$

where:

$$u = e^{v \sqrt{D/T}}$$

$$d = 1/u$$

$$p = (f-d)/(u-d)$$

$$f = 1 + (rD/T)$$

The pseudo-code for simulating  $V_T$  is as follows:

```
Obtain  $V_0, D, T$ 
Calculate  $p, u, d$ 
 $t \leftarrow 0$ 
While  $t < T$ 
  Generate a random sample  $\tau = U[0,1]$ 
  If  $\tau \leq p$ 
     $V_{t+1} = V_t u$ 
  Else
     $V_{t+1} = V_t d$ 
   $t \leftarrow t+1$ 
```

The binomial tree model can be modified (Rubinstein, 1994) to simulate two assets with correlation  $\rho$ . For asset 1,  $u$  and  $d$  are redefined:

$$u_1 = \exp \left[ \left( r - \frac{1}{2} v_1^2 \right) \frac{D}{T} + v_1 \sqrt{\frac{D}{T}} \right]$$

$$d_1 = \exp \left[ \left( r - \frac{1}{2} v_1^2 \right) \frac{D}{T} - v_1 \sqrt{\frac{D}{T}} \right]$$

$$p_1 = (f-d_1)/(u_1-d_1)$$

For asset 2, A and B are defined as the time-step jumps  $u_2$  and  $d_2$  if asset 1 moved to  $u_i$ ; C and D are the time-step jumps  $u_2$  and  $d_2$  if asset 1 moved to  $d_i$ :

$$A = \exp \left[ \left( r - \frac{1}{2} v_2^2 \right) \frac{D}{T} + v_2 \sqrt{\frac{D}{T}} \left( \rho + \sqrt{1 - \rho^2} \right) \right]$$

$$B = \exp \left[ \left( r - \frac{1}{2} v_2^2 \right) \frac{D}{T} + v_2 \sqrt{\frac{D}{T}} \left( \rho - \sqrt{1 - \rho^2} \right) \right]$$

$$C = \exp \left[ \left( r - \frac{1}{2} v_2^2 \right) \frac{D}{T} - v_2 \sqrt{\frac{D}{T}} \left( \rho - \sqrt{1 - \rho^2} \right) \right]$$

$$D = \exp \left[ \left( r - \frac{1}{2} v_2^2 \right) \frac{D}{T} - v_2 \sqrt{\frac{D}{T}} \left( \rho + \sqrt{1 - \rho^2} \right) \right]$$

$$p_2 = 0.5$$

The pseudo-code to simulate the two assets is:

```

Obtain  $V_{1,0}, V_{2,0}, \rho, D, T$ 
Calculate  $p_1, u_1, d_1, A, B, C, D$ 
 $t \leftarrow 0$ 
While  $t < T$ 
  Generate a sample  $\tau_1 = U[0,1]$ 
  If  $\tau_1 \leq p_1$ 
     $V_{1,n+1} = V_{1,n} u$ 
    Generate a sample  $\tau_2 = U[0,1]$ 
    If  $\tau_2 \leq 0.5$ 
       $V_{2,n+1} = V_{2,n} A$ 
    Else
       $V_{2,n+1} = V_{2,n} B$ 
  Else
     $V_{1,n+1} = V_{1,n} d$ 
    Generate a sample  $\tau_2 = U[0,1]$ 
    If  $\tau_2 \leq 0.5$ 
       $V_{2,n+1} = V_{2,n} C$ 
    Else
       $V_{2,n+1} = V_{2,n} D$ 
   $t \leftarrow t+1$ 

```

## PROPOSED MODEL

The models introduced in the previous section are useful for two assets, but they cannot be generalized for three or more variables. We now introduce the proposed model that can be generalized for two or more assets.

### Proposed Two-Asset Model

Because asset returns often exhibit a high degree of correlation in practice, we have to simulate correlated normal random variables to proxy for the correlated returns.

Let  $X_i$  and  $X_j$  be two random variables. The covariance of  $X_i$  and  $X_j$  is defined to be:

$$Cov(X_i, X_j) = E[X_i X_j] - E[X_i] E[X_j]$$

and the correlation of  $X_i$  and  $X_j$  is then defined to be

$$Corr(X_i, X_j) = \rho(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}}$$

If  $X_i$  and  $X_j$  are independent, then  $\rho = 0$ . The proposed algorithm is formulated in part with a correlation-based lattice. If we have a correlation of  $\rho = 1$  between the two assets, both assets will have identical co-movements (e.g., if asset  $i$  moves to the upper bifurcation in the binomial tree, asset  $j$  will move in the same direction).

If  $\rho = -1$ , the movements per time step will be exactly the opposite for both series. And when  $\rho = 0$ , there is no relationship between the two series (i.e., they are independent).

Simulating the lattice for a two-variable (asset 1 and asset 2) case based on the above exposition is straightforward: In time period  $t$ , we first simulate the path for asset 1. If it experiences an upward movement, asset 2 must follow a downward movement if  $\rho = -1$ , a random movement if  $\rho = 0$  and a corresponding upward movement if  $\rho = 1$ . In other words, the probability of matching co-movements is  $\beta = 0\%$  in the first case,  $\beta = 50\%$  in the second, and  $\beta = 100\%$  in the third case.

The structural relationship between the coefficient of correlation  $\rho$  and the probability of co-movement ( $\beta$ ), might then be set over the entire range of possible values for  $\rho$  ( $-1 \leq \rho \leq 1$ ) with a simple linear transformation:

$$\beta = 0.50 + 0.50 \rho \quad \forall -1 \leq \rho \leq 1$$

The pseudo-code based on the proposed co-movement probability model is as follows:

```

Obtain  $V_{1,0}, V_{2,0}, \rho_{12}, D, T$ 
Calculate  $u_1, d_1, p_1, u_2, d_2, \beta_{1,2}$ 
 $t \leftarrow 0$ 
While  $t < T$ 
  Generate a random sample  $\tau_1 = U[0,1]$ 
  If  $\tau_1 \leq p_1$ 
     $V_{1,t+1} = V_{1,t} u_1$ 
    Generate a random sample  $\tau_2 = U[0,1]$ 
    If  $\tau_2 \leq \beta_{1,2}$ 
       $V_{2,t+1} = V_{2,t} u_2$ 
    Else
       $V_{2,t+1} = V_{2,t} d_2$ 
  Else
     $V_{1,t+1} = V_{1,t} d_1$ 
    Generate a random sample  $\tau_2 = U[0,1]$ 
    If  $\tau_2 \leq (1 - \beta_{1,2})$ 
       $V_{2,t+1} = V_{2,t} u_2$ 
    Else
       $V_{2,t+1} = V_{2,t} d_2$ 
   $t \leftarrow t+1$ 

```

where  $\beta_{1,2}$  is the probability of co-movement between the assets 1 and 2 - obtained from the structural relationship between  $\rho$  and  $\beta$  specified above ( $\beta = 0.50 + 0.50 \rho \quad \forall -1 \leq \rho \leq 1$ ).

### Validation of the Proposed Model

To demonstrate that the asset value at maturity for the proposed model follows a lognormal distribution (and that the correlation between the two assets is  $\rho$ ), we perform two separate simulations: A Monte Carlo Brownian process and the proposed model based on the following parameters:

	Asset 1	Asset 2
Initial Value ( $V_0$ )	9	14
Correlation ( $\rho$ )	0.70	0.70
Annual Interest ( $r$ )	4%	4%
Interval ( $T$ )	63	63
Time ( $D$ )	0.25	0.25
Volatility ( $v$ )	39%	25%

The frequency distributions for assets 1 and 2 are shown in Figure 1 below:

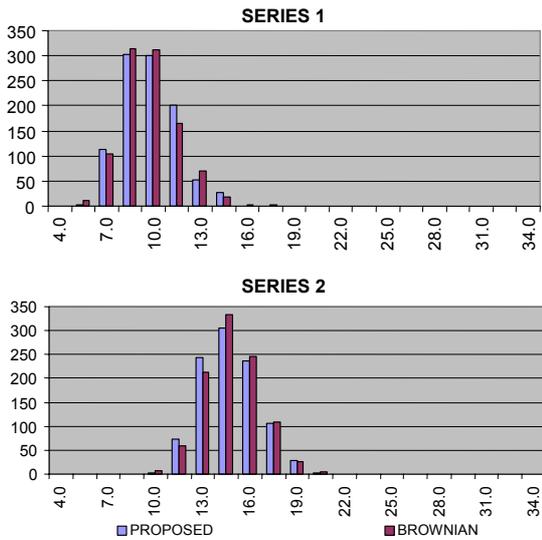


Figure 1. Frequency Distribution for Asset 1 (Series 1) and Asset 2 (Series 2)

A  $\chi^2$  goodness-of-fit test (see Law, 1991 for more on statistical tests for simulations) is used to test the null hypothesis that both models yield identical results. Even with intervals  $T$  as small as 63, we fail to reject the null with a confidence level of 99.73% (a total of 2000 simulation runs were performed).

We also test to see if the correlation at maturity is what it should be (see Ruiz-Maya and Martín Pliego, 1995 for example):

$$\begin{aligned} H_0: \rho &= 0.70 \\ H_1: \rho &> 0.70 \end{aligned}$$

The results are displayed in Table 1 (with a 99.73% confidence level).

Table 1. Confidence Intervals for the Correlation Coefficients

	BROWNIAN MODEL	PROPOSED MODEL
	1-2	1-2
Minimum	-0.6751	-0.6976
Confidence Interval (Inferior)	0.5966	0.5934
Mean	0.6542	0.6514
Confidence Interval (Superior)	0.7052	0.7027
Maximum	0.9898	0.9929
	PASS	PASS

As both confidence intervals contain the theoretical values 0.70, the null hypothesis cannot be rejected.

### PROPOSED MODEL FOR THREE ASSETS

#### Formulation

The generalization of a three-asset bifurcation is shown in Figure 2. Note that as a function of co-movements between asset 1 ( $C_1=2$ ) and asset 2 ( $C_2=4$ ), asset 3 contains eight possible outcomes at each time period ( $C_3=8$ ):

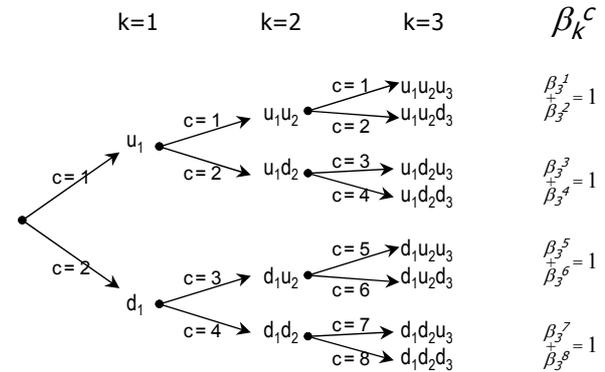


Figure 2. Probability Generalization for  $k$  Assets at Time  $t$  ( $\beta_k^c$  is the probability of co-movement for asset  $k$  and the bifurcation ( $c$ ))

Because the eight outcomes are incompatible in general, the sum of the probabilities for each outcome is 1. However, with regard to the probability of co-movement between assets ( $\beta$ ), we have pair-wise incompatibility. To generate the value of asset 3, a starting point is required (after generating asset 2). If for example, the starting bifurcation ( $c$ ) is at the first branch for asset 2 ( $u_1u_2$ ), we can generate bifurcation  $c=1$  ( $u_1u_2u_3$ ) or bifurcation  $c=2$  ( $u_1u_2d_3$ ) for asset 3. As a result, the co-movement probabilities  $\beta_3^1$  and  $\beta_3^2$  sum up to unity (1). Likewise, the same transpires in pairs:  $\beta_3^3$  and  $\beta_3^4$ ,  $\beta_3^5$  and  $\beta_3^6$ ,  $\beta_3^7$  and  $\beta_3^8$ .

The probabilities  $\beta_3^1$  and  $\beta_3^8$  are the identical since they represent the occurrence of bifurcation  $c=1$  and

bifurcation  $c=8$ , which happens to be the case where the co-movements for all three assets match. Based on the same argument, we observe that  $\beta_3^3 = \beta_3^6$  since there is co-movement between asset 1 and asset 3, but not with asset 2.

Consequently, before simulating asset 3, we know *a priori* that:

$$\begin{array}{ll} \beta_3^1 = \beta_3^1 & \beta_3^3 = \beta_3^3 \\ \beta_3^5 = \beta_3^4 = 1 - \beta_3^3 & \beta_3^7 = \beta_3^2 = 1 - \beta_3^1 \\ \beta_3^2 = 1 - \beta_3^1 & \beta_3^4 = 1 - \beta_3^3 \\ \beta_3^6 = \beta_3^3 & \beta_3^8 = \beta_3^1 \end{array}$$

The crucial values left are  $\beta_3^1$  and  $\beta_3^3$ . They can be obtained using simple probability theory (with the definition of classical probability).

$\beta_3^1$  is the probability of asset 3 moving to  $u_3$  divided by the probability of moving to the upper bifurcation or lower bifurcation ( $u_3$  or  $d_3$ ); if assets 1 and 2 have moved to  $u_1$  and  $u_2$  respectively. The jump to  $u_3$  is equivalent to the condition that the correlations of  $\rho_{13}$  and  $\rho_{23}$  exist and that it has been converted to the probability of co-movement using the structural relationship between  $\beta$  and  $\rho$  ( $\beta_{1,3}$  and  $\beta_{2,3}$ ). The movement to  $d_3$  can also be derived from  $(1-\beta_{1,3})$  and  $(1-\beta_{2,3})$ :

$$\beta_3^1 = \frac{\hat{\beta}_{1,3} \hat{\beta}_{2,3}}{(1-\hat{\beta}_{1,3})(1-\hat{\beta}_{2,3}) + \hat{\beta}_{1,3} \hat{\beta}_{2,3}}$$

In the case of  $\beta_3^3$ , the outcome for  $u_1 d_2 u_3$  is:

$$\beta_3^3 = \frac{\hat{\beta}_{1,3} (1-\hat{\beta}_{2,3})}{(1-\hat{\beta}_{1,3})\hat{\beta}_{2,3} + \hat{\beta}_{1,3} (1-\hat{\beta}_{2,3})}$$

The pseudo-code for a three-asset approximation described above is:

```

Obtain  $V_{1,0}, V_{2,0}, V_{3,0}, \rho_{12}, \rho_{13}, \rho_{23}, D, T$ 
Calculate  $u_1, d_1, p_1, u_2, d_2, u_3, d_3$ 
Calculate  $\beta_{1,2}, \beta_{1,3}, \beta_{2,3}, \beta_3^1, \dots, \beta_3^8$ 
 $t \leftarrow 0$ 
While  $t < T$ 
  Generate a random sample  $\tau_1 = U[0,1]$ 
  If  $\tau_1 \leq p_1$ 
     $V_{1,t+1} = V_{1,t} u_1$ 
    Generate a random sample  $\tau_2 = U[0,1]$ 
    If  $\tau_2 \leq \hat{\beta}_{1,2}$ 
       $V_{2,t+1} = V_{2,t} u_2$ 
      Generate a random sample  $\tau_3 = U[0,1]$ 
      If  $\tau_3 \leq \beta_3^1$ 
         $V_{3,t+1} = V_{3,t} u_3$ 
      Else
         $V_{3,t+1} = V_{3,t} d_3$ 
    Else
       $V_{2,t+1} = V_{2,t} d_2$ 
      Generate a random sample  $\tau_3 = U[0,1]$ 
      If  $\tau_3 \leq \beta_3^3$ 
         $V_{3,t+1} = V_{3,t} u_3$ 
      Else
         $V_{3,t+1} = V_{3,t} d_3$ 
    Else
       $V_{1,t+1} = V_{1,t} d_1$ 
      Generate a random sample  $\tau_2 = U[0,1]$ 
      If  $\tau_2 \leq (1-\hat{\beta}_{1,2})$ 
         $V_{2,t+1} = V_{2,t} u_2$ 
        Generate a random sample  $\tau_3 = U[0,1]$ 
        If  $\tau_3 \leq \beta_3^5$ 
           $V_{3,t+1} = V_{3,t} u_3$ 
        Else
           $V_{3,t+1} = V_{3,t} d_3$ 
      Else
         $V_{2,t+1} = V_{2,t} d_2$ 
        Generate a random sample  $\tau_3 = U[0,1]$ 
        If  $\tau_3 \leq \beta_3^7$ 
           $V_{3,t+1} = V_{3,t} u_3$ 
        Else
           $V_{3,t+1} = V_{3,t} d_3$ 
     $t \leftarrow t+1$ 

```

Note that the correlation matrix may not always be valid: A congruent matrix is required (a positive definite or positive semi-definite correlation matrix - Morrison, 1990). For example, it is unlikely that we will obtain a perfect correlation between asset 1 and asset 2 ( $\rho=1$ ). If the correlation between asset 1 and 2 is perfect, the correlation between assets 1 and 3, and assets 2 and 3 will be zero.

### Proposed Model Validation

A third asset is added to the two-asset example that was used previously for validation of a two-asset case (the data is the same from the previous section for the first two assets - the correlation matrix is positive definite (Schmidt (1981)):

Initial value ( $V_0$ ) = 18  
Volatility ( $v$ ) = 28%  
Correlation ( $\rho$ ) between asset 1&3 = 0%

Correlation ( $\rho$ ) between asset 2&3 = -5%

The frequency histogram for asset 3 is depicted in Figure 3.

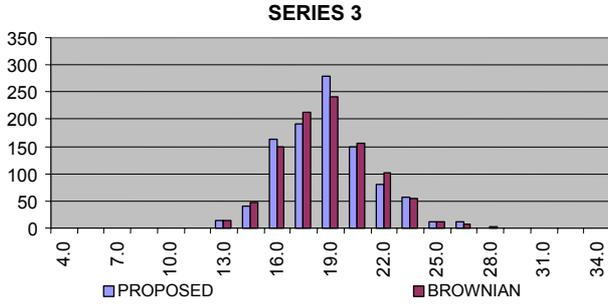


Figure 3. Frequency Distribution for Asset 3 (Series 3)

The results with a 99.73% confidence level show that the proposed model follows a lognormal process. Table 2 confirms that the correlations between assets are valid.

Table 2. Asset Correlation

	BROWNIAN MODEL		PROPOSED MODEL	
	1-2	1-2	1-3	2-3
Minimum	-0.6751	-0.6976	-0.9396	-0.9571
Confidence Interval (Inferior)	0.5966	0.5934	-0.1408	-0.1419
Mean	0.6542	0.6514	-0.0468	-0.0480
Confidence Interval (Superior)	0.7052	0.7027	0.0481	0.0469
Maximum	0.9898	0.9929	0.9377	0.9378
	PASS	PASS	PASS	PASS

It is clear at this point that the proposed model can be generalized for two or more assets.

### PROPOSED MODEL FOR $K$ SERIES

In this section, we present the generalization for three assets or more. If  $K$  is the total number of assets, an algorithm can be developed to simulate asset  $k$  as a function of the  $k-1$  series previously simulated. As in the two- and three-asset models, the bifurcations are generated for each of the asset and the probabilities of co-movements obtained. The assets are further simulated for each time period till the maturity period.

### Incorporated Lead-Lag Structure

To capture the lead-lag (i.e., the order of  $k$  asset simulation) transmission between assets in the proposed model, we first define the changes in daily returns:

$$\text{Let } R_t^k = \text{Log}\left(\frac{P_t^k}{P_{t-1}^k}\right)$$

where  $P_t^k$  ( $P_{t-1}^k$ ) represents the closing price of asset  $k$  on day  $t$  ( $t-1$ ).

For distinguishing the specific lead-lag effects (the number of possibilities is  $P_k = k!$ ), we use the following system of equations (for conciseness, we use a two-asset example):

$$R_t^i = a_i + b_{ij}R_{t-1}^j + b_{ii}R_{t-1}^i + \varepsilon_t^i$$

$$R_t^j = a_j + b_{ji}R_{t-1}^i + b_{jj}R_{t-1}^j + \varepsilon_t^j$$

where coefficients  $a_k$  are intercepts,  $b_k$  are slope coefficients and  $\varepsilon_k$  are error terms. To capture the serial correlation of returns on each asset, the preceding returns at time  $t-1$  are included as explanatory variables (the inclusion of this term will not affect the estimation process in any substantial way).

The coefficient  $b_{ij}$  ( $b_{ji}$ ) measures the effect of the daily returns of asset  $i$  ( $j$ ) on the following returns of asset  $j$  ( $i$ ). If the returns on asset  $i$  ( $j$ ) influence the subsequent returns on asset  $j$  ( $i$ ),  $b_{ji}$  ( $b_{ij}$ ) should be positive.

If the signal quality of asset  $i$  ( $j$ ) is better (in terms of its ability to signal a "lead" on the returns of asset  $j$  ( $i$ )) than that of asset  $j$  ( $i$ ), we will have  $b_{ji} > b_{ij}$ .

The proposed model uses iterated seemingly unrelated regressions (ITSUR) to estimate the above system of equations for all  $k$  assets simultaneously. By allowing cross correlations between error terms of all  $k$  assets, this procedure delivers more efficient estimates of coefficients than ordinary least squares (OLS) regressions in large samples. If the error terms are assumed to be normal the ITSUR technique will yield equivalent maximum likelihood estimators.

### Proposed Model Formulation

The number of bifurcations for individual asset  $k$  is  $C_k=2^k$ . They are generated as follows:

```

I1←1
While I1≤2
  I2←1
  While I2≤2
    ...
    Ik←1
    While Ik≤2
      m11, m12, ..., m1k
      Ik← Ik+1
    ...
  I2← I2+1
I1← I1+1

```

where  $m_{1i}=u_i$  if  $I_i=1$  and  $m_{1i}=d_i$  if  $I_i=2$ .

The  $\beta_k^c$  variables are equal in pairs:

$$\beta_k^c = \beta_k^{C_k-c+1} \quad \forall c = 1, \dots, C_k \text{ cases}$$

The pairs  $\beta_s^c$  and  $\beta_s^{c+1}$  (when  $c$  is odd) are calculated as follows:

$$\beta_k^c = \frac{\prod_{i=1}^{k-1} H_{i,k}}{\prod_{i=1}^{k-1} H_{i,k} + \prod_{i=1}^{k-1} (1 - H_{i,k})}$$

$$\beta_k^c = 1 - \beta_k^{c+1}$$

where:

$$H_{i,k} = \beta_{i,k} \quad \forall u_i$$

$$H_{i,k} = 1 - \beta_{i,k} \quad \forall d_i$$

For  $K = 4$  assets, the  $2^4 = 16$  bifurcations are:

- c=1  $u_1 u_2 u_3 u_4$
- c=2  $u_1 u_2 u_3 d_4$
- c=3  $u_1 u_2 d_3 u_4$
- c=4  $u_1 u_2 d_3 d_4$
- c=5  $u_1 d_2 u_3 u_4$
- c=6  $u_1 d_2 u_3 d_4$
- c=7  $u_1 d_2 d_3 u_4$
- c=8  $u_1 d_2 d_3 d_4$
- c=9  $d_1 u_2 u_3 u_4$
- c=10  $d_1 u_2 u_3 d_4$
- c=11  $d_1 u_2 d_3 u_4$
- c=12  $d_1 u_2 d_3 d_4$
- c=13  $d_1 d_2 u_3 u_4$
- c=14  $d_1 d_2 u_3 d_4$
- c=15  $d_1 d_2 d_3 u_4$
- c=16  $d_1 d_2 d_3 d_4$

To obtain  $\beta_4^5$  which corresponds to  $u_1 d_2 u_3 u_4$ , we have to assess  $H_{1,4}$ ,  $H_{2,4}$  and  $H_{3,4}$ :

$$\beta_4^5 = \{H_{1,4}(1-H_{2,4})H_{3,4}\} / \{[H_{1,4}(1-H_{2,4})H_{3,4}] + [(1-H_{1,4})H_{2,4}(1-H_{3,4})]\}$$

$$\beta_4^6 = 1 - \beta_4^5$$

$$\beta_4^{11} = \beta_4^6$$

$$\beta_4^{12} = \beta_4^5$$

Once the co-movement probabilities  $\beta_k^c$  are calculated, a general algorithm can be applied:

```

Obtain  $V_{k,0}$ ,  $\rho$ ,  $D$ ,  $T$ 
Calculate  $u_k$ ,  $d_k$ ,  $p_1$ 
Calculate  $\beta_k^c$ 
 $t \leftarrow 0$ 
While  $t < T$ 
   $k \leftarrow 2$ 
  Generate a random sample  $\tau_1 = U[0,1]$ 
  If  $\tau_1 \leq p_1$ 
     $c \leftarrow 1$ 
     $V_{1,t+1} = V_{1,t} u_1$ 
    While  $k \leq K$ 
      Generate a random sample  $\tau_k = U[0,1]$ 
      If  $\tau_k \leq \beta_k^{2c-1}$ 
         $V_{k,t+1} = V_{k,t} u_k$ 
         $c \leftarrow 2c - 1$ 
      Else
         $V_{k,t+1} = V_{k,t} d_k$ 
         $c \leftarrow 2c$ 
     $k \leftarrow k + 1$ 
  Else
     $c \leftarrow 2$ 
     $V_{1,t+1} = V_{1,t} d_1$ 
    While  $k \leq K$ 
      Generate a random sample  $\tau_k = U[0,1]$ 
      If  $\tau_k \leq \beta_k^{2c-1}$ 
         $V_{k,t+1} = V_{k,t} u_k$ 
         $c \leftarrow 2c - 1$ 
      Else
         $V_{k,t+1} = V_{k,t} d_k$ 
         $c \leftarrow 2c$ 
     $k \leftarrow k + 1$ 
   $t \leftarrow t + 1$ 

```

## CASE STUDY

In the sections to follow, we present the simulation results for the proposed joint lead-lag binomial lattice-Monte Carlo approximation. The real assets were selected arbitrarily from the New York Stock Exchange (NYSE) and the National Association of Securities Dealers (NASDAQ): Boeing Company (NYSE: BA), International Business Machines (NYSE: IBM), Hewlett-Packard (NYSE: HPQ) and Microsoft Corporation (NASDAQ: MSFT). The assets are simulated independently but with an imbedded correlation and lead-lag algorithm. The terminal values  $V_{ik}$  are aggregated according to the weights  $w_k$  to calculate the value of the portfolio  $V_T$ .

To benchmark the proposed model, we perform an independent Monte Carlo simulation with no correlation or lead-lag structure (i.e., each asset path is simulated independently) where the asset values at maturity  $V_{ik}$  are aggregated according to weights  $w_k$  to obtain the value of the portfolio  $V_T$ .

The sample data spans from 23 August 2002 to 2 April 2003. The initial time  $t=0$  is set at 2 January 2003. Before the actual approximation, we perform the lead-lag ITSUR analysis to obtain the order of simulation: MSFT -> IBM -> BA -> HPQ.

The summary statistics 90 days prior to simulation are as follows:

	$S_0$	$\nu$	Correlation Matrix			
<b>MSFT</b>	25.62	0.3931	1.00	0.91	-0.36	0.88
<b>IBM</b>	76.79	0.4853	0.91	1.00	-0.10	0.88
<b>BA</b>	32.18	0.4377	-0.36	-0.10	1.00	-0.10
<b>HPQ</b>	17.08	0.6115	0.88	0.88	-0.10	1.00

$r$  = annual interest rate = 0.04  
 $D$  = time horizon (expressed in years) = 0.25  
 $T$  = number of intervals per time period  $D=63$   
 $N=200$  repetitions

The following parameters are used to compare the estimations from simulations: 1) The average value of assets  $k = \overline{V_{Tk}} = \frac{1}{N} \sum_{n=1}^N V_{Tkn}$ ; 2) the average value of the portfolio  $= \overline{V_T}$ ; 3) the Value-at-Risk (VaR) of each stock  $(VaR_k) = \alpha$ -th percentile of the distribution of simulated values of  $V_{Tk}$ ; and 4)  $VaR$  of the portfolio =  $\alpha$ -th percentile of the distribution of simulated values of  $V_T$ . For added robustness, we employ three different measures of volatility: 1) constant volatility; 2) variable volatility (lag = 90 days); and 3) exponential weighted moving average (EWMA) (with decay factor  $\lambda = 0.94$ ).

The graphical outputs are as follows:

### Case Study

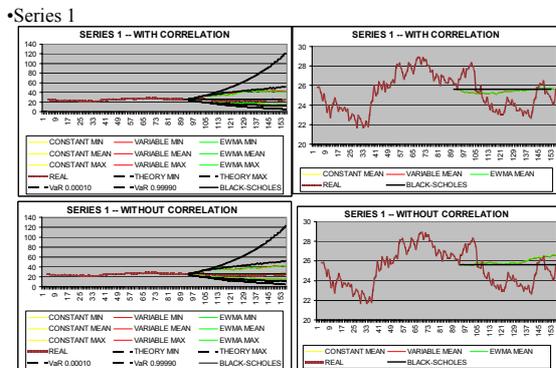


Figure 4. Microsoft (MSFT)

### Case Study

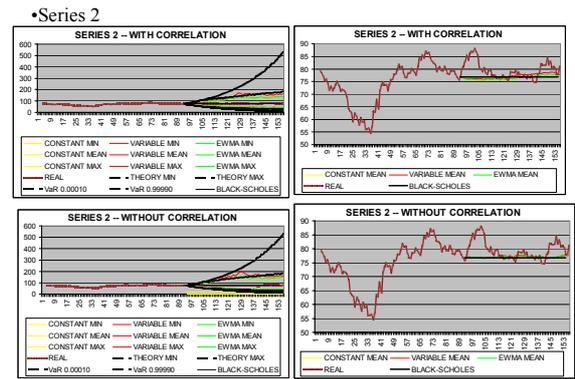


Figure 5. International Business Machines (IBM)

### Case Study

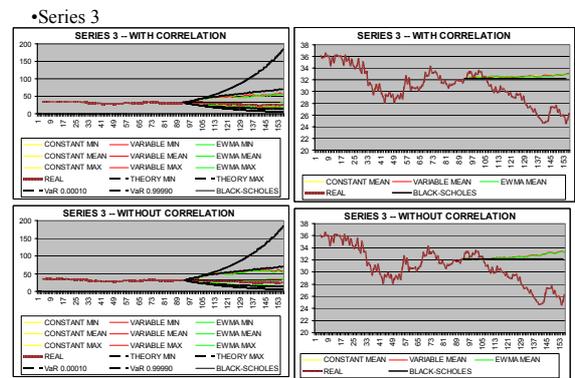


Figure 6. Boeing (BA)

### Case Study

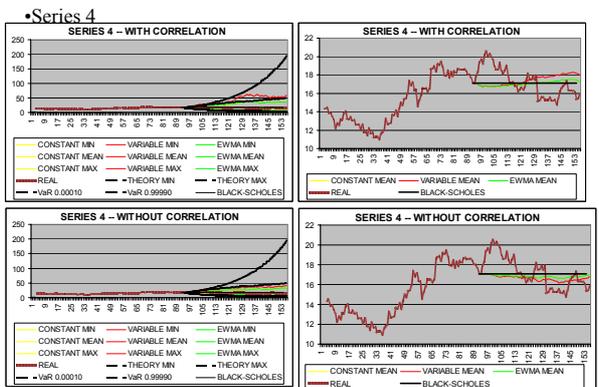


Figure 7. Hewlett-Packard (HPQ)

## Case Study

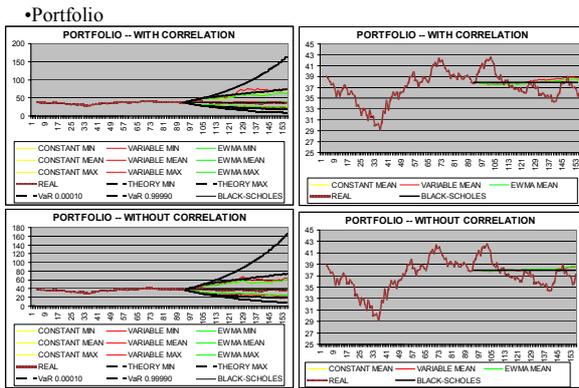


Figure 8. Portfolio

The results are summarized in Table 3 and Table 4:

Table 3. Simulation Results with Correlation and Lead-Lag (Proposed Model)

PROPOSED MODEL						
Volatility ( $v$ )	$V_t$	MSFT	IBM	BA	HPQ	PORTFOLIO
CONSTANT	D	38.44	-181.86	233.35	-9.68	20.05
VARIABLE	D	38.51	-149.37	233.02	12.79	33.74
EWMA	D	38.53	-186.42	234.83	-9.24	19.43
CONSTANT	AVERAGE DI	0.61	-2.89	3.70	-0.15	0.32
VARIABLE	AVERAGE DI	0.61	-2.37	3.70	0.20	0.54
EWMA	AVERAGE DI	<b>0.61</b>	<b>-2.96</b>	<b>3.73</b>	<b>-0.15</b>	<b>0.31</b>
CONSTANT	VARIANCE DI	2.39	11.61	8.13	3.12	4.78
VARIABLE	VARIANCE DI	2.40	12.74	8.33	4.03	5.33
EWMA	VARIANCE DI	2.39	11.58	8.03	3.10	4.77
CONSTANT	LOWER CI DI	0.09	-4.03	2.75	-0.74	-0.41
VARIABLE	LOWER CI DI	0.09	-3.57	2.73	-0.47	-0.24
EWMA	LOWER CI DI	0.09	-4.10	2.78	-0.74	-0.42
CONSTANT	UPPER CI DI	1.13	-1.75	4.66	0.44	1.05
VARIABLE	UPPER CI DI	1.13	-1.18	4.67	0.88	1.31
EWMA	UPPER CI DI	1.13	-1.82	4.68	0.44	1.04
CONSTANT	ABS(D)	91.02	202.07	245.51	93.93	112.13
VARIABLE	ABS(D)	91.27	189.65	247.65	107.98	123.95
EWMA	ABS(D)	91.01	205.21	246.97	93.75	111.56
CONSTANT	AVERAGE ABS(D)	1.44	3.21	3.90	1.49	1.78
VARIABLE	AVERAGE ABS(D)	1.45	3.01	3.93	1.71	1.97
EWMA	AVERAGE ABS(D)	1.44	3.26	3.92	1.49	1.77
CONSTANT	VARIANCE ABS(D)	0.65	9.63	6.64	0.88	1.66
VARIABLE	VARIANCE ABS(D)	0.65	9.25	6.53	1.09	1.68
EWMA	VARIANCE ABS(D)	0.65	9.70	6.53	0.87	1.68
CONSTANT	LOWER CI ABS(D)	1.18	2.17	3.03	1.18	1.35
VARIABLE	LOWER CI ABS(D)	1.18	1.99	3.08	1.37	1.53
EWMA	LOWER CI ABS(D)	1.18	2.21	3.06	1.18	1.34
CONSTANT	UPPER CI ABS(D)	1.71	4.25	4.76	1.81	2.21
VARIABLE	UPPER CI ABS(D)	1.72	4.03	4.79	2.06	2.40
EWMA	UPPER CI ABS(D)	1.71	4.30	4.78	1.80	2.20

Table 4. Simulation Results with No Correlation and No Lead-Lag (Independent Model)

INDEPENDENT MODEL						
Volatility ( $v$ )	$V_t$	MSFT	IBM	BA	HPQ	PORTFOLIO
CONSTANT	D	72.34	-156.84	231.78	-23.67	30.90
VARIABLE	D	70.87	-156.56	233.30	-36.35	27.82
EWMA	D	71.98	-155.51	230.58	-23.39	30.91
CONSTANT	AVERAGE DI	1.15	-2.49	3.68	-0.38	0.49
VARIABLE	AVERAGE DI	1.12	-2.49	3.70	-0.58	0.44
EWMA	AVERAGE DI	<b>1.14</b>	<b>-2.47</b>	<b>3.66</b>	<b>-0.37</b>	<b>0.49</b>
CONSTANT	VARIANCE DI	2.60	11.62	9.36	2.26	4.55
VARIABLE	VARIANCE DI	2.57	11.75	9.46	1.98	4.44
EWMA	VARIANCE DI	2.60	11.66	9.28	2.27	4.55
CONSTANT	LOWER CI DI	0.61	-3.63	2.65	-0.88	-0.22
VARIABLE	LOWER CI DI	0.59	-3.63	2.67	-1.05	-0.26
EWMA	LOWER CI DI	0.60	-3.61	2.64	-0.88	-0.22
CONSTANT	UPPER CI DI	1.69	-1.35	4.70	0.13	1.20
VARIABLE	UPPER CI DI	1.66	-1.34	4.73	-0.11	1.15
EWMA	UPPER CI DI	1.68	-1.32	4.68	0.13	1.20
CONSTANT	ABS(D)	112.37	186.93	250.69	80.17	113.88
VARIABLE	ABS(D)	110.91	187.85	252.16	76.46	111.42
EWMA	ABS(D)	112.02	186.38	249.53	80.37	113.94
CONSTANT	AVERAGE ABS(D)	1.78	2.97	3.98	1.27	1.81
VARIABLE	AVERAGE ABS(D)	1.76	2.98	4.00	1.21	1.77
EWMA	AVERAGE ABS(D)	1.78	2.96	3.96	1.28	1.81
CONSTANT	VARIANCE ABS(D)	0.71	8.97	7.02	0.76	1.47
VARIABLE	VARIANCE ABS(D)	0.71	8.99	7.12	0.82	1.46
EWMA	VARIANCE ABS(D)	0.71	8.96	6.95	0.76	1.47
CONSTANT	LOWER CI ABS(D)	1.50	1.96	3.09	0.98	1.40
VARIABLE	LOWER CI ABS(D)	1.48	1.98	3.11	0.91	1.36
EWMA	LOWER CI ABS(D)	1.50	1.96	3.08	0.98	1.40
CONSTANT	UPPER CI ABS(D)	2.07	3.97	4.87	1.56	2.21
VARIABLE	UPPER CI ABS(D)	2.04	3.99	4.90	1.52	2.17
EWMA	UPPER CI ABS(D)	2.06	3.96	4.84	1.57	2.21

From the absolute errors  $Di$  (real value at maturity less average simulated value), it is clear that the optimal procedure is the proposed joint lead-lag binomial lattice-Monte Carlo model approximated with a EWMA volatility.

## SUMMARY AND CONCLUSIONS

In this article, we employ a new numerical approximation technique for valuing multiple assets that are correlated, using jointly binomial lattices and Monte Carlo simulation. We show that when applied to a series of correlated assets (BA, IBM, MSFT, HPQ), the proposed algorithm offers a higher level of consistent, efficient and stable results.

## REFERENCES

- Black, F., and M. Scholes, 1973, "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, 81, 637-659
- Boyle, P., 1977, "Options: A Monte Carlo Approach," *Journal of Financial Economics*, 4, 383-405
- Boyle, P., 1988, "A Lattice Framework for Option Pricing with Two State Variables," *Journal of Financial and Quantitative Analysis*, 23, 1-12
- Boyle, P., Evnine, J., and S. Gibbs, 1989, "Numerical Evaluation of Multivariate Contingent Claims," *The Review of Financial Studies*, 2, 241-250

- Brennan, M., and E. Schwartz, "The Valuation of American Put Options," *Journal of Finance*, 32, 449-462
- Cox, J., S. Ross; and M. Rubinstein. 1979. "Option Pricing: A Simplified Approach." *Journal of Financial Economics* 7, 229-63.
- Haug, E. 1998. *The Complete Guide to Options Pricing Formulas*. McGraw-Hill, New York
- Hull, J., 2000, *Options, Futures and Other Derivatives*, Prentice Hall, New Jersey.
- Kamrad, B., and P. Ritchken, 1991, "Multinomial Approximating Models for Options with K State Variables," *Management Science*, 37, 1640-1652
- Jarrow, R., and A. Rudd, 1983, *Option Pricing*, Richard D. Irwin, Englewood Cliffs, New Jersey
- Law, A. and D. Kelton. 1991. *Simulation Modeling and Analysis*. McGraw-Hill, New York
- Leisen, D., and M. Reimer, 1996, "Binomial Models for Option Valuation-Examining and Improving Convergence," *Applied Mathematical Finance*, 3, 319-346
- Morrison, D. 1990. *Multivariate Statistical Methods*. McGraw-Hill, New York
- Otamendi, J., and M. Hon, 2004, "Approximating Lognormal Lattices with Multiple Correlated Assets," *IABT Working Paper WP0402*, Saint Louis University
- Rubinstein, M. 1994. "Return to Oz." *Risk Magazine* 7, 11
- Ruiz-Maya, L. and J. Martín Pliego. 1995. *Estadística II: Inferencia*. Editorial AC, Madrid
- Schmidt, J. W., and R. Davis. 1981. *Foundations of Analysis in Operations Research*. Academic Press, New York

Risk Consultant for corporate clients in the banking and telecommunication industries. Dr. Hon currently directs the IABT as well as the Department of Business Administration and Economics at Saint Louis University in Spain.

## AUTHOR BIOGRAPHIES



Javier Otamendi Fernández de la Puebla received the B.S. and M.S. degrees in Industrial Engineering at Oklahoma State University, where he developed his interests in Simulation and Total Quality Management. Back in his home country of Spain, he received a B.S. in Business Administration

and a Ph.D. in Industrial Engineering. He is currently a simulation and statistics consultant and professor.



Mark Hon obtained his bachelor's degree in finance, investment and banking from the University of Wisconsin-Madison in the U.S. He acquired his master's degree in business administration, master's program in economics

and finance and a doctorate in financial economics at the University of Bristol in the U.K. He has served as a

# SIMULATING INFORMATION POLICY MAKING

Zoltán Szabó, Erika Sudár and Dr. András Gábor  
Department of Information Systems  
Budapest University of Economic Sciences and Public Administration  
H-1093, Budapest, Fővám tér 8., Hungary  
E-mail: szabo@informatika.bke.hu, esudar@informatika.bke.hu, gabor@informatika.bke.hu

## KEYWORDS

Simulation, information society, policy making

## ABSTRACT

Recently it has become obvious that the creation of an information society in accordance with the eEurope objectives presumes governmental support. The programs and action plans, that are to support and accelerate the transition, have been developed in almost every country. However, at developing the strategy the policy-makers have to consider the complexity of the environment and the diverseness of potential effects. This complicates the forecast of the impact of the different measures. The scope of the research was to set up a model in order to support decision- and policy-makers. With the model the ones responsible will be able to estimate the impact of different measures, and the alternatives can be compared with each other.

## INTRODUCTION

Since the possible advantages of the development of Information Society and the disadvantages of a delay have been understood the decision-makers have tried to support the development. The decision-makers can use many tools to support the development of Information Society. Some of these have a straight effect on the economic processes. These are the direct economy stimulating steps, the effects of which are easily measurable and can be simply evaluated. Nowadays these direct influencing tools are being replaced with indirect solutions. Their effects usually last longer and affect every actor of the economy, but the evaluation of the results is highly complicated. Such indirect tools in building the information society are:

- Citizen friendly administration
- Improvement of education and health care
- Improvement of administration

The fact that the decision makers cannot foresee the necessary resources for the actions can raise difficulties when forming the specific plans. The complexity of the problem comes from the chain reaction possibility of the effects of our interventions. An accelerator effect is also present because of the connections of the economic

participants. This acceleration drives the effect of the action even further and causes further changes. This way of acting has an impact on every actor but the results can only be observed later.

This accelerator effect and its impacts on the economy must be estimated and it must be built into the model in order to achieve a more effective decision-making. Nevertheless the governmental actions bring results that can hardly be measured. Such result is for example the higher sufficiency of the customers caused by a possible improvement of public services. Despite the difficulties in their evaluation the impacts of these actions cannot be disregarded.

Another important problem is the estimation of the time factor during the development of the action plans and measurements. After some time supplementary support could be necessary in order to achieve our original goal. However, the final goal is to establish a self-sustaining system by making self-financing possible. An indispensable condition for this is the creation of an economical balance where the supply and demand of the services are more or less equal. As long as the society cannot achieve this aim, supplementary actions are necessary. We can see that forecasting whether the necessary balance can be achieved with the current preconditions and predicting the time when information society will reach the state of sustainability have a high importance.

On basis of the problems mentioned above, the need for a method to determine the effectiveness of governmental actions in different areas is apparent. Neither the EU strategies nor the national initiatives provide a measurable technique to define the desirable points of intervention and to evaluate the effectiveness of the actions, although they do contain indicators for estimations. With Hungary's accession to the EU a proper method to measure the effectiveness of spending is even more substantial. To be able to use the EU funds in an efficient way a weighting system for the different areas is essential. To solve these difficult issues a modeling approach is very suitable. Our research has been aiming at these questions.

Recently our team have come across with the very active Information Society related information policy making, due to the above the average government

commitment of speeding up internet and internet related services penetration. Under the framework of the National Research and Development Program a two years research project were launched in 2001 aiming the investigation of the sociological, economic, administrative and technical preconditions of the more intensive development and progress. Our team focused on the macroeconomic modeling and forecasting. The model have been developed is suitable for testing and experimenting several type of policies, the scale goes from the orthodox Keynesian approach to the ultra liberal type of decisions.

The System Dynamics is one of the most powerful and well-acknowledged simulation tool for modeling - amongst other - macroeconomic decisions. Applying System Dynamics in the modeling of information society is also quite unique in our research. With this methodology it is possible to give advise for the decision-makers about the most effective ways and the most effective points where spending can have the most beneficial effect on creating the information society.

### MODEL DEVELOPMENT

The action plans are made to help obtaining the main goals by creating actions to draw the real situation towards them. However, we have encountered a few difficulties already. During the creation of action plans we have to face some problems, as during the decomposition one has to pay attention not to lose the original focus. It can often be experienced that our action plans are defocused and they are not able to support us in answering the original question. Another problem factor is getting realistic feedback on the process to achieve our goals.

Naturally, the development of the actions must be continuously observed in order to bring in the necessary corrections. These results however, cannot be defined in all cases and the collection of data can cause difficulties.

For this purpose we used the aspects of the eEurope Benchmark 2005 (eEurope 2005, 2002).

The European Commission has drawn up a list of twenty basic public services to be benchmarked. These include twelve public services aimed at individual citizens and eight for businesses. By identifying the main factors of the Information Society, these benchmarks offer help in the structure for data collection and can be used as starting points for model creation. The benchmark indicators ensure the international comparison as well.

To minimize the foreseeable difficulties we decided to use System Dynamics (SD) tools for modeling, research and interpretation of the results (John D. Sterman, 2000). System Dynamics – a technique created almost half a century ago (Paulré 1980) – made it possible to represent the factors of Information Society in all their detailed complexity. SD also helped to form the model in a way that it helps in creating deductions regarding the operation of real-life processes. The model itself was created with the help of the SD modeling software VenSim. Developing a useful model is difficult enough; using modeling to help implement new policies is even harder. As in our case the basic data were incomplete and the actions of the sub-systems were hardly predictable the deployment of a well-formed framework was a must. Since effective modeling rests on a strong base of data and understanding of the issues, the rules and instructions of framework inspired us to disclose the problems thoroughly.

### MODEL DESCRIPTION

The model reflects to the findings of a recent done nationwide research in which the relevant influencing factors of the penetration of the information society (IS) paradigm were thoroughly investigated. The sub-models are working with statistical variables, mainly defined by the EU Benchmark 2005. The logic of the model can be demonstrated with the following figure.

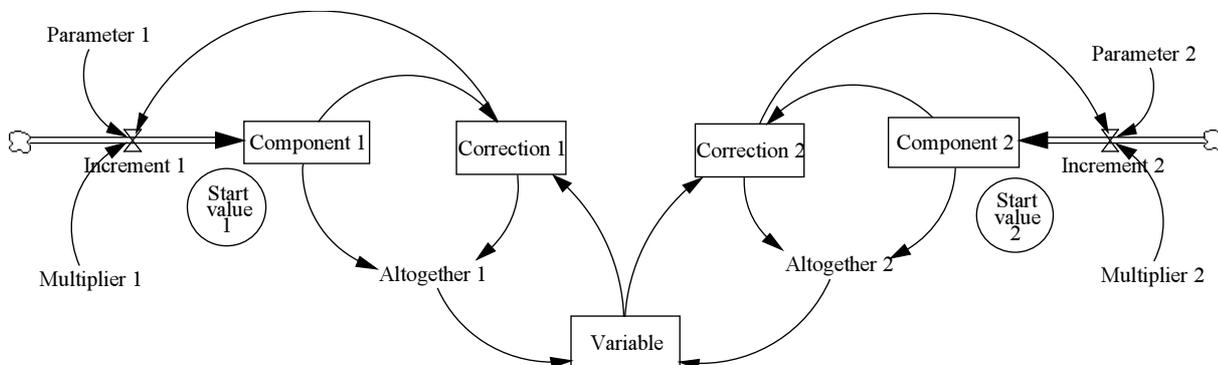


Figure 1: The Logic of the Model

In the model each variable is made up of some components. During the examination of the simpler parts we only have to consider the effects of two components at the same time, but in the more complicated parts of the model the development of the variable is influenced by more than ten components.

We have statistical data for the value of the variables to start with that later can be regarded as given values. During the simulation, these start values are increased by the multiplier that can be established with statistical data series or with professional estimate. The development of the components can be influenced by the acceleration rate through the alteration of other variables.

By taking these factors into consideration the values of the components valid for the given period and determine the value of the desired variable. The appropriate weights of components permit the differentiated consideration of the effects. The components have to match the value of the developed variables. The corrections reduce the potential failure of the model, as without correction the value of the components can secede from the weight of the variable. With the

correction we can ensure the consistence of the simulation model.

The complex model is built up from of seven sub-models:

- E-economy (B2B, B2C)
- Quality of Access (Expenditures on Security)
- Infrastructure Development
- Public Service / Administration (services for individuals, services for enterprises)
- Corporate Access and Use
- Individual Access and Use
- Tariff, which are connected to each other.

The scheme of the model can be seen on the following figure.

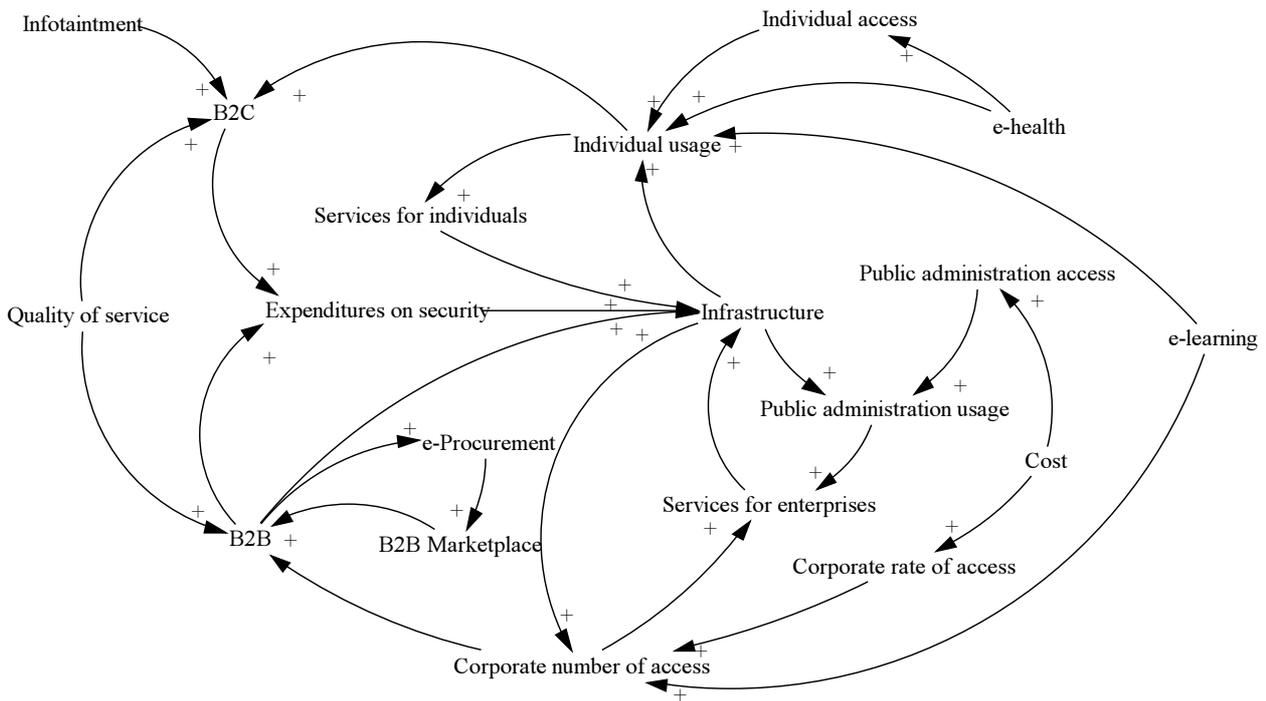


Figure 2: Overview of the Information Society Penetration Model

Each sub-model contains 2 to 12 components, connected through the growth ratio. The connections between the sub-models are inserted with specifiable parameters, therewith assuring that the strength of cross-effects can be regulated. The accelerator effect influences the growth of the individual areas in a direct

or indirect way. As the growths of the different partitions are not the same, the indicators' value will lead to several combinations. The model starts with realistic statistical values and the different strategies and policies are translated into the concrete values of the growth parameters.

The SD model is implemented in Vensim 3.0.

### MODEL SCENARIOS AND RESULTS

The model can be used to conduct several simulations depending on the scenarios and partly on the decision makers' attitude (more market-oriented or committed to the public expenditure).

These scenarios are constructed on the basis of different preconditions – supplemented with the examination of sensitivity – and they are giving the most important results of the modeling. The results of the model were examined in seven scenarios: (1) a realistic situation, (2) Public service dominated scenarios a) Public services for individuals and enterprises b) Infrastructure

development c) Public e-procurement enhancement, (3) Competitive sector dominated service scenarios a) B2B b) B2C c) E-health.

Applying the methods detailed above a multitude of simulations can be made and various conclusions can be drawn. In this paper we can only illustrate the most important implications regarding the two basic concepts: public service dominated scenarios and competitive sector dominated service scenarios Further analysis still can be made. The interpretation of data is always partly affected by convictions; however in this case the results of simulations create a well-defined and determinate basis.

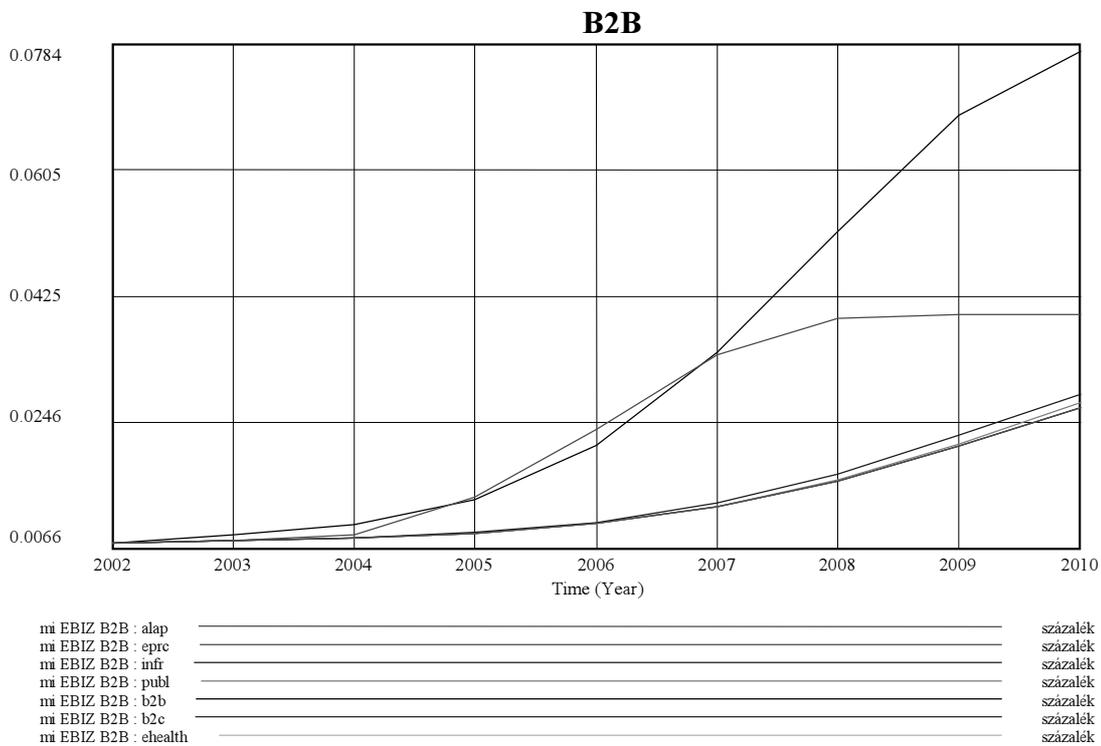


Figure 3: Business-to-Business EC

The main question of our research was whether the development of internet-penetration is the task of the government or public organizations. In order to answer the research question, the impacts of 6 examined approaches have been compared in selected key areas. The results are the following:

- In order to develop B2C commerce market solutions and the improvement of market organizations are more adequate comparing to other approaches, which require governmental sponsorship.
- In the area of B2B commerce both market and governmental intervention can be successful.

Beside the market based development of B2B, the introduction of e-procurement can support this area, which is strong governmental role.

- The development of infrastructure can be based both on governmental and market resource. In short term the direct development (infrastructure investments, subventions) can be the most successful, but in long term the other approaches have approximately the same effect. All approaches based on government service improve the state of infrastructure, while among the market-based approaches the development of B2B has the most powerful effect.

- The individual and corporate usage of Internet can be improved by both approaches. The familiarization of users to Internet by means of other services has the most powerful effect in this area (services of e-health or public services

for individuals or enterprises). The development of the infrastructure can improve the usage of Internet, as well.

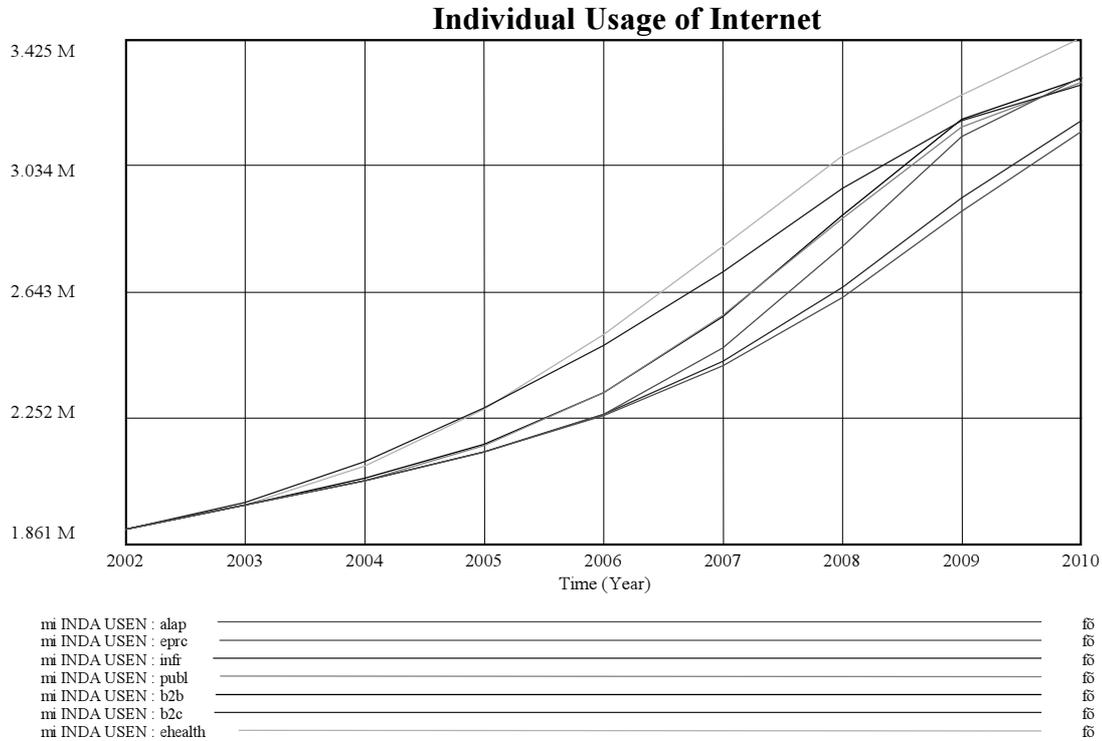


Figure 4: Individual usage of Internet

All this concludes to the implication that Information Society building requires the intervention of the government in the short terms, however the penetration is expected to grow by itself – or rather by the actions of the participants – from a point onwards. The Information Society can become a self-sustaining system.

#### REFERENCES

eEurope 2005: An information society for all, 2002. Commission of the European Communities, Brussels

E. Paulré (ed.), 1980. System Dynamics and the Analysis of Change, North-Holland Publishing Company, Amsterdam

John D. Sterman, 2000. Business Dynamics – Systems Thinking and Modeling for a Complex World, McGraw-Hill Higher Education

#### AUTHOR BIOGRAPHIES

**ANDRÁS GÁBOR** is an economist, graduated from the then Karl Marx University of Economics. He has a second degree in Computer Science (1979) and earned his Ph.D. in 1983, CISA (Certified Information Systems Auditor) since 1999. He is Associate Professor, the Head of the Department of Information Systems at the Budapest University of Economic Sciences and Public

Administration. His research field includes systems design, information management, intelligent systems, simulation and knowledge management. He was a visiting scholar in Harvard Business School 1995, the University of Amsterdam, (1990-1995), the Imperial College of Science and Technology, 1986, the DePaul University, Department of Computer Science and Information Systems, Chicago, USA, 1985, and the Imperial Chemical Industries, Pharmaceutical Division, UK 1975. He is the holder of the Award for Excellence of the President of the Hungarian Academy of Sciences.

**ZOLTÁN SZABÓ** is an economist, graduated from the Budapest University of Economic Sciences, specialized in Information Management in 1994. He has earned his Ph.D. in 2001 and a managerial level ITIL certificate from the British ISEB since 1997. He is a senior lecturer and responsible for several graduate and postgraduate courses of the Budapest University of Economic Sciences and Public Administration. His research field includes strategic information system planning, management of IT services, IT governance, e-commerce. He was expert consultant for Hungarian and international organizations (Ministry of Education, Hungarian World Bank Higher Education Project, T-Mobile, Budapest Public Transport Company, etc.).

**ERIKA SUDÁR** is an economist, graduated from the Budapest University of Economic Sciences, specialized in Information Management in 2002. She is a Ph.D

student at the Department of Information Systems, BUESPA. Her research field includes e-government, simulation and information society.



**SIMULATION IN  
ELECTRONICS,  
COMPUTERS AND  
TELECOM**



# FULL-SCALE COMPLEX ANALYZERS FOR PROCESSES OF FUNCTIONING POWER PLANTS' ELECTRICAL EQUIPMENT

Alexey I. Poydo, electric stations' chair assistant professor, Moscow Power Engineering Institute, Russia  
Vladimir A. Rubashkin, "Power Plants Simulators", Russia, Moscow, Semenovskiy per., 15, 224,  
[pps@edunet.ru](mailto:pps@edunet.ru), [www.fpps.ru](http://www.fpps.ru)

## KEYWORDS

Simulation technologies, electric circuit, power plant.

## ABSTRACT

For Russia for the first time a fully digitized complete engineering model of a power plant electric circuit that is able to work in the real time in a frame of a computer simulator has been developed. The model and the simulator can be used by the power plant operational personnel to predict what operation mode of the power plant electric circuit shall happen at the real equipment as a result of a required change over.

## INTRODUCTION

According to the existed circumstances the main facilities of Russian Federation power plants were put into operation in a short time of the second half of XX century. The same situation was arisen as well in the West European countries, the USA, Canada and some other countries.

From this it follows that the equipment with finished service life should be substituted by the new one in the comparatively short time and in large quantities. It is necessary to attract for it the huge funds. At the same time the investigations showed that the service life could be continued, if the equipment is maintained under soft conditions.

The new software system, which is able to help in providing of the soft conditions of electrical equipment operation, is described in this article.

## STATEMENT OF A PROBLEM AND GENERAL TOPICS

When a power utility dispatcher makes an order for a power plant to output certain power along the requested transmission lines, neither this dispatcher nor the power plant specialists beforehand know well, which regime of electrical equipment operation would be installed as a result of operating this command. The assessment of obtained result can be made only after operating the order. However the necessary volume or the amount of measuring devices is traditionally not installed at some electric circuits, and therefore even after operating the dispatcher order the personnel has not the full and accurate estimation about the formed regime of electric equipment operation. But if even after operating the dispatcher command it became evident that one of the elements of electrical

equipment works with overload, it is still not clear, how it should be necessary to change the regime to avoid this overload and at the same time to fulfill the dispatcher order.

Until recently there was no technical means to forecast on the fly the mode of operation of a power plant electric circuit as a result of a planned switching over. The available systems for calculating the parameters of electrical regimes are oriented for the long preparation of initial data in a certain format, they have not a convenient user friendly interface, and in addition the time of calculating one regime by means of such software is still quite large. All these facts limit their utilization at the real power plants. Another systems (the simulators of switches) that have a user friendly interface are not able for the most part to calculate the operating conditions of electrical equipment.

The Russian company "Power plants simulators" managed together with the electric stations' chair of Moscow Energy Institute (MEI) to solve first the pointed problem in the full scale.

The technical problems, which had to be solved, concern the following areas:

- Algorithms of functioning the separate elements of electric circuit;
- Optimization of the rate of solution;
- Organization of software structure;
- Complex debugging of complicated systems.

As a result the "Complex analyzer of the processes of power plants' electrical equipment functioning" (CAEE) was developed, while the high accurate model of electrical equipment functioning lays in the basis of it. Being created for a particular power plant with taking into account all its distinctive features, this model allows to calculate the main parameters of the power plant electric circuit under any switching over.

It is evident that in order to successfully forecast the operating conditions of electrical equipments of a power plant, CAEE must manage to simulate in full-fledged manner:

- generators of any kind including their excitation systems,
- transformers,
- switches
- and so on.

It is extremely important that the developed software system allows to execute the calculations of processes occurring in the electrical equipment in the real time.

CAEE doesn't demand the exhausting preparation of initial data for calculations. It has a user friendly interface that allows easy to learn how to use it.

If we have the high accurate model, it is possible not only to forecast the operating conditions of power plant's electric part, but to analyze as well the occurred emergency situations. A function of creating the oscillograms of "fast" transient processes was included in CAEE for facilitating the analysis of emergency situations by means of the CAEE. It is supposed to use as a "oscillograph", for example, Microsoft Excel, while CAEE is able to create the data that are necessary for it.

Moreover, if the additional means for training the personnel are included in CAEE, it is possible to obtain a high quality simulator for training of electricians, which is totally adopted with the special features of given power plant. The CAEE development engineers envisaged a possibility of its equipment with the subsystem supporting the personnel training. In such configuration the CAEE can be equipped with the following features:

- additional protection and blocking, which are absent at the real power plant, but which prejudice the trained person about approaching to the emergency situation in a few steps before it;
- subsystem of calculating the residual reliability of electric circuit during any switches over and an integrated estimation of the whole transient process from one regime to another in accordance with the conditions of reliability;
- the traditional servicing functions of training simulators such as:
  - loading an initial state;
  - saving the current state;
  - real/accelerated time modes;
  - run/freeze modes;
  - a feature that a previously performed exercise can be later repeated in automatic mode,
  - and so on;
- the CAEE can be run on a few computers interconnected to a LAN

The Russian company "Power plant simulators" has already more than 10 years occupied the leading positions in Russia in the area of development of simulators-analyzers of thermo mechanical processes of thermal power units. In this period the company delivered to the customers in Russia and abroad more than 20 simulators of different type. The joining of its advanced simulation technologies with the advanced scientific research of the leading specialists of MEI electric stations' chair resulted in the development of the software product, which has no analogues not only in Russia, but in our opinion also abroad. The software product can allow the electric power plants and utilities to reduce substantially the expenditures for running repair and updating of electrical equipment.

A development of the first CAEE for Moscow power plant number 26 is completed in April 2004. The main circuit of electrical connections at the power plant reproduced in CAEE consists of the following elements:

- 32 nodal points (points of electrical connection of the groups of elements),

- more than 300 switching elements (switches and disconnectors),
- ten 220-500kV-transmission lines,
- 14 transformers with power from 32 to 400 MVA
- 7 generators with power from 120 to 320 MWt

## REGIMES THE CAEE REPRODUCES

The regimes of possible overloads of equipment on current and voltage are most important in the context of preserving the service life of electrical equipment.

Based on the accurate mathematical modeling, CAEE allows to reproduce practically all steady-state and transient regimes of generators, transformers, switching equipment etc. including both the normal and abnormal regimes.

In this chapter we will enumerate some most frequently met and most important regimes of electrical equipment operation, which are reproduced in CAEE.

### Transformers

The overload of coupling transformers (generator switchgear - GSG) of generator voltage and OSG (outdoor switchgear) of HV (high voltage), OSG of MV (middle voltage) is possible with the generators operation along the schedule of heat production and the reduction of 6-10 kV-local consumer load. In this case the disconnection of one coupling transformer will result in the overload of another ones. The duration of this overload is determined either by the duration of transformer disconnection or the time, which is necessary for the readjustment of technological conditions of heating system and the unloading of generators required. The level of overload can be significant and exceeds the systematic overloads admissible by Russian State Standards (GOST) as well as the admissible short-term overloads permitted by the "Rules of technical maintenance" (RTM). For the transformers, for which the service life is about over, such overload can result either in the final malfunction or in the failure of devices for VCL (voltage control under the load) or for SWE (switch without the excitation). The disconnection of the transformer can be caused both by the failure of the transformer itself and the failures in the work of switching equipment in its circuit. CAEE allows to the operating personnel to analyze all possible ways going out the formed critical situation with the full control of operating conditions.

The overload of coupling autotransformers HVSG (high-voltage switchgear) – MVSG (middle-voltage switchgear) in the power systems with the relatively short transmission lines are caused, as a rule, by the power system regime. Taking into account, that the relation of inductive reactance's in the transformers of power units and transmission lines constitutes 5-10 in such systems, the capacity redistribution at the generators of power units effects poorly on the transit flow through the coupling autotransformer. The solution of a problem is possible only by the combined actions of operating personnel of the power plant and the power utility. For the electrically remote power plants the mentioned regime can be corrected by means of load redistribution between units switched for HVSG and MVSG. In all these cases CAEE

gives the exhaustive information on the possible consequences of operating personnel actions and the operating conditions.

The overload of unit transformers at the units of Russian power plant is unlikely, because their rated power exceeds the power of unit generators. The overload of reserve transformer for auxiliaries in the regimes, when it works simultaneously for a few units, is possible, if there is a failure of some working transformer of auxiliaries. In this case the overload duration is determined by the possible overswitching in the circuit of reserve power supply of power plant auxiliaries or the duration of unit outage. CAEE covers the pointed regimes and allows to forecast the optimal actions of operating personnel.

### Generators

The steady-state regimes of generators' operation at real power plants are also conducted in some cases by the overloads exceeding the permissible ones. We enumerate the most often met regimes, which are successfully implemented by CAEE.

- Separation of a part of power plant to a stand-alone system with the formed shortage of active power of the separated part. The pointed regime is accompanied by the reduction of the separated part frequency, the increase of generators' excitation with the possible overload of rotor and stator on current. The many ways are known for going out the critical regime including: the disconnection of a number of consumers; the attempt of resynchronization with frequency difference exceeding the permissible one for the conditions of accurate synchronization; the transfer of a local consumer feeding that doesn't require the uninterrupted power supply to the reserve system of buses with temporary disconnection of its power supply, etc. CAEE is modeling both the development of situation and the possible ways of going out and returning to a allowed regime.
- The voltage reduction in network is a special phenomenon in the deficit power systems, when a part of the system is separated in a stand-alone system. In this case depending on the adjustment of voltage deviation channel in ACE (automatic controller of excitation), a series of the alternative regimes of generator operation is possible: from the overload on rotor current to going out the synchronism under conditions of static stability. CAEE allows to predict in advance the consequences, to install the desired distribution of reactive power between generators and to select the optimal ACE adjustment by means of pointed channel.
- The loss of generating facilities is dangerous in the deficit power utilities. Therefore, when in such power utilities a generator comes to the asynchronous regime, there are the attempts to keep it in operation in spite of some difficulties to keep its thermal regime. In excess power utilities, the generator that felt to asynchronous regime is, as a rule, switched off. In the same time the rules of technical maintenance don't forbid for some generators to work in the asynchronous regime during a limited time. When the generator works in the asynchronous

regime without excitation it consumes a large amount of reactive power, it compensates the generators, which are adjacent with it and are working with excitation. In this case the generators, which have the largest excitation control factor for the voltage deviation, are overloaded for rotor current and stator current by 20-30%. The pointed regime is successfully implemented in CAEE. The power plant personnel obtains a possibility to set up beforehand the limitation of overexcitation of adjacent generators, which is switched on in the asynchronous regime of one of them.

- At the present time the existence of prolonged 500 kV-networks results in the voltage growth both in 500 kV-networks as well as in the 220 and 110 kV-networks connected with them in the case of excess generation of reactive power in the pointed networks. The considerable differences in the consumption of reactive power in daytime and at night promote to it. When the consumption of reactive power in system is insufficient, some generators must be transferred in the regime of reactive power consumption. In this case they are working with the considerable underexcitation. It can result and results in the loss of dynamic stability of generator, if the lengthy short current coincides with this regime. In addition the overexcitation of transformers and autotransformers as well as the increase of voltage for auxiliaries are observed in such regimes. CAEE allows to forecast the pointed emergency regimes and to find the optimal regimes of generators' operation for the reactive power output within the frameworks of dispatching active load curve.
- Some power systems yet currently use the method of switching on the generators in network by means of self-synchronization in the emergency regimes. It accelerates the process of putting the facility into operation if there is a power deficit but creates the dangerous situations both for the generator and the power system. CAEE implements the pointed processes in full as well as the regimes appearing with failure of autosynchronizers and not synchronous connection of generators to network.
- The regime of transferring from the working excitation to the reserve one for electric machines is of great importance for the generators that have not 100%-reserve in the thyristor excitation systems. There were the emergencies with the difficult consequences in some power plants. CAEE implements all these regimes appearing during the transfer to the reserve excitation.

### Switching Devices

The current overload of switching devices and current transformers, which is possible with the different switches in the ring circuits, is an abnormal phenomenon because it is not predictable. In the ring circuits with 3/2 and 4/3 of switches for connection, the current distribution takes place on the basis of inductive impedances of buses' sections and resistances of contacts. The repair and after-emergency regimes are possible in such circuits, and during these regimes the separate switches will be over-

loaded on current. In the case of current overload the superheat of contact system occurs, and that is undesirable. A software block, which allows to calculate continuously the pointed regime and correspondingly to forecast it, was included in CAEE.

The similar phenomenon takes place in the switching devices of generators, which are equipped by sectional current-limiting reactor. The concrete current-limiting reactors for the rated current 4000 A are not produced in Russia since 1990. The service life of pointed reactors is 25 years, and this is determined by the period of service life of concrete. At the present time the majority of pointed devices exhausted their service life. In the same time the overload of pointed reactors is limited similar to the dry-type transformers. The serious emergencies connected with distortion of pointed devices took place at some power plants. The initial analysis of operating regimes of switchgears' sections allows in the case of existence of reserve system of buses to provide the optimal operating conditions for the sectional current-limiting reactor, which eliminate the overloads.

### MATHEMATICAL BASES OF CAEE

The special features of block diagrams of the simulated electric circuits, which are typical for power plants and in many cases also for modern power systems, were used during development of CAEE.

For the most part the structure of electric circuit at power plants is represented by the nodes of this electric circuit with series connection. These circuits are begun from the nodes embodying the external power system, and they exceed greatly in power the modeled generators connected into other nodes. A number of cross linkages between such circuits is negligible. As the investigations of networks existing in Russia showed, the similar structures are typical for the circuits adjacent to the power plants. The networks with voltage 110 kV and in some cases also with 220 kV are radial even in the compact power systems. Such structure is used, because the short-circuit current are too large in the complicated loop networks, where there are many jumpers.

Under such conditions the modeling of network like quasi complicated loop circuit, especially for the inside circuit of power plant, seems to be not expedient.

The use of methods for calculating the ladder circuits allowed to avoid in modeling power plant circuit the difficulties, which appear with using the methods that are characteristic to the complicated loop circuits.

In our mind the complex technology of processing the sparse matrixes is unnecessary in this case.

The presentation of general circuit in a form of ladder elements with actively inductive couplings allowed also to decline the use of global coordinate system. In this case the ladder circuit is modeled by the power flows from a node to a node by means of solving the system of nonlinear algebraic constraint equations.

The methods for solving the pointed system of equations was developed. They give the steady-state solution, which is calculated at the modern computers in real time scale.

The advantage of used method is a presentation of a branch of a network within the frameworks of general

iterative process as a separate element, in which the power flows between the nodes of electric circuit are determined on the basis of vectors of voltages in nodes and without connection with the general system of coordinates.

The use of global coordinate system seems to be necessary only for modeling the cross-linkages. They include the coupling autotransformers of switchgears with different voltage and the sectional reactors.

The determination of magnitudes and phases of voltage vectors in the nodes by means of power balance allows to use locally the diacoptical method for finding the currents in any element of circuit.

The generators are modeled in the orthogonal system of coordinates  $d, q$  by means of the full Gorev-Park's equations. In this case the modeling of network by a balance of powers in nodes allows to avoid the so called coordinate converters.

If the ac generator with rectifier is used in the generator excitation system, this generator is modeled as well by means of the full Gorev-Park's equations, while the rectifier is modeled by means of its characteristics.

The turbines is modeled in CAEE as a complex of automatic frequency control of generator rotation without taking into account the mechanical processes in details.

Thus, the application of above mentioned methods in CAEE allowed to obtain:

- Full modeling adequacy in all operating regimes of power plant's electric circuit for the periodic component of forward sequence currents,
- Full modeling adequacy in all operating regimes of power plant's electric circuit for the flows of active and reactive power,
- Full modeling adequacy in all operating regimes of power plant's electric circuit for the electromagnetic moments in generators,
- Full modeling adequacy in all operating regimes of power plant's electric circuit for the voltages in nodes of an electric circuit,
- Determination with insignificant error, which is admissible by Guiding documents in calculating the short-circuit currents, of aperiodic components in the currents of any element,
- Periodic currents and voltages for the regimes of not symmetric short-circuits by using the equivalent circuit for local section and the equivalent reaction of another part of circuit.

The mathematic methods used in CAEE allowed to avoid the difficulties, which are inherent to the solution of a problem in the global formulation. On the other hand, they allowed to realize the all-regime model of electric equipment operation in the modeling system, in which practically all variables of modeling have the concrete physical meaning. It allowed to lighten greatly the debugging of model, because the diacoptical method allows to fulfill the adjustment of each element model separately.

The model debugging and testing are greatly complicated with the global problem statement, when that is the problem of modeling electric circuit of power plant. It is connected with the fact that the separate element can not be allocated from the general model in the known computa-

tional techniques with the global problem statement, while its parameters are the part of generalized parameters of global model, which can't have a concrete physical meaning.

The means of looking for errors are especially important for the system, which has a few ten thousand of variables. It is extremely important for the solution of a problem of forecasting the power plant, because there is not standard solution for the majority of regimes.

The structure of mathematical provision and software of CAEE allows to develop and to adjust the models of electrical equipment operation for the circuits of practically any complexity.

## **CONCLUSIONS**

At the present time the prolongation of electrical equipment service life represents the serious problem for electric power industry of all the country, and this problem needs the comprehensive solution. The software complex CAEE described in this article can help to solve the problem due to its unique capabilities.

## **AUTHOR BIOGRAPHIES**

**VLADIMIR A. RUBASHKIN** was born in 1963 in Moscow. He graduated from the Moscow institute of railway transport with "computer science" specialization. After the graduation for 6 years he had been working in the Moscow academic institute of control problems. From 1992 he has been working in the "Power plant simulation" company on the technical director position.

**ALEXEY I. POYDO** was born in 1938 in Moscow. He graduated from the Moscow Power Engineering Institute in 1960 with "electric stations" specialization. For more 40 years he has been teaching in the Moscow Power Engineering Institute to the students the different branches of the electric station science. He is an author of more than 120 papers in many scientific magazines. He is a co-author of a few textbooks. The main field of his scientific interests is the mathematical modeling of the power plant electrical equipment.

# USING STATIC PROGRAM ANALYSIS TO COMPILE FAST EXECUTION-DRIVEN SIMULATORS

Vesa Hirvisalo  
Helsinki University of Technology  
Laboratory of Information Processing Science  
P.O. Box 5400, FIN-02015 HUT, Finland  
E-mail: Vesa.Hirvisalo@hut.fi

## KEYWORDS

Simulation methodology, cache simulation, simulation optimization, execution-driven simulation.

## ABSTRACT

This paper presents a generic approach for compiling fast execution-driven simulators, and applies the approach to simulating the effects of program execution in computer hardware. Our approach is based on using static program analysis to guide partial evaluation and slicing of simulators. Because results of some simulation operations are known before execution, a cache simulator program can be partially evaluated during its compilation. Program slicing can be used to remove the computations that have no effect on the simulation result.

Our experimental work with cache analysis shows that our approach significantly speeds up simulations. Fast cache simulation is needed in development of both computer software and hardware. To properly understand the cache behavior caused by a computer program, simulations must be done with sufficiently many inputs. Traditional simulation of memory operations caused by a computer program can be orders of magnitude slower than execution of the program. Our approach reduces the time needed in cache performance evaluations without losing accuracy of the results.

## INTRODUCTION

This paper discusses how static program analysis can be used to compile fast execution-driven simulators. In an execution-driven simulation, the execution of a computer program is interleaved with the simulation that describes behavior implied by the computer program.

Execution-driven simulation is a straightforward approach to analysis of computer programs. It is performed by simply executing a subject program with instrumentations that collect the analysis data and simulating the behavior that we are interested in.

Static program analysis is an other approach to analysis of computer programs. In static analysis, we try to understand

the run-time behavior of a program without executing it with a specific input. Static analysis is usually motivated by its ability to simultaneously give results for a set of inputs, often for all inputs of a program.

For some analysis tasks, simulation methods are slow, and any speed-up of simulation is useful. Our approach is to use partial evaluation and program slicing guided by static analysis to compile fast simulators.

In our presentation, we concentrate on a specific application: simulating the cache memory behavior that is caused by an execution of a computer program. Such cache performance evaluation is a demanding program analysis task, and thus a good test of the applicability of our approach.

Understanding memory performance of computer programs is hard. The steps executed and the related memory references can be seen from the code of a program. However, cache misses and the related executions stalls cannot be seen. Typical hardware does not support analysis of memory operations (Horowitz et al. 1996). Therefore, the memory operations caused by program execution are often simulated.

The traditional cache analysis method is trace-driven simulation (Uhlig and Mudge 1997). A trace-driven simulation has two main phases. In the first phase, an access trace is collected. Because hardware support for tracing is rare (Horowitz et al. 1996), the collection is typically done by augmenting the subject program with trace emitting code. In the second phase, a memory simulation is executed using the collected trace as the input.

Execution-driven simulation can yield more accurate results than trace-driven simulation. In an execution-driven simulation, the execution of a software is interleaved with the simulation of the underlying hardware. Thus, execution-driven simulation allows feed-back from the hardware simulator to the software. Such a simulation technique is useful in program performance analysis and simulating parallel and distributed systems.

Simulation is a flexible and accurate method, but simulating memory operations of a program can be orders of mag-

nitude slower than execution of the program (Uhlig and Mudge 1997). To properly understand the memory behavior of the program, simulation must be done with sufficiently many inputs. This leads to simulation times that are often infeasible in system development.

Because of the central role of simulation in cache performance analysis, a variety of methods have been developed for speeding up simulations. Many of the methods are designed for trace-driven simulation. The traditional speed up methods operate between the trace generation and the trace consumption (i.e., simulation). They modify the trace in a way that makes the processing faster. Such methods include, for example, packing methods (Ha and Johnson 1994, Samples 1989), stack deletion methods (Smith 1977), cache filtering methods (Puzak 1985), and spatial blocking (Agarwal and Huffman 1990).

Recent methods for speeding up cache simulations often apply parallel and distributed techniques. Such methods divide the simulation task into a number of smaller subtasks that can be simulated in parallel by a number of processors. Such methods include, for example, set and time partitioning (Heidelberg and Stone 1990), stack distance methods (Nicol and Greenberg 1992), and methods based on applying generic parallel discrete event simulation techniques (Fujimoto 1999).

Our method for speeding up simulations is based on static program analysis. Our static analysis finds out memory references that always cause cache hits or always cause cache misses. Based on the static analysis, parts of the simulation task can be solved during compilation of an execution-driven simulation program. This is done by omitting simulation at references, whose effects are statically known, and using simplified simulation at references, whose effects are partially known.

The structure of this paper is the following. The second section discusses the problem of cache performance analysis of programs. The third section presents a model of execution driven simulation. The fourth section discusses static cache analysis, and the fifth section describes how it can be used to compile fast cache simulators. The sixth section presents our experimental results, which show that significant speed up can be achieved by using the method. In the last section, we draw some conclusions.

## CACHE MEMORIES

Cache memories (see e.g., Przybylski 1990) improve memory access times. They reduce the number of cycles a processor is waiting for data; in the best case, the processor can continue its operation without any stall. Present day first level caches can give access to data two orders of magnitude faster than main memory. Thus, memory can become a major performance bottleneck, and careful design can significantly improve performance of systems using cache mem-

ory.

Cache memories consist of blocks called *cache lines*, which are used to store frequently used blocks of memory. We denote the length of a cache line by  $L$ . Cache lines are organized into *cache sets* of equal size. The size of a cache set is also called *associativity* of the cache (typically, it is 1–16 lines), which we denote by  $A$ . We denote the number of cache sets by  $N$ . Thus, the size of a cache is  $N \cdot A \cdot L$ .

A *memory line* is an aligned block in memory that is of the size of a cache line. Each memory line (and thus also each memory address) is uniquely mapped to a set. Two memory accesses (or references<sup>1</sup>) *conflict* if their addresses are mapped to the same cache set.

Before accessing the main memory, the computer hardware checks whether the addressed datum is stored in a cache line (in the cache set of the memory address). If the datum is in the cache, then a *hit* occurs. Otherwise, a *miss* occurs and a block in the cache is replaced for the new one. The misses can be categorized into three:

- *Compulsory misses*, the first access to a line causes always a miss.
- *Capacity misses* occur when the cache is too small to hold all of the lines needed during an execution of a program.
- *Conflict misses* occur when the cache has sufficient space, but the organization of the cache does not allow the data to be kept in the cache.

Because of the complexity of the memory hardware, interactions of memory references are complex. To improve performance, we must understand the cache behavior that the references cause. Typical hardware does not support analysis of memory operations (Horowitz et al. 1996). Therefore, the memory operations of such programs are usually simulated.

Simulation of memory operations of a program can be orders of magnitude slower than execution of the program (Uhlig and Mudge 1997). Traditional simulation of the effect caused by a single access includes several operations:

- pass the accessed address to the simulator
- break up the address into tag, block number, and block offset
- compute the set number
- search the block in the corresponding set
- update set status and performance metrics

Updating the set status typically consists of several operations, which depend on the cache replacement algorithm that is used in the simulated hardware. In the following, we will assume LRU (Least Recently Used) replacement.

---

<sup>1</sup>A static read or write in a program is a *reference*. An execution of the read or write at runtime is an *access*.

## EXECUTION-DRIVEN SIMULATION

In an execution-driven simulation, the execution of a program is interleaved with the simulation of the underlying hardware. In the rest of this paper, we will use an abstract model of execution-driven simulation to explain our method of building fast simulators.

Let  $P$  be a program and  $O$  its output corresponding to input  $I$ , i.e., the program computes:

$$\llbracket P \rrbracket \langle I \rangle = O \quad (1)$$

To build a simulator that is driven by  $P$ , we must instrument  $P$  with code that simulates the hardware. The instrumented program  $P^A$  computes:

$$\llbracket P^A \rrbracket \langle I, I^+ \rangle = \langle O, O^+ \rangle \quad (2)$$

where  $I^+$  is the input for the instrumentation and  $O^+$  is the analysis output that is measured by the instrumentation.

In an execution-driven cache simulator, each memory reference in the original program code must be instrumented with cache simulation code. Instead of explicitly giving the instrumentation code that is needed, we define a short abstract instrumentation for cache analysis.

The abstract instrumentation uses a mapping `incache` and a set of queues  $Q_t$  (one queue per each cache set). For each cached line  $t$ , `incache( $t$ )` is 1. Otherwise `incache( $t$ )` is 0. For each cache line  $t$ ,  $Q_t$  is the replacement queue of the set of the line. The queues represent the total order needed by LRU management – they are last-in-first-out-queues. For each cache set, the head of its queue is the least recently used cache line.

Analysis instrumentation for a memory reference addressing line  $l$  is:

```
t = line(l, N, A, L);
if not incache(t) then
  set_not_incache(remove_head(Qt))
  set_incache(t)
else
  remove(Qt, t)
end
insert_tail(Qt, t)
```

In the instrumentation, `line` returns the memory line referred to, `remove_head` removes the head of a queue and returns it, `insert_tail` inserts a cache line to the tail of a queue, and `remove` removes a cache line from the inside of a queue.

## STATIC ANALYSIS

In static analysis, we estimate the execution state of a program without actually executing the program. Our static analysis follows the concept of abstract interpretation presented by (Cousot and Cousot 1977). The concept is a formalization of flow analyses used in many optimizing compilers.

To understand the hits and misses, our static cache analysis approximates concrete cache states, which are simulated by the instrumentation presented in the previous section. In the approximation, we use *cache ages* that describe, how recently a data element has been referred to. The age of an element can be  $\{1, \dots, A, \top\}$ , where  $\top$  means that the element is not in cache.

At each program point, we statically compute the upper and lower bounds for cache ages of data elements. I.e., we approximately compute the position of each data element in the  $Q_t$  queues described in the previous section. In *must* analysis of cache states, data elements are mapped to their maximum cache age. In *may* analysis of cache states, data elements are mapped to their minimum cache age. Memory references that always hit are found by *must* analysis and memory references that always miss are found by *may* analysis:

- If the age of a referred to data element is always less than or equal to  $A$ , then the reference is always a hit.
- If the age of a referred to data element is newer less than or equal to  $A$ , then the reference is always a miss.
- Otherwise, we do not know.

For example, let associativity of a cache be two. Consider the following piece of code using four pointer variables  $a$ ,  $b$ ,  $c$ , and  $d$  that point to different memory lines belonging to a single cache set.

```
y := *b
x := *a
if x > y then
  x := *c
if x > y then
  y := *a
else
  y := *d
```

Consider state of the cache before the second `if` statement. In our *must* analysis the abstract state is  $\{(\hat{a}, 2), (\hat{b}, \top), (\hat{c}, \top), (\hat{d}, \top)\}$ . Thus, the maximum age of the data element pointed by  $a$  is 2. For other locations, it is unlimited, i.e.,  $\top$ . In our *may* analysis, the abstract state is  $\{(\hat{a}, 1), (\hat{b}, 2), (\hat{c}, 1), (\hat{d}, \top)\}$ , i.e., three locations may be cached. Thus in the second `if` statement, the memory reference using pointer  $a$  is always a hit and the memory reference using pointer  $d$  is always a miss.

## COMPILING FAST SIMULATORS

Our combined analysis has three phases: a compilation phase, an execution phase, and a summary phase. In the compilation phase, a subject program is statically analyzed and a simulator is built. In the execution phase, the simulator is executed (typically with several inputs). In the summary phase, the analysis information of the compilation phase and the analysis information of the simulation phase are combined.

The execution phase follows the typical procedures of simulation studies. The summary phase simply merges the results of the static analysis and simulation. The compilation phase is special, therefore, we will describe it in detail.

The compilation phase consists of three program transformation steps: program instrumentation, partial evaluation, and program slicing. The first step creates an execution-driven simulator and the last two are program specializations, which make it faster and more compact. The two specialization steps need static analysis information (i.e., results of the analysis explained in the previous section).

### Partial evaluation

Partial evaluation (see e.g., Jones et al. 1993) is a program transformation that is given a subject program with part of its input data. It constructs a new program that, when given the remaining input, will yield the same result that the original subject program would have produced given both inputs.

Consider the program  $P^A$  computing  $\llbracket P^A \rrbracket \langle I, I^+ \rangle = \langle O, O^+ \rangle$ . Let  $peval$  denote the partial evaluator, then

$$\llbracket peval \rrbracket \langle P^A, I^+ \rangle = P_{I^+}^A \Rightarrow \llbracket P^A \rrbracket \langle I, I^+ \rangle = \llbracket P_{I^+}^A \rrbracket \langle I \rangle$$

for all  $I$ . Thus, we fix the analysis initialization (input) and evaluate statically the part of the analysis that we can.

Our partial evaluation is based on static analysis of the original program  $P$ . The static analysis is done for the same task as the preceding instrumentation for the simulation (which produced  $P^A$  from  $P$ ). Static analysis can give us static values of both the original program and its instrumentation.

Instead of using the complete analysis instrumentation given in the third section, we can use simplified instrumentations at references that are known to hit or miss, i.e., only one branch of the **if** statement is used and the condition is omitted. Further, we can use a faster operation  $line'(l)$  instead of the complete operation  $line(l, N, A, L)$ , because  $N$ ,  $A$ , and  $L$  are constants.

### Slicing

Program slicing (see e.g., Tip 1995) is an operation that identifies semantically meaningful decompositions of programs. Usually, two kinds of slices are identified:

- A *backward slice* of a program  $P$  with set of program elements  $S$  consists of all program elements that might affect the values computed by  $S$ .
- A *forward slice* of program  $P$  with set of program elements  $S$  consists of all program elements that might be affected by the values computed by members of  $S$ .

$S$  is called the *slicing criterion*. In our analysis, the slicing criterion consists of input statements of the original program, and output statements of the analysis instrumentation. We use our slicing criterion to compute a backward slice of the simulator.

As described in the preceding, a partially evaluated analyzer computes both the original output and the analysis output:

$$\llbracket P_{I^+}^A \rrbracket \langle I \rangle = \langle O, O^+ \rangle$$

We do not need the original output. Further, we do not need the program elements that do not affect our analysis result.

We use slicing to implement a program transformation that yields a program  $P_{I^+S}^A$  computing only the analysis output. Let  $slicer$  be the transformation, then for all  $I$ :

$$\begin{aligned} \llbracket slicer \rrbracket \langle P_{I^+}^A, S \rangle &= P_{I^+S}^A \Rightarrow \\ \llbracket P_{I^+S}^A \rrbracket \langle I \rangle &\stackrel{O^+}{=} \llbracket P_{I^+}^A \rrbracket \langle I \rangle \ \& \ \llbracket P_{I^+S}^A \rrbracket \langle I \rangle \stackrel{O}{=} \emptyset \end{aligned}$$

where  $S$  is the slicing criterion,  $\stackrel{O^+}{=}$  denotes equality of analysis output, and  $\stackrel{O}{=}$  denotes equality of original output.

Program slicing is done by analyzing relations between program elements. Program elements do computations by using values to define new ones or to control program flow. If a value or a flow of control is not used, then the elements defining the value or controlling the flow can be removed.

Optimizing compilers typically apply simple forms of program slicing, e.g., dead code elimination. Dead code elimination removes program elements that compute unused values. Analyzing control dependencies caused by jumps and especially subroutine calls is more complex. Compiler rarely do such analysis.

Consider the analysis instrumentation of our cache simulator. A hit removes a cache line from the replacement queue. If we know that a reference is always a hit, then it must be preceded by a reference that places the cache line in the queue. Thus, we have a pair of the form:

```
insert_tail(Q_t, t)
...
remove(Q_t, t)
```

Actually, there is no need to insert the cache line in the replacement queue, because of the following hit. We know that the inserted line will never reach the head of the queue. The insertion and the related removal can be sliced away.

To cope with undetected pointer aliasing we use a counter of pending insertions for each cache line instead of the simple *incache* flag. The slicing can be complicated by the branching structure of a program. We can slice away only those insertions that are definitely followed by a hit. By using such a conservative approach, the sliced simulator computes the same cache behavior for a program than the original simulator.

Such a code transformation may seem insignificant. However, most of references in a typical program are hits, and sequences of hits are common. (That is the reason why cache memories are efficient.) For such re-hits, both the removal and the insertion are sliced away, and thus, also the computation needed to identify the line is not needed. No instrumentation remains after the slicing at such references.

The slicing can proceed even further. If there is no instrumentation at a memory reference, it is possible that there is no need to compute the reference. Thus, it becomes possible to slice away parts of the original program that is driving the simulation.

## EXPERIMENTAL RESULTS

We experimented with our method to show its potential. We found out in our experiments that a relatively simple static analysis is sufficient to yield significant speed up of simulations. In the following, we describe our experiments that were solely based on *must* analysis of abstract cache state.

Instead of considering some specific hardware, we analyzed the operation of a generic load/store architecture machine, which we call SM (Simple Machine). SM is a register machine with a simple instruction set. The main feature of SM is that its memory system can be parameterized to conform with various memory configurations. In the context of cache analysis, only instructions addressing the memory are significant. Therefore, SM is a representative for a large set of machines using the load/store architecture.

In our experiments, we used a tool set called MSE (Memory Simulation Environment) (Hirvisalo 2004). MSE has a compiler that generates code for SM. We used the compiler to compile three applications that were written in C language:

- *di*: A message dispatcher, which receives messages, decodes them and routes them further. The decoding and routing is implemented hard coded. Addresses of most memory references are dynamically computed.
- *da*: A relational database application, whose index is implemented as an unbalanced binary tree. Addresses of most memory references are dynamically computed.
- *co*: A control application, which operates like a device driver. The data structures of the application are mostly static.

We did two kinds of experiments, which we call static and dynamic experiments. In both experiments, we used two tools: one that built traditional simulators (as explained in the third section) and one that built specialized simulators (as explained in the previous section). In static experiments, we measured and compared the simulators. In dynamic experiments, we measured and compared their execution.

appl.	associativity=1		associativity=16	
	static solution	speed incr.	static solution	speed incr.
<i>di</i>	15%	25%	52%	90%
<i>da</i>	60%	110%	62%	125%
<i>co</i>	42%	60%	70%	175%

The performance of the method depends mostly on the target program and associativity of the cache. Our static cache analysis could classify 15% to 70% of the memory references. The performance of the static analysis depended on the dynamism of the addressing and on the interleaving of the memory references. Accesses of *Database* are more local than accesses of *Dispatcher*. They both use dynamic addressing, but *Control* uses static addressing. The actual (i.e., simulated) cache hit ratio for all the applications was typically around 90%.

We used a Pentium computer to run the dynamic experiments. The specialized simulators were 25% to 175% times faster than the original simulators. The speed-up is mostly caused by slicing. The direct effect of partial evaluation is minor, but it makes the slicing possible by removing branching in the instrumentations.

## CONCLUSIONS

This paper presented an approach for building fast simulators. The approach combines simulation and static program analysis: it uses simulation to fill in the results of static analysis and static analysis to speed up simulation.

We use two kinds of program specialization in speeding up simulation: partial evaluation and program slicing. They supplement each other: partial evaluation works forwards and slicing works backwards in the flow of control of a program.

Our approach is an abstract one. This leaves several options available in implementing a specific analysis. The instrumentations needed in simulation depend on the analysis to be done. Also, the instrumentation mechanism can differ. The same is true for the static analysis suggested. Abstract interpretation gives a theoretical framework for static analysis. Several implementations are possible within the framework, e.g., the work-list algorithm (see e.g., Nielson et al. 1999).

Alternatives exists also for implementing partial evaluation and program slicing. For example, simple methods like

constant folding or complex methods like polyvariant specialization (Jones et al. 1993) can be used in partial evaluation. In program slicing, there exists methods, which are based on data-flow equations, information-flow relations, and dependence graphs (Tip 1995).

As an application of the generic approach we discussed cache performance evaluation. Many processors have counters for cache misses and hits. Cache performance can be directly measured by using them. However, it is very hard to link those results with the program code and data structures. Thus, simulation has usually been the choice for performance studies and the related tools giving detailed information (e.g., Martonosi and Ga 1992).

Cache simulations can be speeded up, because most of memory references in a typical program always cause cache hits. The approach can be combined with parallel and distributed simulation methods. Our experimental results show that significantly faster analysis is achieved by using our approach.

## REFERENCES

- A. Agarwal and M. Huffman, 1990. "Blocking: Exploiting Spatial Locality for Trace Compaction." In *Proceedings of 90' ACM SIGMETRICS*, Performance Evaluation Review, 18(1), pages 48–57.
- P. Cousot and R. Cousot, 1977. "Abstract Interpretation: A Unified Lattice Model for Static Analysis of Programs by Construction or Approximation of Fixpoints." In *Fourth ACM Symposium on Principles of Programming Language*, pages 238–252, Los Angeles, California. ACM Press, New York.
- R. Fujimoto, 1999. "Parallel and Distributed Simulation." In *Proceedings of 1999 Winter Simulation Conference*, pages 122–131.
- J. Ha and E.E. Johnson, 1994. "PDATS: Lossless Address Trace Compression for Reducing File Size and Access Time." In *Proceedings of 1994 IEEE International Conference on Computers and Communications*. IEEE.
- P. Heidelberg and H. Stone, 1990. "Parallel Trace-Driven Simulation by Time Partitioning." In *Proceedings of 1990 Winter Simulation Conference*, pages 734–737.
- V. Hirvisalo, 2004. *Using Static Program Analysis to Compile Fast Cache Simulators*. Helsinki University of Technology, Laboratory of Information Processing Science, Espoo, Otaniemi. PhD thesis.
- M. Horowitz, M. Martonosi, T.C. Mowry, and M.D. Smith, 1996. "Informing Memory Operations: Providing Memory Performance Feedback in Modern Processors." In *Proceedings of the 23rd Annual International Symposium on Computer Architecture*, pages 260–270.
- N.D. Jones, C.K. Gomard, and P. Sestoft, 1993. *Partial Evaluation and Automatic Program Generation*. Prentice Hall, New York.
- M. Martonosi, A. Gupta, and T. Anderson, 1992. "MemSpy: Analyzing Memory System Bottlenecks in Programs." In *Proceedings of the 1992 SIGMETRICS Conference on the Measurement and Modeling of Computer Systems*, pages 1–12.
- D.M. Nicol, A.G. Greenberg, and B.D. Lubachevsky, 1994. "Massively Parallel Algorithms for Trace-Driven Cache Simulations." *IEEE Transactions on Parallel and Distributed Systems*, 5(8):849–859.
- F. Nielson, H.R. Nielson, and C. Hankin, 1999. *Principles of Program Analysis*. Springer.
- S.A. Przybylski, 1990. *Cache and Memory Hierarchy Design*. Morgan Kaufmann Publishers, Inc., Palo Alto.
- T.R. Puzak, 1985. *Analysis of Cache Replacement Algorithms*. University of Massachusetts, Department of Electrical and Computer Engineering. PhD thesis.
- A.D. Samples, 1989. "Mache: No-Loss Trace Compaction." In *Proceedings of the Sigmetrics 89 Conference*, pages 89–97.
- A.J. Smith, 1977. "Two Methods for the Efficient Analysis of Memory Address Trace Data." *IEEE Transaction on Software Engineering*, SE-3(1).
- F. Tip, 1995. "A Survey of Program Slicing Techniques." *Journal of programming languages*, 3(3):121–189.
- R.A. Uhlig and T.N. Mudge, 1997. "Trace-driven Memory Simulations: A Survey." *ACM Computing Surveys*, 29(2):128–170.

# NESTED ON-LINE SIMULATION — A M/M/1-QUEUEING EXAMPLE

Thomas Bessey

Helena Szczerbicka

Simulation & Modeling, Department of Computer Science, University of Hannover

Welfengarten 1, 30167 Hannover, Germany

{tby,hsz}@sim.uni-hannover.de

## KEYWORDS

On-Line Simulation, Queueing Discipline, Nesting, Recursion

## ABSTRACT

For a M/M/1-queueing system, we study a novel queueing discipline based on on-line simulation which seeks to optimize future performance of the system by present decisions. Every arriving job has a due date which should not be exceeded by its actual completion time; in case of such excess, an undesired tardiness of the job is observed. The performance of the system is measured in terms of the overall mean tardiness. Basically, the queueing discipline used is Earliest Due Date. In addition, for every arriving job, its completion time is estimated by means of simulation; the job is rejected to be processed if its estimated completion time exceeds its due date (otherwise it is enqueued according to Earliest Due Date). Since future jobs may be enqueued in front of jobs already waiting in the queue (due to them having earlier due dates), the simulation should include the arrival process of the system for accurate estimations.

However, simulation of the system, particularly of its arrival process, introduces nesting of on-line simulation in itself due to recursive calls. In previous work, it turned out that nesting may have a significant impact on the performance resulting from operation of a queueing discipline based on on-line simulation. In this paper, we compare the performance of the system as it is controlled by on-line simulation with and without inclusion of the arrival process as well as with and without nesting of on-line simulation. It turns out that the results differ significantly not with respect to the overall mean tardiness (as in previous work) but with respect to the percentage of rejected jobs, which may have a considerable impact on overall customer satisfaction in real-world scenarios.

## INTRODUCTION

Today, semiconductor manufacturing becomes more and more customer-driven, i.e., instead of production "to the stock", chips are manufactured "just in time" (JIT) of customer demands. For such systems, optimality of their operation largely depends on their ability to respond rapidly to customer demands as they change over time (dynamic demands). In this context, novel approaches for real-time scheduling of jobs as well as injection of jobs into the system have been proposed, e.g. [7, 5, 8, 11] and [10], respectively. Such approaches apply simulation in order to simulate scheduling or injection alternatives and to select the alternative which leads to best estimated performance based on the current state of the system; as the simulation is executed during actual system operation and initialized to the current state, it is referred to as "on-line simulation" [4].

### On-Line Simulation

The idea is to optimize the system's performance on-line by repeatedly adjusting the system's parameters properly. This is referred to as adaptive control. In general, adaptive control is either reactive or proactive. The former type is characterized by adjusting the system's parameters only after a considerable performance drop is observed.

Proactive adaptive control tries to adjust the system's parameters before a performance drop is to occur, in order to avoid this drop. To this end, the system's future evolution is assessed in advance repeatedly. The instants of time at which the assessment of the further evolution and adjustment of the parameters are done are called decision points. The assessment is done as follows: First, the system's current state is copied to several identical system models. For each of these models, certain values of the system's parameters are set, according to some appropriate policies that alternatively could control the system. Once the initialization is done, the models are analyzed in order to assess the future evolution

under each policy. As the system under control typically is complex, simulation is the only feasible analysis method. This method is referred to as on-line simulation. With the results, the policy that leads to the optimal future performance of the system under control is chosen to be implemented next, that is, the system is controlled by the chosen policy until the next decision point. This process is referred to as decision making.

There are several problems encountered with this approach, such as setting of the decision points, repeated validation of the system model (the search space for the parameters may vary over time) and proper analysis of the simulation results [1]. As simulation runs consume much time, the number and the length of the simulation runs become crucial. Since the system under control continues to evolve while the next policy is sought by on-line simulation, further problems arise [1]. For a detailed discussion of on-line simulation and associated proactive adaptive control, see [4].

For some applications of this approach for traffic systems and (flexible) manufacturing systems, see [6] and [11, 5, 9], respectively. However, these applications are by far not suitable for widely adoption to the real world; they merely employ classical off-line simulation techniques for on-line usage. By now, no strict theoretical research has been done.

In this work, we focus on the application of on-line simulation to the control of injection of jobs into the system.

We have introduced on-line simulation as a means for evaluation of different alternative system configurations that may be operated for the time between two subsequent decision points in response to certain characteristics of the environment in order to avoid performance drops. There is some more specific usage of on-line simulation that has attracted attention recently: Novel approaches based on on-line simulation are suggested for the adaptive control of arrival processes into a (sub)system. For example, the injection of jobs (release of work) into a manufacturing system could be controlled in an adaptive manner in that jobs may be re-ordered or re-scheduled (e.g., grouped in batches) or even rejected in order to optimize overall job processing with respect to certain performance measures such as minimal tardiness. On-line simulation is used to simulate injection alternatives. The key difference of this approach to what we introduced before is that on-line simulation is executed for every job arriving at the system in order to make a decision (regarding ordering, scheduling and acceptance/rejection).

Instead of choosing between different configurations (injection policies) that can alternatively be operated for some time in order to make decisions regarding injection (such as First Come First Served, Earliest Due Date, Least Slack First, Shortest Job First etc.), on-line simulation is used for every job individually. Thus, the resulting injection policy itself is continuously adapting to the job injection process (it is the only configuration of the system for its entire operational lifetime). This way, it is expected to get even better results than when applying the "classical" approach since subsequent decision points have minimal distance regarding the external cause of state changings of the system (as there is not any arrival event between two subsequent decision points).

In [10], simulation-based job acceptance/rejection is presented as a novel approach for performance optimization (of a manufacturing system). However, the cited work is based on considerable simplifying assumptions such as exclusion of any arrival process from on-line simulation, questioning its applicability.

As inclusion of the arrival process leads to nesting of on-line simulation in itself due to recursive calls (the arrival process in turn is controlled by on-line simulation) which have to be terminated at some depth, the impact of different depths on the success of the approach certainly is of great interest.

## This Work

For a M/M/1-queueing system, we study a novel queueing discipline based on on-line simulation which seeks to optimize future performance of the system by present decisions. As in the case of semiconductor manufacturing, every arriving job has a due date which should not be exceeded by its actual completion time; in case of such excess, an undesired tardiness of the job is observed. The performance of the system is measured in terms of the overall mean tardiness  $\bar{T}$ . This measure is of major importance in real-world scenarios as it has direct impact on satisfaction of customer needs resulting from the JIT-paradigm. Basically, the queueing discipline used is Earliest Due Date (EDD). The basic idea of EDD is to order jobs in the queue according to their due dates; it gives jobs having earlier due dates higher priority than jobs having later due dates, potentially reducing overall mean tardiness. In addition, for every arriving job, its completion time is estimated by means of simulation; the job is rejected to be processed if its estimated completion time exceeds its due date (otherwise it is enqueued according to Earliest Due Date). Since future jobs may be enqueued in front of jobs already waiting in the queue (due to them having earlier due dates), the simulation should

include the arrival process of the system for accurate estimations.

Of course, a trade-off between mean tardiness according to all jobs and rejection of single jobs in order to optimize overall customer satisfaction should be found, which is out of scope of this paper. However, we show that the percentage of rejected jobs differs significantly depending on whether the arrival process is included in the on-line simulation or not and whether the on-line simulation is nested or not.

This study is intended as providing further insight into the application of on-line simulation to the control of injection of jobs; it is a supplement of previous work [2].

The remainder of this paper is organized as follows: In the next section, the system together with the novel queueing discipline based on on-line simulation is introduced in detail; following this, some experimental results according to performance of the system as it operates under the queueing discipline with and without inclusion of the arrival process as well as with and without nesting of on-line simulation are given. Finally, in the last sections, we outline future work and give an overall conclusion, respectively.

## THE SYSTEM

The system is an open queueing system with one single job arrival process and one single resource or server for the processing of the jobs. Whenever the resource is busy processing a job, arriving jobs wait in an unlimited queue (if not rejected); each job visits the resource at most once, i.e., there is not any loop nor any preemption. Both the interarrival times and the processing times are assumed to be exponentially distributed, i.e. the processes in question are Poissonian. The queueing system described is referred to as M/M/1 [3]. With  $\lambda$  being the arrival rate and  $\mu$  being the service rate ( $\lambda, \mu > 0$ ), the system is stable, i.e., it reaches a steady state, if and only if  $\lambda < \mu$ .

We consider  $K = 3$  job classes that may differ in their mean processing times  $\mu_k^{-1}$  ( $k = 1, 2, \dots, K$ ) as well as in their mean relative due dates, where we define the mean relative due date of a job class as a multiple of its mean processing time, denoted by the factor  $f_k$ . The due date of an actual job is its arrival time advanced by the relative due date of its class. The weight  $w_k$  of job class  $k$  is defined as the probability that an arriving job is of class  $k$ , i.e.  $\sum_{k=1}^K w_k = 1$ . This way, the arrival process with rate  $\lambda$  is logically split into  $K$  subprocesses with rates  $\lambda_k$

Table 1: Parameter Values of this Study

<b>k</b>	<b>w<sub>k</sub></b>	<b><math>\lambda_k^{-1}</math></b>	<b><math>\mu_k^{-1}</math></b>	<b>f<sub>k</sub></b>
<b>1</b>	0.75	$0.75^{-1}$	0.90	11
<b>2</b>	0.20	$0.20^{-1}$	0.95	13
<b>3</b>	0.05	$0.05^{-1}$	1.50	15

respectively; it can be shown that these subprocesses are in turn Poissonian, with rates  $\lambda_k = w_k \lambda$  [3].

The parameter values used in this study are given in table 1, where the mean interarrival time over all job classes is  $\lambda^{-1} = \left( \sum_{k=1}^K \lambda_k \right)^{-1} = \underline{1}$ . The system is stable regardless of the queueing discipline used, as the mean processing time over all job classes is  $\mu^{-1} = \sum_{k=1}^K w_k \mu_k^{-1} = \underline{0.94}$ . The rather high utilization of the system, which is  $\sigma = \lambda/\mu = \underline{0.94}$ , provides a sufficiently filled system for significant observations.

## On-Line Simulation (OSIM)

For every job, the following algorithm is applied:

```

Copy current system state  $q \rightarrow$  queue  $q'$ 
Insert job into  $q'$  according to EDD
Simulate system with initial state  $q'$ 
 $\rightarrow$  Let  $c$  be completion time of job

```

The simulation introduces parameters such as number and length of replications, confidence level, relative error and warmup period. Their values will be given when presenting the results as these are dependent from that values. In any case, the simulation is executed fast enough to avoid actual job arrivals during simulation, so jobs are never rejected because of inavailability of the queueing discipline OSIM being busy.

If the length of replications is not large enough, a job might not be completed during simulation; in such case,  $c := \infty$ .

The job is rejected to be processed if and only if  $c > Due$ , where  $Due$  is the due date of the job.

With the given algorithm, OSIM may reject jobs for optimization purposes, as opposed to other queueing disciplines such as EDD. More specifically, OSIM prevents any estimated tardiness of newly arriving jobs. However, as a job is waiting in the queue, it may still experience tardiness due to enqueueing of new jobs in front of it, so the policy **does not** aim at

estimated performance  $\tilde{T} = 0$  (optimality). While a modified algorithm could easily dequeue any waiting job predicted to experience tardiness (as estimated by simulation), such approach of dropping jobs initially accepted is of no general use in real-world scenarios as it is likely to reduce customer satisfaction.

The simulation based queueing discipline introduces another parameter, which denotes the queueing discipline assumed during simulation. As job arrivals occur during simulation, they have to be enqueued according to some discipline, just like in case of the real system. (Though simulated in this study, we may refer to the actually operated system as *real system*.) This way, we may even nest OSIM in itself. The Java code we have written for this study allows us in a very simple way to do such nesting up to any depth (for statistical correctness, with separate random number streams for any arrival process and any service process). The key question is whether and to what depth such nesting may sufficiently improve performance of the system. There is an intuitive idea of such improvement by nesting as performance estimation by on-line simulation should be much better when assuming the actual queueing discipline (OSIM) of the real system instead of some approximation (such as EDD, which is without any job rejection). The nesting of OSIM is terminated at some depth by assuming First Come, First Served (FCFS) as queueing discipline of the innermost OSIM; this way, the arrival process is effectively excluded from on-line simulation at this depth.

The major issue will be to treat nesting of on-line simulation analytically or numerically and to find that way applicable approaches for the computation of feasible and reasonable nesting depths before actual operation. The need for such approaches becomes evident facing the complexity of nested OSIM with respect to execution time, which increases exponentially according to the nesting depth  $d$ . Clearly, we seek for a trade-off between the expected performance gain, as achieved by application of nested OSIM, and the effort for executing such nesting.

For real-world complex systems, execution of nested on-line simulation is likely to become a tremendous task even for  $d = 2$ , questioning the purpose of any research on nesting. However, in any case where on-line simulation may be executed sufficiently fast (by means of existing or future speed-up techniques), the issue of nesting as well as the need for a solution becomes inevitable.

Table 2: Results of OSIM without Inclusion of Arrival Process

$r_1$	$r_2$	$r_3$	$\bar{T}_1$	$\bar{T}_2$	$\bar{T}_3$	$\bar{T}$
0.06	0.02	0.0	0.3	0.3	0.6	0.3

## EXPERIMENTAL RESULTS

We study the results observed with and without inclusion of the arrival process in the on-line simulation as well as with and without nesting of on-line simulation in order to discuss the impact of nesting on the results.

For an analytical treatment of OSIM, we need the distribution of the response time, as any response time less than the relative due date has to be excluded from consideration; however, there is no analytical solution yet even for the M/M/1-system, so we use simulation instead.

We verified our simulator of the real system with the help of some simple analytical results according to FCFS; as OSIM in turn is implemented using the very same simulator, all simulation results are considered as correct.

In any case, simulation of the real system is executed at confidence level 99% with a relative error limited to 10% (with respect to the performance measure). Its simulated operational lifetime is always large enough (fix at 10,000) to reach steady state, with the warmup period being 10% of that lifetime.

### EDD

Using EDD without any job rejection (without any on-line simulation) as queueing discipline, we get a performance of about 7.3.

### OSIM without Inclusion of Arrival Process

As any experiment involving OSIM takes much time to be completed, we limit the number of replications of any on-line simulation to 10; however, given the simplicity of the system, this limitation is not crucial to the results. The length of every replication of the on-line simulation is fix at 100; as the traces show, large lengths are unnecessary.

The results are given in table 2, where  $r_k$  is the percentage of rejected jobs of class  $k$  related to all arriving jobs of that class.

**Result:** OSIM leads to dramatically improved performance as compared to EDD, with an overall im-

Table 3: Results of OSIM with Inclusion of Arrival Process Assuming EDD

$r_1$	$r_2$	$r_3$	$\bar{T}_1$	$\bar{T}_2$	$\bar{T}_3$	$\bar{T}$
0.04	0.07	0.12	0.3	0.3	0.4	0.3

provement of about 96%.

In order to ensure that the improved performance is not just a consequence of random (uniform) rejection of jobs, we compare the results of OSIM with those of a modified version of EDD which is enhanced with uniform job rejection, where the probabilities of rejection are given by the percentages of rejection as observed when operating OSIM. The modified EDD results in a performance of about 2.3; we still note an improvement of performance when operating OSIM of about 87%.

### OSIM with Inclusion of Arrival Process

As discussed earlier, inclusion of the arrival process requires that the on-line simulation assumes some queueing discipline in order to enqueue the arriving jobs virtually. We present some results observed when assuming EDD as well as OSIM without inclusion of the arrival process, i.e., OSIM nested once.

#### ... Assuming EDD

The results are given in table 3.

**Result:** Interestingly, the performance remains unchanged (as compared to OSIM without inclusion of the arrival process), while the percentages of job rejection now are significantly higher (except for the main job class); in particular, while jobs of the third class are never rejected in the former case, we now note a considerable percentage of rejection, resulting in a reduced tardiness of that job class. While this has no effect on the overall tardiness (since jobs of the third class are rare), the result shows a significant difference between exclusion and inclusion of the arrival process with respect to job rejection, which can be crucial in real-world scenarios.

#### ... Assuming OSIM without Inclusion of Arrival Process

The results are given in table 4.

**Result:** In comparison to the former case of assuming EDD, the percentages of job rejection now are significantly lower among all job classes, resulting in a tardiness being reduced by 33%. The result shows a significant difference between assuming EDD (which

Table 4: Results of OSIM with Inclusion of Arrival Process Assuming OSIM without Inclusion of Arrival Process

$r_1$	$r_2$	$r_3$	$\bar{T}_1$	$\bar{T}_2$	$\bar{T}_3$	$\bar{T}$
0.03	0.03	0.01	0.2	0.2	0.2	0.2

is just an approximation for the actual queueing discipline) and assuming OSIM without inclusion of the arrival process (which is also just an approximation, however a better one).

### Preliminary Conclusion

The above results show that the overall tardiness depends not only on the percentages of job rejection but also on **which** jobs are rejected; OSIM nested once leads to a reduced tardiness while providing still fairly low percentages of job rejection.

### FUTURE WORK

We seek to find applicable approaches for the computation of feasible and reasonable nesting depths before actual operation. Promising solutions include analytical treatment of nesting by means of abstract models, numerical approximation by means of iteration schemes as well as metamodeling by means of self-referenced neural networks. We are currently working on these issues.

### CONCLUSION

On-line simulation is a promising approach for novel job acceptance/rejection policies; however, its success largely depends on the accuracy of approximation of the actual policy by the policy assumed during on-line simulation. Nesting of on-line simulation in itself due to recursive calls can dramatically improve accuracy, as we have shown in previous work. More generally speaking, nesting has a significant impact on the operation of a simulation-based policy, depending on the definition of performance measures and decision criteria, as we have illustrated in this work by means of a simple yet surprising example. However, as the effort for executing nesting of on-line simulation becomes tremendous for real-world complex systems, solutions must be found providing either feasible and reasonable bounds for the nesting depth or sufficient metamodels. This is considered to be a major challenge.

## REFERENCES

- [1] T. Bessey. On-line simulation: Towards new statistical approaches. In *Proc. Summer Computer Simulation Conference (SCSC)*, 2003.
- [2] T. Bessey. Application of on-line simulation to M/M/1-priority queueing. In *Proc. Summer Computer Simulation Conference (SCSC)*, 2004.
- [3] G. Bolch, S. Greiner, H. d. Meer, and K. S. Trivedi. *Queueing Networks and Markov Chains*. Wiley, 1998.
- [4] W. J. Davis. On-line simulation: Need and evolving research requirements. In J. Banks, editor, *Handbook of simulation*, chapter 13. Wiley, New York, 1998.
- [5] G. R. Drake and J. S. Smith. Simulation system for real-time planning, scheduling, and control. In *Proc. Winter Simulation Conference (WSC)*, 1996.
- [6] J. Esser, L. Neubert, J. Wahle, and M. Schreckenberg. Microscopic online simulation of urban traffic. In *Proc. 14th International Symposium on Transportation and Traffic Theory*, 1999.
- [7] C. M. Harmonosky. Implementation issues using simulation for real-time scheduling, control, and monitoring. In *Proc. Winter Simulation Conference (WSC)*, 1990.
- [8] C. M. Harmonosky, R. H. Farr, and M.-C. Ni. Selective rerouting using simulated steady state system data. In *Proc. Winter Simulation Conference (WSC)*, 1997.
- [9] S. Manivannan and J. Banks. Real-time control of a manufacturing cell using knowledge-based simulation. In *Proc. Winter Simulation Conference (WSC)*, 1991.
- [10] A. Nandi and P. Rogers. Simulation-based order acceptance in make-to-order manufacturing systems. In *Proc. Summer Computer Simulation Conference (SCSC)*, 2003.
- [11] A. I. Sivakumar. Optimization of cycle time & utilization in semiconductor test manufacturing using simulation based, on-line near-real-time scheduling system. In *Proc. Winter Simulation Conference (WSC)*, 1999.

# SOLUTION APPROACHES FOR THE CLUSTER TOOL SCHEDULING PROBLEM IN SEMICONDUCTOR MANUFACTURING

Heiko Niedermayer  
Computer Networks and Internet  
Wilhelm-Schickard-Institute for Computer Science  
University of Tübingen  
D-72076 Tübingen, Germany  
E-mail: niedermayer@informatik.uni-tuebingen.de

Oliver Rose  
Distributed Systems  
Department of Computer Science  
University of Würzburg  
D-97074 Würzburg, Germany  
E-mail: rose@informatik.uni-wuerzburg.de

## KEYWORDS

simulation, manufacturing, semiconductor, cluster tools, scheduling

## ABSTRACT

With the increase in computer performance scheduling complex manufacturing plants with global heuristics is becoming a realistic option. The complete factory scheduling problem consists of the global problem and its subproblems for each work center and machine. In semiconductor manufacturing a lot of processing is done using cluster tools. Cluster tools are a special kind of machine that can be described as a small factory. We discuss solutions for the optimization of schedules for cluster tools which is a subproblem of the factory scheduling problem. Our main idea is to use rules based on slow-down factors or search approaches based on the use of slow-down factors for predicting the cycle times. We use simulation to evaluate the schedules.

## INTRODUCTION

Semiconductor manufacturing plants are often considered to be the most complex factories that currently exist. Even ignoring the size of these manufacturing plants the scheduling problem is extremely complex. It is a job shop scheduling problem that includes batch machines, sequence-dependent setups, recirculating flows, and cluster tools. While batch machines and setups are widely studied in the literature (Pinedo 2001) the problem with cluster tools is not.

Cluster tools are machines that consist of loadlocks, handlers and other machines to process the wafers of the lots in the loadlocks. The advantage of cluster tools is that the processing of the wafers is pipelined. Thus compared with the cycle time of a lot using consecutive machines the cycle time of a lot in a cluster tool is reduced. As cluster tools are usually pumped to vacuum the yield may also be improved and less clean-room space may be required.

Since cluster tools are small factories themselves that may process more than one lot at a time, the behavior of

cluster tools with respect to the cycle time is complex. In this paper we usually consider cluster tools to have two loadlocks which is a valid assumption for most cluster tools currently in use.

## RELATED WORK

Typical work center problems that are addressed in the literature are single machine work centers, parallel machines, batch machines, and machines with sequence-dependent setups. Solutions and approaches for many of these problems can be found in the book by Pinedo (Pinedo 2001).

An important objective of the optimization problem is the total weighted tardiness (TWT). The FORCe scheduling project aims at globally scheduling a complete factory using the Shifting Bottleneck heuristic (Fowler et al. 2002a, Fowler et al. 2002b). For the solution of the work center problems they use the BATCS rule which is an adaptation of the ATC rule (Apparent Tardiness Cost) for batch machines and setups (Fowler et al. 2002b, Pabst et al. 2002).

For rather simple cluster tools Perkinson et al (Perkinson et al. 1994, Perkinson et al. 1996) analyzed their throughput and cycle time behavior analytically. Others used simulation and optimized schedules with genetic algorithms (Dümmler 1999). Dümmler already introduces the notion of a slow-down factor, but did not use it for scheduling. We analyzed and simulated cluster tools to study slow-down factors ourselves (Niedermayer and Rose 2003, Niedermayer and Rose 2004).

## FACTORY SCHEDULING

Scheduling deals with allocating resources for tasks over time. The goal is to find a solution that is optimal or near-optimal with respect to given objectives. Initially makespan was the main objective. The focus of manufacturing today is often more on customer orders and on-time delivery, thus it is most important to satisfy all or at least the most important due dates.

Finding an optimal schedule is not always easy. In fact, most scheduling problems are NP-hard.

Definition: *Lateness*

Let  $c_i$  be the completion time of lot  $i$  and  $d_i$  its due date, then the lateness  $L_i$  of lot  $i$  is  $L_i = c_i - d_i$ .

Definition: *Tardiness*

Let  $c_i$  be the completion time of lot  $i$  and  $d_i$  its due date, then the tardiness  $T_i$  of lot  $i$  is  $T_i = \max\{0, c_i - d_i\} = \max\{0, L_i\}$ .

Definition: *Total Weighted Tardiness (TWT)*

In addition to the due date each lot  $i$  is assigned a weight  $w_i$  that specifies its importance. Then the total weighted tardiness of all lots is defined as  $TWT = \sum_i w_i T_i$ .

In manufacturing it is common to use dispatching rules at the work centers or machines. Dispatching rules are myopic methods. Hence, they are usually not optimal for the solution of the global problem. It is likely that they are also not optimal regarding the local objective of interest.

### PROBLEM GRAPH AND SCHEDULE GRAPH

Scheduling problems can be transformed into graph problems. Additionally, graphs give a visual representation of the problem.

The problem graph is a directed graph. It has a source (o) and sink (\*) node, and there is a node for each operation. A detailed description of problem and schedule graphs can be found in the book by Ovacik and Uzoy (1997). Figure 1 shows the problem graph. It includes all potential arcs for determining a schedule. Operations at one machine or work center build a clique, except for operations of one job.

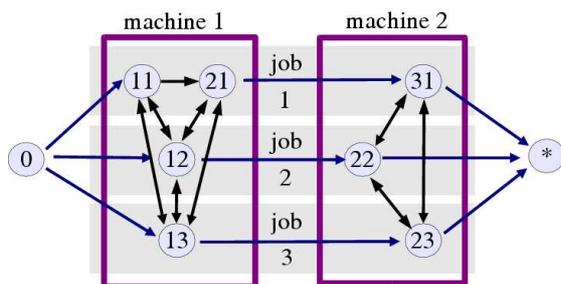


Figure 1: Problem graph

The schedule graph (Figure 2) is a directed acyclic graph. The arcs indicate the order of the operations. The longest path to an operation is the time at which the operation can start. The longest path from source to sink is the makespan. As one can see the scheduling problem quite naturally decomposes into subproblems at the work centers or machines. The graph specifies the release dates for the operations of the subproblem. When a one machine or work center is scheduled, these release dates for other work centers change. Hence, one

solution for the scheduling problem can be to iteratively solve work center subproblems, adapt the graph and solve the next subproblem, and so on.

In this text we focus on the solution of such subproblems.

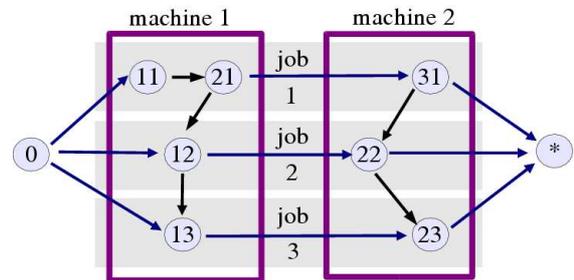


Figure 2: Schedule Graph

### WORK CENTER SUB PROBLEMS

The jobs for each work center or machine have to be scheduled. A decomposition of the global scheduling problem usually consists of such problems. The same subproblem may be optimized several times with different release dates as arcs in the schedule graph change. The result is a sequence of operations in the input queue of the work center or machine. This may also include the assignment of an operation to a specific resource, e.g. to a particular machine in parallel machine problems.

Typical problems are single machine problems, parallel machine problems, batch machine problems, problems with sequence-dependent setups, and, as in our case, cluster tool problems.

### CLUSTER TOOLS

Since the 1990s cluster tools are becoming a more and more integral element of wafer processing in semiconductor manufacturing. A cluster tool consists of a mainframe with several machines (chambers). A machine usually processes one wafer. A cluster tool has handlers to move the wafers from one chamber to another. The lots are stored in the loadlocks. A handler takes a wafer from a lot in the loadlock and moves it to the processing chamber as indicated by its recipe. A wafer may visit several machines until it is completed and brought back to its lot. A lot is completed and can leave the cluster tool when all its wafers are completed. Since there is vacuum inside the tool the loadlock has to be pumped after a lot enters and vented before the lot leaves the tool.

Figure 3 shows the model of an Endura cluster tool with 2 hand-over chambers. A typical configuration could be that the chambers in the section close to the loadlocks are used for pre-processing steps (alignment, heating) and post-processing steps (cooling) and the chambers of the other section are the machines for the long processing steps (Seidel 2001).

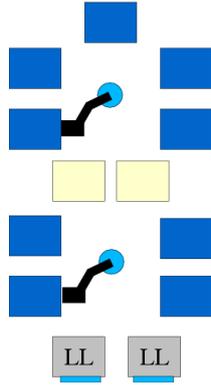


Figure 3: Model of an Endura cluster tool

For our scheduling problem cluster tools with one loadlock behave exactly like single machines (with lot-dependent processing times). Thus, we focus on parallel mode cluster tools with two loadlocks. Here, lots can overlap and since they use common resources they slow each other down during their overlap. Depending on the compatibility of the recipes the slow-downs can range from small (roughly 1) to large (more than 2). The result is a rather complex behavior with respect to the lot cycle time.

Definition: *Slow-down factor*

Let  $CT(A, f)$  be the time needed for the fraction  $f$  of the work for lot  $A$  in the cluster tool when lot  $A$  is alone. Let  $f$  be the fraction of the work done for lot  $A$  during its overlap with lot  $B$  and  $CT(A, A + B, f)$  be the length of this overlap. Then the slow-down factor is

$$SDF_{AB} = \frac{CT(A, A + B, f)}{CT(A, f)}$$

In previous papers we studied this slow-down factor and studied how to predict it (Niedermayer and Rose 2003, Niedermayer and Rose 2004). The slow-down factor may be different for different kind of overlaps, e.g. our simulator tends to prefer the lot that entered the tool first.

The particular tool cycle times for lots in a schedule can only be predicted for the complete schedule from the beginning to the time of interest. This can be done using simulation or approximation. An approximation can be computed as follows. We assume that we know the cycle times of all lots while they use the cluster tool exclusively (in single mode). This can be estimated once for each recipe by simulation or analytically. Additionally, we assume that we know the slow-down factors for the recipe combinations of interest. If a lot is processed completely during an overlap with another lot its cycle time can be approximated by

$$CT[A, A + B, 100\%] = SDF_{AB} CT[A, 100\%]$$

For lots that overlap with more than one lot or only partially overlap with one lot, the amount of work that was done during each overlap has to be determined and the cycle time can then be computed accordingly. Figure

4 shows the basic algorithm in pseudo code. The function `TimeToNextEvent` and the estimation of the amount of work done during an overlap need the slow-down factors and the single mode cycle time predictions for the lots.

```

ALGORITHM
INPUT: Queue with lots in the order given by the
schedule, Time
WHILE Queue.notEmpty() and Clustertool.notEmpty()
DO
  FOR all empty loadlocks DO
    IF Queue.nextLotReady(Time) THEN
      Add Q.next() to loadlock and set its start time.
    ENDIF
  ENDFOR
  LengthOfOverlap = TimeToNextEvent(all lots in
  clustertool);
  Time = Time + LengthOfOverlap
  FOR all lots in cluster tool DO
    Determine the amount of work done for lot during
    the overlap.
    IF lot is completed THEN Remove lot from
    loadlock and set its completion time.
  ENDFOR
ENDWHILE

```

Figure 4: Basic algorithm for computing lot cycle times of schedule for cluster tools

Compared to simulation this algorithm is very efficient because it only needs a few floating-point operations per lot.

From this description of cluster tools in parallel mode it becomes obvious that parallel mode cluster tool scheduling is a problem that is different from standard scheduling problems with batch machines or sequence-dependent setups, etc.

## CLUSTER TOOL SIMULATION

For our studies we used the cluster tool simulator `CluSim` that was developed by Dümmler et al. (Dümmler 1999, Schmid 1999) at the University of Würzburg and is also used at Infineon Technologies for cluster tool optimization.

We used simulation for two purposes. The first one was to determine slow-down factors for all combinations of lots. Since the slow-down factor may vary depending on how lots overlap and which lot is processed first, we simulated different overlaps. We discussed this in more detail in a previous paper (Niedermayer and Rose 2003). This is important because in the next section we will use these slow-down factors for scheduling.

The second purpose of simulation is to evaluate the schedules computed by the different optimization approaches. For the evaluation scenarios were created, for each test set 20 scenarios with 20 lots, each lot with

weight, release date and due date. The lots were released in a way that the cluster tool did not run out of work. The results of the simulation runs are used to determine cycle times, makespan and tardiness. The cluster tool input queue in CluSim is a FIFO queue (First In First Out). To evaluate a particular schedule the release dates in the input file of the simulator have to be adapted to ensure the correct order. The release date of a lot following another lot has to be higher than the release date of its predecessor even though it is actually released before its predecessor.

Simulation is also important for the overall scheduling problem. Once the operations for a cluster tool are scheduled, the schedule can be simulated and the results can be used to specify the processing times in the schedule graph.

## CLUSTER TOOL SCHEDULING

In this section we describe two approaches to schedule cluster tools. We assume that we have one input queue and that we can modify the order of the lots in the queue.

Our first approach is a dispatching rule based on the slow-down factors introduced in the last two sections.

The dispatching rule SDFavg works as follows. Let lot A be in the cluster tool. From all lots that are currently released and available in the input queue take the lot B

that minimizes  $\frac{SDF_{AB} + SDF_{BA}}{2}$ .

To implement this rule we need to predict the cycle times of the lots already scheduled to determine the current time. This is necessary to identify the lots that are released, but not yet scheduled.

Our second approach is a Random Search approach. It is actually a variation of a genetic algorithm. It uses a population of individuals, elitism, and the mutation is based on permutation of lots. Schedules are evaluated with cycle time predictions using slow-down factors for the lot combination and single mode cycle time predictions for the lots. We did not search for the fastest search algorithm. We simply wanted to study which solution such a search algorithm can find given a considerable amount of time. Since the predictions differ from the actual cycle times an optimum of the Random Search is not necessarily an optimum of the real scheduling problem.

## COMPARISON

In this section we compare SDFavg, the Random Search based on predictions with slow-down factors, FIFO and EDD (Earliest Due Date).

First, we examined the quality of the schedulers for the objective makespan. Table 1 shows that SDFavg outperforms FIFO and that SDFavg is roughly as good as the Random Search.

Table 1: Objective Makespan

Optimizer	Test set	Normalized Makespan
FIFO	Dresden 1	1.152
SDFavg	Dresden 1	1.036
Random Search	Dresden 1	1.032
FIFO	Dresden 2	1.129
SDFavg	Dresden 2	1.015
Random Search	Dresden 2	1.033
FIFO	Villach	1.135
SDFavg	Villach	1.045
Random Search	Villach	1.037
FIFO	Simple	1.293
SDFavg	Simple	1.031
Random Search	Simple	1.060

For the test sets Simple and Villach we also computed lower bounds for the makespan for each scenario using a work distribution algorithm. Since precedence constraints are ignored these bounds are true lower bounds. Any schedule has to be worse. Table 2 gives the details. On average SDFavg is rather close to this lower bound.

Table 2: Comparing the schedules with a true lower bound

Test set	Lower Bound (unrealistic)	SDFavg	FIFO
Villach	47396s	54749s	61822s
Simple	67169s	78687s	101641s

As second objective we used the total weighted tardiness (TWT). Table 3 shows that the cluster tool throughput is sequence-dependent and that in the schedules produced by EDD the throughput is reduced. As a consequence, a lot of lots are not completed on time. Our rule SDFavg on the other hand does not know anything about due dates, but since its schedules result in high throughput most lots are completed in time. The Random Search with objective TWT was usually slightly better.

We can conclude that dispatching rules that ignore cluster tool behavior produce poor throughput and may therefore fail to achieve their primary objective. Thus, combining them with SDFavg or similar approaches is necessary.

Table 3: Objective Total Weighted Tardiness

Optimizer	Test set	Normalized TWT
EDD	Dresden 1	1.70
SDFavg	Dresden 1	1.18
Random Search	Dresden 1	1.28
EDD	Dresden 2	1.53
SDFavg	Dresden 2	1.19
Random Search	Dresden 2	1.08
EDD	Villach	1.35
SDFavg	Villach	1.32
Random Search	Villach	1.06
EDD	Simple	1.46
SDFavg	Simple	1.15
Random Search	Simple	1.10

### LOCAL IMPROVEMENTS

To avoid long individual cycle times or to avoid poor overall throughput it is useful to have the option to let one lot wait and thus to avoid certain combinations of lots.

To do this it is necessary to have combination characteristics that indicate for the scheduler whether a combination should be avoided. Again, slow-down factors are a solution. To avoid long individual cycle times one could avoid lot combinations for which the slow-down factor of one lot is larger than a given threshold, say 2.5. To avoid poor overall performance one could avoid a lot combination when the average slow-down factors are larger than a particular threshold, say 2.0.

### CONCLUSIONS

We demonstrated that dispatching rules based on slow-down factors are an interesting and promising approach for scheduling cluster tools. This approach takes the special characteristics of cluster tools into account and avoids poor throughput and long cycle times. It is also efficient and schedules can be computed quickly.

Larger studies are to be made. In particular, the impact of prediction errors has to be analyzed more deeply. Future work may also include the adaptation of the ATC rule for the use with cluster tools and to include slow-down factors. One might also think of combining EDD or Critical Ratio with a rule based on slow-down factors for local optimization. Another option is to analyze and use other, maybe simpler, lot compatibility measures than slow-down factors. Another field of research could be to find other problems where such an approach might be useful.

### REFERENCES

- Dümmler, M. 1999. "Using simulation and genetic algorithms to improve cluster tool performance." In Proceedings of the 1999 Winter Simulation Conference. 875-879.
- Fowler, J., Carlyle, M., Runger, G., Gel, E., Mason, S., Rose, O. "A New Approach for Scheduling Semiconductor

Wafer Fabs." Semiconductor Fabtech, 15th Edition, pp. 39-41, 2002.

- Fowler, J., Brown, S., Carlyle, M., Gel, E., Mason, S., Mönch, L., Rose, O., Runger, G., Sturm, R. "A Modified Shifting Bottleneck Heuristic for Scheduling Wafer Fabrication Facilities." In Proceedings of the FAIM 2002, July 15-17, Dresden, Germany, pp. 1231-1236, 2002.
- Niedermayer, H. and Rose, O. "A Simulation-based Analysis of the Cycle Time of Cluster Tools in Semiconductor Manufacturing" In Proceedings of the 15th European Simulation Symposium, Delft, Netherlands, 2003.
- Niedermayer, H. and Rose, O. "Approximation of Cycle Time of Cluster Tools in Semiconductor Manufacturing" In Proceedings of the Annual IIE Industrial Engineering Research Conference, Houston, Texas, 2004.
- Ovacik, I.M. and Uzsoy, R. 1997. "Decomposition Methods for Complex Factory Scheduling Problems" Kluwer Academic Publishers.
- Pabst, D., Fowler, J., Pfund, M., Mason, S., Rose, O., Mönch, L., Sturm, R. "Deterministic Scheduling of Wafer Fab Operations." In Proceedings of the Brooks Worldwide Automation Symposium 2003, Oct 2003.
- Perkinson, T.L., McLarty P.K., Gyurcsik, R.S., Cavin III., R.K. 1994. "Single-Wafer Cluster Tool Performance: An Analysis of Throughput." In IEEE Transactions on Semiconductor Manufacturing Vol. 7, August 1994.
- Perkinson, T.L., McLarty P.K., Gyurcsik, R.S., Cavin III., R.K. 1996. "Single-Wafer Cluster Tool Performance: An Analysis of the Effects of Redundant Chambers and Revisitation Sequences on Throughput." In IEEE Transactions on Semiconductor Manufacturing Vol. 9, August 1996.
- Pinedo, M. 2001. "Scheduling. Theory, Algorithms, and Systems. 2<sup>nd</sup> edition. Prentice-Hall.
- Schmid, M. 1999. "Modellierung und Simulation von Cluster Tools in der Halbleiterfertigung." Master's thesis. Department of Computer Science. University of Würzburg, Germany.
- Seidel, G. 2001. "Simulation und Optimierung von Cluster Tools in der Halbleiterfertigung". Master's thesis, Institute of Mathematics. Technical University of Graz, Austria.

### AUTHOR BIOGRAPHIES

**HEIKO NIEDERMAYER** is Ph.D. student at the Department for Computer Networks and Internet at the University of Tübingen. He received an M.S. degree in Computer Science from the University of Würzburg. His e-mail address is:

niedermayer@informatik.uni-tuebingen.de .

**OLIVER ROSE** is assistant professor in the Department of Computer Science at the University of Würzburg, Germany. He received an M.S. degree in applied mathematics and a Ph.D. degree in computer science from the same university. He has a strong background in the modeling and performance evaluation of high-speed communication networks. Currently, his research focuses on the analysis of semiconductor and car manufacturing facilities. He is a member of IEEE, ASIM, and SCS. His web address is:

www3.informatik.uni-wuerzburg.de/~rose .

# PERFORMANCE COMPARISON OF TRADITIONAL SCHEDULERS IN DIFFSERV ARCHITECTURE USING NS

Miklós Lengyel  
János Sztrik  
Department of Informatics Systems and Networks  
University of Debrecen  
H-4010 Debrecen, P.O. Box 10, Hungary  
E-mail: [mlengyel@inf.unideb.hu](mailto:mlengyel@inf.unideb.hu)

## KEYWORDS

AF PHB, differentiated service, EF PHB, queue management, scheduler, UDP

## ABSTRACT

The aim of our investigation is to consider a simple dumbbell Diffserv network topology in which performance comparison (in terms of throughput, delay and queue length) is made between the traditional traffic scheduling algorithms: Priority (PRI), Weighted Round Robin (WRR) and Weighted Interleaved Round Robin (WIRR) schedulers. Random Early Detection (RED) is used as active queue management algorithm. An earlier version of this paper can be found in [6]. In this paper we investigate how the above mentioned performance measures vary if we change the packet size. In the core of the network there is a bottleneck link and the consideration is performed on that node. All of our traffic generators are Constant Bit Rate (CBR), the transport protocol is User Datagram Protocol (UDP). We used Network Simulator (NS, version 2) for our simulation experiments.

## 1. INTRODUCTION

The history of the Internet has been of continuous growth in the number of hosts, the number and variety of applications, and the capacity of the network infrastructure. A scalable architecture for service differentiation must be able to accommodate this continuous growth. The Differentiated Services (Diffserv or DS) architecture [1] provides a more flexible, scalable architecture than the existing models of service differentiation. The specification of Diffserv architecture appeared in 1998, but the current research is still expanding it. The architecture is based on a simple model where traffic entering a network is classified and possibly conditioned at the boundaries of the network, and assigned to different DS codepoints. Within the core of the network, packets are forwarded according to the per-hop behaviour associated with the DS codepoint. A per-hop behaviour (PHB) is a description of the externally observable forwarding behaviour of a DS node applied to packets with a particular DS codepoint. PHBs are implemented in

nodes by means of some buffer management and packet scheduling mechanisms. Two different PHBs were developed: the Assured Forwarding (AF) PHB [2] and the Expedited Forwarding (EF) PHB [3]. The AF PHB group provides delivery of IP packets in four independently forwarded AF classes. Within each AF class, an IP packet can be assigned to one of three different levels of drop precedence. EF PHB is intended to provide low delay, low jitter and low loss services by ensuring that the EF packets are served at a certain configured rate.

The Diffserv architecture achieves scalability by implementing complex classification and conditioning functions only at network boundary nodes, and by applying per-hop behaviours to aggregates of traffic which have been appropriately marked using the DS field in the IPv4 or IPv6 headers. This architecture only provides service differentiation in one direction of traffic flow and is therefore asymmetric.

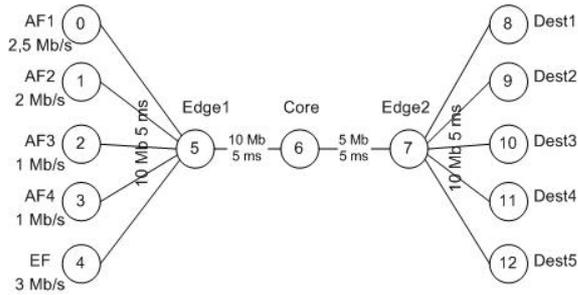
While many studies have addressed issues on the Diffserv architecture (e.g., dropper, marker, classifier and shaper), there have been few attempts to analytically understand a flow's behavior in a diffserv network.

In this paper we enhance our earlier paper [6], in which a performance comparison was made between the traditional traffic scheduling algorithms (PRI, WRR, WIRR) in a dumbbell Diffserv topology using packets with size of 1000 bytes. We consider how the performance measures vary if we use packets with the size of 500 bytes in the same environment.

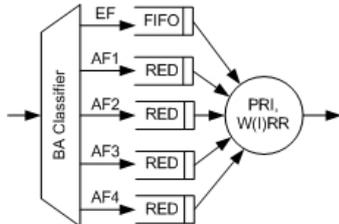
Section 2 presents the results of the simulations. Conclusions are drawn in Section 3.

## 2. SIMULATION RESULTS

Simulations were performed using Network Simulator [7] (NS, version 2.1b9a), which was developed at the University of California. NS is an event-driven network simulator, which is implemented in C++ and uses OTcl (Object Tool Command Language) as the command and configuration interface. We considered the simple dumbbell topology shown in Fig. 1/a. All links have the same fixed delay of 5 ms. The consideration is performed on the Core node where there is the bottleneck link. The structure of the output interface of the core node is shown in Fig. 1/b.



a.) Simulated network topology



b.) Simulated output interface

Figure 1: Simulation scenario

We ran each simulation for 60 seconds. The traffic generators are CBRs over UDP. The nodes 0-3 generate AF1-AF4 traffic, while 4. node generates EF. The  $i$ -th node sends packets to  $i+8$ -th node,  $i = 0,1,2,3,4$ . An AF class is implemented in the nodes as a RED physical queue with three virtual queue, while EF as a droptail (FIFO) queue. Figure 2,3 show the Sent packet ratio and the Received packet ratio, which was set up such that they are equal in case of the three schedulers. We make throughput, delay and queue length comparison between the scheduling algorithms PRI, WRR and WIRR and we confront it with our earlier results [6].

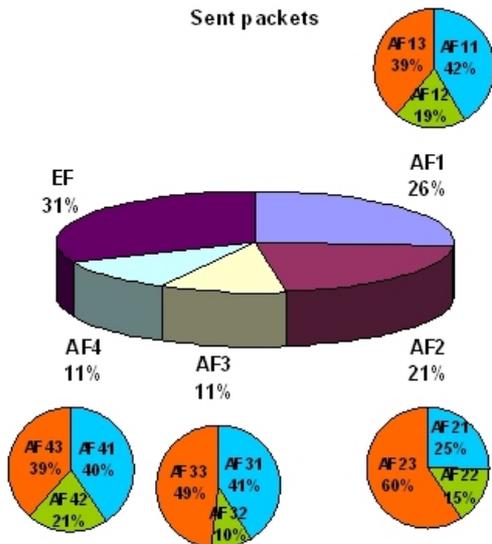


Figure 2. Sent packet ratio

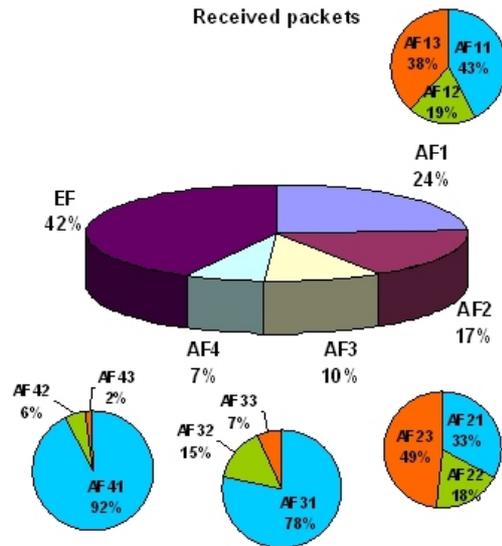


Figure 3. Received packet ratio

The whole simulation scenario is the same like in our previous investigation [6], the only difference is that we use packets with size of 500 byte instead of 1000 byte. This means that the link buffer must have a capacity (in packets) which is equal with the buffer length (in packets) in original simulation multiplied by two. Since the RED algorithm in our simulation works in “packet mode” (not in “byte mode”) we have to change the adequate RED parameters in the simulation script regarding to new packet size.

First of all we consider the queue length. The next three figures show how the queue length varies in case of the three schedulers. Currently the maximum queue length can be 100 packets, because we use packet size which is equal with the packet length in the original simulation divided by two.

All the observations (see [6] for details) which were taken in case of the original simulation are relevant to this simulation also. We can see that in case of 500 byte packet size simulation the queue length variation density is twice time bigger than in case of 1000 byte packet size simulation. This is because in case of 500 byte packet size simulation the number of generated packets (by source nodes) is twice time bigger than the number of generated packets in case of 1000 byte packet size simulation.

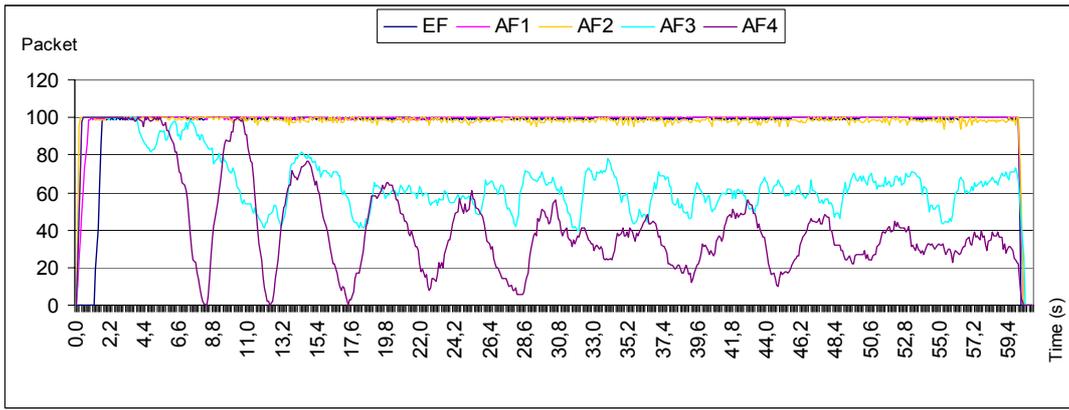


Figure 4. Queue length in case of PRI scheduler

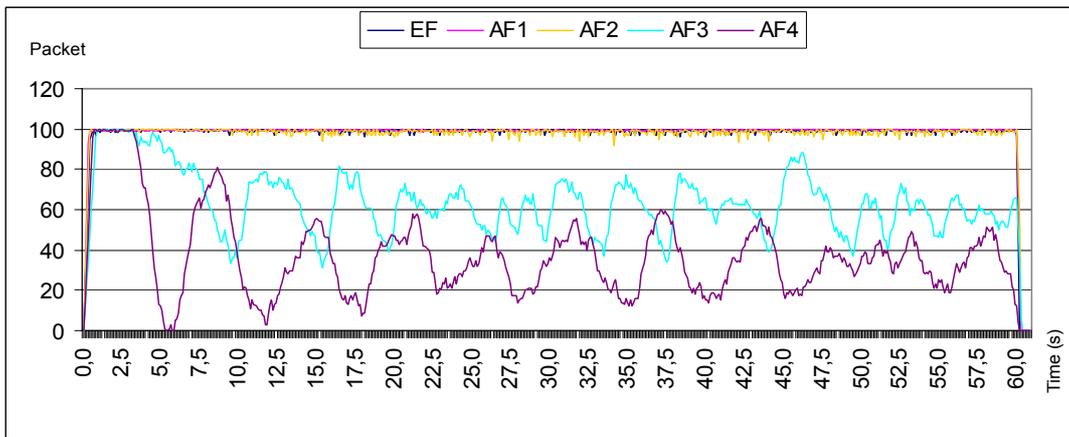


Figure 5. Queue length in case of WRR scheduler

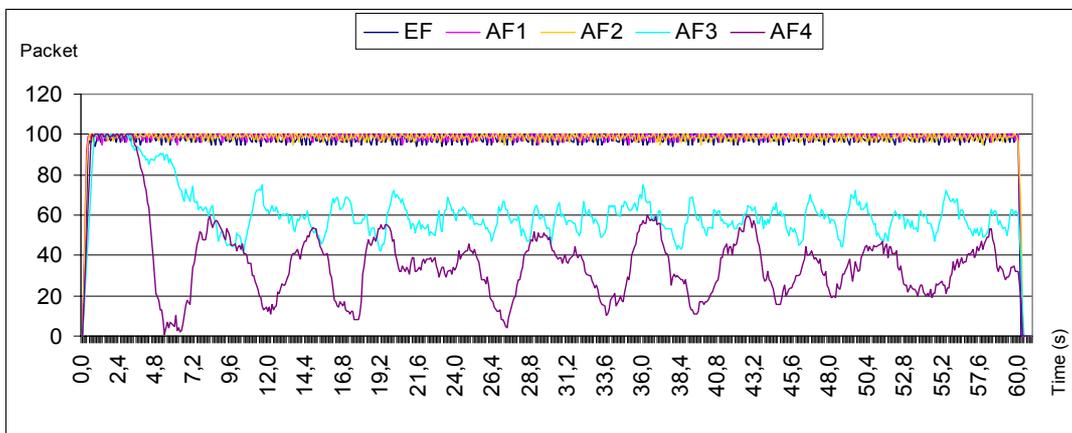
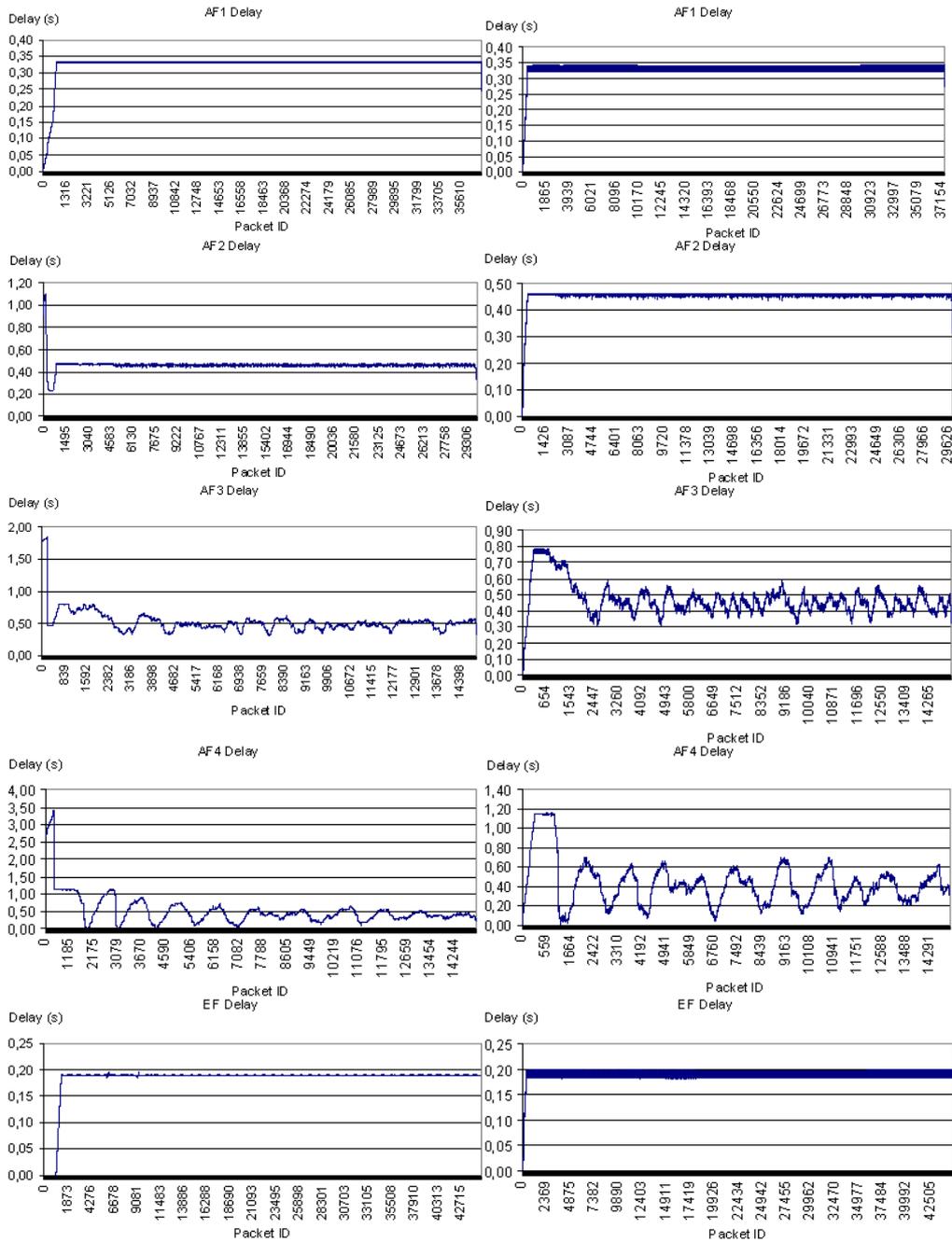


Figure 6. Queue length in case of WRR scheduler

Figure 7 shows the delay variation of packets. The delay varies exactly as the queue size varies, conform to the well known Little-formula ( $Q = \lambda * W$ ). This

means that the delay variation density is also twice time bigger than in the case of original (1000 byte packet size) simulation [6].



a.) Delay in case of PRI scheduler

b.) Delay in case of WRR scheduler

Figure 7. Delay of packets

Because of page number limitations we only present the delay of packets in case of PRI and WRR

scheduler, but the WRR scheduler also holds the above criteria.

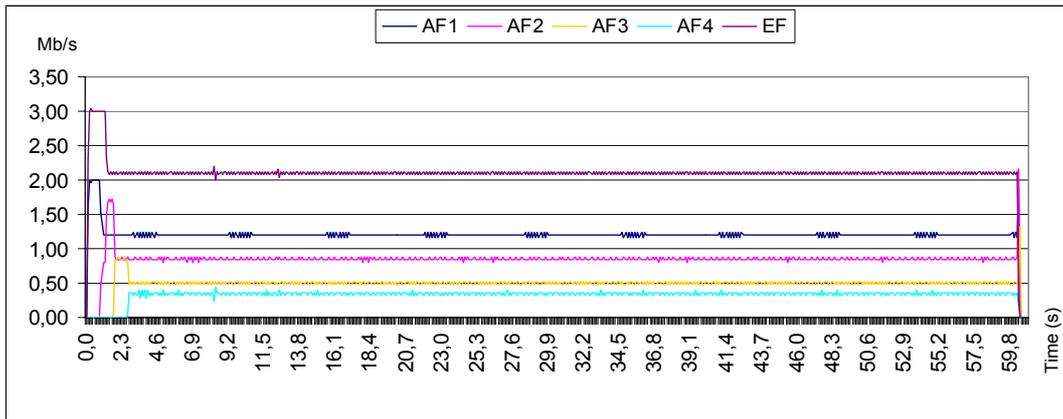


Figure 8. Throughput in case of PRI scheduler

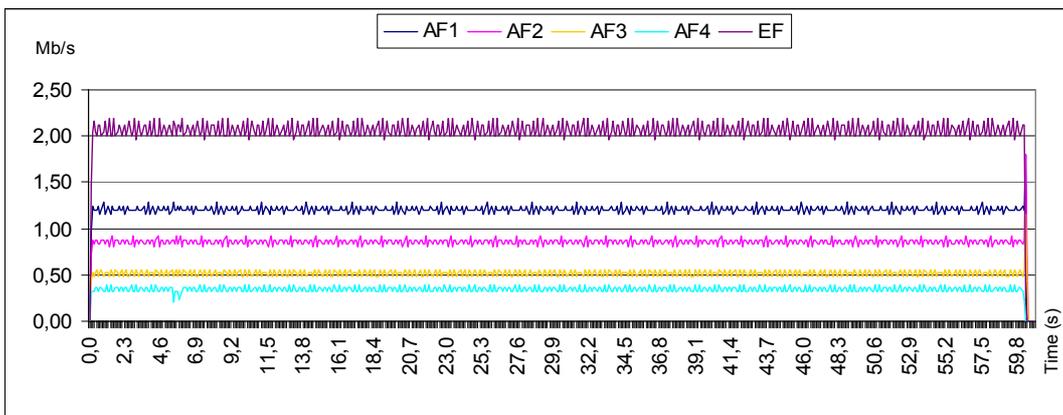


Figure 9. Throughput in case of WIRR scheduler

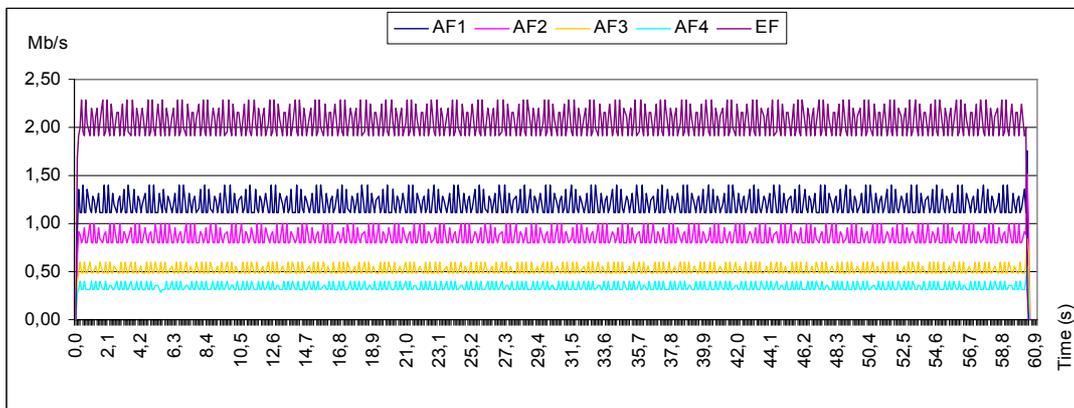


Figure 10. Throughput in case of WRR scheduler

Figure 8, 9, 10 show the throughput variation, which has the same characteristic like queue size (or delay), namely that the variation density is bigger (twice time) than in the case of 1000 byte packet size simulation.

Similar to the original simulation, the average realized throughput per class is the same for all schedulers, but the deviation (jitter) from the mean is the smallest in case of PRI scheduler and the greatest in case of WRR (while WIRR is between them).

The next figures show a comparison between the original (1000 byte) and the actual (500 byte) simulation in terms of arithmetic mean delay per class in case of the three schedulers.

It can be observed that the packet size changing does not have significant effect to the average delay.

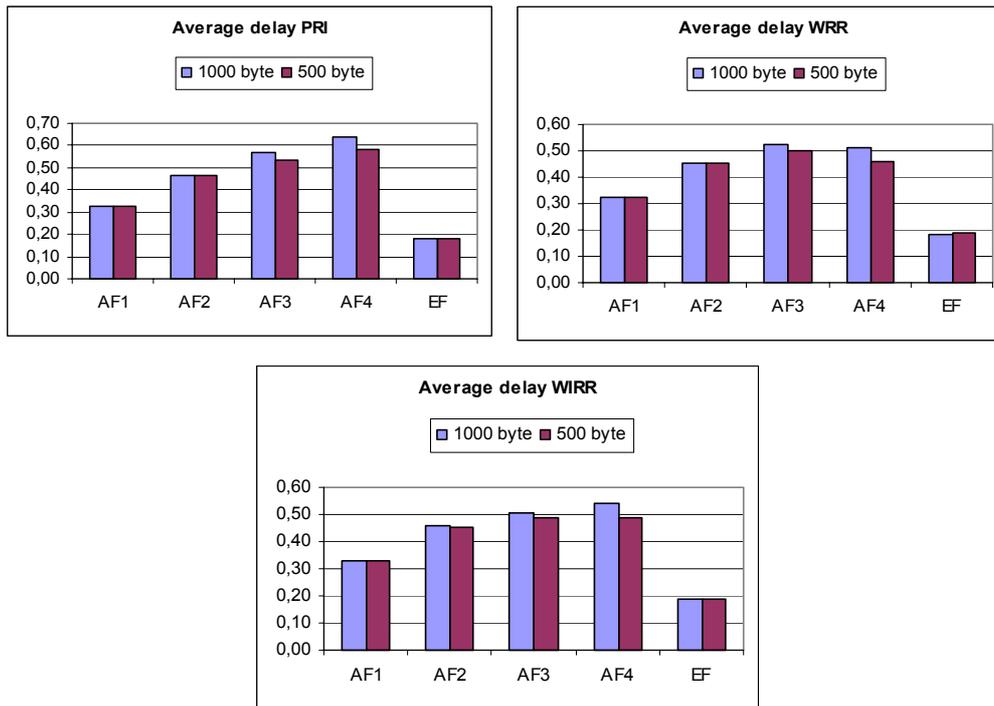


Figure 11. Arithmetic mean delays

### 3. CONCLUSIONS

A performance comparison was made between the traditional traffic scheduling algorithms in a simple dumbbell Diffserv topology. The novelty of the paper is these comparisons since to the best knowledge of the authors in the earlier papers only individual schedulers were analyzed. We enhanced our earlier paper [6] and we investigated how the performance measures vary if we change the packet size. Future extensions of this work will include the validation of our results with an analytical (Markovian) model.

### REFERENCES

- [1] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, "An Architecture for Differentiated Services", *IETF RFC 2475*, December 1998.
- [2] J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, "Assured Forwarding PHB Group", *IETF RFC 2597*, June 1999.
- [3] B. Davie, A. Charny, J. Bennett, K. Benson, J. Boudec, W. Courtney, S. Davari, "An Expedited Forwarding PHB", *IETF RFC 3246*, March 2002.
- [4] B. Braden, D. Clark, B. Davie, S. Floyd, V. Jacobson, J. Wroclawski, L. Zhang, "Recommendations on Queue Management and Congestion Avoidance in the Internet", *IETF RFC 2309*, April 1998.
- [5] S. Floyd, V. Jacobson, "Random Early Detection gateways for Congestion Avoidance", *IEEE/ACM Transactions on Networking*, V.1 N.4, August 1993, pp. 397-413.
- [6] M. Lengyel, J. Sztrik, "Simulation of Differentiated Services in Network Simulator", *Annales Universitatis*

*Scientiarum Budapestinensis de Rolando Eötvös Nominatae, Sectio Computatorica*, 2003.

- [7] K. Fall, K. Varadhan, "The ns Manual", available via <http://www.isi.edu/nsnam/ns/ns-documentation.html>, April 2002.
- [8] R. Jain, "The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation and Modelling", *John Wiley and Sons, Inc.*, New York, 1991.

### AUTHOR BIOGRAPHIES



**MIKLOS LENGYEL** received M.Sc. degree in Computer Science from the University of Debrecen, Debrecen, Hungary, where he is currently corresponding Ph.D. student. His research interests include Quality of Service architectures, Differentiated Services model and performance evaluation in computer networks. His Web-page can be found at <http://www.inf.unideb.hu/~mlengyel>.



**JANOS SZTRIK** is a full professor of Faculty of Informatics at the University of Debrecen. His research interests include stochastic modelling, reliability theory and queueing theory. He has published more than 90 journal articles and refereed conference papers

in these areas. He is a Doctor of the Hungarian Academy of Sciences. Dr. Sztrik has served and continues to serve on the executive and technical program committees of many international conferences. He is head (or project leader) of different research groups and member (or formal member) of european mathematical and computer societies. His Web-page can be found at <http://irh.inf.unideb.hu/user/jsztrik>.

# **SIMULATION-BASED SYNTHESIS OF COMPOSITE**

## **DISPATCHING RULES**

**Martin Schickmair**

**Martin Graml**

PROFACTOR Produktionsforschungs GmbH

Im Stadtgut A2

4407 Steyr-Gleink · Austria

E-Mail: martin.schickmair@profactor.at

**Christoph Pichler**

ATENSOR Engineering and Technology

Systems GmbH & CoKG

Im Stadtgut A2

4407 Steyr-Gleink · Austria

E-Mail: christoph.pichler@atensor.com

**Walter Laure**

Infineon Technologies Austria AG

Siemensstraße 2

9500 Villach · Austria

E-Mail: walter.laure@infineon.com

### **KEYWORDS**

Dispatching, scheduling, simulation, optimisation, lead time, model, semiconductor manufacturing, wafer test, fab, eM-Plant.

### **ABSTRACT**

This paper describes the simulation-based synthesis of optimum dispatching rules for the wafer test area of a semiconductor production plant. The target of the optimisation task is the minimum average lead time under the constraint of maintaining on-time delivery. Dispatching is a very fast and robust technique for sequence-planning even when cycle times are not perfectly controlled. A broad range of standard dispatching rules can be found in the literature, each with its own strengths and weaknesses. Less information is available on the use of combinations of these rules. An analysis of the properties that influence the lead times and delivery dates of the lots in the facility shows that composite dispatching rules need to be used to meet the specified target.

In this paper we describe how simulation is used to explore the design space of composite dispatching rules. Based on a simulation model of the entire wafer test area, we have tested various combinations of dispatching rules

under realistic operating conditions. The optimum rule combination thus found leads to an average lead time reduction of 15% while maintaining, or improving, on-time delivery.

### **INTRODUCTION**

The design and fabrication of an integrated circuit is a complicated endeavor (Boning 1991). Modern semiconductor technology uses fabrication process sequences that consist of several hundred process steps. An integrated circuit (IC) is created from a raw semiconductor wafer by sequentially applying a specified processing sequence to the wafer. Wafers are moved through the fabrication facility (fab) in lots of 20 to 100 units. In general, more than one processing station is qualified to perform a given processing step in a product's processing sequence. Product lots are routed from one processing station to the next according to station availability. Process flows are re-entrant, i. e., a lot may pass through the same group of processing stations up to more than 20 times according to the process technology used.

At the end of the processing sequence, which may take several weeks to complete, basic device functionality is verified at the parameter control measurement (PCM) step. Finally the ICs are electrically tested against product specifications on the test floor. Those circuits that do not

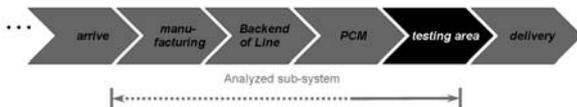
pass the test are flagged and the tested wafer is shipped to the customer. Due to the large number of ICs per wafer and the complexity of the circuitry, the final testing of an entire product lot may take more than a week.

The development of new fabrication processes and new product generations continuously pushes the limits of knowledge of the physics involved. Fabrication facilities and equipment involve large investments that are amortized by fabricating a large number of different product families at a given facility. Fabrication facilities, in general, are not purpose-built for a specific product. Rather, product and process technologies are developed for existing facilities.

The facility considered in this paper comprises more than 1000 processing and testing stations and handles between 500 and 1000 different products and several hundred process flows in several process technologies simultaneously. Total output is about 30000 wafers per week. Work-in-progress inventory (WIP) is a significant cost factor, as is equipment utilisation. Demand is extremely volatile both in total volume and in the mix of the products. An effective dispatching strategy is crucial for keeping WIP low, meeting delivery schedules, and maintaining responsiveness to order fluctuations. The average run time of product lots through the facility is an excellent indicator of how well these criteria are met.

## SITUATION ANALYSIS

In Figure 1 the complete process that is performed in a production line is shown. Beginning with the arriving of a job in the plant it makes its way through the manufacturing area and the backend of line to the parameter control measurement (PCM).



**Figure 1:** Principle process flow: from silicon to a finished wafer

When a lot passes the PCM it is carried to the wafer test area (WTA). The emphasis of this work is on the WTA at the end of the production cycle; this area is therefore discussed in more detail.

The WTA consists of more than 100 testers. Dependent on the product type of the specific lot it can be processed on one to about 30 different testers. This number of variability is called the “flexibility” of the lot.

Whenever a tester finishes processing the current lot a green light on top of the machine indicates to the operator (skilled worker in the WTA) that it is ready for the next lot. The operator then uses a terminal placed in the vicinity of the tester to determine which of the lots that are cur-

rently waiting can be processed on this specific machine. He is presented with a sorted list of lots and he is supposed to take the one on the top. The sorting of the lots is done by the dispatch software based on specific rules (dispatching rules).

Based on these rules the sequence of the lots through the WTA is decided. This has enormous effect on the average lead times and the adherence to delivery dates.

As we will show, a broad range of rules and combination of rules are principally applicable, each with its specific strengths and weaknesses. To identify which rules contribute most to the given objective is quite a complex and delicate problem. Thus the problem was solved using a detailed simulation model of the testing area as test-bed for assessing and optimizing different rules.

## OBJECTIVES

The overall objective of our project was to identify, develop and optimise new dispatching rules for the testing area. Target of the optimisation process was to reduce the average lead times of the lots through the WTA under the constraint that the adherence to delivery dates has to be at least the same as today.

The lead time of a lot (time span from entering the WTA until leaving it) is the sum of all waiting times before processing and all physical processing times. It is obvious that the physical processing times cannot be reduced by dispatching rules, so only waiting times can be affected by. In fact dispatching rules may change lot sequences which yield in shorter waiting times.

Most time several lots are waiting for the same resource, respectively a tester. These lots differ in their properties, for example:

- their flexibility (on how many testers in the whole WTA can it theoretically be processed).
- their process time (the time it takes to perform all the necessary tests).
- their delivery date – some lots have to be delivered earlier than others.
- set-up: Does the specific machine need to be set-up, or is the lot tested before of the same type as one of the currently waiting ones?

Because of these lot-specific properties the sequence of lots does affect the average lead time and the adherence to delivery date. Therefore an ideal dispatching rule has to consider all this information in the right way. How the properties influence the objective and how the “right” balance can be found is shown in the following sections.

## APPROACH AND ASSUMPTIONS

Now we know that our rule consists of the right combination of the lot-specific properties; dynamic, discrete com-

puter simulation is the right way to meet this challenge. Why?

- We have to regard that the lots in the WTA affect each other - lots are fighting for the same resource at the same time – the WTA is a very dynamic system.
- Static analysis has already been shown to be insufficient to meet our goals.
- Similarly without any reliable test-bed human intelligence and expertise turns out to be insufficient to deal with that amount of complexity.
- The product mix of the fab (which lots are in the WTA present at the same time) needs to be considered for finding the rule – exact quantification of the effects can only be done with simulation.
- The real process is not available for experiments as it operates 24 hours 7 days a week. Any application of dispatching rules that are not fully tested would bear too much risk.

As we need only *relative* comparisons of different dispatching rules, we can reduce the effort for data collecting and modelling by allowing some simplifications and abstractions between real world and the model. It is not the aim of the project to predict absolute production values as well funded predictions of future orders are not available anyway. It is expected that rules showing better performance with past data will also promise a performance increase for future orders. The model will be based on the following assumptions:

- The availability of the testers is not affected by dispatching rules, so we don't have to consider machine break-downs in our model.
- The same applies to transportation times of the lots from one machine to another – in the simulation model the transport times are zero.
- In reality the plant is operated by a limited number of machine operators. So if several machines need to be set-up at the same time additional waiting times may occur due to resource bottlenecks. In our perfect simulation-world we ignore these operator based bottlenecks and assume to have an unlimited amount of operators available. The relative comparison is nevertheless exact.

The following chapter describes how the model was implemented.

## THE SIMULATION MODEL

The model was built using the discrete event simulator eM-Plant™ in version 7.0. EM-Plant was developed as SIMPLE++ (Simulation in Production Logistics and Engineering programmed in C++) by Tecnomatix, Stuttgart (Germany). Today eM-Plant is a standard software in the

automotive industry for object-oriented graphic and integrated modeling, simulation and optimisation. Complex systems and business processes can be represented in a realistic way. The advantages of conventional concepts such as modules, language and lists are integrated in eM-Plant. EM-Plant is a single simulation system for production, logistics and engineering; these are the reasons for choosing it.

Following some details about using eM-Plant for our problem are described. Our model consists of entities, modules, tables and methods. Entities represent the lots travelling through the plant. All characteristics of a lot are assigned to the entity using custom attributes. For the intermediate buffers between the machines the standard module “storage” from eM-Plant is used; the “station” module is used to model the testers.

The main task of model implementation was to reproduce the correct flow of the lots through the wafer test area, which is controlled by the specific dispatching rules. Each lot is attributed with a product type (over 1000 different products) and the number of wafers within the lot (between 1 and 50). Lots are generated by an order source randomly based on distributions retrieved from historical production data. Each lot generated waits in a storage until a tester becomes available and the dispatcher clears the lot for processing. The wafers are tested for a determined time which usually varies according to the product type. If no further inspection is needed the lot exits the testing area and leaves for shipping.

In our model the interval for order arrival is based on an exponential distribution. The values for the product type and the number of wafers are sampled from empirical distributions based on data analysis of the real WTA. Controlled by these distributions the entities (lots) are created and their attributes are assigned. Typical attributes are product type, entering-time, number of wafers and a table of the process sequence specified for the specific product type (see window in figure 2). The entries for a process step in the sequence table are process time, the code for this process step and a table of suitable working stations (testers).

After assigning the basic attributes the entity is sent to a storage, where an entrance control calls a user defined method (methods see figure 2 on the left side). This method writes the type of the entity and some other relevant information into a custom attribute at each station where the specific lot can be processed. This attribute is called the “virtual-buffer” and is realised as a table containing all lots that can be processed on this specific station. Doing so for each station the “virtual-buffer” contains all actual lots that are waiting for processing.

Another method searches for all potential testers (see figure 2 middle) waiting for a lot and moves the lot the specific station.

Whenever a station finishes processing another method is called which chooses the next lot from its virtual buffer based on the actual dispatching rules.

The tested lot is moved to the exit method as the new one starts processing. In the exit method all relevant statistics are written to internal tables. At the end of a simulation experiment the collected data are written to Microsoft-Excel for further data processing.

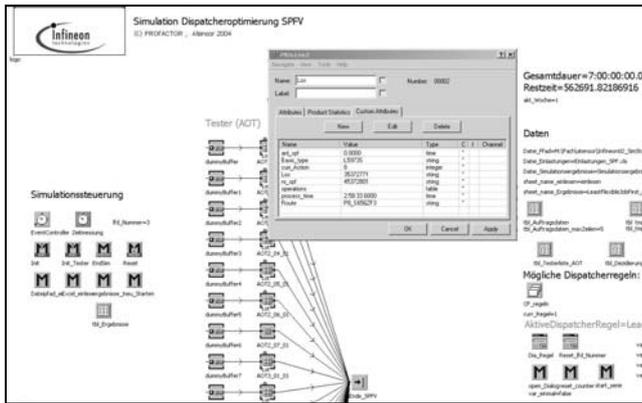


Figure 2: Screenshot: Part of the simulation-model

To execute series of simulation runs we have implemented an auto-experimenter. This module refreshes experiment-dependent variables between every simulation run (experiment) and writes the results via ActiveX in a predefined Excel table, from which several charts and experiment summaries are automatically generated. Statistics of interest are for example minimal-, maximal- and average door-to-door time, maximal- and average WIP, utilisation for each station, number of tested lots per station, maximal numbers of entries in the virtual buffer per station or calculation time per run.

Based on the simulation results we developed and optimised the new dispatching rules for choosing the next lots from the virtual buffer.

**RESULTS**

First we want to show the possibly not so obvious positive effects that the consideration of the processing time of a lot in the dispatching rules can have. Figure 3 shows a scenario where two lots are competing for a tester. Let us assume that lot 1 has a processing time of 1 time unit, lot 2 of 10 time units. In this simple case two sequences are possible: If we process lot 2 first and then lot 1 the sum-lead time of the both lots calculates to 21 time units. For the second possible sequence – first lot 1 and then lot 2 the sum is only 12 time units. The difference is 43% with the higher value as base! This example shows clearly that the optimised dispatching rule has to regard among other parameters the processing times in that way that the lot with the shortest test time should be processed first (Shortest-

ProcessingTimeFirst – SPTF). Another and probably the most important positive effect of SPTF is that the average number of lots waiting in the WTA is reduced (see also (Rose 2001)).

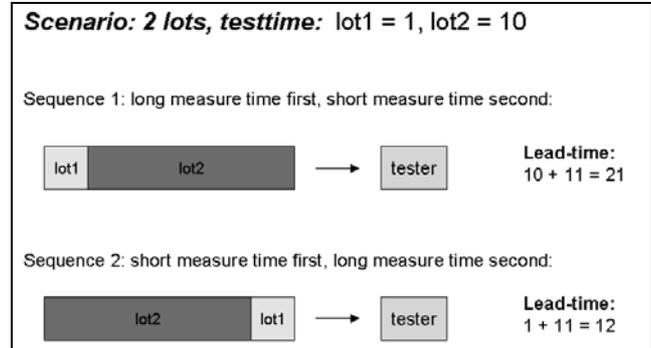


Figure 3: Effects of ShortestProcessingTimeFirst (SPTF)

But unfortunately the situation is not that easy: Without considering other lot-specific properties SPTF performs rather poor. It does not optimise the adherence to delivery dates and leads to a very unbalanced tester-utilisation that in turn results in increasing waiting times.

As already mentioned, the combination of the lot-specific properties to a single rule is the key for success. This proofed to be true after a large number of simulation experiments. Based on heuristic optimisation and far reaching discussions with operators and process experts we achieved our aim. A weighted linear combination of the lot-specific properties provided the best results.

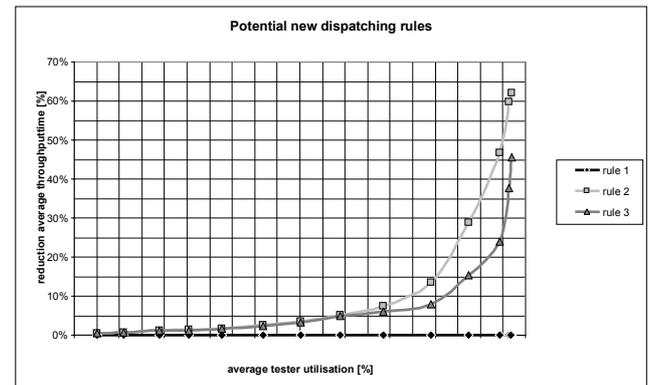


Figure 4: Lead time reduction with optimised dispatching rules

In figure 4 we show three graphs, each representing a specific dispatching rule. On Y-axis we have put the lead-time reduction in percent, on X-axis the average tester utilisation. Unfortunately but understandably we are not allowed to present the exact dispatching rules and values of utilisation. But the effects are clear: Compared to the

“old” rules currently implemented the optimised rules perform better, the higher the average tester utilisation is. The reason for this behaviour is found in the fact, that an increasing plant utilisation leads to a greater average number of waiting lots in the intermediate buffers and therefore to an increased spectrum of possible solutions for choosing an “ideal” lot. For every rule the average lead-times grow with increasing values of utilisation, but using the optimised rules with a lower gradient. The positive effects of the optimised dispatching rules for the real WTA can be summarized as following:

#### Lead time:

- at the average tester-utilisation of the real WTA a reduction of about 15% is achieved!
- the relative reduction increases heavily with higher tester utilisations (see Figure 4)

#### Adherence to delivery

- guaranteed at least equal to today’s values
- in average better because of shorter lead times

#### Average number of lots in WTA (WIP)

- 15% less WIP caused by 15% shorter lead times

#### Investment necessary to implement

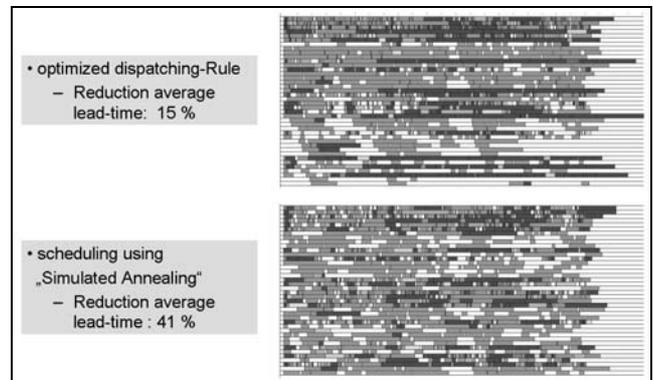
- no invest in new hardware necessary
- it is estimated that the implementation of the new dispatching rules in the WTA will need an effort of about one personnel-day

However, one interesting question is still not answered so far: How near to the theoretical optimum perform the new dispatching rules? Or the other way round, how much potential is still available for better rules?

To answer this question we have made the following estimation: We took a representative lot-arrival-sequence to the WTA with a duration of three weeks and fed it into our simulation model. During simulation execution the sequence of the lots through the WTA were logged using the old dispatching rules. Then the new dispatching rules and a heuristic optimisation algorithm based on Simulated Annealing were used to optimise the order sequence in order to reduce the average lead time to a minimum. Because of the long calculating time we allowed the heuristic algorithm to perform, it is quite sure that the found sequence is very close to the theoretical optimum. The results are shown in figure 5.

The figure shows two Gantt-Charts representing the sequence of the lots on a section of some stations. One line in the chart stands for one tester, each rectangle for a specific lot. The darker a rectangle on the chart represents the longer the waiting time of the specific lot. As expected we see that the sequence based on dispatching produces more “dark-lots” than scheduling does and it takes longer time to finish processing all orders.

The formerly developed optimised dispatching rules minimise the average lead times by 15%, the scheduling even up to 41%!



**Figure 5:** Dispatching vs. Scheduling (Simulated annealing)

Now we know that our sequences generated by our dispatching algorithms are far away from the theoretical optimum, but what are the reasons for that?

The answer is quite clear: Dispatching chooses the lot to be processed next under the number of lots that are actually waiting in front of the specific tester and has to decide its solution based on general rules considering only local information. Simulated Annealing does not decide on specific rules, but it tests a vast number of sequences and takes the best it finds. Furthermore the scheduling algorithm was programmed and optimised to reach the theoretical optimum in a way to consider the lots that will arrive in future too. Dispatching cannot use this information because it is impossible to predict exact arriving times of lots in the real WTA.

On the other hand in the real WTA the scheduling algorithm cannot be implemented because of the limitations found in real processes like machine break-downs, unknown or not exactly known process times and long calculation times. Anyway the result is a good estimation on how good our new dispatching rules perform.

Finally the project-team could convince the plant manger responsible for the WTA to implement the new dispatching rules in the real facility.

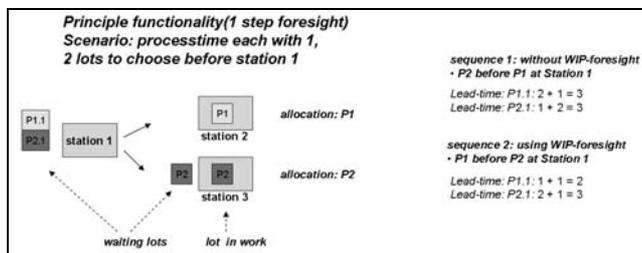
## CONCLUSION AND FUTURE WORK

Dispatching is a very fast and robust technique for sequence-planning. A broad range of dispatching rules can be found in literature, each with its own strengths and weaknesses.

The most attractive rule for a given problem has always to be customized to the special facility and the given objectives and requirements. Our project has shown that all lot- and facility-properties that influence the lead-time and the

adherence to delivery have to be combined in the right way to reach the theoretical optimum as near as possible. Human intelligence and expertise is insufficient for dealing with that amount of complexity. Even simple combinations of rules have surprising effects. It has been shown that computer-simulation is a very comfortable, efficient and maybe the only tool for finding the optimal solutions for a real facility. The return on investment of the presented project is excellent.

The next logical step for the project team is to optimise the dispatching rules beyond the wafer test area, for the whole fab. To find the optimum for this problem it seems to be necessary to include one more dynamic property of the facility, namely the actual number of waiting lots on the downstream stations (see figure 6).



**Figure 6:** Lead time reduction with optimised dispatching rules

The figure above shows a simple scenario of two lots waiting before station 1. Lot “P1.1” has successor “Station 2” and “Lot 1.2” the “Station 3”. If we proceed lot “P1.1” the average lead times are shorter than otherwise. This effect has to be considered by the dispatching rules.

First simulation experiments have shown that finding the optimal combination is very critical but not impossible. The project-team is confident to reach the goal again.

## REFERENCES

- Atherton, L.F. and R.W. Atherton. 1995. Wafer Fabrication: Factory Performance and Analysis. Kluwer.
- Atherton, R.W.; F.T. Turner; L.F. Atherton; and M.A. Pool. 1990. “Performance Analysis of Multi-Process Semiconductor Manufacturing Equipment.” In Proceedings of the IEEE/SEMI Advanced Semiconductor Conference 1990.
- Banks, J. (ed): Handbook of Simulation – Principles, Methodology, Advances, Applications, and Practice; John Wiley and Sons, 1998; ISBN 0-471-13403-1.
- Benecke, C.: „Simulation von Materialfluss- und Lagerprozessen“. Zeitschrift für Logistik, Heft: 5, 1990, S. 30 - 32
- Boning, D. S. 1991. Semiconductor Process Design: Representation, Tools, and Methodologies. PhD thesis, Massachusetts Institute of Technology, 1991.
- Domschke W. Armin Scholl, and Stefan Voß. Produktionsplanung. Springer, 1997.
- Fowler, J. and J. Robinson. 1995. Measurement and improvement of manufacturing capacities (MIMAC): Final re-

port. Technical Report 95062861A-TR, SEMATECH, Austin, TX.

- Kosturiak, J., Gregor, M.: Simulation von Produktionssystemen. Springer Verlag, 1995
- McKiddie, R., Brown, S., and Neacy, E. 1994. Predicting the Impact of Short-Term Increases in Wafer Starts on a Constant Start-Rate Semiconductor Factory: Applications of SEMATECH's Future Factory Analysis Methodology. SEMATECH Technology Transfer 9402223A-XFR..
- Pichler, C. 1997. Integrated Semiconductor Technology Analysis. Österreichischer Kunst- und Kulturverlag.
- Rose, O. 2001. The Shortest Processing Time First (SPTF) Dispatch Rule and Some Variants in Semiconductor Manufacturing. In Proceedings of the 2001 Winter Simulation Conference, pp. 1220-1224.
- Zeichen, G., Fürst K.: Automatisierte Industrieprozesse. Springer Verlag, Wien, New York 2000

## AUTHOR BIOGRAPHIES

**MARTIN SCHICKMAIR**, born in Wels, Austria studied electrical engineering at the Technical University Vienna and received a degree of a Dipl.-Ing. He joined the simulation based design and optimisation department at Profactor Produktionsforschungs GmbH in 1999. His special interests are the integrated simulation of technical and business-processes as well as simulation based optimisation of production logistics.

**MARTIN GRAML**, born in Linz, Austria studied Mechatronics at the Johannes Kepler University Linz. He is with Profactor Produktionsforschungs GmbH since 2003. His special interests are in the field of simulation based optimisation and software development.

**CHRISTOPH PICHLER** heads ATENSOR's Holistic Engineering department, focusing on the optimum design and operation of production facilities in the automotive and electronics sectors. Before joining ATENSOR, he was with National Semiconductor Corporation in Santa Clara, where he led the company's concurrent engineering efforts group. He holds a Ph.D. degree from Vienna University of Technology and an M.B.A. from INSEAD. His professional interests include the organisation and operation of manufacturing companies and the deployment and integration of digital methods across the enterprise.

**WALTER LAURE** joined Infineon Technologies as a Software Project Engineer in 1996. For four years he has been responsible for line simulation at Infineon Technologies. His career started with Alcatel as a software developer in the High Speed Network area. He has a diploma in Technical Mathematics awarded by the Technical University in Graz, Austria. His professional interests include wafer fab simulation, line logistics and developing and integrating software in semiconductor wafer fabs.

# **SIMULATION IN LOGISTICS, TRAFFIC AND TRANSPORT**



# DOUBLE-SIMULATION METHODOLOGY TO DETERMINE THE NUMBER OF SERVERS IN SITUATIONS WITH PEAKS

Javier Otamendi  
Manuel Cano Espinosa  
TYPESA – Técnica y Proyectos, S.A.  
Plaza del Liceo, 3  
28043 Madrid, Spain  
E-mail: [jotamendi@typsa.es](mailto:jotamendi@typsa.es)

## KEYWORDS

Traffic simulation, queueing systems, system design, flows with peaks, discrete event modelling.

## ABSTRACT

A simulation model that represents a multichannel situation with a time-dependent input flow distribution is presented. The objective of the study of the system is to assess its performance under different configurations of the service stations needed to handle the peaks. A methodology is developed based on the execution of a simulation model in two steps. First, to look for the satisficing alternatives, a set of constant flow simulations is run. Second, to obtain the optimum alternative, the model is run with real flows. The presentation is based on standard queueing theory terminology and includes an example of the design of a paytoll booth in a highway in which the incoming traffic is highly seasonal.

## INTRODUCTION

A system is a combination of units that look for service. An airplane (unit) trying to take-off in a given runway (service station), a vehicle trying to refuel at a gas station, a customer trying to make a deposit... They are all examples of real systems, in which if the service station does not have enough capacity, queues are formed with the corresponding hassle for the customers. Therefore, when designing the system (Rubinstein 1986), there must be a compromise between cost (or built capacity) and customer service (waiting times in queues).

The decision is specially difficult when the arrival of units is not smooth or constant but following peaks that depend on time. More planes take off early in the morning or late in the afternoon than in the middle hours of the day, more vehicles stop at a gas station early morning or late afternoon, banks usually have a more stable arrival of customers...

The system that is the focus of this article is a paytoll station in the highway at the entrance of a big city. The arrival of vehicles is clearly seasonal, with tremendous peaks almost every Sunday afternoon and the day of the return from a regional holiday. Two or three-hour

peaks in which the hourly arrival rate is sometimes 100 times more than on a week day. A decision has to be taken about the number of booths to install to cover most of the incoming flow but without overdimensioning the system (Gross and Harris 1985).

Queueing models are analytical models that are used to analyse and quantify the performance of a service station. They are analytical models that, for specific combinations of input flow, service times, number of servers and queue disciplines, estimate the queue lengths and times in the system (Taha 1987). The requisite, however, is that the capacity installed must be enough to handle all the incoming flow.

Even if that condition holds, not all the situations might be represented with analytical models (for example, moving from queue to queue, not constant incoming service rate) and other tools that provide for a good abstraction of the system have to be used.

Simulation models are descriptive models that might reflect complicated queue disciplines and any distribution for the input parameters, especially when they vary with time.

A combination of queueing theory and simulation modelling is proposed to analyze the seasonal incoming flow at the service stations and to design the system to reach a satisfactory compromise between cost and customer service.

In order to determine the capacity of the system to be installed, it is necessary to follow a reasonable methodology that searches through the available alternatives and selects the satisficing one. This methodology has to be based on the fact that both queueing and simulation models are descriptive models and not optimization models, since they are not used to optimize a given situation but to describe it. That means that, for a given set of input parameters, the model is executed and the performance estimated. Since the objective of the study is to search for a satisficing alternative, then, the only way is to vary the values of the input parameters and estimate the objective function for each combination, with a

posterior step to compare and select the alternative with the best values for the set of input parameters.

This search process is depicted in the following figure, which includes a four-step procedure:

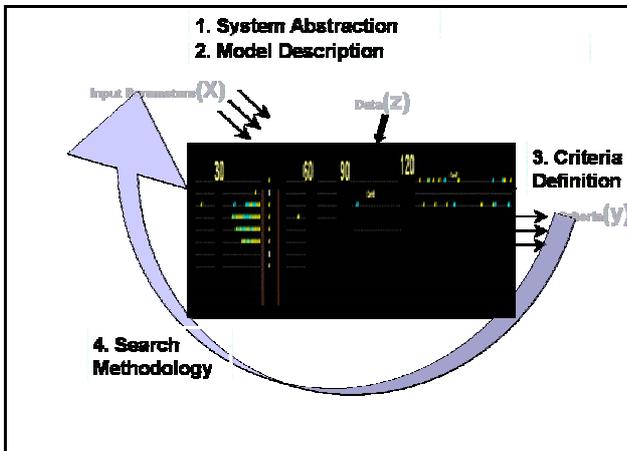


Figure 1. Search Process

The first and second steps are used to describe the actual situation, including the possible input variables. The third is to develop a suitable objective function to compare and select among alternatives. The final step is the iteration process in which values for the input parameters are changed.

### DESCRIPTION OF THE SYSTEM/MODEL

Let us use this section to describe the system under consideration. It is just a multichannel service station in which customers arrive and ask for service.

A common way used to describe a queueing system is Kendall's notation in which six parameters must be specified (Taha 1987):

$$(a/b/c):(d/e/f)$$

where

- a: input flow distribution with average  $\lambda$
- b: service time distribution with average  $\mu$
- c: number of service stations
- d: queue discipline (FIFO, LIFO, ...)
- e: available queue length
- f: maximum number of arrivals.

### Input Flow

The main characteristic of peaks is that the input flow varies with time. There are periods in which the number of customers that arrive is significantly greater than the average as to determine that the distribution is not the same along the time horizon of the study.

Let us define, then, a distribution function  $f(x)$ , measured in arrivals per unit time, and whose expected value is  $\lambda$ . Its cumulative function is  $F(x)$ . Figure 2 shows how the real input flows along time might be converted into a probability distribution function.

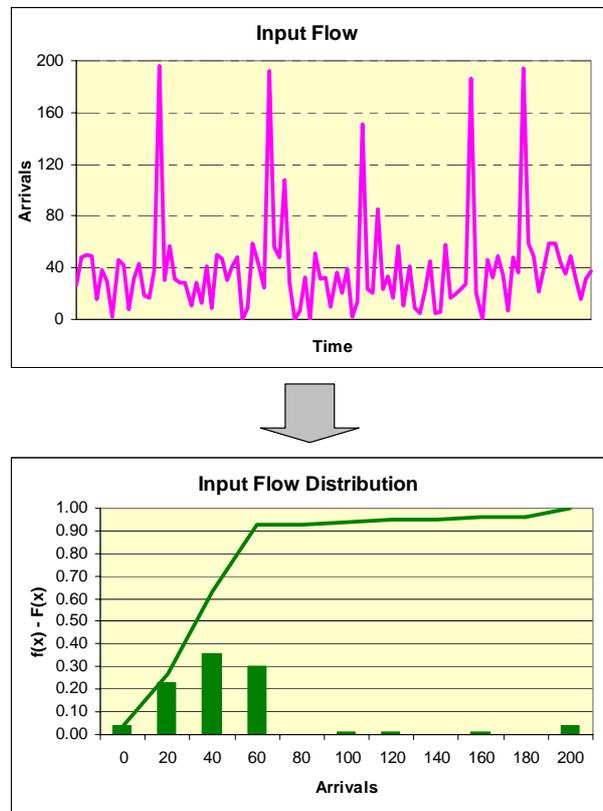


Figure 2: From Real Flows to Input Flow Distribution

The average is in this case meaningless since it is not constant with time. What matters is the maximum number that can arrive per unit time.

### Service Time

Let us define throughput as the number of customers that a single station can serve per unit time. This value is usually not dependent on time, since the service rate is the same regardless of the number of customers waiting or the number of servers available. The distribution is usually multimodal, since the customers can be differentiated in types, each with a given distribution. Let's call this distribution  $f(y)$ , its expected value  $\mu$  and its cumulative distribution  $F(y)$ . Figure 3 shows an example of a distribution of this type:

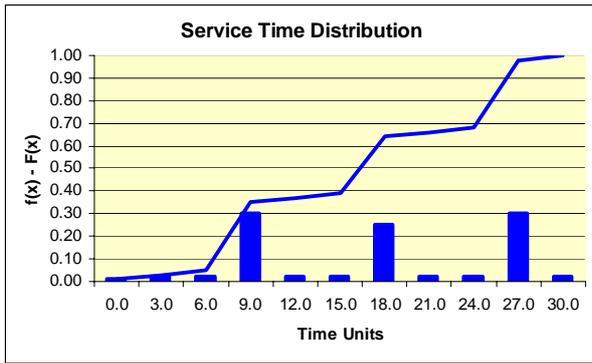


Figure 3: Service Time Distribution

### Service Stations

The service stations are set in parallel, each with its own queue. The space in front of the stations is arbitrarily long. The number of service stations are  $S$ , the quantity that needs to be calculated. Notice that usually this number has an upper bound,  $S'$ , which depends on the space available and on budget.

The number of servers directly determines the capacity of the system, CAP, or the average number of customers that the whole set of stations can serve per unit time, which is calculated as:

$$CAP = \mu * S$$

It is also the maximum input flow that the system can handle without the queues growing infinitely.

Figure 4 shows this relationship in graphical form:

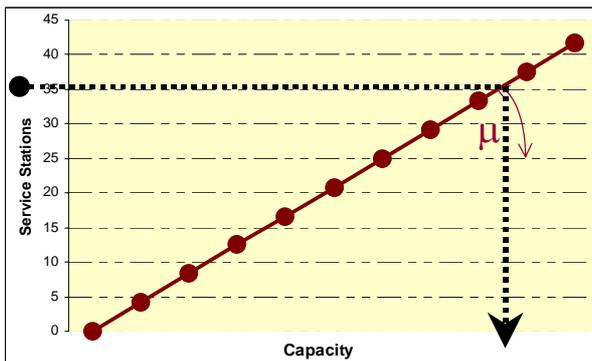


Figure 4. Capacity as a Function of  $S$

For any given  $S$ , capacity can be calculated. The slope of the line depends on the value of the mean service time  $\mu$ .

### Queue Discipline

The customers are served on a first-come first-served basis in each of the servers. However, changing queues is allowed. The vehicles cannot leave the system.

### Available queue length

It is the available space for queues to be formed. In the situation under study, there is no limit since the incoming flow waits for service even in the highway.

### Size of Source

It is the total number of possible clients that access the service stations. In the situation in hand, the total number of input customers is known in an hourly basis.

## MEASURES OF PERFORMANCE

In this section, the most important measures of performance are defined, quantified and related mathematically. Queueing models are used to calculate queue lengths and times in the system, as well as percent utilization. They are all related in terms of  $\lambda$ ,  $\mu$  and  $S$ . In this case, let us take the queue length as the driver and percent utilization as a summary measure. Also, a cost criteria and idea of flow coverage are included.

### Cost

Let's assume that the cost increases linearly with the number of servers:

$$COST = S * \$$$

where  $\$$  is the cost per service station.

### Coverage

The idea of flow coverage specifically applies to situations with peaks. Analytical queueing models usually require that the input flow is less than the capacity for the system to be stationary or stable.

In situations with peaks is economically infeasible to design the system to cover all situations. The design capacity will cover most of the income flows but not all the peak flows. Even in these busy systems, sometimes it is desired to install less capacity to control the flow intensity downstream (Huang and Huang 2002).

Let's define then coverage (COV) as the percentage of income flow that corresponds with the capacity of the system:

$$F(x=CAP) = COV$$

In Figure 5, the actual coverage, that is, the one that corresponds to the capacity of the system, is calculated using the input flow distribution.

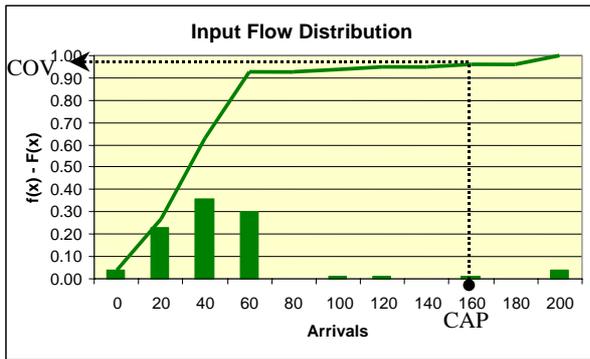


Figure 5: From Capacity to Coverage

The relationship between CAP and COV is then crucial in the design of the system. It is obvious that the higher the desired coverage the higher the capacity required. Also, it must be mentioned that when the coverage is already high, small increases in coverage require a large increase in capacity.

Since CAP depends directly on S, there is also a direct relationship between the number of servers and the coverage, which is depicted in Figure 6.

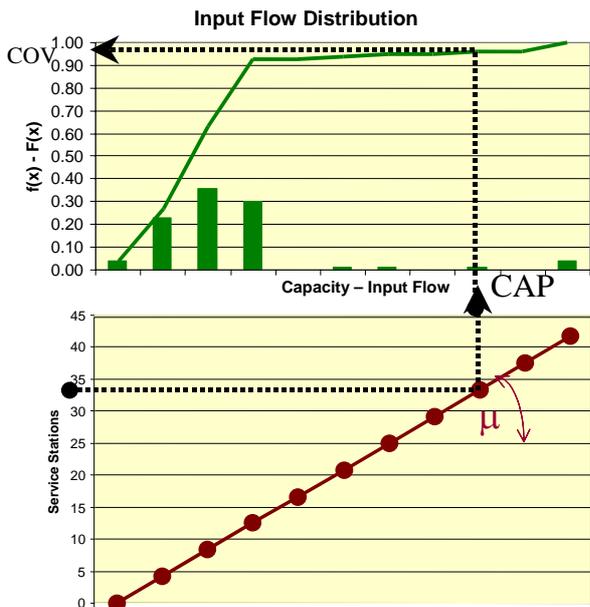


Figure 6. From Service Stations to Coverage

The number of servers sets the capacity, which is equivalent to input flow. The corresponding coverage might then be calculated.

The lack of coverage (1-COV) is also an important measure since it is the percentage of time in which customers arrive at a higher rate than the service rate and large queues will be formed. That percentage corresponds to the peaks.

## Utilization

Utilization, %UTIL, is the percentage of time the servers are attending customers. It also might be defined as the ratio between what arrives to the system and what might be served:

$$\%UTIL = \lambda / CAP$$

where  $\lambda$  has already been defined as the average input flow and CAP as the average capacity of the system.

Therefore, the utilization is to be calculated as:

$$\%UTIL = \lambda / CAP = \lambda / (\mu * s) \quad (1)$$

Then, coverage might also be defined as the percentage of time in which the queues are manageable or the percentage of time in which  $\%UTIL \leq 100\%$ .

## Queue Length

The length of the waiting queues (LQ) is going to vary significantly between the normal hours of operation and the peak hours. The relationship that must be studied is between Utilization and LQ, which will be something similar to the one shown in Figure 7. The graph has been obtained using the M/M/S theoretical queueing model (Taha 1987).

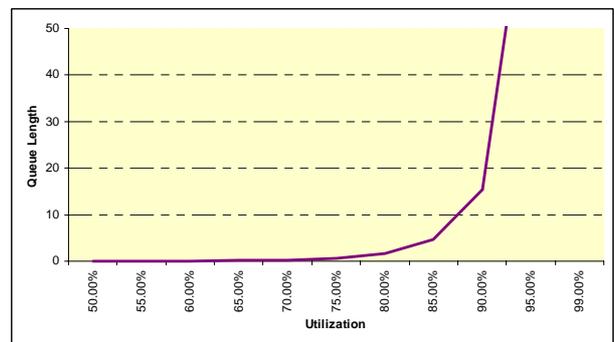


Figure 7. Relationship Between Utilization and Queue Length

If the utilization of the system is low (around 60%), the queues are non-existent, whereas if the utilization is greater than 1, the queues grow to infinity. There is a grey area, that corresponding to a %UTIL close but lower than 1, in which the queues are manageable but usually longer than desired.

The range of the grey area depends on the nature of both the input flow and the service time distributions, as well as the number of service station. An upper limit for the %OCUP must be investigated, as it will definitely influence on the design decision, since it is a function of the permissible queue length.

## SEARCH METHODOLOGY

The objective of the methodology is to calculate the optimum number of servers needed to balance the different criteria involved in the design of the system. The more servers are included, the higher the capacity, the higher the coverage, the smaller the queue length, but the higher the cost.

With the graphs included in the previous section, a trial-and-error procedure might be implemented so that for any given value of  $S$ , the performance measures are calculated, the alternatives compared against each other and one of them selected.

However, this procedure can be very tedious if the possible number of servers is too large. For that reason, what is proposed is to perform a first step to reduce the number of feasible alternatives to a manageable set.

The idea is to use the performance criteria also as restrictions by setting bounds on their values. The procedure is therefore to work on an optimization mode, and instead of calculating the performance measures for a given value of  $S$ , determine the number of servers that correspond to a particular limiting value of each criteria.

What must be a priori provided then is a set of satisficing values on each of the measures that will help do the balancing. Let's define those limits as:

- **BUDGET** = Maximum cost that might be invested in servers. Obviously, if money is not a factor, the service station could be built with as many servers as desired and no queues will ever be formed.
- **ALQ** = Maximum average queue length. It is the satisficing value for the average number of customers waiting for service.
- **S%UTIL** = Satisficing level for the utilization of the servers. It is defined as a level in which the average queue length is kept below its satisficing limit.
- **mCOV** = Minimum coverage. It is the minimum value of income flow that must be attended without forming queues.

The search methodology follows a five-step procedure that includes obtaining input data and a double simulation. The result is a range of feasible values for the number of servers needed to satisfy the utilization, the coverage, the queue length and the cost requirements.

### Step 1: Obtain data

#### 1a. Input flow distribution

Historic data of flows per unit time is obtained and converted into an absolute frequency distribution, its

corresponding relative frequency or probability distribution  $f(x)$  and the cumulative probability distribution  $F(x)$  (see Figure 1).

#### 1b. Expected value of the service time distribution, $\mu$

Historic data of time spent in the service station per unit is obtained and its average calculated.

### Step 2: Determine the admissible input flow

#### 2a. Define minimum coverage

The value of mCOV is defined subjectively.

#### 2b. Calculate the minimum coverage flow

The input flow value corresponding to that coverage value might then be calculated as:

$$F_{\text{mCOV}} = F^{-1}(\text{mCOV})$$

Figure 8 shows the calculation in graphical form:

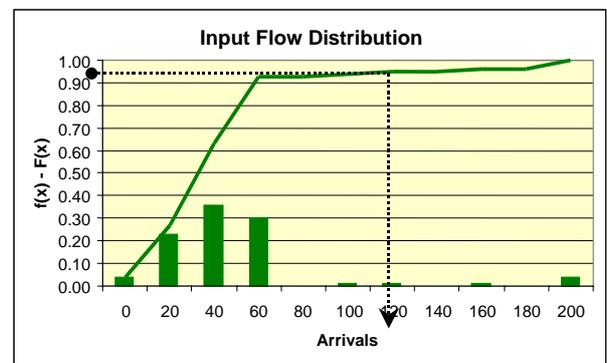


Figure 8. Calculation of coverage flow

#### 2c. Calculate the admissible input flow

The coverage flow might be denominated also as the admissible input flow  $F_{\text{ADM}}$ :

$$F_{\text{ADM}} = F_{\text{mCOV}}$$

It is then the absolute minimum flow that must be attended without forming queues.

### Step 3: Calculate the feasible range on the number of servers

A feasible range for the number of servers can now be calculated. For the upper bound,  $S'$ , the BUDGET constraint is used:

$$S' = \text{BUDGET}/\$$$

For the lower bound, 'S, the value is estimated knowing the minimum coverage and its desirable input flow (calculated in Step 2c) and the satisficing utilization, using Equation (1) as follows:

$$S = \text{sup} (F_{\text{ADM}} / (\% \text{OCUP} * \mu)) \quad (2)$$

where  $\text{sup}()$  indicates rounding to the higher integer.

Figure 9 shows the calculation of the lower bound for different values of %OCUP:

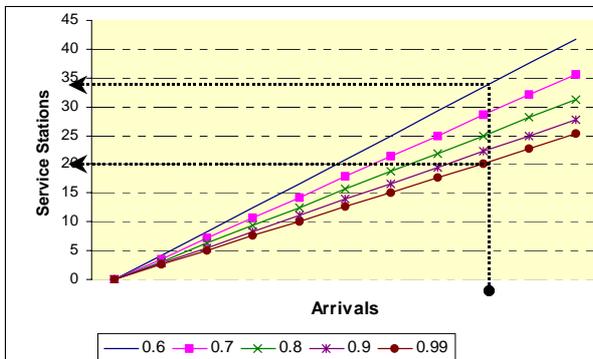


Figure 9. Calculation of 'S'

Several feasible values for S are now remaining, each providing a satisfying compromise between cost and service (queue length). Among these values, the greater the S, the better the service but the higher the cost.

**Step 4: Obtain the satisfying occupation level via constant-input simulation.**

The satisfying %OCUP has been defined as the one that corresponds with a reasonable queue length, ALQ. This S%OCUP will vary with ALQ, but also with number of servers S and the mean service time  $\mu$ .

As it has been already demonstrated, in general, if the %UTIL is close to 1, the queues will be too long, and if %UTIL is close to 0, there will be no queues. But the relationship in the rest of the range cannot be quantified easily. A value of 60% is considered as a lower bound below which no queues are developed.

Here is where simulation or queueing models come into play. If an analytical model corresponds to the situation in hand, which is not usually the case, queueing models are to be used to calculate the average queue length. If the situation cannot be analytically modelled, a simulation study can be performed. This is the case in which the input flow distribution presents heavy peaks and the service distribution is multimodal.

In this step, the simulation model is to be used to evaluate the mean queue length under varying conditions of utilization. For that reason, it is not worthwhile to execute the model using the real flow with peaks, but with a constant average input flow,  $\lambda$ . If the simulation model is run for different values of  $\lambda$  and S, but for the real values of the service distribution, a feel for the value of the average queue length can be obtained:

$$LQ = f(\%OCUP) = f(\lambda, S) \text{ for a given } \mu$$

A satisfying value of S%OCUP might then be estimated. This value will correspond to the maximum %OCUP that satisfies the ALQ requirement:

$$S\%OCUP = \max \%OCUP \mid LQ \leq ALQ$$

With this value, 'S' is to be more closely estimated.

**Step 5: Compare the feasible alternatives via real-flow simulation**

But obviously, the study cannot forget about the actual flows, that is, the evaluation of the flow with respect to time. It's then the appropriate moment to use the simulation model for the second time.

With real flows, the simulation model can be executed for the satisfying values of the variable S and the relationship between cost and service quantified. The influence of the peaks is then dynamically studied and the final decision taken accordingly.

**Graphical tools for calculating the feasible range of servers**

The sequential procedure taken in the first three steps might be easily performed with the use of a pair of simple graphs. The double-graph tool is shown in Figure 10.

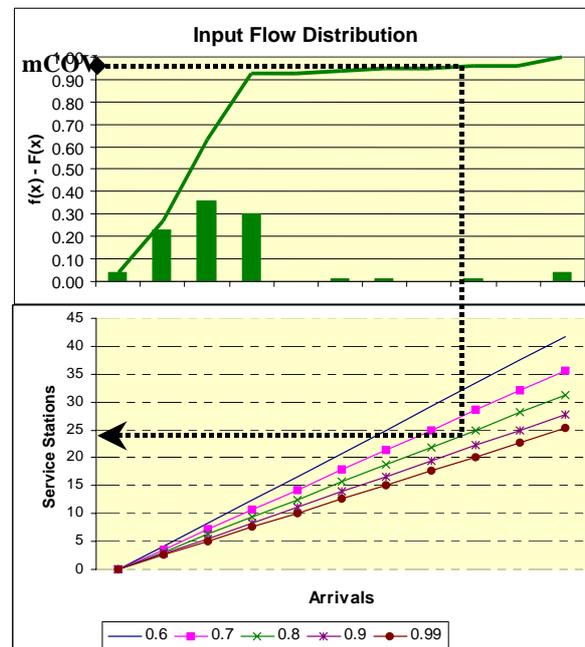


Figure10. Double-Graph Analysis

The first one uses the minimum coverage as the input value to calculate the admissible input flow (Steps 1-2). This value is used in the second graph with the calculated satisfying occupation value (Step 4) to

provide a lower bound for the number of servers (Step 3).

An equivalent single-graph tool is shown in Figure 11. The results obtained are the same, but information concerning the allowable input flow is lost.

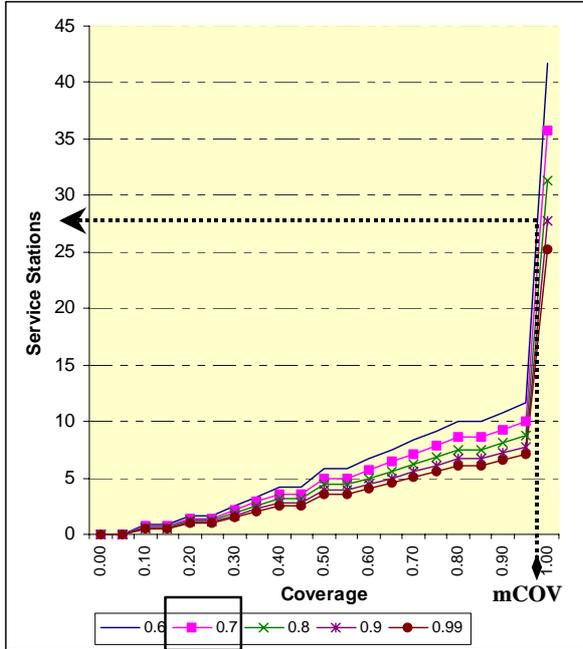


Figure 11. Single-Graph Analysis

The shape of the graph, and therefore the final decision, depends directly on the input flow distribution and the number of classes that have been used to develop its frequency distribution.

**PAYTOLL EXAMPLE**

One of the real-life systems in which significant queue lengths are formed caused by an input flow distribution with large peaks is the paytoll system in highways. The design of these systems is not easy since a large percentage of the time the input flow is not important, whereas a small percentage of time the input flow is tremendous. In a country like Spain, there exists a huge agglomeration of people going in and out of the big cities during a holiday period.

The difference in flow between a high and a low period makes it economically impossible to design the paytoll system to cover all the possible incoming flows. A design tool is necessary to balance the counteracting effects of cost and service while representing the busy situation over long periods of time.

What follows is a description of the analysis that was performed to analyze one of these paytolls, using the methodology that is the focus of this article. The input flow and service distributions have been altered to preserve proprietary information.

**Step 1a: Obtain the input flow distribution**

The flow was obtained for a period of 6 months, counting the number of vehicles that had entered the highway per hour. Figure 12 is the histogram that has been obtained after treating the data.

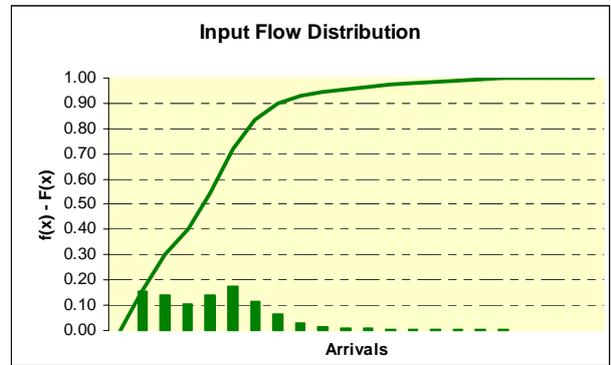


Figure 12. Input Flow Distribution

About 1/3 of the possible input values account for 90% of the data. Higher cumulative frequency is only achieved then if the arrival rate is increased significantly, showing that few but high peaks appear. In fact, the peaks correspond to Sundays when people return to the big city from the mountain or the beach.

**Step 1b: Calculate the expected value of the service time distribution,  $\mu$**

Historic data of time spent in the service station per unit is obtained and represented in Figure 13. The distribution shows two modes, corresponding to payment using credit cards or cash.

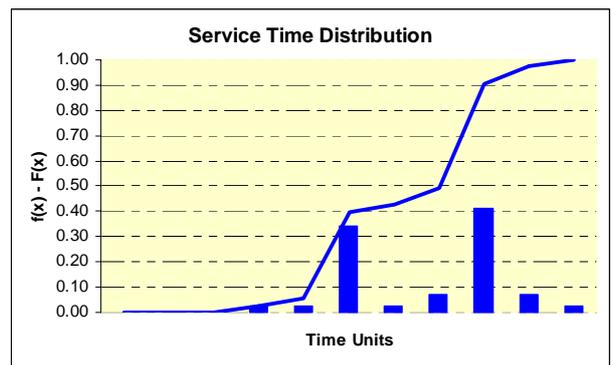


Figure 13. Service Time Distribution

**Step 2 & 3: Calculate the feasible range of number of servers**

The value of mCOV is first defined subjectively to 98%.

The use of the double-graph analytical tool is then used to determine the feasible lower bound on the number of service station. Figure 14 shows the process.

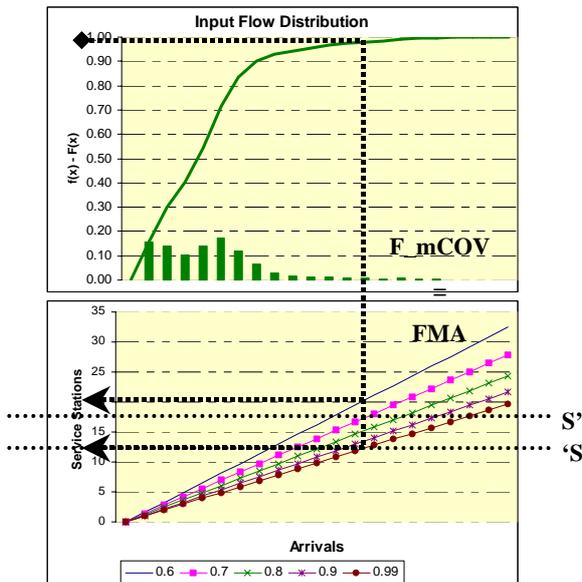


Figure 14. Calculation of the Feasible Limits

For a subjective mCOV value of 98%, the corresponding minimum coverage flow F\_mCOV is calculated, value which is also equal to the admissible flow FMA.

If no queues are to be formed (60% occupation) there is a need for 20 service stations, whereas if the occupation strived for is almost 100%, with 12 servers is enough, this level becoming the lower limit 'S'.

The budget however imposes an upper limit  $S' = 17$  service stations. Therefore,

$$12 \leq S \leq 17$$

**Step 4: Obtain the satisficing occupation level via constant-input simulation.**

The aim of this important step is to determine the admissible queue length for the admissible coverage flow.

A set of constant flow simulation runs is performed, varying the number of servers in the feasible range between 12 and 17 and the occupation level between 60% and 100% in increments of 10%, measuring in each case the average queue length. Figure 15 summarizes the results.

Maximum average queue length					
Servers	Occupation				
	60%	70%	80%	90%	100%
12	2	6	8	9	Saturation
13	2	5	7	10	
14	2	4	7	10	
15	2	4	6	11	
16	2	4	6	11	
17	2	3	6	11	

Figure 15. Queue length versus Occupation

The results show that saturation is important once the occupation is above 90% and that the queue length decreases slightly as the number of servers increases, except for an occupation of 90%. The reason in this last case is that the total number of vehicles or admissible input flow becomes too large (heavy traffic conditions) when the number of servers grows (Huang and Huang 2002).

Management then determined also that 8 cars had to be the acceptable maximum and that 6 was probably the target, fixing therefore the occupation level between 70% and 80%.

Translating these results into the single-graph tool (Figure 16), the range of the feasible number of servers can be narrowed.

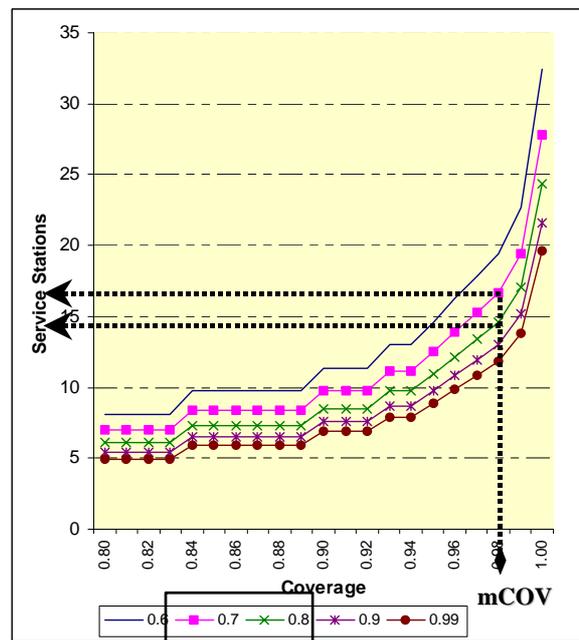


Figure 16. Calculation of the Tight Feasible Limits

The new feasible range is between 15 and 17 service stations:

$$15 \leq S \leq 17$$

### Step 5: Compare the feasible alternatives via real-flow simulation

The simulation model with real flows over a six-month period is run for the three available alternatives, measuring queue length.

For the alternative with 15 service stations, there are 17 days in which the system is collapsed when the users return from holidays (Figure 17). Almost every Sunday the queue length is larger than the desired level. The rest of the days, the queue is basically non-existent.

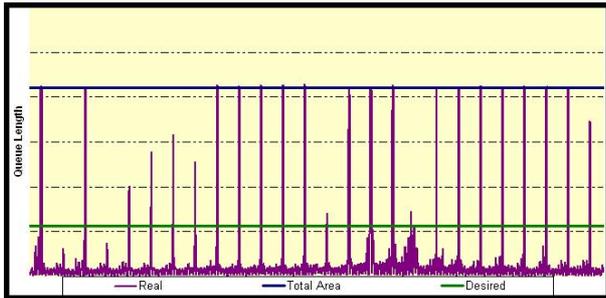


Figure 17. Simulation with Real Traffic and  $S = 15$

In Figures 17, 18 and 19, the total available pay toll area is also included to show that in certain days the queues that are formed go beyond the pay toll area and even collapse the highway.

For the alternative with 16 service stations (Figure 18), the performance is obviously improved, with only 10 days collapsing the highway.

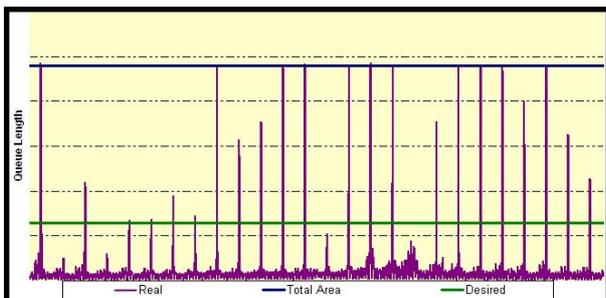


Figure 18. Simulation with Real Traffic and  $S = 16$

Finally, with 17 service stations (Figure 19), only 1 day shows problems and 16 Sundays are above the desired level.

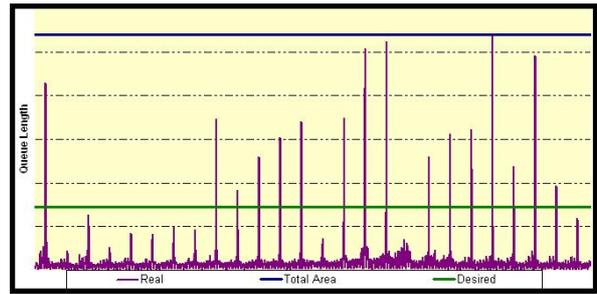


Figure 19. Simulation with Real Traffic and  $S = 17$

At this point, management decides to eliminate the alternative with 15 servers. Besides the sub par performance, it presents side effects like blockage at some entry points into the highway several kilometres back.

The final decision was then between an increase in performance (17 booths) and the save in money (16 booths).

### SUMMARY AND CONCLUSIONS

A methodology is presented to design systems that suffer short periods of very high peaks based on both queueing theory and simulation modelling. The proposed methodology follows a five-step procedure that includes obtaining the input flow and the service time distributions, a simulation model that is run both with constant and real flows and a double-graph summary tool.

This double use of simulation has been proven very successful since the time spent in running the model was very small compared to the time spent in collecting the data and developing the model. Once the model was validated, it was used for a thorough experimentation period that led to a strong-based decision.

A real-life example of a pay toll system in a highway is presented to validate the methodology. The limits on the number of booths are set quickly and the simulation runs help to make the final decision.

Therefore, the combination of experimentation via simulation models with quantitative queueing analysis has been proven as an interesting tool to treat situations with peaks that incorporates both management input and intensive collection of data.

### REFERENCES

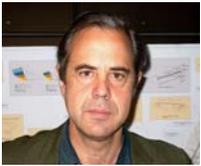
- Gross, D. and C. Harris. 1985. *Fundamentals of Queueing Theory*. John Wiley and Sons, New York.
- Huang, D. and W. Huang. 2002. "The effects of tollbooths on highway traffic." *Physica A* 312, No.3-4, 587-608.
- Rubinstein, R. 1986. *Monte Carlo Optimization, Simulation, and Sensitivity of Queueing Networks*. John Wiley and Sons, New York.

Taha, H.A. 1987. *Operations Research*. Macmillan, New York.

#### **AUTHOR BIOGRAPHIES**



**JAVIER OTAMENDI** received the B.S. and M.S. degrees in Industrial Engineering at Oklahoma State University, where he developed his interests in Simulation and Total Quality Management. Back in his home country of Spain, he received a B.S. in Business Administration and a Ph.D. in Industrial Engineering. He is currently a simulation and statistics consultant and university professor. His e-mail address is: [jotamendi@typsa.es](mailto:jotamendi@typsa.es)



**MANUEL CANO ESPINOSA** received the Civil Engineering degree from the Universidad Politécnica de Madrid in 1983. He has worked as a supervising director in road and highway design projects, both in Spain and in Europe and Latin America. He is now the Project Manager of the Road Construction Division at the consulting firm Técnica y Proyectos S.A. (TYPESA). His e-mail address is: [mcano@typsa.es](mailto:mcano@typsa.es)

# PROCESS SIMULATION APPROACH TO DESIGN AND EVALUATION OF TOLL PLAZA WITH ETC GATES

Teruaki Ito  
Universty of Tokushima  
2-1 Minami-Josanjima  
Tokushima 770-8506, Japan  
E-mail: ito@me.tokushima-u.ac.jp

## KEYWORDS

Process simulation, ETC system, toll plaza design, gate management, model verification

## ABSTRACT

The study proposes a process simulation-based approach to traffic jam issues in toll plaza of expressways with ETC gates. First, the paper describes overview of ETC system around the world as well as in Japan, and clarifies the issue of traffic jams, which occurs in toll plaza of Japanese expressways and prohibits the progress of ETC system. Then the paper describes the process simulation model which has been developed in this study, and also shows its internal procedures to simulate the traffic jams. The result of simulation provides two kinds of solution to traffic jam issues. One is to determine the appropriate time in gate change for combination gates. The other one is to study the layout redesign of toll plaza to achieve more efficient performance. Using the result of simulation, feasibility of the approach will be discussed. Since the goal of the study is to propose a practical solution to the traffic jam issues which prohibits the progress of ETC system, this paper also shows the results of simulation of a case study and discusses what the simulation-based approach can provide.

## INTRODUCTION

Electrical Toll Collection (ETC) system was started at Tokyo Metropolitan Expressway in the year of 2001, which was the beginning of ETC systems in Japan. Since then, the number of toll plaza equipped with ETC gates has been increased gradually and steadily, by additional construction of ETC gates or updating conventional toll gates to ETC gates. By the end of fiscal 2003, ECT gate will be available at every 1,300 toll plazas of expressways around the country. As a part of Intelligent Transportation System (ITS) in Japan, ETC system is expected to provide several great advantages; improvement of usability for vehicle drivers, drastic counter-measures to reduce traffic jams, a solution to environmental issues such as air pollution or noise pollution by reducing traffic jams, a crucial technologies to realize integrated toll collection system, an important infrastructure for smart intersections to enhance regional activities, and so on.

As one of the promising advantage, ETC system is expected to make great contribution to drastically reduce traffic jams over toll expressways, which are mainly derived from the conventional toll collection at toll gates, where each vehicle is required to stop to pay toll fees. According to the survey, 30% of traffic jams in expressway is occurred around the toll plaza, which means that non-stop toll payment of ETC system can drastically eliminate the traffic jams. However, advantage of ETC has not been fully utilized so far for several reasons. If vehicle drivers do not encounter any traffic jams in a less traffic volume condition, ETC vehicles can enjoy the advantage of non-stop driving to go through toll gates. Since some numbers of toll gates are assigned specific to ETC vehicles in each toll plaza, the number of available toll gates for non-ETC vehicles has been reduced. However, relatively decreased number of non-ETC toll gates does not give any disadvantages over non-ETC vehicles, although non-ETC vehicles remain the same as before the introduction of ETC system. In the mean time, if the traffic volume becomes higher, the situation changes worse. On one hand, ETC vehicles need to go through heavy traffic jams towards one of the available ETC gates at the end of toll plaza. On the other hand, less number of available toll gates for non-ETC vehicles makes traffic jams worse. Furthermore, non-ETC vehicle drivers are discouraged by the fact that an open way to ETC gate is not available for non-ETC vehicles. As a result, overall total travel time for both ETC/non-ETC vehicles around toll plaza becomes longer because of the introduction of ETC systems.

Simulation-based approach has been applied to various areas related to transportation systems and its feasibilities have been reported (Abbas-Turki et al., 2001; Fernandes et al., 1998; Gale et al., 2002; Krajzewicz, et al., 2002; Lucjan et al., 1999). Based on a simulation-based analysis of traffic jams at toll plaza with ETC gates (Horiguchi et al., 2000), the study focuses on two issues to consider counter-measures to fully utilize the advantages of ETC system; Estimation of gate change timing for combination gate and redesign of toll plaza. For one thing, most of the toll plaza had already been designed and operated before the introduction of ETC system. Some gates were replaced with ETC gate, or some ETC gates were

additionally installed in toll plaza, both of which were based on the conventional layout design of the plaza. Considering the non-conventional flow of traffic which is generated in combination with ETC and non-ETC vehicles, various kind of layout design of toll plaza could be studied using a simulation-based analysis (Ito, 2004; Ito and Hiramoto, 2004). Because of the high construction cost, however, redesign of toll plaza is not always possible. Therefore, some gate converted to be used as combination gate for both ETC and non-ETC vehicles, where gate change timing is very critical but it is not always easy to determine the most appropriate time for gate change in actual situations. The simulation model also helps to determine the appropriate time of gate change.

Outlining the ETC systems around the world as well as in Japan, the paper describes the process simulation model which has been developed in this study, and also shows its internal procedures to simulate the traffic jams. The result of simulation provides two kinds of solution to traffic jam issues. One is to determine the appropriate time in gate change for combination gates. The other one is to study the layout redesign of toll plaza to achieve more efficient performance. Using the result of simulation, feasibility of the approach will be discussed. Since the goal of the study is to propose a practical solution (Biacandi et al., 2000; Schwentke, 2000) to the traffic jam issues which prohibits the progress of ETC system, this paper also shows the results of simulation of a case study and discusses what the simulation-based approach can provide.

## **ETC SYSTEMS IN THE WORLD**

More than 30 countries have introduced electrical toll collection system practically or tentatively, in most case, to collect a flat fee. In Europe, ETC was started in Norway in 1987, followed by Italy, France, Spain, and Portuguese in 90's. Germany made an experimental trial of ETC between 1994 and 1995 to collect fees from large vehicles. Since the introduction of ETC in Texas in 1989, most of the toll ways have been equipped with ETC system in US, which is one of the most ETC prevailing countries in the world. More than 70% of vehicles in commuting time enjoy the benefit of ETC in New York area, sweeping away the traffic jams. Canada and Mexico are also introducing ETC systems just like US. ETC was introduced in Hong Kong in 1993, followed by Malaysia in 1995.

In some countries, ETC system is not compatible each other among neighboring countries, which gives an obstacle to prevail ETC system. International standardization may need to be considered in the nearest future. As for Japan, the history and its current situation has been described in the previous section, to collect non-flat fees.

Electrical Toll Collection (ETC) system was started at Tokyo Metropolitan Expressway in the year of 2001, which was the beginning of ETC systems in Japan. Since then, the number of toll plaza equipped with ETC gates has been increased gradually and steadily, by additional construction of ETC gates or updating conventional toll gates to ETC gates. By the end of fiscal 2003, ETC gates have become available at every 1,300 toll plazas of expressways around the country. The average penetration rate of ETC vehicles in Japan is gradually increasing and now reaches to 12.5% as of January 2004. However, low penetration ratio of ETC vehicles is raising several critical issues to be solved as pointed out in the introduction section.

## **PROCESS SIMULATION MODEL FOR TOLL PLAZA WITH ETC GATE**

This section describes process simulation model and its internal process for traffic jams at expressway toll plaza with ETC gates simulation. The model in this study has been developed based on these modules.

### **Simulation model development**

Using the following 6 modules based on several software modules used in this study, 6 procedures are defined for toll fee payment by vehicles at toll plaza as follows.

**CREATE module:** to generate entities (objects for simulation) based on a predefined schedule or adequate intervals.

**PROCESS module:** to represent major processes in simulation, utilizing resources.

**DECIDE module:** to determine appropriate decision in the model.

**ASSIGN module:** to set appropriate values into system parameters such as entity properties and types.

**RECORD module:** to collect statistical values.

**DISPOSE module:** to represent the final point of simulation. Statistical values are recorded before the destruction of entities.

(1) Generation of vehicles to toll gate: CREATE module generates two types of entities, or general and ETC. ASSIGN module defines toll collection time according to the time and type in each entity generation.

(2) Lane selection: DECIDE module branches the way towards which each entity take based on arbitrary probability, entity type, traffic conditions, etc.

(3) Travel on lane: PROCESS module gives travel time to each entity based on the traffic condition

(4) Gate selection: DECIDE module branches the gate to each entity based on the traffic condition.

(5) Toll collection: Entity is captured by resources based on the predefined time for payment in PROCESS module, the entity is released.

(6) Leave gate: After time stamp by RECORD module, entity is destroyed by dispose module.

### Definition of parameters and internal procedure in the model

Figure 1 shows a sample of process simulation model for toll plaza with ETC gate with 3 gates for general vehicles and 1 gate for ETC vehicles, which represents a typical toll plaza of expressway in Japan. The letter A represents generation point for entity, the letter B represents expansion point of drive lane, the letter C represents gate point. Queue.i is defined as waiting queue between A-B points, and Queue.j is defined as waiting queue between B-C points.

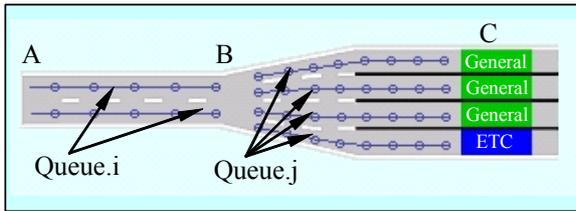


Figure 1: A Process Simulation Model

Number of incoming vehicles per hour is defined as Scheme (1) for non-ETC vehicles and Scheme (2) for ETC vehicles, respectively. ETC mixing ration in overall vehicles is defined as Scheme (3).

$$t_g = \frac{3600}{x_g} \quad (1)$$

$$t_e = \frac{3600}{x_e} \quad (2)$$

$$p = \frac{x_e}{x_g + x_e} \quad (3)$$

$t_g$  : time interval for incoming non-ETC vehicles [sec]

$x_g$  : incoming hourly volume of non-ETC vehicles [no/hr]

$t_e$  : time interval for incoming ETC vehicles [sec]

$x_e$  : incoming hourly volume of ETC vehicles [no/hr]

$p$  : ETC ratio in all vehicles [%]

Entity moves from point A to point C on the queue lines at a certain time interval as defined in Scheme (4).

$$T_d = [L - (q_i + q_j)l - L_s]S_c + L_s S_s \quad (4)$$

$T_d$  : time delay for move [sec]

$L$  : distance between start and goal (A-C) [m]

$q_i$  : number of vehicles in Queue.i

$q_j$  : number of vehicles in Queue.j

$l$  : ave. distance between two vehicles in queue [m]

$L_s$  : distance from speed-down and halt points [m]

$S_c$  : ave. time for unit distance [sec/m]

$S_s$  : ave. time for unit distance at speed-down [sec/m]

$T_g$  and  $T_e$  represent required time from the queue-end to toll gate for non-ETC and ETC vehicles, respectively as shown in Scheme (5) and (6).

$$T_g = \frac{q_{ig} t_{pg}}{m} + q_j t_{pg} \quad (5)$$

$$T_e = \frac{q_{ie} t_{pe}}{m} + q_j t_{pe} \quad (6)$$

$T_g$  : time from queue-end to toll gate (non-ETC) [sec]

$q_{ig}$  : total number of non-ETC in Queue.i (i=1,2)

$t_{pg}$  : time to pay at toll gate (non-ETC) [sec]

$T_e$  : time from queue-end to toll gate (ETC) [sec]

$t_{pe}$  : time to pay toll gate (ETC) [sec]

$m$  : number of gate for non-ETC vehicles

The model was developed using process simulation software ARENA (Kelton, R. et al. 1998).

### PROCESS SIMULATION AND ITS VERIFICATION BASED ON MATHEMATICAL DATA

To analyze traffic jams at toll plaza with ETC gates, some comparative simulation was conducted using the process simulation model. Feasibility of the model is also studied using layout design consideration around the toll gate plaza. In each simulation, entity generation is started at 500m before the toll gate with 2 driving lanes in Queue.i, with 8 entities in each Queue.j towards 4 gates to drive through. Triangular distribution to calculate required time at toll payment is used for ETC as (3, 4, 5) and non-ETC as (14, 16, 18), respectively. Total simulation time is 1 hour.

Table 1: Conditions in 3 Cases for Simulation

Case	The gate type to be used
Case-1	General gate*4
Case-2	General gate*3 ETC gate*1
Case-3	General gate*3 Combined use gate*1

### Basic model

Table 1 shows 3 types case studies in this simulation. All 4 gates are installed with conventional non-ETC in

Case-1, 1 out of the 4 is converted to ETC in Case-2, and the ETC gate is modified to combined gate which can be used for both ETC and non-ETC in Case-3. 10% of ETC penetration rate was used in Case-2 and Case-3. A non-ETC vehicle takes the lane in the shortest queue length when traffic jam is occurred, whereas an ETC vehicle takes the closest lane to ETC gate.

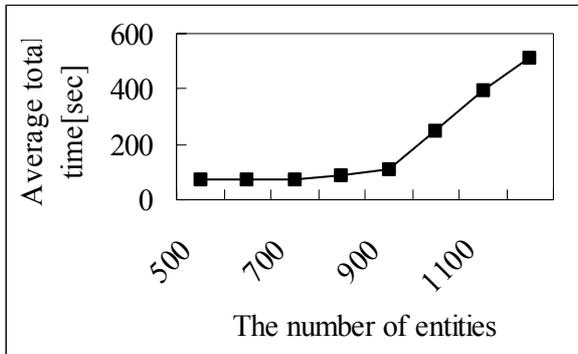


Figure 2: Gate-through Time in each traffic volume

Figure 2 shows the required time to pass the toll gate, or drive-through time in Case-1 according to the hourly traffic volume ranging from 500 to 1200 vehicles per hour. The time stays around 100 seconds up to 900 vehicles, but it drastically increases when the number of incoming vehicles exceeds over 1000, which means that the capacity of toll gates in this simulation is about 900 vehicles per hour.

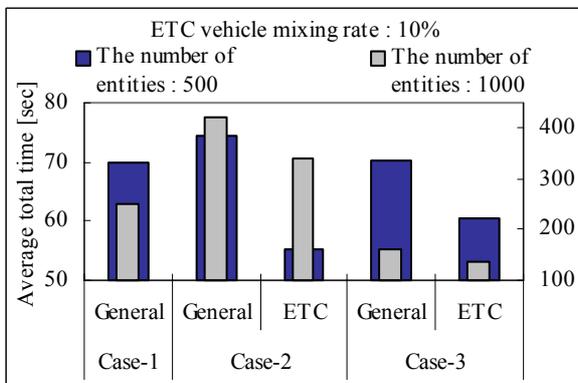


Figure 3: Comparison of Gate-through time for ETC/non-ETC vehicles in Each Case

Introduction of ETC system can provide a solution to solve the problem of traffic jam at toll gates. Figure 3 shows the comparison of drive-through time at toll gates when the hourly traffic volume is 500/1000 vehicles. The drive-through time of ETC vehicle in Case-2 becomes by-far less than that in Case-1, whereas drive-through time of non-ETC vehicle becomes longer. In Case-3, or as a result of combination gate management, drive-through time of non-ETC vehicle becomes the same level of Case-1, although ETC-vehicle takes more time than that in the dedicated ETC toll gate in Case-2. When the traffic

volume becomes up to 1000 as shown in the right side-bar scale in Figure 7, traffic jams in Case-2 cannot be taken care of without the management of combination gate in Case-3. The model well represents the current situation of ETC gate traffic as a general case.

### Estimation of appropriate time in combination gate management

To estimate the appropriate time of gate change in combination gate management between ETC dedicated use and combination use, Case-2 and Case-3 are compared.

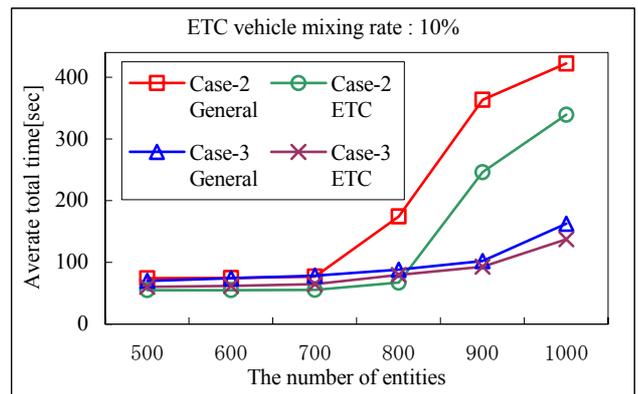


Figure 4: Comparison of Gate-through Time in each Traffic Volume

Figure 4 shows the comparison of average time for the number of entities. Both cases are almost identical if the hourly traffic volume is 800 or less, and the ETC gate provides the highest performance in all of the 4 gates. However, when the traffic volume is increased up to 800 per hour or more, average total time in Case-2 is increased for both ETC/non-ETC vehicles, which means that around 800 would be the capacity of the gates. The results suggests that when the traffic volume reaches this level, it would be the appropriate time of gate change from ETC dedicated use to combination-gate in order to avoid traffic jams, and maintain the high performance of ETC gate.

### Study on layout design in toll plaza to achieve higher performance

Management of combination gate may give a solution to traffic jams, however, redesign of toll plaza is sometimes required for consideration to achieve better performance. Process simulation model in this study can also be used to consider and to study those designs. Figure 9 shows 2 types of simple layout examples in different designs, in which Layout-2 is modified on the side of ETC gate to make wider lane before the ETC gate, whereas Layout-1 is identical to the layout used in the basic model as shown in Figure 4. Figure 10 shows the comparison of total average time to pass the toll plaza. As show in Figure 10, performance of

Layout-2 is better than that of Layout-1 when the hourly traffic volume is over 800.

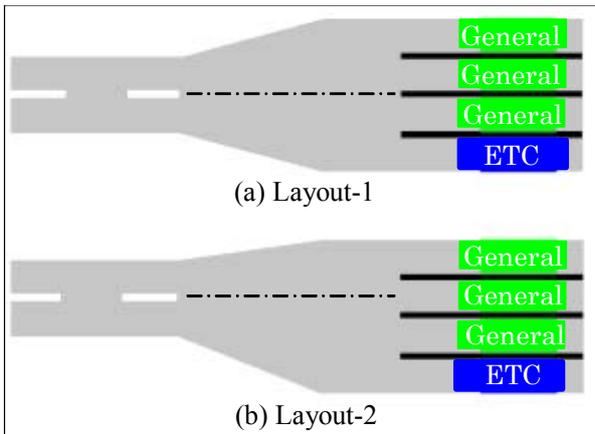


Figure 5: Layout Design Examples

### SIMULATION-BASED ANALYSIS AND SOLUTION FOR ACTUAL TOLL PLAZA

Section 3 describes the basic ideas for simulation-based approach based on mathematical data, and shows its feasibility of this approach. Since the goal of the study is to propose a practical solution to traffic jam issues actually occurred in the expressway, we have designed and developed a simulation model for Kochi IC in Japan as a case study. Physical layout data and measured data are applied in this model and internal procedure is defined as well. This section shows the result of verification on the model, and proposes some solutions derived from the simulation of the model, which is on the direction of promoting ETC users.

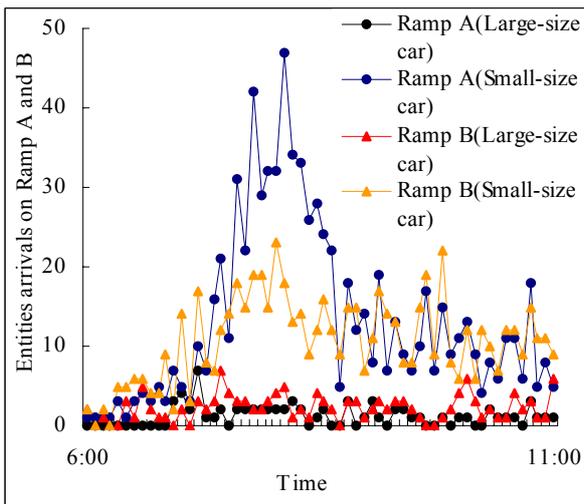


Figure 6: Time Interval Data on Entities Arrivals from Ramp A and Ramp B

### Model verification

Quality of simulation model plays a very important role to apply it to actual data. To verify the model which we have developed in this study, several comparisons were conducted between collected data and simulated data. The measurement for data collection was carried out between 6:00 am and 11:00 am on a single day at several measurement sites. During the measurement, vehicles were assigned to either small category or large category with or without ETC equipment in the measurement. Figure 6 shows the raw data of entities which came to the Kochi IC toll gate from the Ramp A and Ramp B direction.

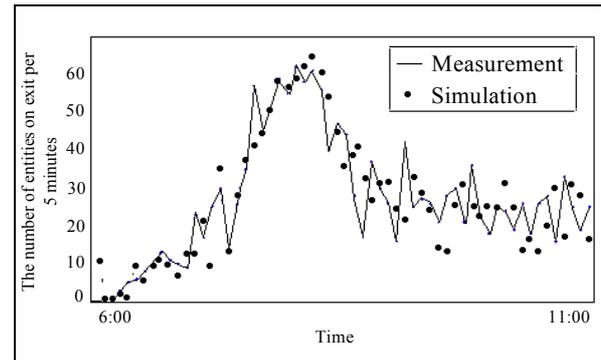


Figure 7: Comparison of Measured and Simulated Data for Out-going Traffic

To verify the model, several kinds of comparison were conducted using measured data and simulated data, including in-coming traffic, out-going traffic, traffic queue time and physical length of the queue, etc. As an example, Figure 7 shows the comparison between measured data and simulated data for out-going vehicles, of which graphs show almost identical results. According to the results of these comparisons, the simulation model in this case study shows good reproducibility.

### Solution for actual toll plaza

This section presents some practical solutions which the simulation model can provide in the case of Kochi IC. The model can provide some practical solution to actual problems. For example, the estimation on the effect of combined gate use can be made. According to the simulation using this model, the traffic of 200 vehicles length is supposed to reach 800 vehicles length if the traffic volume goes up to 20%. As a counter-measure to this traffic, the traffic length can be drastically reduced by the introduction of combined gate use and be remained almost the same level even if the traffic volume goes up to 20%. Although gate management with combined use can provide a short-term solution, it decreases the benefits of ETC system for ETC vehicles. To study the effect of ETC mixing rate over the traffic jam length, the simulation-based approach can also provide an appropriate reference data. Figure 8 shows that the difference between combined gate use and ETC specific gate use will be

almost identical if the mixing rate is increased by 25%. This means that combined gate usage as a countermeasure for traffic jams may not be required if the ETC penetration ratio reaches 12.5% or more from the current level of 10%.

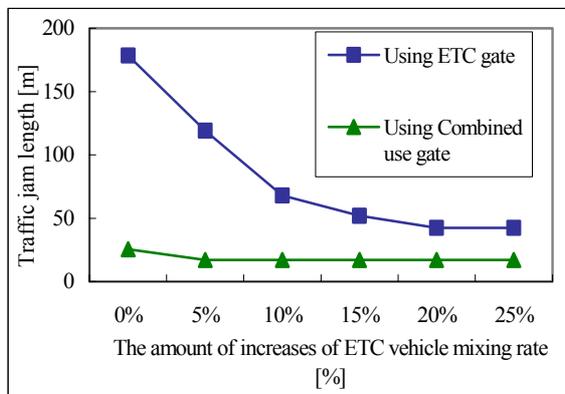


Figure 8: Effect of ETC Mixing Ratio on the Length of Traffic Jams

The result of simulation also indicates that if the number of ETC vehicles increases 5% or more in the Kochi IC case, the gate-through time of ETC vehicles becomes drastically less and it does not matter whether the ETC gate is either combined or dedicated use.

## CONCLUDING REMARKS

This paper describes the overview of ETC system in Japan, and pointed out critical issues of traffic jams around expressway toll plaza to be solved to achieve the better performance of toll gates with ETC system. The paper showed the process simulation-based approach, describing the definition and internal procedures, and presented some result of process simulation to analyze the traffic jams which occur in toll plaza of expressways with ETC gates. The paper also showed that the model developed in this study can support estimation of appropriate time in gate change in combination gate, and also consideration of layout redesign of toll plaza to achieve better performance. Since the goal of the study is to propose a practical solution to the traffic jam issue which prohibits the progress of ETC system, this paper also showed the results of simulation of a case study and presents what the simulation-based approach can provide. Conducting much more case studies in actual toll plaza, we would like to work on further studies to increase the quality of the model, and to propose more practical solutions to enhance the progress of ETC systems.

## ACKNOWLEDGEMENT

The author would like to thank Shikoku Branch Office of Japan Highway Public Corporation (JH), or Nihon Doro Kodan for their assistance in data collection and

for their valuable advice to our research. The author would also like to thank Mr. Tomoyuki Hiramoto, a graduate student of Graduate School of Mechanical Engineering at the University of Tokushima for his assisting with the programming.

## REFERENCES

- Abbas-Turki, A., O. Grunder and A. Elmoudni. 2001. "Simulation and optimization of the public transportation connection system", In *Proceedings of the 13 th European Simulation Symposium*, October 18-20, Marseille, France, pp.435-439.
- Biancadi, A., R. De Lotto and A. Ferrara. 2000. "A multi-modal transport simulation tool to solve urban location problems", *Proceedings of the 12 th European Simulation Symposium*, September 28-30, Humburg, Germany, pp.437-442.
- Fernandes, R.J. and S. Bampi. 1998. "A software environment to integrate urban traffic simulation tasks", *Proceedings of the 10 th European Simulation Symposium*, October 26-28, Nottingham, United Kingdom, pp.371-377.
- Gale, C. and M.J.Oliveron and G.Silvan. 2002. "Simulation tool for managing a non-automated distribution warehouse", *Proceedings of the 14 th European Simulation Symposium*, October 23-26, Dresden, Germany, pp.266-270.
- Horiguchi, R. and M. Kuwahara. 2000. "A theoretical analysis for the capacity of toll plaza partially with ETC tollgates", *J. of Japan Society of Civil Engineers*, No.653/IV-48, pp.29-38. (in Japanese)
- Ito, T. 2004. "Simulation-based analysis of traffic jams at toll plaza with ETC gates", *Proceedings of the Japan/USA Symposium on Flexible Automation*, Denver, CO, July 19-21.
- Ito, T. and T. Hiramoto. 2004. "Process simulation model towards analysis of traffic jams around toll gates", *Information Technology Letters, Forum on Information Technology 2004*, LO-002. (in Japanese)
- Kelton, W.D., R.P.Sadowski, and D.A.Sadowski. 1998. *Simulation with arena*, WCB/McGraw-Hill.
- Krajzewicz, D., G. Hertkorn and P. Wagner. 2002. "An example of microscopic car models validation using the open source traffic simulation SUMO", *Proceedings of the 14 th European Simulation Symposium*, October 23-26, Dresden, Germany, pp.318-322.
- Lucjan, G. and O. Jozef. 1999. "Modeling of public transport commuters flow at urban interchange centers", *Proceedings of the 11 th European Simulation Symposium*, October 26-28, Erlangen, Germany, pp.217-219.
- Schwentke, R. 2000. "Integrated training system for traffic control", *Proceedings of the 12 th European Simulation Symposium*, September 28-30, Hamburg, Germany, pp.451-455.

# SIMULATION OF TURNING RATES IN TRAFFIC SYSTEMS

Balázs KULCSÁR

István VARGA

*Department of Transport Automation,  
Budapest University of Technology and Economics  
Budapest, H-1111, Bertalan L. u. 2., Hungary  
e-mail: kulcsar@kaut.kka.bme.hu, ivarga@sztaki.hu*

## KEYWORDS

traffic simulation, split rate estimation, Kalman-filter, Moving Horizon Estimation

## ABSTRACT

Certain variables of dynamic system can not be measured, however they play an important role for control system strategies. In such a situation the approximation, computer based simulation of these variables could be useful for further techniques. There are some estimation methods which can determine the traffic flow in a traffic network. Based on the knowledge of these values, the simulation of the turning rates can be made. The paper treats the simulation of split rates in the traffic systems modelled in terms of linear time varying system using different filtering approaches.

The paper proposes three methods for simulation of turning rates in a basic traffic network. First the unconstrained Kalman filtering, and secondly the algorithm that has been developed for traffic systems is based on the unconstrained and constrained Moving Horizon Estimation(MHE) is presented. The constrained MHE problem for traffic systems is modelled in terms of linear time varying system and solves the split rate estimation process. The estimation is subjected to equality and inequality constraints. A numerical example is solved to demonstrate the Moving Horizon Estimation of split variables.

## INTRODUCTION

Simulation is a reliable tool that one uses for model based design techniques of a real or an abstract system and to conduct experiments with, in order to understand the behavior and to evaluate various theoretic strategies for operating the system.

Applications of traffic simulation can be classified in several cases. Some basic classifications are the division between microscopic, mesoscopic and macroscopic, and between continuous and discrete time approach. According to the problem area we can separate intersection, road section and network simulations. Special fields are traffic

safety and the effects of advanced traffic information and control systems.

A newly emerged area is the demand estimation through microscopic simulation. The dynamic aspect of traffic simulation in a traffic system needs the previously measured or estimated volumes of vehicles. Though the measurement of certain variables in the dynamic description are rather costly, one tries to estimate them. The observation of permanently varying turning rates, in a simple intersection, are rather costly, however the amount of the turning vehicle could be applied for traffic light harmonization, generally speaking for control.

One divides the intersection into three parts such as entry, exit and internal flows. The measurement of both the entry and the exit flows might be assumed. Traffic density cannot be measured without error, so the idealized flow plays role only in theoretical aspects. A model setup of entry-exit travel demands regarding an intersection allows estimation methods to determine the internal link flows. The key of the model buildup is the split parameter ratios. The split rate determines the turning percentage of the vehicles entering a traffic system. If one assumes that these turning rates are slowly varying split probabilities, the methods to determine probabilities are called split ratio methods ([4],[10]). The split rates define a turning proportion. The stochastic view on creating the model was elaborated in [11].

There exist many estimation techniques, for giving reliable estimation on dynamic OD matrix, their results, however, could be different. The short review on OD estimation begins with the Least Squares (constrained or not), statically based methods such as Likelihood methods([11]) and Kalman filtering ([2],[3]), or Bayesian estimator ([20]). Sometimes, combined estimators, using constraints or apriori knowledge about the intersection can be applied.

Constraints must be taken into account in course of dynamic OD estimation. A class of optimal state estimation methods are called Moving Horizon Estimation (MHE) methods([6],[13],[16]). The MHE can be concerned as the dual of the Model Predictive Control, though some special assumptions must be given for filter stability. Another advantage of Moving Horizon Estimation can be

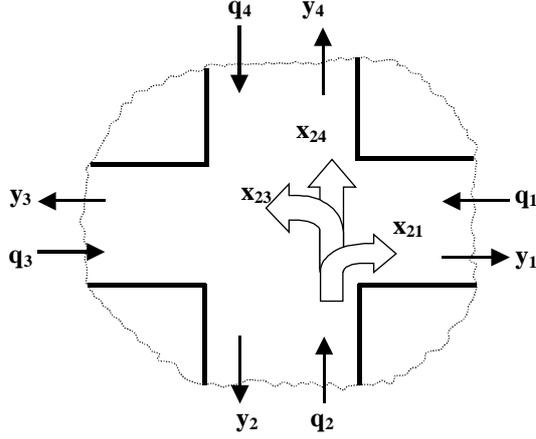


Figure 1: A simple intersection

the fact that constraints assumption can be combined in the estimation process. In the following space the Moving Horizon state estimation method is applied in intersection model.

The paper is divided into 5 chapters. After a short introduction, the problem is setup in the first section. The second section briefly summarizes the estimation techniques for split rate approximation and shows how to apply them for a basic traffic system. The third part gives a numerical example. The conclusion contains further research problems.

## PROBLEM STATEMENT

One of the basic elements in traffic network systems is the intersection. A basic intersection is given in Figure 1. Let us denote the volumes occurring in a simple intersection.

To show the problem the following variables are defined:

- $q_i(k)$  the traffic volume (the number of vehicles) entering the intersection from entrance  $i$ , during time interval  $k = 1, 2, \dots, N$
- $y_j(k)$  the traffic volume (the number of vehicles) leaving the intersection from exit  $j$ , during time interval  $k = 1, 2, \dots, N$
- $x_{ij}(k)$  the percentage of  $q_i(k)$  (turning rate) that is destined to exit  $j$ ,  $k = 1, 2, \dots, N$ .

At the intersection there are no traffic light and the right of way is not regularized, since from point of view estimation it does not take into account, only for control purpose has importance.

Let us consider the following intersection model

$$y_j(k) = \sum_{i=1}^m q_i(k)x_{ij}(k) + v_j(k), \quad (1)$$

where  $i = 1, \dots, n$  and  $j = 1, \dots, m$ .  $v_j(k)$  is a zero mean noise term. The input measurement is a noisy term, since  $q_i(k) = \tilde{q}_i(k) + \zeta_i(k)$ , with the same assumption for the noise  $\zeta_i(k)$  as above.

Split variables are independent trials. The model and its constraints are given by

$$x_{ij}(k+1) = x_{ij}(k) + w_{ij} \quad (2)$$

$$1 \geq x_{ij} \geq 0 \quad (3)$$

$$\sum_{j=1}^m x_{ij}(k) = 1. \quad (4)$$

The random variation in split parameter is small, and the  $w_{ij}(k)$  is a zero mean random component. All random components  $\zeta$ ,  $v$ ,  $w$  are mutually independent terms.

For the sake of simplicity, let us arrange all elements of the OD matrix in a single vector and use the following notations:

$$x_k = [x_{ij}(k)]^T$$

$$w_k = [w_{ij}(k)]^T$$

$$v_k = [v_j(k)]^T$$

The problem is to observe the  $x_{ij}$  states under certain condition. The latest estimation of the split probabilities can be treated as a filtering problem. In the following section one tries to emphasize the effectiveness of the constraint Moving Horizon Estimation (MHE) method as a reliable state observer of split ratios. To understand the difference between them Kalman filtering and unconstrained MHE is computed and simulated as well.

## TURNING RATE ESTIMATION METHODS

Usually for real systems the states can not be measured since the output map is only a subset of the whole state space. In many application the knowledge of the states has particular importance for implementation, for state feedback control problems. State estimation for control purpose becomes primordial and many estimation technique has been developed already.

State estimation gives us the possibility to observe via output the unmeasured states, and for stochastic systems to reduce the state and measurement noise occurring as well. Though for linear stochastic system different estimation methods provide quite a good approximation of the real state, but the realization of nonlinear state estimation under stochastic noise causes problems.

In our case the turning rate estimation of a simple intersection has many control objectives, i.e. the turning amount of the vehicle in a direction could modify the optimal traffic light control. Certain class of noisy linear discrete time system can be described is as follows:

$$x_{k+1} = A_k x_k + B_k u_k + G_k w_k \quad (5)$$

$$y_k = C_k x_k + D_k u_k + v_k \quad (6)$$

One can neglect the control input, since the split rate dynamic can be assumed as a

$$x_{k+1} = Ax_k + Gw_k \quad (7)$$

$$y_k = C_k x_k + v_k \quad (8)$$

with  $x_0$  given and with  $G_k = A_k = I_n$ .

where  $A_k$  shows the propagation of the states from  $x_k$  to  $x_{k+1}$ , the control input  $u_k$  affects the dynamic system through  $B_k$  input direction map, and  $G_k w_k$  is the weighted state noise with zero mean random signal  $w_k$ .  $G_k$  is called noise distribution matrix and colors the white noise  $w_k$ . Usually noisy systems are described with an additional disturbance term in dynamic equation.

The output map  $C_k$  can be a time or parameter varying map deciding about measured outputs, under output noise. When stochastic noise are presents, such as  $w_k$ , respectively  $v_k$ , the resulting state estimator one is often called filter. Although it provides optimal estimation under the noisy, measured output information.

Stability of filter can be shown by computing the error system for nominal case, i.e. the simulation of noiseless filter and real system from a different initial condition. The filter design methods needs the probabilistic description of the noisy term, such as probability density function. As an example of Kalman filter, which gives optimal solution with minimizing covariance of state errors.

Stochastic programming framework for estimation exists as well. Stochastic optimization methods consider zero mean random noise with Gaussian distribution.

When comparing optimal filter design methods, one usually starts with general Least Square(LS) method, which has been first presented in the second parts of 60's. Since computer based numerical solution was unable to gain the estimation results, practical implementation was impossible at the time. The reason why nowadays it has been so successful is the possibility of including explicit information about estimation processes. The implementation of limitation concerning state or measurement noise can be understood from the Nature, since unlimited disturbance can not be interpreted.

General outlook of stochastic state estimation process can be seen on Figure 2.

From point of view numerical realization of state estimation recursive algorithm formulation is necessary. The most general formulation of LS for dynamic systems is the Batch or Full Information Estimation (BE,FIE). It offers the possibility to maintain equality or inequality constraints in an infinite horizon. The main drawback of this solution of state estimation is the computational requirements, because when applying, the entire past behavior of estimated process is familiar for estimation.

The above estimation problem becomes a time variant case for (8).

The batch estimator is given by

$$\begin{aligned} & \min_{(\bar{x}_0, \hat{w}_{-1|k}, \dots, \hat{w}_{k-1|k})} \Psi_k \\ \Psi_k &= \hat{w}_{-1|k}^T Q_0^{-1} \hat{w}_{-1|k} + \\ &+ \sum_{j=0}^{k-1} \hat{w}_{j|k}^T Q^{-1} \hat{w}_{j|k} + \\ &+ \sum_{j=0}^k \hat{v}_{j|k}^T R^{-1} \hat{v}_{j|k}, \end{aligned}$$

subject to:

$$\begin{aligned} \hat{x}_{0|k} &= \bar{x}_0 + \hat{w}_{-1|k} \\ \hat{x}_{j+1|k} &= A \hat{x}_{j|k} + G \hat{w}_{j|k} \\ y_j &= C \hat{x}_{j|k} + \hat{v}_{j|k} \end{aligned}$$

with  $R^{-1}, Q^{-1}$  which are symmetric positive semi-definite noise weighting matrices. While  $Q_0$  penalizes the  $\bar{x}_{k-N}$  initial state,  $R^{-1}$  weights the output prediction error and  $Q^{-1}$  penalizes all estimated state noise.

The optimization problem grows step-by-step and becomes intractable even for small dynamic systems.

The general stability property of unconstrained batch estimation can be applied for constrained with some special assumptions. The feasibility of the constrained optimization comes, over all, from the initial values of the estimated state and noise. Naturally, for constrained batch an unstable  $A$  could cause unstable estimator, since the unstable true system trajectories cannot be observed with constraints on estimator states. However, general (nominal) asymptotic stability even for unstable  $A$  can be ensured.

The Moving Horizon Estimation scheme can be seen in Figure 3.

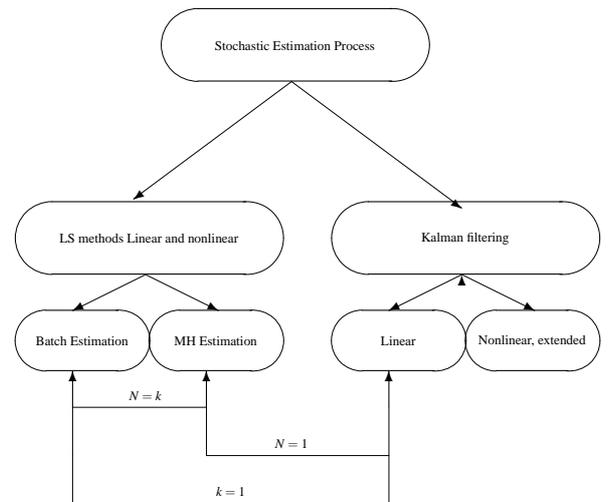


Figure 2: Stochastic Estimation Process

Let the generalized MHE optimization criteria be defined by the following functional

$$\begin{aligned} & \min_{(\bar{x}_0, \hat{w}_{k-N-1|k}, \dots, \hat{w}_{k-1|k})} \Psi_k \\ \Psi_k = & \hat{w}_{k-N-1|k}^T Q_N^{-1} \hat{w}_{-1|k} + \\ & + \sum_{j=k-N}^{k-1} \hat{w}_{j|k}^T Q^{-1} \hat{w}_{j|k} + \\ & + \sum_{j=k-N}^k \hat{v}_{j|k}^T R^{-1} \hat{v}_{j|k} + \Psi_{k-N}^*, \end{aligned}$$

subject to:

$$\begin{aligned} \hat{x}_{k-N|k} &= \bar{x}_{k-N} + \hat{w}_{k-N-1|k} \\ \hat{x}_{j+1|k} &= A\hat{x}_{j|k} + G\hat{w}_{j|k} \quad j = k-N-1, \dots, k-1 \\ y_j &= C\hat{x}_{j|k} + \hat{v}_{j|k} \quad j = k-N-1, \dots, k \end{aligned}$$

If the expected output is small,  $R^{-1}$  has to be chosen large, compared to  $Q^{-1}$ , and the resulting sensor noise vector becomes small, compared to  $\hat{w}_{j|k}$ . On the other hand, if our measurements are not reliable,  $Q^{-1}$  should be chosen large, compared to  $R^{-1}$ .

$\Psi_{k-N}^*$  is the so-called arrival cost, which is analogue to the *cost to go* in MPC technique. The arrival cost summarizes all knowledge about the best estimation before the  $N$ -th step. For unconstrained linear case, the arrival cost can be expressed explicitly. If state or noise inequality constraints, or nonlinearities are present, we have no analytic expression to generate the arrival cost. Though analytic approach is unavailable, an *approximate* cost may

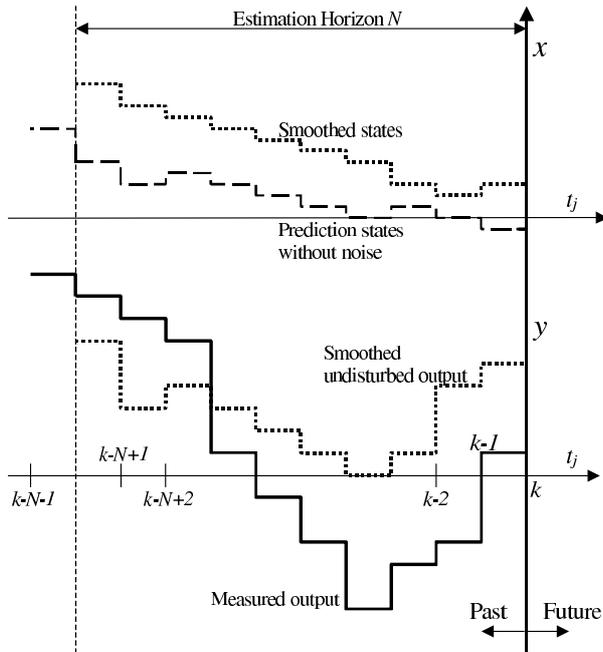


Figure 3: General Moving Horizon Estimation process

be given. When inequality constraints are inactive, the approximation is exact. Therefore, the poor choice of the arrival cost leads to the filter's instability. To find the initial condition of General MHE, one uses a batch estimation for the first  $N-1$  step estimates.

To slide between windows the filtered estimate update is preferred.

Another possibility of state estimation is the use of Kalman filter. The Kalman filter could be applied widely in traffic systems. This has been published in numerous papers ([4],[?]). This method, based on Gaussian distributions of random variables, is defined on a probability framework of the unknown split parameters. Kalman filter equations ([8]) can be formulated as recursive ones started with an initial condition. The optimal estimation depends on the choice of state noise covariance ( $Q$ ) and on the output noise covariance ( $R$ ) weights. The Kalman estimator can be applied subject to inequality constraints by using stochastic programming([19]). The connection between Kalman filtering and full information estimation is known.

#### EXAMPLE

To show the difference between estimation techniques let us consider the following solution of estimation of split ratio. The solution is based on standard Kalman filtering approach, on general unconstrained MH Estimation, and constrained MH Estimation with the equality constraints (4) and dynamic inequality constraint.

One returns to the intersection model which is now given by:

$$\begin{aligned} x_{k+1} &= x_k + u_k + w_k \\ y_k &= C_k x_k + v_k, \end{aligned}$$

where  $C_k$  contains the elements of  $q_i$ , a time varying output map. The structure of  $C_k$  depends upon the layout of the intersection.

$$\begin{aligned} C_k &= \begin{bmatrix} q_1 & 0 & q_2 & 0 \\ 0 & q_1 & 0 & q_2 \end{bmatrix} \\ q_{k+1} &= \tilde{q}_k + \zeta_k \end{aligned}$$

and where  $q_k$  is the noisy input volume of the vehicles entering the intersection, with the sensor noise  $\zeta_k$  a zero mean random signal with the appropriate dimension.

The number of split parameter needs to be well chosen in our situation the  $x = [x_{13} \ x_{14} \ x_{23} \ x_{24}]^T$ , the number of outputs  $y_k = [y_3 \ y_4]^T$ . One should denote that in split parameters two types of variation are present: the permanent systematic component, with a zero average during a time period, and the the random component which is assumed to be small respect to 1.

Let us suppose to have 1 sample in every second. Let the horizon comprise 1 sample, and by applying diagonals  $R$ ,  $Q$  and  $Q_0$ , the following results are gained. The

simulation covered take 1 hour, which can be seen on the Figure 4.

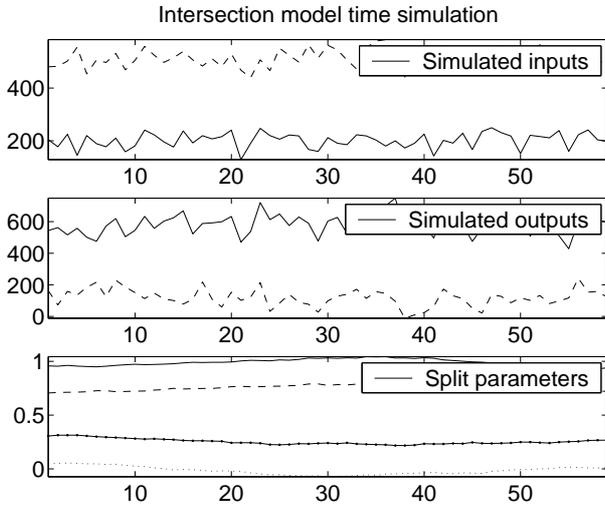


Figure 4: Variables of the basic intersection in time domain

For numerical computation and simulation of the estimated split states one uses quadratic programming either for unconstrained or for constrained MH. The Kalman filtered states are computed by a recursive algorithm.

For solving MHE numerically, either one may use a recursive algorithm, which can be derived from Lagrange multiplier method([6]), or quadratic programming for unconstrained, respectively for constrained MH. Simulating the split parameter for intersection (Figure 1.), the result can be seen in Figure 5.

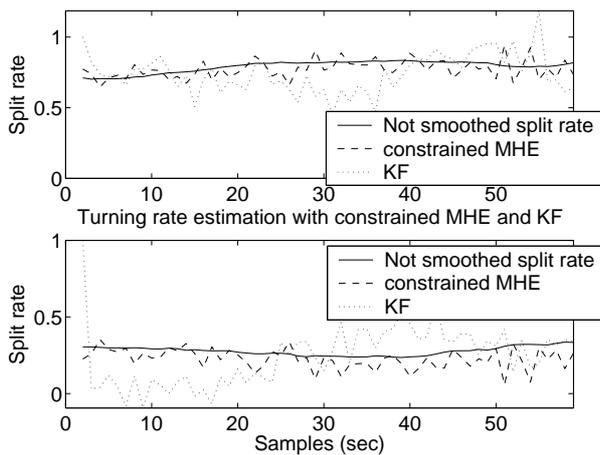


Figure 5:  $x_{12}$ ,  $x_{23}$ ,  $x_{32}$  split parameter time behaviour

In the Figure 5 only the Kalman filter and the constrained MHE is plotted. Since for  $N = 1$  the unconstrained MHE and the Kalman filtering problem gives almost the same result (numerical computational error).

The only difference is the estimation idea, because while for Kalman filtering the prediction is always a forward, the unconstrained MH approach is defined as backward calculation.

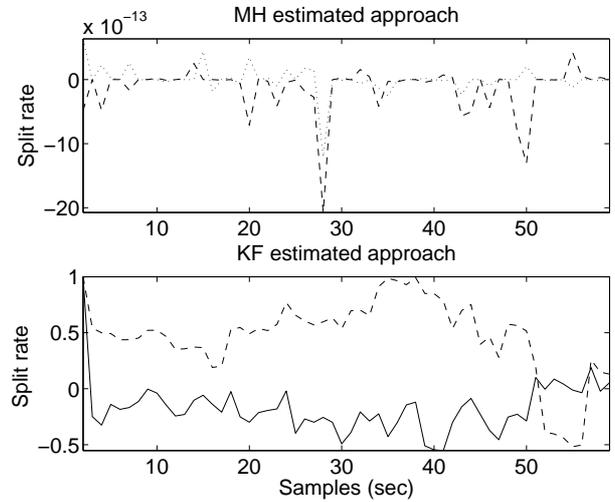


Figure 6: Equality constraints  $x_{13} + x_{14} - 1 = 0$  and  $x_{23} + x_{24} - 1 = 0$  time behaviour

As it has been shown in Figure 6, the simulated turning rates (supposed to be real), are estimated with the MHE process.

## CONCLUSION

The article summarizes the Moving Horizon Estimation approach for a simple traffic system, an intersection. In traffic engineering the estimation of split variables is important for further optimal traffic light control strategies.

The MHE optimal estimation method shows a possible way for including constraints into the design procedure. One could possibly extend the state estimation, based on MHE algorithm with some additional constraints in inequality form on states, noise or other variables. The selection of weighting matrices and estimation horizon and the good approximation of arrival cost influence the performance of the estimation.

A numerical example has been shown to demonstrate how to apply the Moving Horizon technique for split rate observation.

The general MHE technique could be applied to nonlinear processes which will be in the focus of our traffic system estimation research.

## ACKNOWLEDGEMENT

The authors gratefully acknowledge the contribution of Hungarian National Science Foundation (OTKA, Grant number T046220)

## References

- [1] Bemporad, A., D. Mignone, M. Morari "Moving Horizon Estimation for Hybrid Systems and Fault Detection" *ACC San Diego, California, 1999*
- [2] Cremer, M. and Keller, H. "Dynamic Identification of Flows from Traffic Counts at Complex Intersections" *Proc. 8th Int. Symposium on Transportation and Traffic Theory, University of Toronto Press, Toronto Canada, 1981, pp 199-209.*
- [3] Cremer, M. "Determining the Time-Dependent Trip Distribution in a Complex Intersection for Traffic Responsive Control" *IFAC Control in Transportation Systems, Baden-Baden, Germany 1983, pp. 214-222.*
- [4] Cremer, M. and Keller, H. "A New Class of Dynamic Methods for the Identification of Origin-Destination Flows" *Transportation Research-B, 1987, Vol. 21B, No. 2, pp. 117-132.*
- [5] Douglas, G. R., J. H. Lee, "On the use of constraints in least squares estimation and control" *Automatica, 2002, Vol 38 1113-1123*
- [6] Findeisen, P. K. "Moving Horizon State Estimation of Discrete Time Systems" *MS Thesis, U. of Wisconsin-Madison.*
- [7] Hannah Michalska and David Q. Mayne "Moving Horizon Observers and Observer-Based" *IEEE Transactions on Automatic Control, Vol. 40, No. 6, 1995, pp. 995-1006.*
- [8] Kalman, R.E. "A New Approach to Linear Filtering and Prediction" *Journal of Basic Engineering (ASME), 1960, 82D, pp. 35-45.*
- [9] Kenneth R. Muske and James B. Rawlings, Jay H. Lee, "Receding Horizon Recursive State Estimation" *American Control Conferences, 1993, Vol 1. pp. 900-904.*
- [10] Nihan, N.L. and Davis, G.A. "Recursive Estimation of Origin-Destination Matrices from Input/Output Counts" *Transportation Research B, 1987, Vol. 21B, No. 2, pp. 149-163.*
- [11] Nihan, N.L. and Davis, G.A. "Application of Prediction-Error Minimization and Maximum Likelihood to Estimate Intersection O-D Matrices from Traffic Counts" *Transportation Science, 1989, Vol 23, No. 2.*
- [12] Papageorgiou, M. "Concise Encyclopedia of Traffic and Transportation Systems" *Pergamon Press, 1991.*
- [13] Rao, V. C. "Moving Horizon strategies for the constrained Monitoring and Control of Nonlinear Discrete-Time Systems" *PhD Thesis U. of Wisconsin-Madison, 2000*
- [14] Rao, V. C., J. B. Rawlings, "Constrained Process Monitoring: Moving-Horizon Approach" *AICHE journal, 2002, Vol 48. pp. 97-109.*
- [15] Rao, V. C., J. B. Rawlings, D. Q. Mayne, "Constrained State Estimation for Nonlinear Discrete-Time Systems: Stability and Moving Horizon Approximations" *IEEE Transaction on Automatic Control, 2003, Vol 48. pp. 246-258.*
- [16] Tyler, M., M. Morari, "Stability of Constrained Moving Horizon Estimation Schemes" *Automatica, 1996, Vol 17. pp. 1410-1425.*
- [17] Tyler, M., K. Asano, M. Morari, "Application of Moving Horizon Estimation Based Fault Detection to Cold Tandem Steel Mill" *ETH Technical Report AUT96-06*
- [18] Zijpp, N.J. Van Der, R. Hamerslag, "An Improved Kalman Filtering Approach to Estimate Origin - Destination Matrices for Freeway Corridors" *Transportation Research Records, 1994, No. 1443, pp. 54-64.*
- [19] Zijpp, N.J. van der "Dynamic origin-destination matrix estimation from traffic counts and automated vehicle identification data" *Transportation Research Record No. 1607. TRB, Washington, DC, 1997, pp. 87-94.*
- [20] Zijpp, N.J. van der "Comparison of methods for dynamic origin-destination matrix estimation" *IFAC Transportation Systems Symposium, Chania, Greece, 1997, pp. 1445-1450.*

## AUTHOR BIOGRAPHIES

Balázs KULCSÁR was born in Budapest, Hungary. After graduating as traffic engineer in 1999 at the Budapest University of Technology and Economics, he started his PhD study at the Department of Transport Automation. Presently he works as research assistant at the same department in half time. He is working for the Computer and Automation Research Institute of the Hungarian Academy of Sciences. Any complementary information can be found on <http://www.kka.bme.hu/oktkut/munkatarsak.htm>.

István VARGA was born in Budapest, Hungary. He graduated in traffic engineering in 1998 at Budapest University of Technology and Economics (BUTE). Presently, he is doing PhD studies at the same university. Further information can be found at [www.sztaki.hu/scl](http://www.sztaki.hu/scl)

# HEAD LEADING ALGORITHM FOR URBAN TRAFFIC MODELING

David Hartman  
Department of Computer Science and Engineering  
University of West Bohemia  
Univerzitní 22, Plzeň, 360 14, Czech Republic  
Email: tazman@kiv.zcu.cz

## KEYWORDS

Simulation, Urban Traffic, Cellular Automata Model, Nagel-Schreckenberg model, Head Leading Algorithm, JUTS project.

## ABSTRACT

This paper describes construction of discrete urban traffic simulation model based on Nagel-Schreckenberg's cellular automata model (NaSch model) and its modification. These NaSch models have been constructed for simulation of traffic on highways and freeways. We try to modify this approach to obtain a model of urban traffic. The development is based on an object oriented approach.

## GOALS OF THE MODEL

Our objective was construction of a model that could be used for analysis of traffic situation in urban areas, i.e. jams, queues of vehicles, public transport programmes, junction lights switching algorithms, etc. Another very important requirement was object oriented character of the designed model, because we wanted to have the possibility of easy or half-automatic creation of simulation map. It also means that simulation map should be divided into parts (representing urban traffic entities), that are connected together and the whole simulation is performed by communication among them.

## CHOOSSED MODEL TYPE

According to the great study of traffic flow modeling in TBR report (Gartner et al. 1997) and the great study of micro-simulation models in European traffic simulation project SMARTTEST (Alergs et al. 1998), we decided to use a micro-simulation model as our basic model (instead of macroscopic or mesoscopic types). We choosed the Nagel-Schreckenberg cellular automata model, because of its detailed simulation and also easy implementation and natural character of vehicle move modelling. But there was one disadvantage. This model was originally designed for highway or freeway traffic (Nagel and Schreckenberg 1992). So it was important to make some adjustment to ensure that the simulation of urban traffic will be also possible. There were other approaches to do that. An example can be found in the Simon and Nagel simplified model (Nagel and Simon 1998). They used a simplified version of street network

and just single-lane simulation. A More detailed model is constructed by Esser and Schreckenberg (Esser and Schreckenberg 1997). Other approaches can be found in other papers like (Chrobok et al. 2001), (Schreckenberg et al. 2001) and (Chopard et al. 1997). In our case a strong combination with object oriented approach is used. We modified the basic NaSch traffic model and constructed the object oriented model of traffic network that can use a modified cellular automata model as its basis. And finally we added the leading head algorithm into the global model to make the modification in NaSch and communication in object model possible. So the leading head algorithm is a micro-simulation algorithm and also the algorithm for object oriented model control.

## BASIC NAGEL-SCHRECKENBERG MODEL

In this section we describe the basis of the basic NaSch cellular automata model. More detailed description can be found in other papers (Knospe et al. 2001) or original model definition in paper (Nagel and Schreckenberg 1992).

To ensure the completeness of our paper we explain the basis of this model. Let's have a look on the basic NaSch model concerning a single lane situation only. In this model the road is divided into cells. Each cell has the same length of 7.5m. Each cell can be occupied by a single vehicle. The vehicles are therefore just like items of an array. Each vehicle has a discrete speed that represents the number of cells that the vehicle jumps over during its movement (see Figure 1).

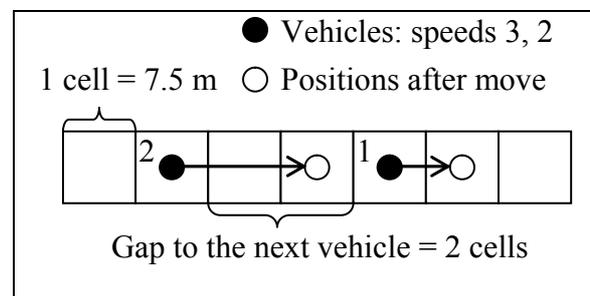


Figure 1: Road Division in NaSch Model

The movement is performed by steps of simulation. During each step the vehicles are shifted about the number of cells that equals their speeds. The positions of the vehicles together with vehicles states constitute the global state of the model at a specific discrete time  $t$ . Let  $i$  be an index of the state (i.e. the value of discrete

time  $t$ ). If we denote the speed of the vehicle as  $v_i$  and gap to the next vehicle as  $g_i$  we can obtain next state  $s_{i+1}$  from the actual  $s_i$  by applying the following basic rules of NaSch model:

1. Acceleration:  $v_i \rightarrow \min(v_i + 1, v_{\max})$
2. Breaking:  $v_i \rightarrow \min(v_i, g_i - 1)$
3. Randomization:  $v_i \rightarrow \max(v_i - 1, 0)$  with  $p$
4. Driving:  $x_i \rightarrow x_i + v_i$

Here we can see the roots of the model. The drivers of the vehicles want to accelerate as much as possible (step n. 1), but they have to slow down to avoid crashes (forbidden in the model). So the speed is restricted by gap to the next vehicle (step n. 2). Then the vehicle can move (step n. 4), but for realistic behaviour of the model the 3<sup>rd</sup> step is added. The adjustment of speed at step n. 3 is performed only with probability  $p$ . This stochastic step (others are deterministic) takes into account the natural velocity fluctuation (human factor or road conditions).

Now we can look at the multi-lane case. We simply take single lanes and place each one alongside the other and add the lane changing rules (see Figure 2) to the dynamic mechanism.

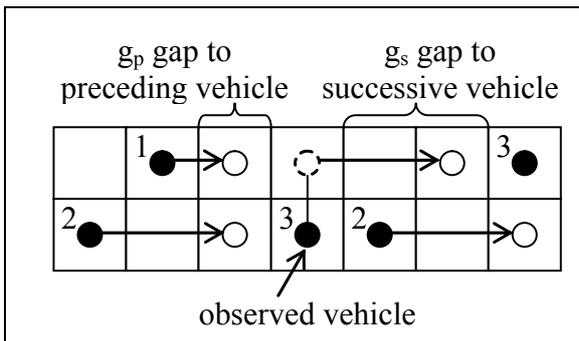


Figure 2: Lane Changing Situation

The lane changing rules differ in used models but their principles are often based on checking the values of  $g_s$  and  $g_p$ . A systematic approach for two-lane rules can be found in (Nagel et al. 1998) and also good information about this problems can be found in (Knospe et al. 1998), but mainly for highways.

When we use these rules, the simulation step is divided into two sub-steps. The first one perform the lane change manoeuvres and the second one perform the NaSch rules. As we see (Figure 2), the vehicle after applying lane change rules is in a position that is the same as in a normal state, so you can easily perform the standard NaSch rules without any problem. There could be some problems when you use the vehicles with length more than one. We discuss it later when we describe our model that uses this approach. For another solution of this problem, see paper (Esser and Schreckenberg 1997). Lane change manoeuvres are also used for modeling overtaking. There could be also problems with deadlocks during lane changing.

Deadlocks occur when the vehicle that wants to change the lane waits until the target lane is free and another vehicle waits for it and blocks it together.

## INCLUDING SLOW-TO-START BEHAVIOUR

A very important part of the model is the stochastic step n.3. The basic NaSch rules use a constant probability for this step, but it seems to be insufficient for modeling metastable states with a very high flow (Barlovic et al. 1998). These metastable states relate with restart behaviour of stopped vehicles. To include this behaviour to simulation in more realistic fashion new models with slow-to-start rules are introduced, e.g. T<sup>2</sup>, BJH and VDR model (Barlovic et al. 1998).

The Velocity-Dependent Randomization (VDR) model is based on the idea of dependent probability. It's simple. The probability in step n.3 from NaSch rules is a function of the vehicle's speed, i.e.

$$p = p(v(t))$$

The parameter  $p$  should be determined before the original 1<sup>st</sup> step from the NaSch rules. Sometimes, often for explanation, is used a very simple case of this model with probability function with this definition:

$$p(v) = \begin{cases} p_0 & \text{for } v = 0 \\ p & \text{for } v > 0 \end{cases}$$

Using one of the mentioned models introduces the restart behaviour into the model and ensures that traffic flow characteristics will better fit realistic values.

## ANTICIPATION IN CELLULAR AUTOMATA

Last modification of NaSch cellular automata model is based on introducing anticipation of the following vehicle's movement. It means that an anticipation rule is added to the original set of rules. It is also useful to add this rule to the VDR model instead of the original NaSch model, because the slow-to-start rule of the VDR model still helps the simulation to be more realistic in modeling restart behaviour and therefore metastable states.

The anticipation rule is a modification of the original adjustment of speed. This new adjustment takes into account the expected behaviour of the leading vehicle (following vehicle in the direction of move). This rule helps the model to simulate different vehicle characteristics like acceleration more realistic. It is also introduced in cases of different vehicle length. In these cases the length that represents one cell is usually set to a smaller value, e.g. 1.5 m (Knospe et al. 2001). The next advantage of anticipation is speed-up effect for lane changing modeling caused by reduction of changing wait times. For systematic approach and more detailed description of this problem including the lane change problem, see (Knospe et al. 1998).

## BASIS OF MODEL AND CELL LENGTH

The model that we constructed is based on NaSch cellular automata as we mention above. We also used a VDR and anticipation modification of the basic algorithm. The original NaSch model is designed for highways and freeways and if we use 1 second as simulation step (as usual) the speeds of the vehicles follow the Table 1.

Table 1: Original NaSch Speeds

Discrete speed [cell/s]	Real speed [km/h]
0	0
1	27
2	54
3	81
4	108
5	135

Our model is intended to be used for urban traffic and therefore these speeds are unusable (few values with respect to permitted speed in cities) and also cell length is very long for city and we want to distinguish individual types of vehicles (mainly vehicle length). These requests result in cell length of 2.5 m. For the permitted speed of 50 km/h used in the Czech Republic, the speed will be as shown in Table 2.

Table 2: Modified Speed Discretization

Discrete speed [cell/s]	Real speed [km/h]
0	0
1	9
2	18
3	27
4	36
5	45
6	54

This length ensures a sufficient number of needed speed values (Table 2) and also represents approximately all known vehicles as shows the Table 3.

Table 3: Representation of Types of Vehicles

Vehicle length [cell]	Real length [m]	Type of Vehicle
1	2.5	motorcycle
2	5	passenger vehicles
3	7.5	van
4	10	minibus
5	12.5	bus and truck
6	15	lorry

These types of vehicle will be used in our discrete model.

## TRAFFIC NETWORK STRUCTURE

We know the type and modifications of the NaSch model and now we can deal with the traffic network structure, because it will be very important for leading head algorithm construction. The whole map definition is strongly object oriented. Most of the traffic network elements are objects.

### Road Segment

We start from the basic element of network. The modified NaSch algorithm is defined for set of traffic lanes in one direction and therefore the road segment is defined as a set of these lanes (Figure 3).

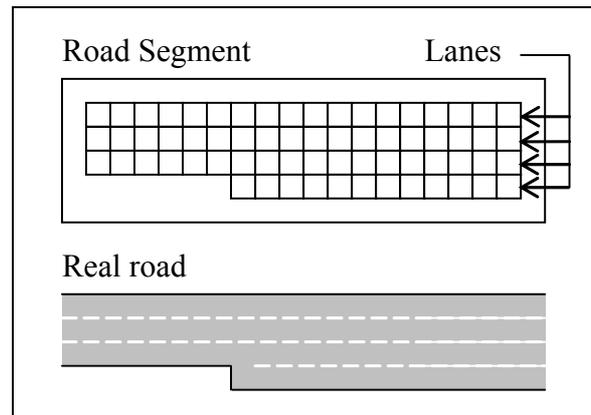


Figure 3: Road Segment

This segment is the basic one, because the main part of simulation is performed here.

### Crossroad Segment

Another important segment is the cross-road. Because the dynamics at these segments are very complicated, there must be an other mechanism than NaSch model to perform vehicle movement. But we also wanted to build up this mechanism on very similar principles. Finally we meet our requirements by introducing special single places that communicate with each other. Through these places the vehicles pass through the crossroads. Figure 4 shows a simple crossroad's structure.

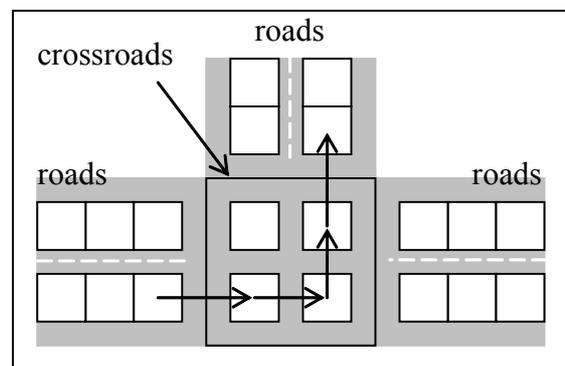


Figure 4: Simple Crossroad Structure

The crossroad solutions are very different in the field of cellular automata models. Sometime just a switching matrix with probabilities is used (Esser and Schreckenberg 1997), or a representation by a rotary's dynamic (Chopard et al. 1997) or by structure from ramps and transfer links (Chrobok et al. 2001).

### Roudabout Segment

Now we have roads and crossroads, so we can make a traffic network as an oriented graph from roads as edges and crossroads as nodes. But there exists another important segment except these two. It is the roundabout and it is also a node in our oriented graph. We have two possibilities how to define this object. The first is to define it same way as a crossroad, but if we construct the model of a large roundabout, the number of places will exceed a reasonable value. Therefore the second possibility will be better and that is to construct the roundabout as a road segment that ends at its beginning. We can do that because we can join the onramps and offramps to a road, so we can also join incoming and outgoing roads to the roundabout. The structure is shown in Figure 5.

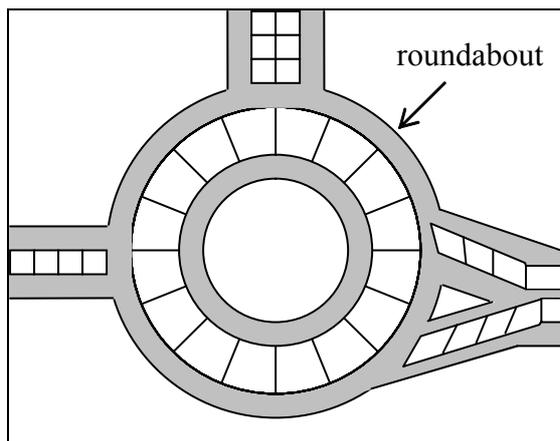


Figure 5: Crossroad structure

Now it seems that we have all segments we need to build up the oriented graph that represents the structure of the traffic network. But there are still one or two segments more to make the network complete.

### Generator and Terminator Segments

To include the environment of the simulated system we need to create segments that will generate vehicles following some stochastic distribution and send those to the network. This distribution differs from place to place (it means from road to road), so there should be more generator segments that should be connected at a specific place to the network.

Except generation of vehicles we probably need a segment to terminate the life of a vehicle. This segment can collect some statistics and check for problems.

### Connecting Segments by Accessplace

We have now defined several segments of the traffic network. We know that these segments should be connected together. A connection can be between a road and a crossroad, a road and a generator, a road and a road, etc. So many different segments can be connected together in a pair. It will be usefull to find the way how to connect these segments by general mechanism. We establish this mechanism as a connection by an Accessplace. An Accessplace is a special object that knows inner structure of the participating segments. With this knowledge it can send vehicles from any segment to any other without condition that these segments must know their neighbours.

The final structure of the traffic network implemented in our model for the case of crossroad shown in Figure 4 is shown in Figure 6.

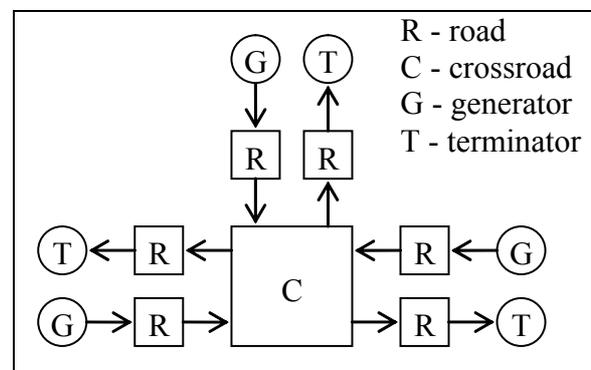


Figure 6: Traffic Network Structure

The structure of the simulation map in Figure 6 is shown without Accessplaces that will be between each pair of connected segments.

### LEADING HEAD ALGORITM

We have defined the structure of the traffic network and we have defined each segment of this structure. Now it's time to describe our leading head algorithm (LHA) for which this structure was built. The basic principle of leading head algorithm is simple. We assume that the vehicles should be with length more than 1. Without this assumption the LHA has no meaning. The movement of the vehicle follows these steps (when the LHA is used):

1. The head of the vehicle moves as a standard vehicle in the NaSch model.
2. Other pieces of the vehicle are shifted over the same way that the head was shifted before.

The principle of the algorithm is shown in Figure 7. The important thing to notice is that testing possible collisions is needed only for the head and the following pieces move automatically without testing. These pieces only occupy the cells on the road. The LHA is also a good mechanism for sending vehicles from one segment to another, because the algorithm only concerns the head and the other parts of vehicle that move over the signed path.

So they need not to pass through the communication object in such a complicated way that the head passes. One can say that the LHA is an algorithm for hiding the length of the vehicle.

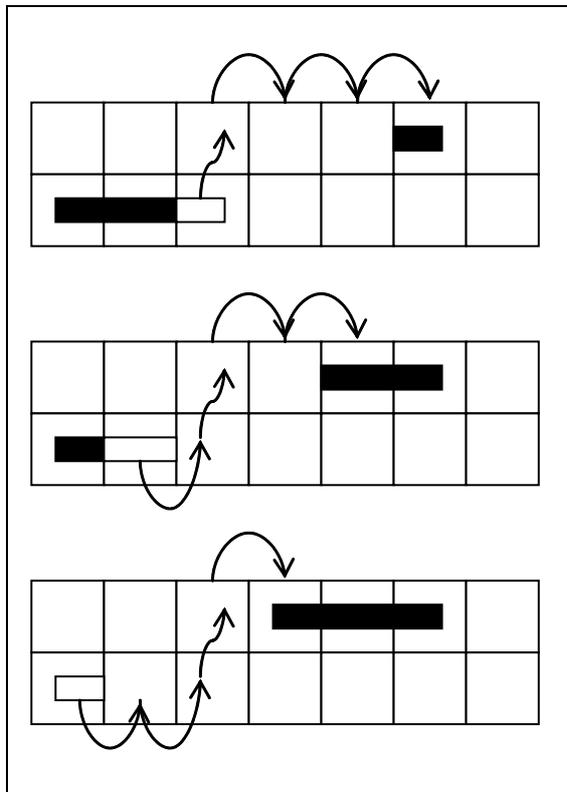


Figure 7: LHA Lane Change

The last advantage of the LHA algorithm together with the designed traffic network is an easy way of a vehicle passage through a crossroad. The pieces of the crossroad are also possible positions of the vehicle, so if the head is going through the path at the crossroad, the other pieces of the vehicle can follow it without problem. This situation with just the head move (the other pieces' moves could be easily added) is shown in Figure 8.

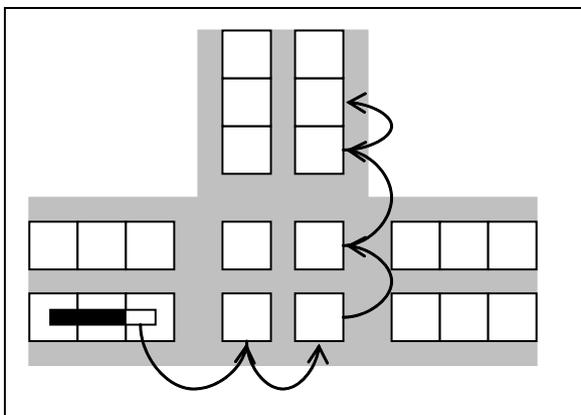


Figure 8: LHA at Crossroad

We have seen the LHA in action during lane change process, we have talked about communication between

segments with a help of LHA and we have seen the LHA solution of crossroad passage. Finally we have to say that this algorithm probably doesn't bring any acceleration of simulation run, but it uses a general method for vehicle moving that could help us to extend possible developed models.

## CASE STUDY

This type of NaSch cellular automata built-in traffic network model, together with lane head algorithm is being developed in the JUTS project (J-Sim Urban Traffic Simulator). This project is written in Java and uses all these principles inside its simulation core. We are trying to create a complex simulation tool with easy methods for editing and generating the simulation map. We are using XML format to store and work with the simulation map and other data.

But until these days the core has not been validated yet, because it is still under development. For more information about this project see (Hartman and Kačer 2003).

We cooperate with Public Transport Department of Pilsen (a Czech city) from which we obtain data from several observers that we want to use for half-automatic simulation map generation. For that purpose the designed model is well suited.

## CONCLUSION

The designed model is based on the Nagel-Schreckenberg cellular automata model with VDR and anticipation that have been tested several times in other projects or smaller research works. So there is quite great guarantee that the simulation based on this model will approach reality.

The solution of crossroad dynamics is one of the most detailed that is used in cellular automata models and at the same time it allows a very simply way to pass the vehicles through the crossroad. It involves a possibility of changing the inner dynamics by changing the permitted directions on the crossroad or you can simply add any new testing of other interesting features during passage through the places at the crossroad.

The leading head algorithm is an easy and flexible way how to move vehicles through a segment and also between segments. The crossroad and road are mostly the same for this algorithm and also the roundabout which is designed as a road in cycle is the same.

Finally we must say that the main goal of the designed model is to make an easily extendable model for big applications that will be able to configure the simulation automatically from the urban traffic data that is obtained from city measurements.

## ACKNOWLEDGEMENT

The research was supported by a grant of the Grant Agency of the Czech Republic - Research of methods and tools for verification of embedded computer systems, no. 102/03/0672.

## USEFULL LINKS

- <http://www.its.leeds.ac.uk/projects/smartest/>  
- SMARTEST project (Simulation Modelling Applied to Road Transport European Scheme Test) review micro-simulations.
- <http://www.trb.org/>  
- TBR home page (Transportation Research Board).
- <http://www.traffic.uni-duisburg.de/>  
- Group "Physics of Transport and Traffic" of prof. M. Schreckenberg.
- <http://www.juts.zcu.cz/>  
- JUTS project (J-Sim Urban traffic Simulator) home page.

## REFERENCES

- Algers S.; E. Bernauer; M. Boero; L. Breheret; C. D. Taranto; M. Dougherty; and K. F. and J.-F. Gabard. 1998 "A Review of Micro-Simulation Models." SMARTEST Project report, Institute of Transport Studies Univerzity of Leeds, UK.
- Barlovic R.; L. Santen; A. Schadschnaider; and M. Schreckenberg. 1998 "Metastable states in cellular automata for traffic flow." *European Phys. Journal*, B 5, 793-801.
- Chopard B.; A. Dupuis; and P. Luthi. 1997 "A cellular automata model for urban traffic and its application to the city of Geneva." *Proceedings of traffic and granular flow '97*, Duisbourg, 154-168.
- Chrobok R.; J. Wahle; and M. Schreckenberg. 2001 "Traffic forecast using simulations of large scale networks." *4<sup>th</sup> International IEEE Conference on Intelligent Transportation Systems*, Oakland, pp. 434-439.
- Esser J. and M. Schreckenberg. 1997 "Microscopic simulation of urban traffic based on cellular automata." *International journal of modern physics*, Vol. 8, 1025-1037.
- Garthner N.; J.C. Messer; and K.A. Rathi. 1997 "Traffic Flow Theory." Update and expansion TRB Special Report 165, Federal Highway Administration, U.S. Department of Transportation.
- Hartman D. and J. Kačer. 2003 "JUTS – J-Sim Urban Traffic Simulator." *Proceeding of the 2<sup>nd</sup> international Conference on the Principles and Practice of Programming in Java*, Kilkenny, Ireland, 113-116.
- Knosp W.; L. Santen; A. Schadschnaider; and M. Schreckenberg. 1998 "Disorder effects in cellular automata for two-lane traffic." *Physica Journal A* 265, 614-634.
- Knosp W.; M. Schreckenberg; R. Barlovic; and H. Klüpfel. 2001 "Statistical physics of cellular automata models for traffic flow." *Computational Statistical Physics*, Springer, 113-126.
- Nagel K. and M. Schreckenberg. 1992 "A cellular automaton model for freeway traffic." *J. Phys I France* 2, 2221.

- Nagel K.; D.E. Wolf; P. Wagner; and P. Simon. 1998 "Two-lane traffic rules for cellular automata: A systematic approach." *Physical Review E*, American Physical Society.
- Schreckenberg M.; L. Neubert; and J. Wahle. 2001 "Simulation of traffic in large road networks." *Future generation computers* 17, 649-657.
- Simon, P.M. and K. Nagel. 1998 "Simplified cellular automaton model for city traffic." *Physical Review E*, The American Physical Society.

## AUTHOR BIOGRAPHIES



**DAVID HARTMAN** was born in Karlovy Vary Czech Republic and went to the University of West Bohemia, where he studied software engineering and obtained his degree in 2003. Then he started PhD. studies at the Department of Computer Science at this university and currently works on the simulation problems. He is a member of the research group DSS (Distributed Systems, Software engineering and Simulations), creator and one of the leading members of the traffic simulation project JUTS. His e-mail address is: [tazman@kiv.zcu.cz](mailto:tazman@kiv.zcu.cz) and Web-page is: <http://www.kiv.zcu.cz/~tazman/>.

# A Framework Combining Cellular Automata and Multi-Agents in a Unified Simulation System for Crowd Control

Robert Signorile  
Department of Computer Science,  
Boston College  
Chestnut Hill, MA – USA  
Phone: 617-552-3936  
Email: signoril@bc.edu

**Keywords:** Complex Systems, Cellular Automata, Agents

## ABSTRACT

Controlling crowds in airports, train terminals, sporting events, etc., is a complex problem. This particular problem has a great deal of interaction between the entities themselves (i.e. among the individual members of the crowd) and the crowd (or individuals) with the environment in which the crowd is placed. This complexity of this system can be described in the much researched area of artificial life. By combining Cellular Automata (CA) with agents, we can construct a system to capture and control the ebb and flow of a crowd, including the particular characteristics of the individuals in the crowd. To this end, we have developed a prototype crowd control simulation system as a test case for this kind of problem. The imbedded CA provides a framework for flow of people, much like traffic models (Nagel, K., and Rasmussen, S. 1994) and (Blue, V. J. and Adler, J.L. 2000a), while the agent reproduces the behavior of a crowd, including subgroup behaviors, interactions, stochastic decisions of single units etc. This work is an extension of the web-based model in (Bruzzone, A. and Signorile 1999).

## CELLULAR AUTOMATA SIMULATIONS AND TRAFFIC SIMULATIONS

Cellular automata (CA) are simple spatial processing models with their origins in the early architecture of digital computers designed in the 1940 and 1950s. CA has close associations with complexity theory and has been employed in the exploration of a diverse range of urban phenomena, generally to investigate ideas about how real urban systems operate, but from a controlled experimental environment within computer software. Urban applications of CA range from

traffic simulation and regional-scale urbanization to land-use dynamics, historical urbanization, and urban development (Center for Advanced Spatial Analysis website). Therefore, CA is particularly useful in simulating complex adaptive systems such as people movement.

There has been a great deal of interest in studying traffic flow with Cellular Automata models. CA models are conceptually simple, thus we can use a set of simple CA rules to produce complex behavior. Using CAs we can capture the complexity of interacting traffic pattern behavior. The basic one-dimensional Cellular Automata model for highway traffic flow is described in (Nagel, K., and Rasmussen, S. 1994). The model describes a one-lane traffic road with sequence of grid points, and each grid point is a square representing one vehicle. There are many variations on the basic model (Blue, V.J. and Adler, J.L. 2000b) that consider the effects of acceleration and delay of vehicles with high speed. The actual speed of the car at each time step depends on the “lambda” value that can be adjusted accordingly. This model captures the realistic traffic situations where the car accelerates and decelerates.

The rules in (Blue, V.J. and Adler, J.L. 2000b) model are very simple, but we get complex behavior out of a population of these rules. This complexity is defined by methods in statistical physics. Such models lead us away from the view of multi-agent traffic models as fundamentally linear where units are treated in isolation, thus motivating us to look into combining agents and CA.

## CELLULAR PEDESTRIAN TRAFFIC SIMULATIONS

The one dimensional car traffic models motivates us to develop more complex models of movement. The area of pedestrian movement has been used as possible application field for the use of cellular automata (Blue, V. J. and Adler, J.L, 2000a ) and (Blue, V.J. and Adler, J.L. 2000b)

These models contain cellular entities that have a forward direction of movement and the idea is to optimize the speed of the agent in a given direction, under a maximum walk speed constraint. Each agent will account for the position of other agents and their direction of forward movement. In the simplest case we could have an environment where each agent is moving in the same direction as every other agent. The next increase in complexity involves flow where two types of agents in the population move in opposing directions. To further increase, complexity, consider moves that cover all possible local moves (Blue, V.J. and Adler, J.L. 2000b). In (Dijkstra, J., A.J. Jessurun, and H.J.P. Timmermans. 2001), using agents as an extension of CA was initially discussed, However, these agents are just extension of the movement rules in the CA, and do not have personable attributes we consider important in crowd behavior.

Most of these models are primarily based on CAs to understand emerging behavior of pedestrian's movement. We are interested in more than just movement models, but also pedestrian behavior models. For example, in a museum setting, each agent/pedestrian makes decisions on moving both on the CA rules defined in (Bruzzone, A. and Signorelle 1999) as well as other internal rules. These internal rules could include the, again for the museum example, the desire to linger at a particular exhibit. This individual pedestrian behavior affects the total crowd behavior in interesting ways.

As modeling of spatial systems improve and develop, systems can be modeled at finer and finer granularity, or scale. This means that activity can be represented in the model at various levels, (for example, at the individual entity in the systems). Understanding complex systems naturally mean injecting individual behavior into the gross systems. Therefore, individual mobility, and state are inevitably woven into the fabric of the complex system. Therefore, we look at developing models

of complex systems that combine cellular automata at the aggregate spatial level and add entity motion to the cells as agents. In this way, we can model gross movement rules (the CS grid rules) and particular individual/group rules (the agents).

## THE PROBLEM OF MODELING CROWD CONTROL

Crowd control can be applied to many different areas, such as police operations (Varner D., Scott D.R. Micheletti J., Aicella G. (1998)), shopping center design, public park re-organization (Bruzzone A.G., Rivarolo D. 1997), rail station re-engineering (Bouvier E., Cohen E. 1995) and epidemic diffusion. In additions, there has been a significant advancement in studying pedestrian traffic patterns in various environments. In (Bierlaire, M., Antonini, G. and Weber, M. 2003), a system based on multiple agents (MAS) was used. The desire of the author was to create a highly flexible system composed of actors that can be modeled individually. In (Helbing, D., and Molnar, Peter, 1998) pedestrian flow is discussed with some features observed. In (Still, K, 2000), the author presents several phenomena about crowd behavior, the most significant being:

- Edge effects
- Finger Effects
- Density Effects
- Shockwave Effects

In (I. Farkas, D. Helbing, T. Vicsek, 2002) the authors developed a model to investigate the Mexican Wave phenomena in a stadium.

To achieve some of the modeling desired by (Bierlaire, M., Antonini, G. and Weber, M. 2003) and the flexibility and simplicity by desired by In (Still, K, 2000), we combined CA and MAS into one system.

To facilitate the pedestrian motion rules for the CA, we combine the following from the discipline of physics:

- Fluid dynamics: This represents a continuous modeling of the crowd by using specific crowd characteristics (i.e. density, speed, etc.) distributed on a grid corresponding to environment in which the crowd is actively moving.

- Particle dynamics: In this approach, each entity in the crowd is represented as a single particle interacting with the other particles by some predetermined attraction/reaction and collisions (Langton C. , 1997) rules.
- Particle queuing: This approach operates in similar way to particle dynamics but doesn't consider attraction/reaction rules and instead stacks the people on queues taking care not to overlap the entities.

We model the individuals as a collection of agents. We also model the agents so that they interact among themselves to maintain consistent groups (i.e. a family or a specified group of people) when applicable. We capture Pedestrian decision making for the agent based on individuality or group membership. These attributes lead to certain decision making on the part of the agent. Decisions are based on Internal Forces (This is related to goal of each entity) and External Forces (This force is related to other entities, layout objects or external force fields and is composed of three parts: (I) collision avoidance, (II) group attraction and (III) external forces).

A relational schema of the above forces is depicted below. For the spatial aspect, consider the need to avoid obstacles (either walls or other individuals) and the need for maintaining the group structure where appropriate.

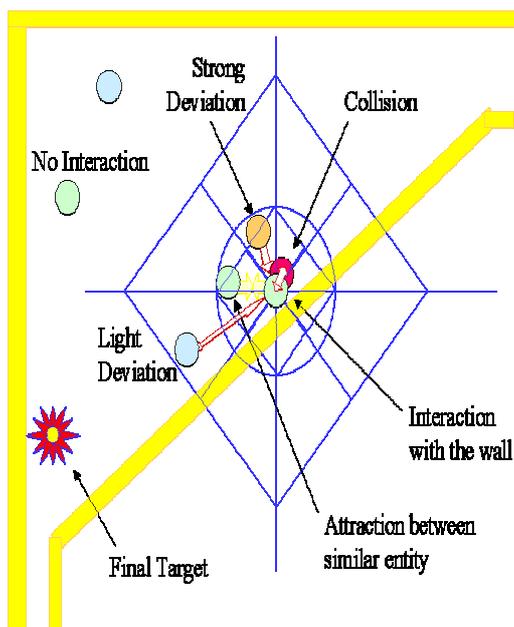


Figure 1: Spatial Relationship of Forces

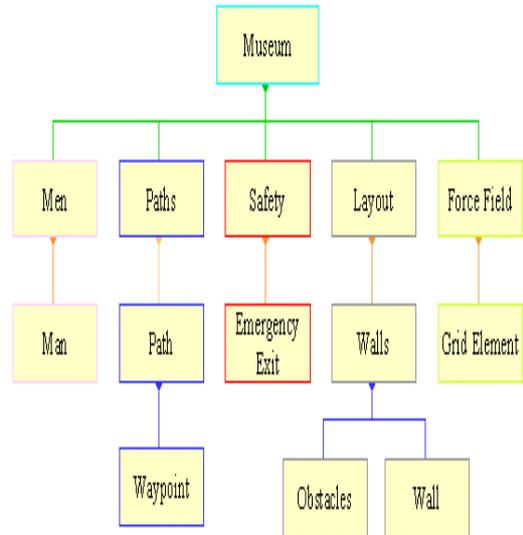


Figure 2. General architecture for objects in the model.

The general architecture of the entities is proposed in Figure 2 above. The figure demonstrates the hierarchical nature of our design. For example, we have an object that is the group of men subject to some specific methods (init, drawing, and alarm for a group of men, changing the behavior of the men instead of the walls or of the attractions).

In Figure 3 below, we see the layout in the application. As you can see, the goal of the individuals, or groups, is to visit the objects in the exhibition. The individuals travel along the two floors (one lower and one upper). There are also emergency exits. We focused, in this paper, primarily on the movement of the individuals/groups among the objects in the exhibit only, and not on the emergency exits. Monitoring agent behavior during emergency exiting will be added to our next version of the system. After viewing the exhibit, the individuals/groups leave.

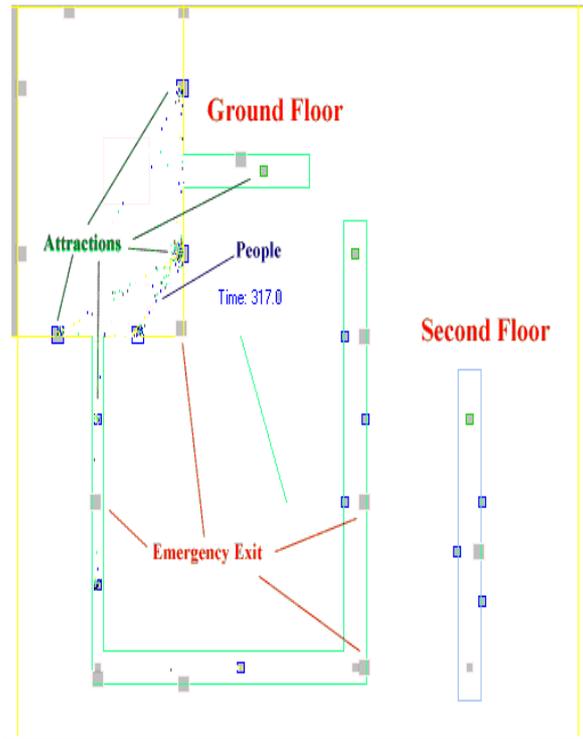


Figure 3: Layout of the environment for the CA and MAS crowd control system.

## SIMULATIONS AND RESULTS

Complex Systems is a field of science studying how parts of a system give rise to the collective behaviors of the system, and how the system interacts with its environment. The field of complex systems cuts across all traditional disciplines of science. It focuses on certain questions about parts, wholes and relationships. The study of complex systems is about understanding indirect effects and observing emergent behavior from these systems. Pushing on a complex system "here" often has effects "over there" because the parts are interdependent. (New England Complex Systems Institute website)

Based on some prior work (Bruzzone, A. and Signorile 1999) with the addition of CA and MAS, we ran simulations to observe the behavior of the agents in a museum. Our agents are modeled (randomly) as individuals, family groups (small group of individuals) and tourist groups (larger groups of individuals). The individual agents have the simple goal of moving through the exhibition, viewing the objects on display and leaving the exhibition. The other two types of agents have the

added goal of maintaining "neighborhood" contact with their group.

We have situations where individuals/groups of agents linger over some specific popular exhibit, where agents move from one room to another then back again, and where there are areas for agents to rest/contemplate exhibits. Some emergent behavior from our simulations where:

- Reverse Edge effects are observed (especially on the edge closest to the exhibit). Thus, this suggests that ample viewing space along the sides is more important than tunnels for the center of the crowd. This is even more prevalent when the percentage of the agents is groups.
- Finger effects (bi-directional crowds moving amongst themselves) are less evident in this environment. However, when individuals are present, and bottlenecks occur, we observed the "wandering" effect for the individuals to be very prevalent.
- Back pressure (especially when the density of groups is large) is observed. As groups approach an obvious stopping point (descriptions on the wall or a particularly popular object), the back pressure increases precipitously. When we moved such stopping points, to either larger viewing areas or areas with more egresses, the back pressure dropped. In addition, the local Density effect was reduced
- Placement of resting areas can adversely affect density in the crowd. If placed too near a popular stopping point, a clustering effect takes control of the crowd, causing severe back pressure.

## CONCLUSIONS AND FUTURE WORK

The study demonstrates the effectiveness of modeling public facilities using both Cellular automata (more basic/complex moving rules) and agents (for behavior attributes) to improve services in a museum setting. Clearly this framework is extendable to modeling many events, such as sporting events, shopping malls, rail stations, airports, etc.

Additionally, we feel that this approach is applicable to other complex systems such as network reliability, supply chain management, and management logistics. Mostly due to the nature of this theory: massive similar entities, internal and external forces to motivate these entities, non-linear interactions between entities and emergent behavior from the system.

We plan to investigate using this theory in the above and other domains.

## REFERENCES

- Bierlaire, M., Antonini, G. and Weber, M. (2003), "Behavioral Dynamics for Pedestrians", Moving Through the Nets: The Physical and Social Dimensions of Travel, K. Axhausen, editor.
- Blue, V. J. and Adler, J.L. (2000a). "Cellular Automata Model Of Emergent Collective Bi-Directional Pedestrian Dynamics." In: Artificial Life VII, Bedau, M.A, McCaskill, J.S, Packard, N.H and Rasmussen, S (MIT Press, 2000)
- Blue, V.J. and Adler, J.L. (2000b) Modeling Four-Directional Pedestrian Movements. The Transportation Research Record, Journal of the Transportation Research Board.
- Bouvier E., Cohen E. (1995) "Simulation of Human Flow with Particle Systems", Proceeding of Simulators International XII, Phoenix, AZ, April 9-13
- Bruzzone A.G., Rivarolo D. (1997) "Virtual Reality as the Starting Point for the Functional Analysis of Public Plant Design Services", Proceedings of High Performance Computing, HPC'97, Atlanta, Georgia, April 6-10
- Bruzzone, A. and Signorile, R. "Crowd Control Simulation in a Java Based Environment", Proceedings of the International Conference on Web-Based Modeling and Simulation, San Francisco, CA, January 1999
- Dijkstra, J., A.J. Jessurun, and H.J.P. Timmermans. 2001. "A Multi-Agent Cellular Automata Model of Pedestrian Movement." In: Pedestrian and Evacuation Dynamics. , M. Schreckenberg and S.D. Sharma, editors, Springer-Verlag, Berlin.pp. 173-181.

I. Farkas, D. Helbing, T. Vicsek,  
Mexican waves in an excitable medium.  
*Nature* 419, 131 (2002).

Helbing, D., and Molnar, Peter, Social Force for Model for Pedestrian Dynamics, 1998

Langton C. Artificial Life an Overview, 1997, MIT Press, Cambridge, MA

Nagel, K., and Rasmussen, S. (1994). "Traffic at the edge of chaos". In : Artificial Life IV, R.A. Brooks and P.Maes (eds.), 222-230. MIT Press (Cambridge, MA, 1994).

Still, K. Crowd Dynamics, 2000

Varner D., Scott D.R. Micheletti J., Aicella G. (1998) "USMC Small Unit Leader Non-Lethal Trainer (SULNT)", Proc. of ITEC98, Lausanne, 28-30 April

New England Complex Systems Institute,  
<http://www.necsi.org/guide/whatis.html>

Center for advanced spatial analysis,  
<http://www.casa.ucl.ac.uk/news/index.htm>

## BIOGRAPHY



Robert Signorile is an Associate Professor in the Computer Science Department of Boston College. His research interests include multimodal simulation, simulation in business, networks and distributed computing. He has published regularly in applied simulation, simulation methodology, distributed systems and networks.

# A DISTRIBUTED SIMULATION APPROACH FOR PROJECT LOGISTICS, MANAGEMENT, AND CONTROL

Romeo Bandinelli  
Dipartimento di Energetica "Sergio Stecco"  
Università di Firenze  
romeo.bandinelli@siti.de.unifi.it

Alessandra Orsoni  
School of Business Information Management  
Kingston University  
KT2 7LB, Kingston upon Thames, UK  
a.orsoni@kingston.ac.uk

## KEYWORDS

Project Logistics, Project Management and Control, Concurrent Engineering, Distributed Simulation.

## ABSTRACT

This paper proposes a distributed simulation approach to address the logistic implications of design and assess their impact on performance during project execution. Relevant areas of application include large construction projects involving the installation of multiple systems and, simultaneously, the interests of multiple parties. Individual models of the installation activities pertaining to each system are built using commercial simulation packages. These models are then linked using the Java RMI® (Remote Method Invocation) standard to provide global measures of project performance for both local and central design decisions. Several uses of this simulation approach can be identified during the different phases of the project life-cycle, for instance to test alternative solutions during the conceptual design phase or to accommodate late design changes with minimal disruptions to the project schedule and budget.

## INTRODUCTION

Distributed simulation has been widely used outside the original military context to address important production and management problems related to the supply chain. This paper extends the use of distributed simulation from mass/batch production to large one-off projects. Examples of suitable projects include for instance the construction of large industrial and occupied facilities. Because such projects involve the interests of multiple parties, it is of the highest importance that any design decisions made either centrally (i.e. at the owner's or at the general contractor's level) or locally (i.e. at the sub-contractor's level) are assessed and discussed using common metrics [1,2]. This ensures that the logistic and performance implications of the relevant design specification are fully understood and appreciated prior to their full-scale implementation. Specifically, the simulation approach aims to establish unbiased criteria and measures for the evaluation of decision alternatives to bridge the gaps among the objectives of the different project stakeholders. By these means distributed simulation supports the application of a concurrent engineering approach as potential issues in the project logistics,

management, and control can be anticipated and addressed since the early stages of design.

## BACKGROUND

The work presented in this paper builds upon prior research on the performance of complex large-scale projects [3,4]. Modelling and simulation techniques, in the form of dynamic process simulation models, were designed to assess the performance implications of change in the presence of inter-system process dependencies. The initial solution worked as a single application package and was meant for the use of the general contractor to evaluate how changes introduced in the design and/or in the construction means or methods could influence project performance, as measured for instance by project duration and cost [3,4]. The implemented simulation tool consisted of a library of system and material specific modules, each one representing the detailed installation/construction activities for a particular system.

Since the early applications of the tool it was observed that each system and material and specific model is better maintained and customised at the sub-contractor's site: the detailed specifications of possible changes are usually implemented directly by the sub-contractors, who are also free to choose the construction means and methods as long as they are able to meet the specified completion deadlines. Based on such considerations, a natural extension of the work involved the realisation of a network of independent system and material specific models based on the Java RMI® (Remote Method Invocation) standard. The paper will discuss the relevant features of the network and will present the results of a case study based on an actual construction project.

## SIMULATION APPROACH

As illustrated in the previous section, a prototype simulation tool was first developed in the SIMPROCESS® environment as a library of system and material specific modules, which enabled the definition of a project by simply dragging-and-dropping the relevant modules. For the purposes of this research the existing modules were extracted and readapted to serve as simulation federates in the distributed architecture. Special attention was devoted the definition of appropriate variables tracking federation-wide the local

progress for the individual systems to ensure that the next installation phases could be scheduled and executed without introducing simulation delays. This is especially important when dealing with the representation of multiple interdependent processes, where technical logical, regulatory and resources constraints tie the relative production rates among the systems. Specifically, the completion of one installation process in a particular zone of the facility, allows for the next crew to start their job in that same zone. If the completion of the previous job is not effectively communicated, the installation process for the next system may be delayed in a way that is not representative of the reality of the process.

The built-in capabilities of SIMPROCESS®, which enable the adoption of the RMI standard greatly simplify the realisation of a network and overcome several of the problems related to the timing of simulation synchronisation. Recent research by the authors has focused on the timing of federation synchronisation, when adopting the “next event” approach within the HLA-RTI standard [5,6]. In such applications it is critical that the frequency of communications among simulators is as reflective as possible of the process characteristics so as to minimise the waste of simulation time caused by the occurrence of asynchronous events.

With the RMI-based approach to distributed simulation, the timing of communication is no longer a problem. A detailed description of the RMI features and functionalities can be found in the relevant website [<http://java.sun.com/products/jdk/rmi/>]. In synthesis, RMI® enables the programmer to create distributed Java technology-based applications, in which the methods of remote Java objects can be invoked from other Java virtual machines, possibly on different hosts. A Java program can make a call on a remote object once it obtains a reference to the remote object itself, either by looking up the remote object in the bootstrap naming service provided by RMI or by receiving the reference to it as an argument or a return value. A client can call a remote object in a server, and that server can also be a client of other remote objects. RMI uses object serialisation to marshal and unmarshal parameters and supports true object-oriented polymorphism.

Since the study represented a first application of the distributed approach to the analysis of the logistic implications of early design decisions in the performance of large projects, significant effort was devoted to the validation of both the individual models and the federation as a whole. The validation of the individual models was highly supported by the availability of historical data from previous projects that enabled cross-checking and validation of the simulated time and performance estimates. The availability of data from past projects allowed for the analogical validation of the entire federation as well, by comparing the

performance measures obtained, such as duration and costs, to their actual values. In effects, project managers typically refer to previous projects to estimate the duration and costs for each aggregate project phase and during project execution they record the actual values for project control purposes. It is important to observe that by looking at processes at the aggregate level (e.g. installation of electrical wiring on the second level, or installation of plumbing fixtures in zone A) it is only possible to derive rough estimates for the start/end dates of each phase. The representation of the logistic implications of design and methods during project execution is far more detailed in the simulation because inter-system process links build their effects at the detailed task and component level and these links cannot be captured by looking at processes at the aggregate level.

### **ASSESSING THE PERFORMANCE OF DESIGN-DRIVEN LOGISTIC CHANGES**

The introduction of design and technological changes impacts the logistics of project execution at three levels: the system, the inter-system, and the whole project level. Specifically, the system level observes the logistic implications of change within the system of introduction. The inter-system level tracks how their secondary effects influence the installation of other facility systems. The whole project level observes their impact on the performance of the overall project. Prior research, based on extensive simulated scenario testing, has demonstrated that the impact of change can be accurately tracked at all three levels, across multiple dimensions of performance [3,4].

Previous studies have shown that the performance of large construction projects is multi-attribute[3,4]. Because the interests of multiple parties are simultaneously involved (e.g. the owner's, the general contractor's and the different specialty contractors'), depending on the particular project and party of perspective, one aspect of performance may appear more important than others towards the success of the project. However, in general, the level of success achieved in a project can only be measured across multiple dimensions of performance [7,8]. Based on these considerations, and on previous testing of the individual models as stand-alone applications, a number of performance measures were selected as suitable indicators of project performance. These include project duration, duration-based cost, cost of utilised resources, percentage resources utilisation, and index of workers' exposure to dangerous conditions (danger index). Specifically, the duration-based cost represents the total cost of the project, assuming that all resources are present on the construction site for the entire duration of their scope of activity within the project. The cost of utilized resources represents the cost of performing project activities and tasks, excluding resource costs of delays and wait times introduced by process interdependencies. The percentage of resource

utilization is the ratio of these two costs. Worker exposure to dangerous condition is measured through an index that builds upon tabulated values of occurrence of injuries during the performance of specific construction tasks [9,10] over the entire duration of the project.

It is important to observe that the choice of performance measures targets the assessment of the impact of inter-system process dependencies. Measures such as duration and duration-based cost in fact are “dynamic” measures of performance, meaning that they account for the actuality of the duration and cost of the construction process because they reflect inter-system process dynamics. Measures of performance such as cost of utilized resources and workers exposure to dangerous conditions, are not dynamic in such a sense. They are direct functions of the number of man hours required to complete the project and do not account for the resource idle time determined by inter-system process dynamics. Any discrepancies in the simulated results for these two sets of performance measures is entirely explained by the effects of inter-system process links.

### **EXAMPLE ANALYSIS**

In order to assess the benefits that may be accrued from the application of the implemented federation architecture to the study of complex projects, an example analysis was performed based on data from an actual construction project. The project involved the realization of a research facility for a large pharmaceutical company, to accommodate both laboratories and office spaces. The study involved the evaluation of design alternatives for the main service systems included in the facility: the electrical and communication systems, the plumbing and fire protection systems, and the heating and ventilation systems. As part of this case study, the simulation tool was used to compare the performance of two pairs of design alternatives during project execution. The first analysis focused on layout alternatives and compared the performance of the centralised and decentralised layout options. The second one focused on system alternatives and compared the performance of air-based versus water based heating.

For the purposes of this analysis the federation had to be customised to accommodate and track the relevant changes in project logistics induced by each design alternative. Typical changes in project logistics are related to changes in the bill of materials for the related sub-systems and to their spatial distribution in the facility. The logistic implications of these changes include the introduction of additional accessibility constraints, which influence the spatial and temporal allocation of crews to different jobs and the cross-system sequencing of installation activities. These combined effects influence the relative production rates among the systems and thereby generate an impact on the overall project performance. Changes in the bill of materials for a particular sub-system involve changes in

the input data for each federate, however their impact on project performance is tracked at the federation level by observing the corresponding changes in the relative installation rates among the systems. Progress is especially difficult to track among different systems because different systems use different spatial units of progress. For instance structural systems use bays, enclosure systems may refer to zones or levels, interior finishing can be tracked by room, and services may be tracked by riser/feeder group, by level, or by zone, depending on the specified facility layout.

### **CENTRALISED VS. DECENTRALISED LAYOUT**

Design changes in the system layout, particularly centralisation versus decentralisation of the vertical risers, were analysed with respect to the overall layout of the water-based service systems (i.e. plumbing, heating and fire protection). Centralising the layout means shifting from a primarily vertical layout, with a number of vertical risers close to the number of usage points, to a largely horizontal layout, where few risers feed the usage points through a long network of horizontal pipes on each level.

Not only does this shift represents a significant design change in terms of number and type of units to be installed, but it also introduces an additional logistic constraint that links the installation processes of the fire protection and plumbing systems. This constraint, driven by spatial and accessibility requirements, ties the rate of installation of the horizontal pipes for the two systems, because the two need to run parallel on the different facility levels and share the same supporting trays.

Moving from a decentralized to a centralized layout has important impacts at the federation control level. The progress status for the installation of the plumbing pipes needs to be tracked at the federation level, and synchronisation events need to be tailored to account for their expected installation time, so that the fire protection units can be timely released for installation on that level.

Changes in layout have no impact on the structure of the individual system federates. The modules are system and material specific, and thus independent of the particular system size and layout. Changes in the number of units to be installed do not influence the type of activities to be performed and, thus, are only reflected in the input quantities (i.e. bill of materials).

The relative changes in project performance are displayed in figure 1. The effects of centralisation measured at the whole project level include major increases in overall project duration (57%) and in duration-based cost (36%). In addition, it is now the installation of the fire protection system, instead of the electrical system, that drives the duration and overall cost of the project. The change in design (number and

type of units to be installed) actually reduces the cost of utilised resources, both at the system and at the whole project level, but also introduces changes in the inter-system installation rates, which impact the overall progress rate and project duration. The combined impact of these two effects is a reduction of resource utilisation (i.e. the cost of utilized resources divided by the duration-based cost) equal to 32%.

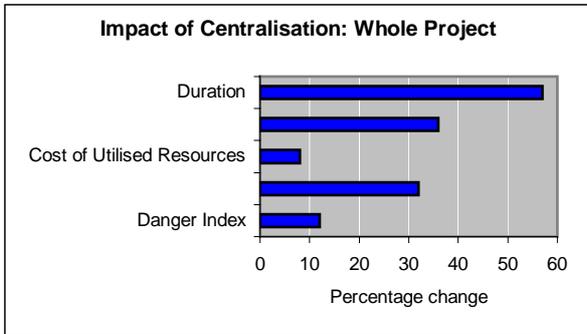


Figure 1: Performance Impact at the Project Level

One of the benefits of a distributed approach in the study of change is the possibility of analysing performance at the system level by looking at the outcomes of each individual federate, when run in the context of the whole federation. In most cases the analysis of the federation outcomes at the system level produces results that are quite far from the expected and can only be explained as the result of inter-system process dependencies. For instance, in this scenario, no change was introduced in the design of the electrical system, which for this project application was fairly centralised to begin with, therefore both the cost of utilised resources and the danger index, which are system specific measures, do not change. However, a decrease in duration-based cost is observed which in turn leads to an increase of resources utilisation. The decrease in duration-based cost for the electrical system is explained in terms of inter-system process links, which constrain the spatial progress on the installation of the heating system to depend on the rate of installation of the heating system.

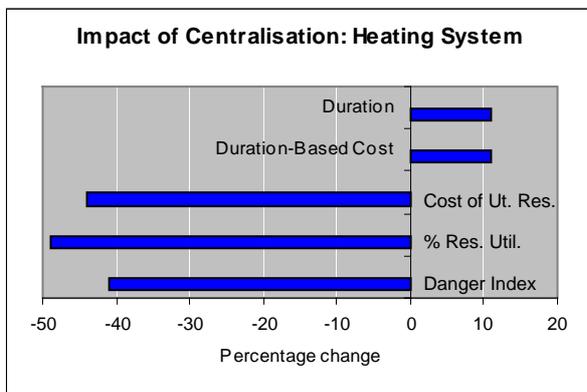


Figure 2: Performance Impact at the System Level

Because the shift to a centralised layout for the heating system makes its installation faster, performance benefits can be observed in the installation of the electrical system due to a reduction in resources idle time (resources are idle while waiting for the installation of the heating pipes in the next zone). Figures 2 and 3 summarise these performance outcomes at the system (heating system) and inter-system (electrical system) levels, respectively.

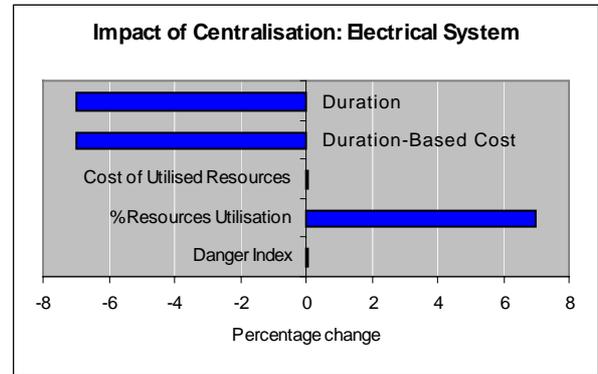


Figure 3: Performance Impact at the Inter-System Level

### AIR-BASED VS. WATER-BASED HEATING

A comparison is made between a water-based and air-based heating solution for a same decentralised layout. The shift from a water based to an air-based system represents a shift from a closed loop type of system, characterised by supply and return pipes, to an open loop type of system, characterised by supply ducts only. The most interesting aspect of this second case is that no significant impacts can be identified at the system level: the time required to install a given length of air ducts is actually longer than the time required to install an equal length of water pipes, however no return line is required in the air-based configuration. For this particular system size and layout, the effects compensate so as to produce negligible changes in the performance measures at the system level. However, important impacts can be observed at the intersystem level: the installation of the electrical system is faster and the associated costs are lower. This effect is mostly determined by the different rate of installation of both the vertical and the horizontal units in the air-based system, as compared to the water-based solution, which overall increases installation efficiency for the electrical system (shorter idle time of resources while waiting for the horizontal heating conduits to be placed). Increased efficiency in the installation of the electrical system has significant impacts on the project as a whole, since any reduction in completion time for the electrical system directly translates in an equivalent reduction in project duration, and consequently decreases project cost.

For the purposes of this comparison the simulation federate representing the installation process for the

water-based heating system had to be substituted with the corresponding air-based one, because the installation activities are substantially different. However, no relevant changes had to be made at the federation control level, since the spatial constraint that links the installation progress of the electrical system to that of the heating system does not vary as the system type is changed (i.e. the air ducts still need to be in place on one level before the installation of electrical conduits and wiring can start on that same level).

As anticipated above, the impacts of this change in the nature of the heating system at the whole project level are a reduction in project duration (-7%) and a corresponding reduction in duration-based cost (-2%). Changes in the cost of utilised resources and in the danger index for the whole project are negligible. Only a slight increase in the percentage resource utilization can be observed (+3%). Figure 4 summarises the relative performance results at the whole project level.

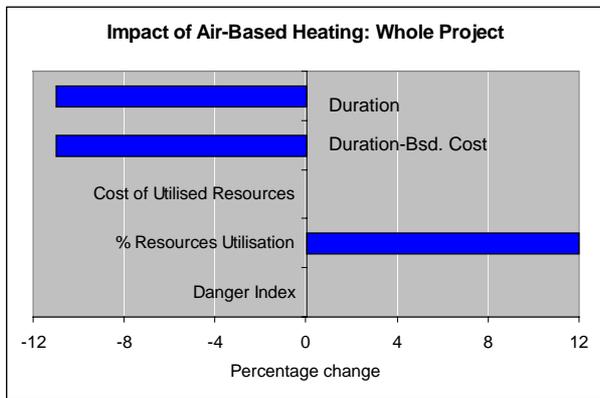


Figure 4: Performance Impact at the Project Level

At the system level, no significant impacts can be observed in the installation of the heating system. As explained at the beginning of this section this finding is rather coincidental, and results from the combination of two effects. The first one is that for this particular layout the total number of man-hours required to install the air-based and the water-based heating are approximately the same. The second one is that the activities required to install air ducts are characterised by the same level of danger as those required to install heating pipes.

At the inter-system level the installation of the electrical system is affected by the change in the type of heating system adopted. Duration and duration-based cost for the electrical system are both lower than the corresponding figures for the water-based heating alternative (-11% for both).

Since no change was introduced specifically in the electrical system itself, both the cost of utilised resources and the danger index remains the same in the two configurations, while the percentage of resource utilization increases (+12%), due to the decrease in

installation time. Figure 5 summarises the relative performance results observed at the inter-system level

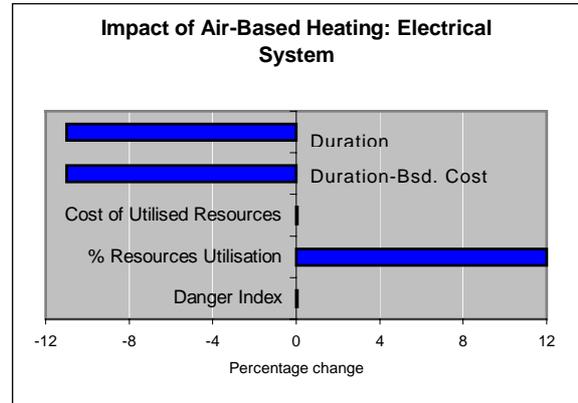


Figure 5: Performance Impact at the Inter-System Level

## CONCLUSION

This paper has analysed the logistic implications of design changes during project execution and their impact on project performance. The analysis of change in complex construction projects requires a distributed simulation approach because many of the logistic implications of change can only be specified at the intersystem level, and only at this level they build their effects on project performance.

A distributed architecture is especially convenient because important construction projects involve multiple sub-contractors who are tied to specified completion deadlines but are otherwise free to choose their own construction means and methods. In this respect it is sensible to have them implement the customisation of their specialty federate models for the project and then observe their effects of their choices at the aggregate level by centrally running the federation. Design changes can be proposed and tested both locally, at the subcontractor's level or centrally are the owner's/general contractor's level, but it is always necessary to test their effects on the performance of the project as a whole. Distributed simulation supports a concurrent engineering approach throughout the different phases of a project's lifecycle, from the early design stages and feasibility studies to the later stage of project execution. The simulation-based methodology presented in the paper favours communication among the different project stakeholders and provides quantitative grounds for discussion before key design decision are made and implemented in an application context where full-scale experiments are financially too risky and time-consuming. The methodology also provides extensive support during project execution to better specify the logistics by which changes can be implemented at such late stages without compromising the existing project schedule and budget.

## REFERENCES

- [1] C.B. Tatum, Integrating Design and Construction to Improve Project Performance, *Project Management Journal*, 21 (2), 1990, 35-42.
- [2] J.T. O'Connor, & R.L. Tucker, Industrial Project Constructability Improvement, *ASCE Journal of Construction Engineering Management*, 117(2), 1986, 259-278.
- [3] A. Orsoni, Assessment of organizational alternatives in complex large scale projects, *Proc. 2001 Summer Computer Simulation Conference (SCSC 2001)*, Orlando, FL, 2001a, 611-616.
- [4] A. Orsoni, Selecting the right technology: costs and benefits of dynamic multi-process simulation in complex, large-scale applications, *Proc. 5<sup>th</sup> World Multiconference on Systemics Cybernetics and Informatics (SCI 2001)*, Orlando, FL, 2001b, 509-514.
- [5] R. Bandinelli, M. Iacono, and A. Orsoni, Anticipating critical events to customise and improve the performance of federation runtime infrastructure, *Proc. 7<sup>th</sup> UKSim Conf., Oxford*, UK, 2004, 59-64.
- [6] R. Bandinelli, M. Rapaccini, S. Terzi, & M. Macchi, Proposal for a framework for production plants remote control: a preliminary test case, *Proc. 15<sup>th</sup> ESS Conf., Delft*, NL, 2003, 343-348.
- [7] C. Hendrickson, & R. Au, *Project Management for Construction* (Englewood Cliffs, New Jersey: Prentice Hall, 1989).
- [8] F.W. Mueller, *Integrated Cost and Schedule Control for Construction* (New York, NY: Van Nostrand Reinhold Co, 1986).
- [9] Occupational Health and Safety Administration. Construction accidents – the worker's compensation database 1985-1988. 1992. U. S. Department of Labor.
- [10] Occupational Health and Safety Administration. Construction industry: OSHA safety and health standards. 1994. U. S. Department of Labor.

## AUTHOR BIOGRAPHIES

**ROMEO BANDINELLI** received his laurea degree in Mechanical Engineering from the University of Florence in 2002 discussing the thesis "Remote Factory Control with Distributed Simulation". He is currently a PhD candidate in the Department of Energetic, Plants and Industrial Technologies at the University of Florence. His research interests include parallel and distributed simulation for industrial and supply chain applications, ICT, and business process re-engineering.

**ALESSANDRA ORSONI** is currently a senior lecturer in the School of Business Information Management at Kingston University (Kingston, UK). She received both her MS in Mechanical Engineering and her ScD in

Engineering Systems Design and Innovation from the Massachusetts Institute of Technology (Cambridge, MA). Prior to joining Kingston University she was a research associate in the Department of Materials Science and Metallurgy at the University of Cambridge (Cambridge, UK) and worked as an assistant professor at in the Department of Production Engineering at the University of Genova (Genova, Italy). Her research interests include modelling, simulation, and AI techniques applied to the design and management of complex industrial systems.



# **DISCRETE SIMULATION LANGUAGES AND TOOLS**



# BIOLOGICALLY INSPIRED DISCRETE EVENT NETWORK MODELING

Ahmet Zengin\*, Hessam Sarjoughian<sup>†</sup>, Huseyin Ekiz\*

\* Technical Education Faculty  
Department of Computer Science Education  
Sakarya University  
Esentepe / Sakarya, TURKEY 54187  
{azengin, ekiz}@sakarya.edu.tr

<sup>†</sup> Arizona Center for Modeling & Simulation  
Computer Science & Engineering Dept.  
Arizona State University  
Tempe, Arizona, USA  
sarjoughian@asu.edu

## KEYWORDS

Beehive, DEVS, Networks, Routing, Scalability.

## ABSTRACT

The simulation study of networks remains attractive due to desire to achieve better important traits such as scalability and performance. This paper describes a biologically inspired discrete-event modelling approach for simulating networks. It introduces a synergistic modelling approach by incorporating key attributes of honeybees and their societal properties into a set of simulation models described in the Discrete Event System Specification. We describe our approach with particular emphasize on how to model the behaviour of the honeybees and their cooperation as discrete event models. The simulation models and their experimental results are presented and discussed.

## 1. INTRODUCTION

The study of complex networked systems especially those that are large-scale are attractive for a variety of reasons such as the analysis and design of transportation systems, supply networks, management of social and ecological systems. Systems that are composed of components share common characteristics such as hierarchy, alternative configurations, patterns of interactions due to varying types of components and behaviour. To model such systems, one can employ a variety of methods to characterize structure and behaviour – e.g., we can use communicating processes or event systems as the basis for modelling components and their interactions. More specifically, from simulation point of view, we may use discrete event, differential equations, or cellular automata to describe behaviour of networks.

Some modelling techniques, not only can be developed based on *artificial computational models* such as *Von Neumann*, but also on *natural phenomena* such as *Ants societies*. One of the advantages of complementing artificial models of computation with their natural counterparts is that we can have a rich laboratory to develop models that can be experimented with. For example, developing biologically inspired models such as Honeybee enjoy from ample scientific and experimental

studies developed based on studying details of the honeybees and their colonies. The knowledge about how honeybees behave and interact offers key insights as how to model inherent complexity of networked systems. Armed with simple yet subtle emerging behaviour of honeybees we may be able to develop models of complex large-scale network systems that can offer desirable performance and scalability qualities (Lunceford and Page 2002).

In the remainder of the paper, we will review Honeybee (Seeley 1995) and DEVS modelling techniques (Zeigler et al. 2000). Based on a general model of Honeybee society, we describe our approach to network modelling where artificial and natural computational models are combined. We then present network model specifications using DEVSJAVA (ACIMS 2004) and associated algorithms followed by describing simulation results and conclusions.

## 2. BACKGROUND

There exist many modelling approaches founded on systems theory (e.g., Mesarovic and Takahara 1989), agent theory (e.g., Wooldridge and Jennings 1995), and object theory (e.g., Abadi and Cardelli 1996) to characterize network systems such as computer networks and natural societies such as honeybees and ants.

### Discrete Event System Specification

Discrete event systems can be described using the Discrete Event Systems Specification (DEVS) formalism (Zeigler et al. 2000) where model behaviour is characterized as events and their processing. This modelling approach supports hierarchical modular model construction, distributed execution, and therefore affords a basis to characterize complex, large-scale systems using formulation of components (atomic and coupled models) and their interactions. *Atomic models* characterize structure and behaviour of individual components via inputs, outputs, states and functions. The internal, external, confluent, output and time advance functions define a component's behaviour over time. Internal and external transition functions describe autonomous behaviour and response to external stimuli, respectively. Confluent transition function is used to account for concurrent occurrences of internal and external transition

functions. Time advance function represents passage of time. Output function is used to generate outputs. Atomic models can be coupled together in a well-defined manner to form more complex models.

A coupled model specifies constructs for composing modular models into hierarchical structures. Behaviour of a coupled model is defined by its constituent atomic (and/or coupled) models. With closure under coupling feature of DEVS, coupled models can be used as atomic models in a larger model. Coupled models can be constructed systematically using the concepts of ports and couplings. When a component sends messages via its output ports, the couplings relay the messages to their designated input ports (Wymore 1993). Upon receipt of messages by atomic models, they immediately process these messages which may result in new states and generation of outputs.

Parallel DEVS is capable of processing multiple input events and provides local control for handling of simultaneous internal and external events. DEVS atomic and coupled models have computational counterparts which may be executed in parallel manner using software engineering concepts (Sarjoughian and Singh 2004). DEVSJAVA is an implementation of the DEVS formalism and its associated simulation protocol (ACIMS 2004). There exist various implementations of the discrete event system specification approach based on single and multiprocessor environments. Parallel and distributed environments have been developed using technologies such as HLA (ACIMS 2004).

### Agents

Agent based approaches are being widely used in distributed network applications. This research area is one of the most attractive and rapidly evolving software technologies of the last decades. A software agent concept has emerged from a specialized class of distributed artificial intelligence and is used to describe the concept of a software entity that automates some of the tasks (Hayzelden and Bigham 1998). Software agents can be defined as autonomous, proactive and reactive computational entities that can exhibit the ability to learn, cooperate and move. To make use of software agents in network management applications, agents must be able to migrate from node to node in network. Furthermore, agents must be able to create new agents, delete themselves, and determine their interaction with their environment (e.g., Uhrmacher et al. 2001).

Agent-based solutions are suitable for management of distributed systems since they are inherently distributed and decentralized (Minar et al. 1999). Decentralization is an efficient way to overcome scalability issues. Network systems are dynamic and highly unpredictable systems and therefore their control and monitoring cannot be readily centralized.

### Honeybees

Biologically inspired modelling approaches have imported some metaphors from biological systems to engineering systems to develop network management frameworks. Particularly distributed, parallel, robustness and fault tolerant nature of some social insects (ants, bees, and termites) have been a source of inspiration for researchers due to their desirable distributed characteristics. Insect societies have advanced mechanisms to maintain colony level survivability against environment conditions. For example, because nectar availability can change rapidly and unpredictably, honeybees are able to cope with such problems and scale themselves to huge number of population (Seeley 1995). Honeybees have sophisticated regulation mechanisms to adapt their capacities against fluctuating and ephemeral resources.

Foraging behaviour in honeybees is a good example to investigate social insect metaphors such as *self-organization*. *Honeybees* collectively decide selection of nectar and pollen resources and allocation of workers among them through self-organization. This selection and allocation of processes among honeybees in their hive is performed by absence of any central authority. In a decentralized and concurrent way, each bee obeys a set of simple rules based on some metrics (e.g., nectar concentration, location of the source, and travel time to the food source). All of the metrics including parameters such as the number of bees responsible for storing food in the hive determine profitability of a nectar source. If colony encounters more than one source of nectar, highest profitable source is preferred by foragers relative to other sources with less profitability. Foragers are distributed among nectar sources using profitability criterion during the course of nectar collecting process. If nectar amount in a certain source changes, then whole colony changes its concerns to that source. Furthermore, the colony deploys certain portion of foragers for searching nectar, namely scouts. Rich sources are found by scouts and nectar availability in environments is monitored by them (Anderson 2001). This assignment of foragers to sources according to profitability is called *scout-recruit system* in honeybees. One of the most well-known mathematical models was developed by Seeley and documented in his Book (Seeley 1995).

### 3. NETWORK MODELLING APPROACH

In order to model a distributed networked system, we have defined a set of *network component models* called *nodes* which communicate with one another via *links* (see Figures 1 and 4). Using node and link capacity assignments, we can develop a variety of complex network configurations. The node and link models are defined as the DEVS atomic components. Other network elements such as packets and scouts are also represented as DEVS models. With this approach, a network model exhibits agent-like behaviour and thus supports decentralized control.

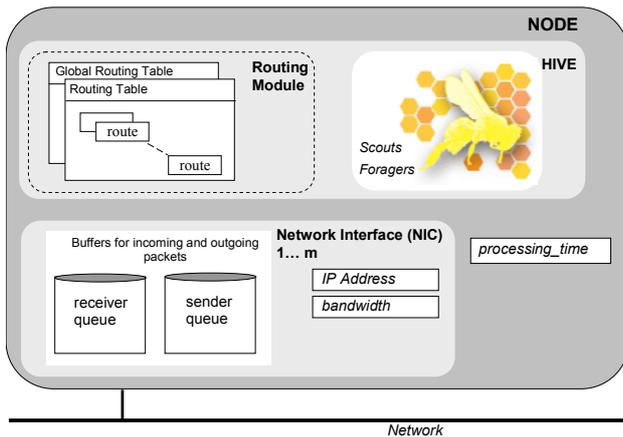


Figure 1: Design of a network node

To realize simulation experiments we utilize the concept of experimental frame where the conditions under which a model can be experimented with and observed are defined. For this work, a typical experimental frame consists of *generator* and *transducer* atomic models. We employ generator in order to create network traffic and to schedule special events such as unavailability of links or nodes. To realize this, a generator model sends messages (see Section 4) to all the appropriate components in the network.

Atomic and coupled models are represented using the parallel DEVS formalism and developed within the DEVSJAVA modelling and simulation environment. In this approach, the dynamics specified within the nodes and links can be used to determine the behaviour (e.g., throughput time) of the network model.

#### 4. NETWORK MODEL DESIGN

In order for modelling a distributed networked system, we have defined a set of basic network simulation model components including nodes and links as detailed next. By coupling these model components in DEVSJAVA, we can develop a variety of network configurations and study network characteristics. Since it is assumed that only nodes and links of a network are able to cause bottleneck, they are modelled as *atomic models* and only their states as well as input and output variables are of interest. Other network components such as packets and routing tables are realised as stateless entities. Network itself is a coupled model. Defined dynamics in node and link atomic models determine the behaviour of network coupled model. All atomic models in our implementation are modelled and defined using the Parallel DEVS formalism (Chow 1996) and realisation in Java (ACIMS 2004).

##### Node

Each node in the network represents a switching unit where it is able to process packets that are described below. Due to a node can be considered as a router. With simple changes, a node can represent other network elements such as hub. Nodes are connected to other

nodes called neighbours via links. To determine the behaviour of a node, we use two parameters: *packet process speed* which directly influences processing time of a node, and *queue* in which incoming and outgoing packets are stored. By toggling these capacities, different kinds of bottlenecks in the network can be modelled.

As shown in Figure 1, one of the main parts of our nodal structure is the *Network Interface*. It provides the fundamental internetworking services such as packet exchanging with neighbouring nodes. *Routing Module* reflects node's routing capability and simple intelligence. At each node, packets are forwarded to their destination nodes by routing module. A routing module includes a local routing table for *local network* as well as a global routing table which can be used to manage the routing between the local network and other parts of the global network. Global routing table fragments the entire network into manageable sizes and therefore it is possible to investigate Internet-like (large-scale) networks. These routing tables reflect state of the network and have resemblance with distance vectors. Also, we have equipped our node model with the *beehive* to implement and test *swarm-based routing algorithms*. In our swarm application, beehive launches scouts or other kind of entities to monitor the network and to reconfigure network resources.

##### Link

All links are communication channels and therefore are viewed as pipes which are characterized with bandwidth (bits/sec) and transmission or propagation delay specified in milliseconds. Each link has a corresponding buffer with finite capacity. The packets that arrive are placed in the buffer and are transmitted to the next node using first-in first-out (FIFO) strategy. Links are modelled as *bidirectional*, thus supporting concurrent bidirectional interactions. Links are able to carry traffic of a certain bandwidth up to the total capacity of the link. Each link atomic model has input and output ports for connecting two nodes in a duplex manner (see for example Link1 atomic model in Figure 4).

##### Data Packets

All packets that are exchanged among components in the form of DEVS messages can be distinguished as *data* and *control* packets. Data packets are basic IP packets which carry information such as *id* and *precedence* (see Figure 2). Control packets allow the node to obtain whole network view and to measure the traffic. For example, they are Routing Information Protocol packets in our distance vector application, while they may be cooperative scouts or ants in swarm based routing. Packets traverse intermediate nodes to go to their destination. As depicted in Figure 2, all packets have a priority field which is used for handling them in some way. For example, while control packets and scouts have high priority, data packets have low priority. The data packets, therefore, are queued and served in FIFO setting. Besides handling data packets in FIFO manner, control

packets have higher priority ranging to 7 by which their queuing order is determined. Packets can be discarded upon arriving at a node because of lack of queue space or expired time to live which limits hop count. In addition, when a packet traverses across a link, if there is no available bandwidth on the link, the packet is lost or dropped.

In our implementation, no arrival acknowledgement or error notification packets are generated back to the source of the packet. Instead, a simple flow control mechanism is devised and implemented. The reason is that we focus on routing algorithms by minimizing the number of interacting components. Passing a packet within a link suffers a delay that can be viewed as transmission delay. Packets may also be subject to the FIFO delay.

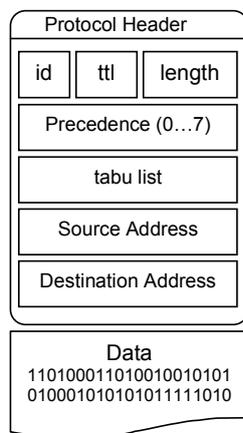


Figure 2: A data packet model with a protocol header

As shown in the Figure 2, we have modelled a packet type with the following fields: *source address*, *destination address*, *source hop address*, *destination hop address*, *packet id*, *precedence*, *total length*, *tabu list*, *TTL (time to live)* and *data*. All these fields excluding data constitute protocol header and are 20 bytes in our model. Data size can vary across applications. Packet id characterizes the packet. In order to avoid a packet to travel around the network for a long time, we restrict the packet with specific hop count, namely TTL value. Total size of packet is stored in the length field of a packet. Packet storing sequence in queue is determined by precedence value ranging from the lowest priority 0 to the highest priority 7. Tabu list allows us to keep track of visited nodes. Source and destination address fields denote packet's origin and final node's IP addresses. Data field is simply used to contain data object.

### Routing Table

By equipping each node model with a routing table, data packets can be systematically routed through the network. The Java implementation of the routing table consists of a collection of *Route* objects where each is an instance variable of the routing module class (see Figure

1). When a node needs to send a packet to a given destination, decisions about which outgoing link (i.e., DEVS atomic output port) to be used are made by means of the information specified in its routing table.

Each node has a routing table for every possible destination in the network, and each table has an entry for every neighbour (see Figure 3). According to the routing algorithm, these routing tables are constructed previously (in static algorithms), dynamically adapted to network load state (in dynamic algorithms) or based on node's (insect's) selection probabilities of the next node to its destination – e.g., using swarm based algorithms. The routing tables are initialised at the simulation start up with routes to directly linked interfaces of cost 1. Routing table can be imagined as a matrix in which rows correspond to destinations and columns to neighbours. During simulation execution new entries may be added to table or current entries may be removed or adjusted according to network traffic. All the values of the entries in the routing table range between 0 and 1, a probabilistic value. We called these entries as *profitability values* through which most profitable routes can be chosen.

### Generator and Transducer Models

In order to experiment with the above network model, it is necessary to model user traffic. To make realisations of network traffic and examine specific scenarios, the experimental frame concept and its DEVSJAVA realisation are employed. In our implementation, a typical experimental frame consists of an event generator and event transducer. The generator generates packets with fixed time intervals by randomly choosing source and destination addresses. As mentioned earlier, generator also can create and schedule specific events in the network such as link down and node congestion events. The transducer observes and analyses the network outputs, and stores these results in trace files. Transducer simply converts data to information which is meaningful for us.

### Coupled Simulation Models

We have developed a discrete event simulation model for networks with varying topologies and structures. As mentioned above, the developed framework is capable of representing the behaviour of different routing algorithms (e.g. shortest-path, distance vector and various swarm algorithms). Hence, the approach serves as a framework to test and evaluate alternative network configurations. By using basic components and tools which have been described above, networks can be built by coupling node and link atomic models in DEVSJAVA simulation viewer (see Figure 4). Furthermore, by coupling these coupled networks, increasingly larger networks can be systematically developed and experimented with.

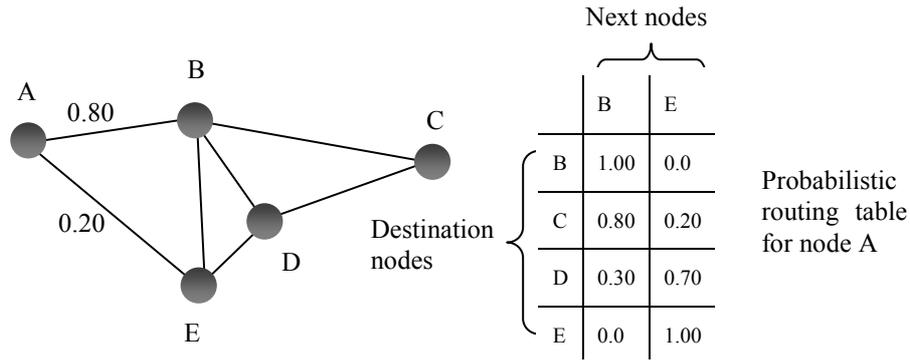


Figure 3: An example of a routing table

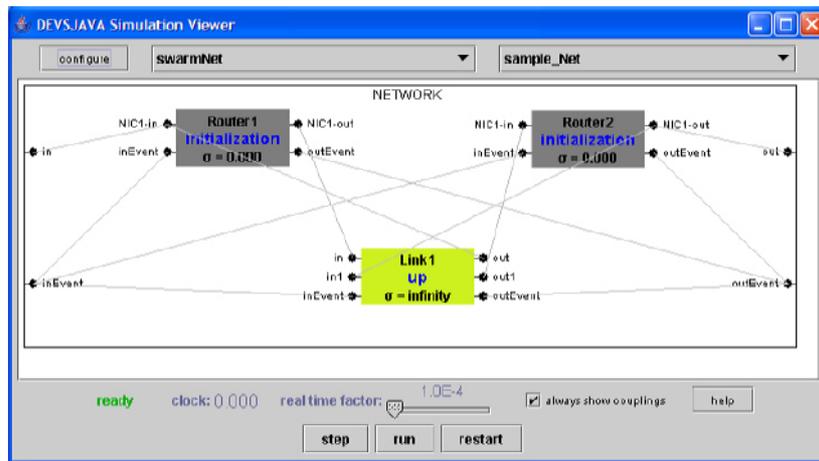


Figure 4: Synthesis of a network model

Some applications have been created in DEVSJAVA to simulate routing algorithms over some network traffic patterns. In our experiments, we have used a simple packet switched network and a set of large-scale networks with increased complexity and connectivity. These networks hereafter are referred to as the simple and complex networks, respectively. They have been designed for testing the model design as well as testing the framework itself (e.g., scalability). The simple network would be sufficient for initial testing of whether routing tables are updated correctly and whether, when links go down or come alive, routing tables are correctly updated. Large-scale and complex networks are used for uncovering dynamics and performance measurements of the models in the DEVSJAVA environment.

### Scaling Coupled Models

To support scalability, we employed and implemented a clustering approach. Clustering provides manageable network sizes by abstracting a subnet to single node in a higher level network. By considering a coupled model as an atomic model, DEVS coupled model concept has a resemblance with clustering. There exists a hierarchy of networks within the total of all nodes and routers. Each coupled model has a number of *border nodes* which are

used for connecting it to other coupled networks. In our approach, clustering is done in *addressing level* of nodes.

Hierarchical and modular structure of DEVS formalism facilitates implementation of clustering approach. Border nodes have an additional routing table consisting of the cluster names. This approach substantially decreases the information stored in routers.

### 5. CREATING HONEYBEE INSPIRED NETWORK MODELS

To show the capability (applicability) of the modelling approach, first we started with well-known routing algorithms. We have implemented static link state algorithm (Dijkstra 1959) to initialize network and distance vector to calculate distances between nodes. In the implementations of these algorithms, we used hop number as a metric, but other metrics such as available link bandwidth may also be used.

As pointed out earlier, our biologically inspired approach was derived based on honeybees and their interactions. For example, the movement of packets (artificial bees) can be used to balance network loads. Focused on biological inspired load balancing mechanism is analogous to honeybee scout-recruit system. In honeybee colonies, a colony deploys certain portion of its foragers

for searching nectar, namely scouts. Scouts find rich nectar sources and monitor their availability. If colony finds additional sources of nectar, the highest profitable source is preferred by foragers relative to other sources with less profitability. Foragers are distributed among nectar sources using profitability criterion during the course of nectar collecting process.

We have developed a set of models which are capable of exhibiting an ensemble of scouts controlling congestion in a distributed environment. In our implementation, analogous to honeybee scout-recruit system, each network node is a beehive. Network corresponds to the world of honeybees who seek rich nectar sources, finding paths with higher capacity to profitable nectar sources, light-weight scout entities searching for nectar, and control packets foraging for information to aid survival of the network (honeybee colonies). Each hive deploys a number of scouts to find the most profitable paths for a given destination. Each router then uses the information received from all the nodes in the network obtained by its scouts to calculate the shortest path to each destination in terms of a chosen metric. Scouts control congestion by making alterations to routing tables in order to route new traffic away from congested nodes. Then, packets are dispatched from a source to a destination according to information gathered by scouts.

The developed approach offers some useful properties such as probabilistic routing, optimal system performance by tuning parameters, event-driven updates based on network flow, low convergence time, low control packet traffic and scalability via clustering which reduces routing information stored in a node's memory. Social insect inspired approaches bring a probabilistic routing method to network routing domain, therefore we use probabilistic cost values in order to represent source profitability. The cost metric can be based on the bandwidth of the link or can be dynamically measured as in the case of delay or load.

The routing table is then updated with the new information. In the networks we have experimented with, initially no apriori knowledge is known about the routes. All routes were computed in parallel during initialization phase. Each route determined for a given destination node based on the Dijkstra shortest path about the minimum hop. When an event occurs, such as a link going down or a node failing, then routing scheme has to be able to handle the situation. An event-driven update is selected for routing information update in response to changes that are detected. The use of event-driven updates rapidly disseminates the data about the failed route, which reduces the change for growing data and route loops to occur. By doing so, the entire procedure can be completed in a less time than periodical updates.

## 6. SIMULATION RESULTS & DISCUSSION

In order to compute the performance of the routing approach, we developed a set of models and experimented with them. Using node and link models

defined above, various network topologies can be formed. Then, developed network models are run under traffic load by using experimental frame model. The simplest network modelled has 11 nodes and 18 bidirectional links (see Figure 5), while larger models has up to 3520 components.

Each node in the network is represented a routing table storing the neighbouring node to which traffic should be routed. Each simulation run consisted of an adaptation to topology (initialisation phase) and test period. During the initialisation phase, system runs without load and initial routing tables are formed according to the number of hops (shortest path estimation). During the test period we measured and recorded the network performance in terms of average packet delay, throughput, convergence time and packet loss ratio.

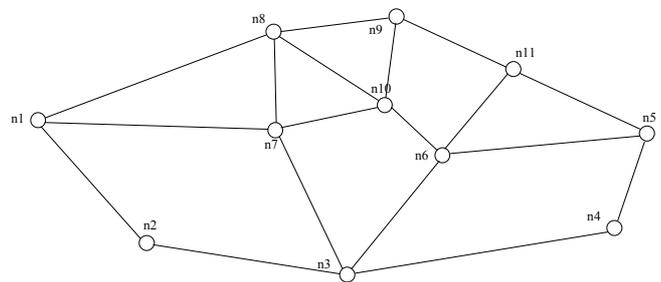


Figure 5: A simple network

### Node and Link parameters

All nodes are designed as routers and each one has own interface(s) equal to their neighbours. Nodes have the same buffer size (1Mbit) but have different IP addresses. Moreover, node packet's processing time is selected as 1 msec.

Links are bidirectional and their bandwidths range between 1.5 to 6 mbps with propagation delays ranging between 1 to 5 msec.

### Traffic model

Traffic flows in the network are simulated by a traffic generator model component. This model generates data packets which are then periodically sent out to the network using uniformly randomly selected source and destination nodes. We observe network for one second (see Figure 7). Generator sends 1000 packets to the network in course of one second which packet sizes varying from 10 bytes to 100 Kbytes.

We used two standard performance metrics: throughput and packet delay. We avoid generation of packets with the same source and destination. The amount of network traffic is determined by the number of packets in the network. Generally, many packets must wait in limited capacity FIFO queue for processing at the nodes.

We compare our approach with a state-of-the-art algorithm, namely RIP (Routing Information Protocol).

RIP is an instance of distance vector algorithm and still being widely used in Internet networks (Steenstrup 1995). In Figure 6, results obtained from both RIP and ecological approach are presented together. As mentioned earlier, average packet delay and throughput are major performance criteria for evaluation.

In Figure 6, it can be shown that ecological approach shows better throughput than RIP. After a short time (~200 msec), the throughput reaches steady values and remains constant to the end of the simulation. This means load balancing is achieved successfully.

Average packet delay values are almost same, 9 msec (see Figure 7). However, bees approach's packet delay remains low up to 0.5 msec and later has greater values than RIP. The reason is that probabilistic routing forwards the packets alternative routes for load balancing, while RIP selects shortest paths. But, ecological approach has better load balancing and lower packet loss ratio.

### Performance of the framework on increased scalability and connectivity

In our experiments, one of the key independent variables was the degree of connectivity and scalability. In order to examine the scalability aspect of our approach, we developed various networks ranging from 29 to 3520 components (see Table 1). These models were executed with acceptable performance in the DEVSJAVA environment. The largest network took less than three hours on a 2.4 GHz processor and 512MB RAM while the simple one took a few minutes.

Table 1: Large-scale network models

Network	Number of component	Number of colonies
NET 1	116	4
NET 2	319	11
NET 3	960	87
NET 4	3520	125

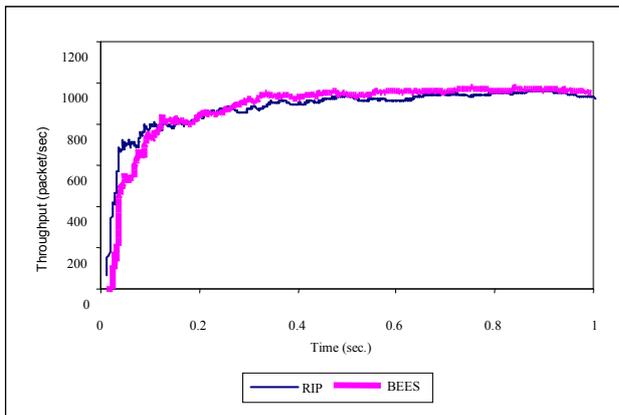


Figure 6: Throughput comparison of different algorithms

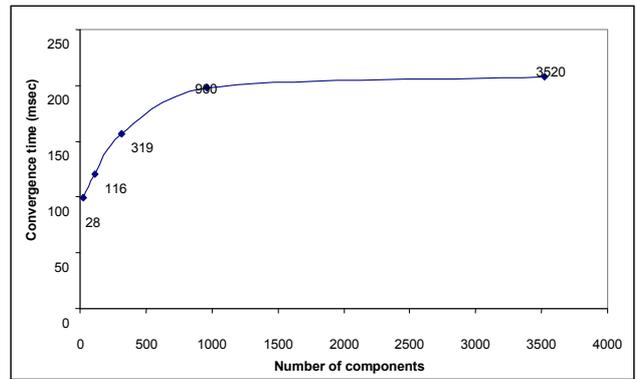


Figure 8: Convergence time of networks with different scales

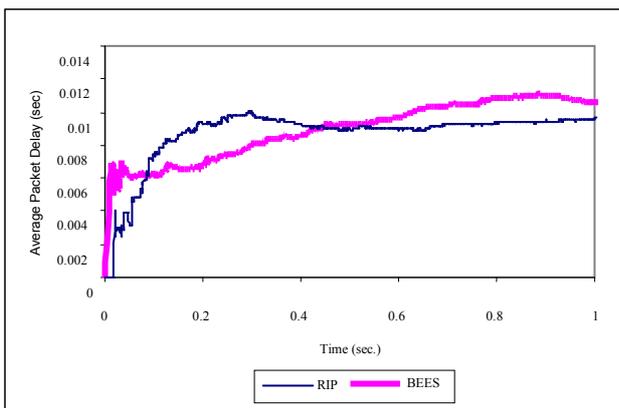


Figure 7: Average packet delay comparison

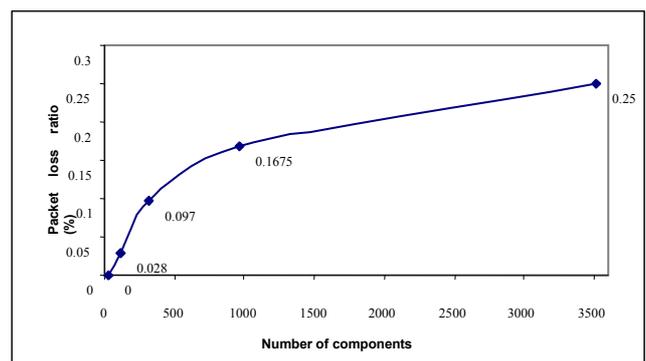


Figure 9: Packet loss ratio of networks with different scales

As shown in Figure 8, larger models exhibited lower than expected time for convergence. This is partially due to the DEVS discrete event modelling paradigm and its implementation in DEVSJAVA. The maximum number of components is partially due to the platform used for executing the simulations. Packet loss ratio gradually increases with the increase in the number of components.

However, packet loss remains acceptable even for large-scale networks (see Figure 9).

## 7. CONCLUSIONS

This paper proposed a discrete-event modelling approach for networks. The models are devised based on biologically-inspired routing mechanisms to tackle the scalability aspect of large-scale networks. The approach and its implementation are promising for handling models composed of hundreds to thousands of components. The routing strategy, which is based on the behaviour of beehives, is robust and exhibits similar or better performance compared to the contemporary routing RIP technique. This research suggests the proposed modelling approach can be used for the design and development of robust and scalable network systems.

## Acknowledgement

This research is partially supported by NSF grant Scaleable Enterprise System (DMI-0122227).

## REFERENCES

- Abadi, M. and L. Cardelli. 1996. *A Theory of Objects*. Springer.
- ACIMS. Arizona Center for Integrative Modeling and Simulation. 2004. <http://www.acims.arizona.edu/SOFTWARE/software.shtml>
- Anderson, C. 2001. "The adaptive value of inactive foragers and the scout-recruit system in honey bee (*Apis mellifera*) colonies". *Behavioral Ecology*. 12, No. 1, p. 111-119.
- Chow, A.C.-H. 1996. "Parallel DEVS: A Parallel, Hierarchical, Modular Modeling Formalism and its Distributed Simulator", *Transactions of the Society for Computer Simulation International*, 13, No. 2, p.55-67.
- Dijkstra, E.W. 1959. "A Note on Two Problems in Connexion with Graphs". *Numerische Mathematik* Vol. 1.
- Hayzelden, A. and J. Bigham. 1998. "Heterogeneous Multi-Agent Architecture for ATM Virtual Path Network Resource Configuration", *Proceedings of Intelligent Agents for Telecommunications Applications*, Springer Verlag. 45-59.
- Lunceford, W.H. and E.H. Page. 2002. Editors. *International Conference on Grand Challenges for Modeling and Simulation*, San Antonio, Texas, USA.
- Mesarovic, M.D. and Y. Takahara. 1989. *Abstract Systems Theory*. Springer Verlag.
- Minar N., Gray M., Roup O., Krikorian R., and Maes P. 1999. "Hive: Distributed Agents for Networking Things", *First Int'l Symp. Agent Systems and Applications and Third Int'l Symp. Mobile Agents*. IEEE Computer Soc. Press.
- Sarjoughian, H.S. and R. Singh. 2004. "Building Simulation Modeling Environments Using Systems Theory and Software Architecture Principles", *Proceedings of the Advanced Simulation Technology Conference*, 99-104, Washington DC (April).
- Seely, T.D. 1995. *The Wisdom of the Hive*. Cambridge, Mass: Harvard University Press.
- Steenstrup, M. E. (Ed.). 1995. *Routing in Communications Network*. Prentice-Hall.
- Uhrmacher, A.M., P.A. Fishwick, and B.P. Zeigler. 2001, Agents in Modeling and Simulation: Exploring the Metaphor (eds.). *IEEE Proceedings*.
- Wooldridge, M. and N.R. Jennings. 1995. Intelligent Agents: Theory and Practice. *The Knowledge Engineering Review*, 10, No. 2. 115-152.
- Wymore, W.A. 1993. *Model-based Systems Engineering: An Introduction to the Mathematical Theory of Discrete Systems and to the Tricotomy Theory of System Design*, Boca Raton, CRC.
- Zeigler, B.P., H. Praehofer, and T.G. Kim. 2000. *Theory of Modeling and Simulation: Integrating Discrete Event and Continuous Complex Dynamic Systems*. Second Edition Academic Press.

## AUTHORS BIOGRAPHIES

**AHMET ZENGİN** is a PhD candidate at Sakarya University, Turkey. His experience with modelling and simulation includes a one-year-stay in ACIMS Lab at the Arizona State University. His research topics include DEVS theory, multi-formalism modelling, parallel and distributed simulation, modelling and simulation of large-scale networks, distributed systems management, biologically-inspired optimisation schemes. His main research interest lies in parallel and distributed simulation and the High Level Architecture.

**HESSAM S. SARJOUGHIAN** is Assistant Professor of Computer Science and Engineering at Arizona State University, Tempe. His research includes modeling theory, collaborative modeling, distributed co-design, intelligent agents, and software architecture. His industrial experience has been with Honeywell and IBM. Visit <http://www.eas.asu.edu/~hsarjou/index.htm> and <http://www.acims.arizona.edu> for more information.

**HUSEYİN EKİZ** is received M.Sc. in 1993 from Gazi University, Turkey, and Ph.D. degree in computer engineering in 1998 from the University of Sussex, England. He is currently Professor of the Department of Computer Systems Education and Dean of the Technical Education Faculty, Sakarya University, Turkey. His research interests are in the fields of network systems, distance education, digital circuit design and microprocessor architectures.

# SIMULATING COMPLEX SYSTEMS IN LABVIEW

György Lipovszki

Department of Production Informatics Management and Control  
Budapest University of Technology and Economics  
H-1111. Budapest, Műegyetem rkp. 3-9 D. 428., HUNGARY  
E-mail: lipovszki@rit.bme.hu

## KEYWORDS

dynamical systems, mathematical modeling, other computational methods

## ABSTRACT

Modeling and simulation of systems, especially in science and engineering can help to reduce risk and cost of design and testing processes. According to Cellier, the established mathematical models can be classified as follows: continuous time, discrete time, quantitative models and discrete event models.

A huge number of simulation software has been developed to support modeling and simulation efforts. All of these software tools support the use of one or more mathematical model classes. Despite all of these efforts, it is hard to find simulation software, which is capable of combining several model classes in a real industry standard environment. The paper presents a series of simulation software products, which have been developed using an industry standard programming environment widely applied to data acquisition, process control and data visualization: National Instruments' LabVIEW.

The first development was the TUBSIM, a continuous time simulation toolbox and it was followed by the discrete event extension called Discrete Event Simulator (DES). The elements of the toolboxes facilitate block oriented modeling using LabVIEW's high level graphical editor and calculating power of. Both LabVIEW-based simulation libraries are widely used in education and research. During the years several additional modules were and are still developed, e.g., fuzzy rule-based systems, optimization using genetic algorithm, compartment modeling systems for pharmacodynamical and pharmacokinetical applications, etc.

Applying these toolboxes one can model and simulate complex systems where continuous and discrete event driven parts are working together. The continuous part can insert events into the discrete event task list on the fulfillment of different continuous state variable conditions. The discrete event system can also change the value of state variables, together with input values in the continuous system.

## INTRODUCTION

One of the most important applications of computers is imitating, or *simulating*, the operation of various kinds of real world facilities and processes. The facility or process of interest is usually called a *system*, and in order to study it scientifically we often have to make a set of assumptions about how it works. The assumptions, which usually take the form of mathematical or logical relationships, constitute a *model* that is used for trying to gain some understanding of how the corresponding system behaves.

If the relationships that compose the model are simple enough, it may be possible to use mathematical methods (such as algebra, calculus, or probability theory) to obtain *exact* information on the question of interest; this is called an *analytic* solution. However, most real-world systems are too complex to allow realistic models to be evaluated analytically, and these models must be studied by means of simulation. In a *simulation* we use a computer to evaluate a model *numerically*, and data are gathered in order to *estimate* the desired true characteristics of the model. Application areas for simulation are numerous and diverse. Below is a list of the most important types of simulation areas:

- Continuous systems simulation formulated by differential equations
- Stochastic discrete event driven systems
- Complex systems, mixture of previous types of systems

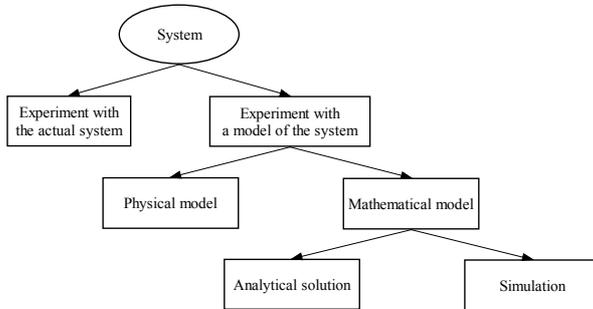
A *continuous system* is one in which the state variables change continuously with respect to time. A forklift-moving trough the workshop is an example of a continuous system, since state variables such as position and velocity can change continuously with respect to time.

A *discrete event driven system* is one in which the state variables change instantaneously at separated points in time. A post office is an example of a discrete system, since state variables – the number of people in the post office – change only when a new person arrives or when a person departs after being served. Few systems in practice are wholly discrete or wholly continuous, but one type of change predominates for most systems.

A *complex system* where the state variables can create events in the system when a given condition

became valid and like a reverse action an event also can cause changing of system input value or a state variable in the continuous part of the model.

At some point in the lifecycle of most systems, there is a need to study them, to try to gain some insight into the relationships among various components, or to predict performance under some new condition. Figures 1 maps different ways in which a system might be studied.



Figures 1 Ways to study a system

## CONTINUOUS-TIME SIMULATION IN LABVIEW

Nowadays, computer-based materials are increasingly used in education. A number of interactive, computer-based course materials in various disciplines and subjects, from all over the world are available through the Internet. These courseware materials have diverse target audiences both in level and type of students, e.g., from primary school to university, from regular weekly lectures to distant-learning.

Systems engineering courses are taught prior to Control Engineering courses at BUTE, to introduce modeling and simulation theory, methods and applications, not only for control, but for other disciplines as well. Modeling and simulation theories, methods and techniques – parallel with and serving the needs of systems and control engineering – have undergone a significant development during the last century. Measurement technology achieved important milestones too. Up-to-date instruments and principles, accurate and fast measurements characterize today's measurement technology. Modern computing resources increase data management speed and capacity. Scientific results of the above-mentioned fields are joined together in interdisciplinary applications. The most recent information should be used in education, while at the same time the basics have to be taught, too.

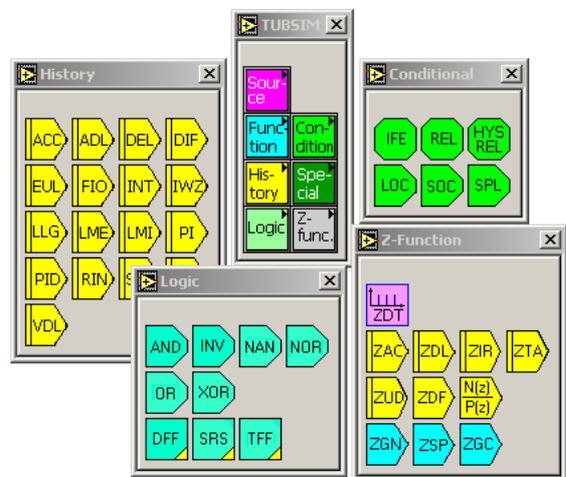
Keeping the "old" knowledge and including the new improvements poses a problem for educators. They have to find the balance between the amount and quality of information relayed to students. In systems and control engineering frequency-domain techniques are still taught, however their importance have been reduced by the improvement in computing and time-domain simulation. They are however

necessary, because of the insight and handy techniques they give to students. At the same time, modern modeling and control knowledge, such as soft computing, adaptive and optimal control has to be included in the curriculum. To satisfy both tasks a number of programs have been developed to help visualize the conventional information (text, diagrams, drawings, and equations) in textbooks and lecture notes. These programs could easily be considered as computer games. It is widely proven that learning-by-doing is a leader among the approaches of learning. Playing with these simulations radically improves the understanding and future application in real world situations.

The author in LabVIEW environment has developed the continuous simulation packages named TUBSIM. The package was initially developed according to the CSSL (Continuous System Simulation Language) recommendations.

## TUBSIM for LabVIEW

TUBSIM is the interpretation of an analogue computer in the LabVIEW graphical programming environment. Analogue computers were widely used for simulation, but they had many disadvantages (e.g., the size of the computer grows with the size of the model).



Figures 2 TUBSIM VI icons in LabVIEW

When digital computers became universally available, analogue computers and analogue simulation were soon replaced with digital simulation. One group of the digital simulation systems uses the same principles as analogue simulation does. TUBSIM belongs to this group. The TUBSIM VI (icons) Library (Figures 2) contains the basic blocks typical to analogue computers (summers, integrators, potentiometers, signal generators). In addition, TUBSIM has different Boolean blocks, typical systems engineering elements (for example first order element, continuous time controllers – like PI, PID –, time delay block)

and sampled time blocks. TUBSIM has successfully been used in various applications from teaching aids to large-scale industrial processes (e.g., the Secondary Side Water Chemistry model of the Nuclear Power Plant in Paks, Hungary).

### Continuous Simulation Applications

A continuously increasing number of TUBSIM simulation applications are used in both introductory and advanced systems and control engineering courses at BUTE. There are other advanced studies (e.g. Computer Controlled Systems, MSc and PhD theses) too, which utilize similar applications.

State-space models are particularly hard to solve with conventional methods, especially when non-linearity and time-delays are involved.

An example of such a complex non-linear system from the area of biomedical engineering is the compartment model of enterohepatic circulation. Biomedical simulations are especially useful when “control systems” have to be designed to compensate the effects of illnesses. One such example is the development of medication regimes for patients with diabetes.

Simulation is a very appropriate way for presenting and comparing different types of controllers. The application of knowledge learnt in basic control theory courses are best tested with simulation programs. Currently the speed of LabVIEW applications, especially the ones with a large need of run-time calculations are considered. When for example a fuzzy controller has to be optimized with a genetic algorithm, DLLs (dynamic link libraries) are used to speed up the calculation.

The advantage of developing native LabVIEW programs lies in the ease and speed of development; however the speed of calculation is best boosted with the use of external function calls from a DLL. DLLs open the world of object-oriented programming to LabVIEW, as well.

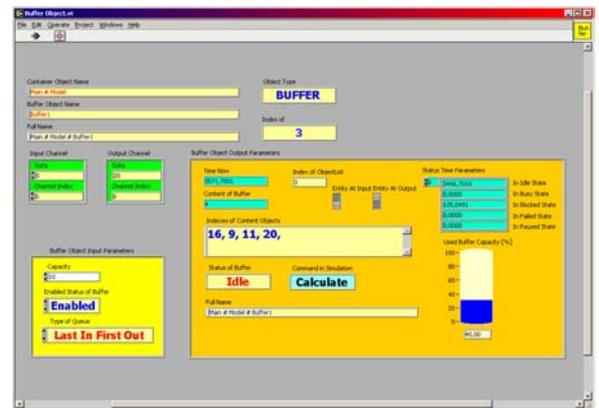
### DISCRETE EVENT SIMULATION IN LABVIEW

The latest application library for LabVIEW is the Discrete Event Simulator (DES) package. This package contains approximately 140 elements, with LabVIEW terminology VIs (virtual instruments). Objects in the DES package for building complex logistic systems are: CONTAINER, ENTITY, SOURCE, BUFFER, MACHINE, SINK, JOIN, SELECTOR, PACK, UNPACK. Beside the main objects, there are so-called work procedures to create and destroy entities, to calculate different distribution functions, to set and get an object’s attributes and to manage the task list – the heart – of the discrete event simulation system. The Discrete Event Simulator

package is used to give an interactive experimental environment to study processes with uncertainties.

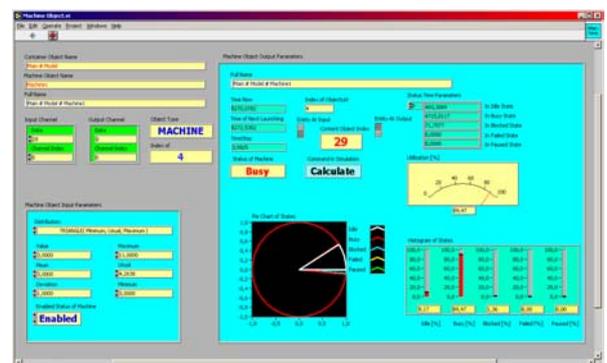
Attributes of existing objects in the DES package are the follows:

- CONTAINER: (parent object) this object allows the development of DES subroutines, which can be copied and renamed, so they can be applied with the same functionality and different inputs.
- ENTITY: this is the “working object” in the simulation system. It could represent information or material flow in the simulation model. It has any number of attributes represented by numerical or string values.
- SOURCE: this object creates new entities. The time duration between the produced entities is given by different types of distributions (Constant, Exponential, Normal, Triangular, Uniform and User Defined).



Figures 3 Front panel of Buffer Object

- BUFFER: this object temporarily stores entities, till the material or information flow of the object connected after the buffer becomes able to send or receive the outgoing entity (Figures 3).



Figures 4 Front panel of Machine Object

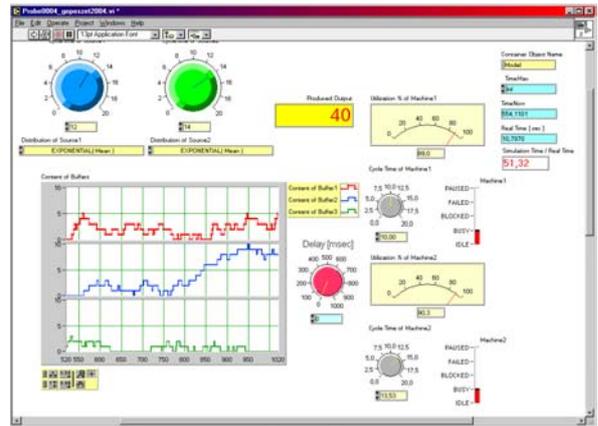
- MACHINE: this object delays the flow of entities with a given type of distribution (Constant, Exponential, Normal, Triangular, Uniform and User defined) (Figures 4).
- SINK: this object destroys “used” entity objects and frees the used part of memory.

- **JOIN:** this object has more than one input channels and has only one output channel. One of the input parameters the “Join Input Channel Index,” is for selecting one of the input channels with a given strategy. In the next event the waiting entity from this channel is “read”, and sends away into the information or material flow.
- **SELECTOR:** this object has one input channel and more than one output channels. With one of the input parameters the “Selector-Channel-Index,” the output channel is selected with a given strategy and an entity (if exists) is sent out on this channel.
- **PACK:** with this object one or more entities can be packed into another entity. The Pack object has two inputs and one output. The first input channel receives the objects to be packed the second input channel receives a package entity. In a packaging process, the package arrives first and waits inside the Pack object till the required amount of entities arrive into the package object. When the process has finished only one entity is sent away, but it contains a given number of other entities.
- **UNPACK:** with this object we can unpack one or more entities from an entity (package entity). The Unpack object has one input and two outputs. The first output channel sends away the unpacked objects; the second output channel is used to send out the package object. The unpack process starts with the arrival of a package object (entity). After the package has arrived, the Unpack object unpacks its content into an inner buffer and sends them out over the first output channel as soon as possible.

Beside the main objects there are so called *work procedures* to create and destroy entities, to calculate different distribution functions, to set and get an object’s attributes and to manage the *task list* – the heart – of the discrete event simulation system. Using this task list the simulator establishes a *next-event time advance* mechanism that always gets the most imminent future event and copies this event’s time part into the simulation system clock. At this time (at the occurrence of the most imminent event), the state of the system is updated, and future event time(s) are determined. The process of advancing the simulation clock from one event time to another is continued and eventually finishes as some specified stopping condition is satisfied.

### Discrete Event Simulation Application

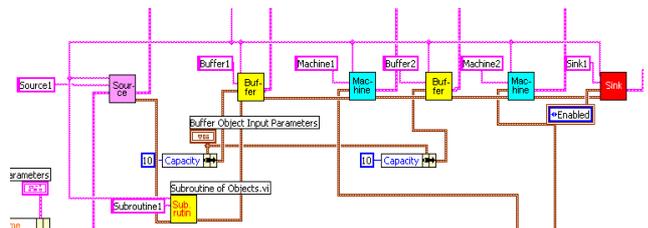
To introduce the possibilities of the DES package, a simple manufacturing problem is solved, where the material flow goes through buffers,



Figures 5 Control surface (Front panel) of DES simulation system

machines and a subroutine that also contains basic DES objects. The control surface (Figures 5) elements (fields, buttons, meters and displays) are changing the parameters and analyzing results. Graphical programming helps us in semantic control too (easy search of lost connecting lines) and in the reorganization of the logistic net.

The Discrete Event Simulator package is able to use any number of continuous simulation systems and to calculate them, because the continuous system clock could send events into task list.



Figures 6 Program surface (Diagram panel) of DES simulation system

Figures 6 shows a simple discrete event driven system architecture in the LabVIEW based graphical environment. Icons of the basic objects and commands can easily be connected together with the color identified “wires”. This wires mean different data structures (record of different data types).

### Complex Simulation Applications

Complex simulation – discrete event and continuous simulation together in the same system model – is becoming more and more important in the education and different industrial applications. The continuous system at every calculation (sampling) time inserts a new “calculation request” into the discrete event simulator’s task list, that is how more than one continuous simulation subsystems can run together with different time steps inside the same simulation system. This is very important because each subsystem in the simulation model requires a

different time step in order to optimize calculation time of the simulation. From view of discrete event simulation, every discrete event simulation system can contain special conditions, which need to be calculated by using continuous simulation sub-models. These complex systems cannot be solved analytically because the stochastic behavior of discrete events and the different nonlinearities of continuous subsystems.

Let us look at a special queuing system of airplanes waiting to land at an airport. It is a retrying system; airplanes retry to get landing permission until it is granted. The overall picture of the system behavior will be highly influenced by numerous small continuous subsystems, which are operating on the airplanes and/or strongly connected to their operation. All of these continuous subsystems can influence the new order of the waiting airplanes. The problem can be modeled as a complex (discrete-continuous) simulation system with dynamic priorities in the waiting lines.

Some modeling considerations are listed as follows:

- One of the most important conditions is the actual content of the fuel tank of the planes, which would make the waiting possible. The fuel consumption of a plane is a typical continuous simulation subsystem function. Under a minimal level of the fuel tank responsible landing is impossible and a queue reorganization request has to appear promptly in the task list of the simulation, otherwise the plane crashes.
- The number of the people on the planes is another important aspect, in other words how many lives are endangered because the landing queue calculation did not use this data. This data (number of people) is an important weighting factor in recalculating the landing queue.
- Required landing circumstances are an important condition for each type of planes. Huge jets can land without disaster only if they use the regular runway; however small airplanes are able to land on a highway or a field if necessary.
- The state of the health of the people on the plane is also a determinative aspect at the calculation of the landing queue. A special event, for example a heart attack, a pregnant woman going into labor, urgent need of a certain medical intervention or medicine could also reorganize the landing queue.

## CONCLUDING REMARKS

This paper reviewed the use of continuous and discrete event simulation as teaching aid for systems and control engineering education. LabVIEW is the mostly used programming environment for these applications, with simulation packages – TUBSIM and DES – developed by the authors of this paper. Current research and development work includes the

development and integration of special purpose DLLs for fuzzy logic, neural network and genetic algorithm applications. The Discrete Event Simulation package and any one of the continuous time packages can be combined to handle continuous-discrete hybrid systems.

## REFERENCES

1. Åström, K.J., Wittenmark B. (1997) *Computer-Controlled Systems*, Prentice-Hall
2. Bequette, B.W. (1998) *Process Dynamics*, Prentice Hall
3. Dorf, R.C., Bishop, R.H. (1998) *Modern Control Systems*, Addison Wesley Longman
4. Gordon, G. (1969) *System Simulation*, Prentice Hall
5. Hartley, T.T., Beale, G.O., Chicatelli, S.P. (1994) *Digital Simulation of Dynamic Systems – A Control Theory Approach*, Prentice Hall
6. Kheir, N.A. (editor) (1995) *Systems Modeling and Computer Simulation*, Marcel Dekker, Inc.
7. Law, A.M., Kelton W.D. (1991) *Simulation Modeling & Analysis*, McGraw-Hill
8. Man, KF, Tang, KS and Kwong, S. (1999) *Genetic Algorithms*, Springer
9. Monsef, Y. (1997) *Modelling and Simulation of Complex Systems. Concepts, Methods and Tools*, Society for Computer Simulation International
10. Wells, L.K.; Travis J. (1995) *LabVIEW for Everyone – Graphical Programming Made Even Easier*, Prentice Hall
11. Zeigler, B.P., Praehofer, H., Kim T.G. (2000) *Theory of Modeling and Simulation*, Academic Press
12. Pidd, M, 1992, *Computer Modeling for Discrete Simulation*, Chichester, England, John Wiley

## AUTHOR BIOGRAPHY



**GYÖRGY LIPOVSZKI** was born in Miskolc, Hungary and went to the Budapest University of Technology and Economics, where he studied electronics and graduated in 1975. He is now Associate Professor at the Department of Department of Production Informatics Management and Control and his research field is development of simulation frame systems. in different programming languages.

His e-mail address is: lipovszki@rit.bme.hu

# .:PSIM:. – A LABVIEW-BASED SIMULATION SYSTEM AS A LEARNING AID

Petra Aradi

Department of Informatics, Faculty of Mechanical Engineering  
Budapest University of Technology and Economics  
Műgyetem rkp. 3., Budapest, H-1111, Hungary  
e-mail: petra@rit.bme.hu

## KEYWORDS

continuous simulation, education, tool, LabVIEW

## ABSTRACT

.:PSim:. a LabVIEW-based simulation system suitable for both educational and industrial purposes is presented. Educational aspects of .:PSim:. are emphasized below, especially in areas of systems and control engineering as taught in various courses at Budapest University of Technology and Economics (BUTE). .:PSim:. is not just another simulation tool in LabVIEW, it is a set of building blocks like LEGO that can be combined into various constructions and can further be extended with new elements.

## INTRODUCTION

.:PSim:. started as a collection of LabVIEW VIs developed by the author for simulating continuous time processes, according to the CSSL (Continuous System Simulation Language) recommendations. Later on the traditional time-domain simulation was extended with frequency domain methods. After implementing these “traditional” techniques, soft computing methods, specifically fuzzy systems and neural networks were added to handle complex non-linear models or models based on measured data. Bondgraph models, discrete-time systems, identification, stability analysis and compartment models were the latest enhancements in .:PSim:..

The main goal while developing .:PSim:. was to create an almost general-purpose simulation tool that is suitable both for educational and industrial purposes. .:PSim:. had to retain the characteristics of CSSL whilst including the emerging techniques like Soft Computing. As educational applications were the first concern an easy to use programming environment had to be used which is suitable both for illustration during lectures and for student projects as well.

## WHY LABVIEW?

LabVIEW from National Instruments was chosen as the programming environment for .:PSim:.. LabVIEW stands for Laboratory Virtual Instrumentation

Engineering Workbench, and is available in various computer platforms, such as Linux, MacOS, Windows, Solaris, HP-Unix. LabVIEW applications are called virtual instruments or VIs for short. VIs have a user interface and a block diagram, where the actual program is built with LabVIEW’s graphical components (Fig. 1). The ease of use and programming, together with LabVIEW’s excellent connection to the outside world through data-acquisition, I/O and network protocols makes it the ideal tool for scientists and engineers. LabVIEW has quite a number of front panel elements that facilitate the quick and easy fabrication of instrument-like user interfaces, naturally leaving the opportunity to create customized elements like buttons and displays. A similarly huge amount of functions – mathematical, I/O, etc. – are also readily available.

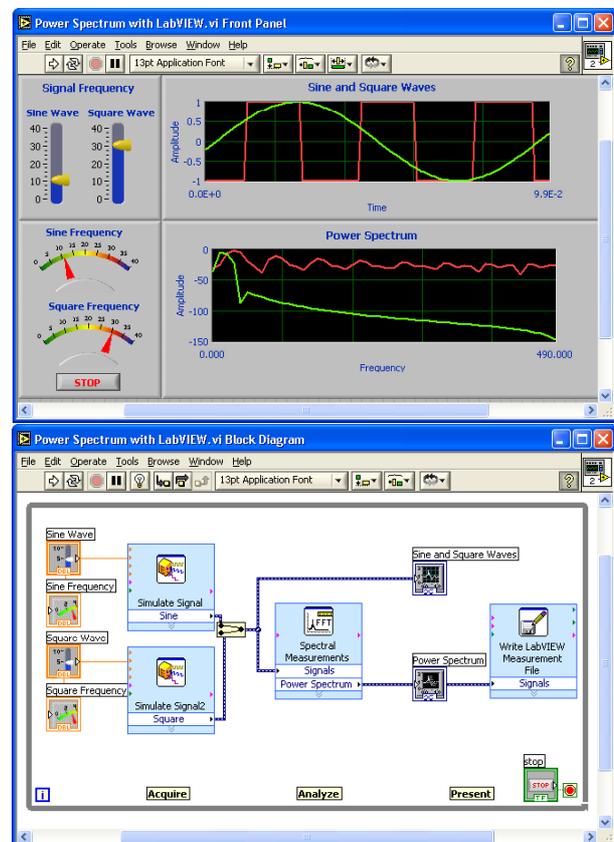


Fig. 1: Front Panel and Diagram Panel in LabVIEW

Years of experience in LabVIEW programming suggested its use as a quick and easy tool to produce spectacular and effective simulation programs to illustrate lectures even on the fly, during a lecture. Students also seem to be interested in just watching the birth of a simulation program with the rather advanced graphical programming capability, they normally do not learn to use. Even using LabVIEW's built-in functionality without any add-ons impresses them, especially when the computer is connected to a real world system, either just to measure it or even to control it.

According to the author's observations students with no previous experience in programming – not even text-oriented languages – could grasp the basics of algorithmic thinking and programming structures after four or five lectures followed by practice at a computer. Afterwards they are capable of solving simple programming and – what is more important – simulation problems with LabVIEW. They really enjoy the power of being able to tell the computer what to do, how to help them solve their tasks.

This explains why LabVIEW was chosen.

## ..PSIM:. FOR LABVIEW

### Numerical Integration

As mentioned above, ..PSIM:. started as a CSSL implementation in LabVIEW with the characteristic blocks of the digital implementation of an analog computer. The most important of these blocks are integrators with four numerical integration methods implemented. It is of course possible to use other formulas as well, however it was so far unnecessary to use higher than second order integrators (e.g. Adams-Basforth) in educational applications. Numerical integration can be accomplished with fixed time-step or with variable time-step. To be in synch with the mathematical knowledge of undergraduate students fixed time-step methods are used in educational applications.

### Beyond CSSL

It very soon became evident, that systems and control engineering courses are yearning for more powerful simulation methods than block oriented simulation with CSSL elements. That is why state space models and basic transfer blocks like proportional, integrator, lead-lag, PID, etc. were also implemented as sub-VIs. Adding these sub-VIs to the block diagram of a LabVIEW program rather complex systems may be built. It is possible to combine all the above mentioned system models within the same application.

As stability is an existential issue in the analysis of dynamic systems, various stability criteria (like Routh's table and Hurwitz's determinant) are also included in ..PSIM:.

Furthermore, as frequency domain methods still have their significance in system analysis, such tools like Bode and Nyquist diagrams were also implemented. As the classical control theory states rather simple approximate schemes to connect frequency and time domain properties to establish the quality of control loops, methods like the Nyquist stability criterions were implemented accordingly.

Mostly, but not exclusively complex system models require the use of logical (Boolean) and non-linear functions (e.g. hysteresis, relay).

Digital controllers can also be modeled with the use of the discrete-time VIs.

Additional blocks are fuzzy rule-based systems that can be used as controllers, neural networks that can be trained to mimic a modeled system, bondgraph elements and a simple discrete event system.

Fig. 2 shows the previously mentioned ..PSIM:. function libraries with additional VIs for file operations, conversion among mathematical models such as differential equation, state-space model, zero-pole-gain representation, as well as the series, parallel and feedback connections blocks.

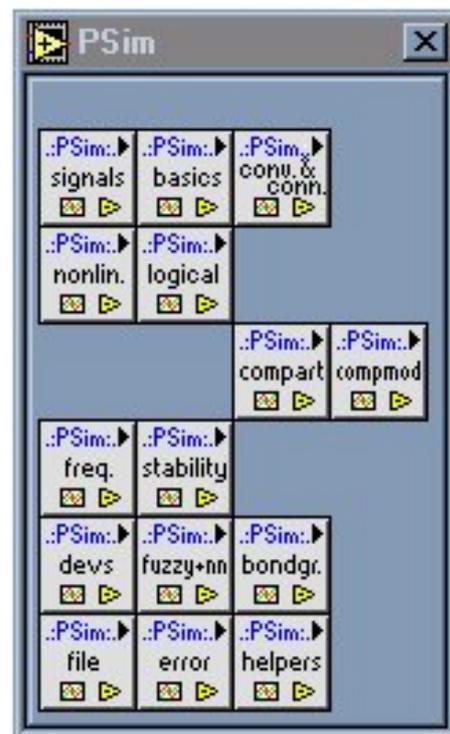


Fig. 2: ..PSIM:. Function Libraries

### ..PSIM:. Compartment

The most recent additions to ..PSIM:. are compartment models that are widely used in pharmacological and physiological modeling. Although compartment models are just a special kind of state-space models, it seemed important to further accentuate the versatility of LabVIEW to practitioners of other, non-engineering disciplines.

There are two compartment libraries available. Compartment model VIs (Fig. 3) are conventional

structures, that are widely used in pharmacokinetics (to investigate how medicines travel and transform in the body from intake to exit). These VIs are very easy to use with just a minimal LabVIEW expertise. Parameters according to the compartment structure represented in the given VI can be set, results both in graphical and numerical format can be studied and stored in files for further processing.

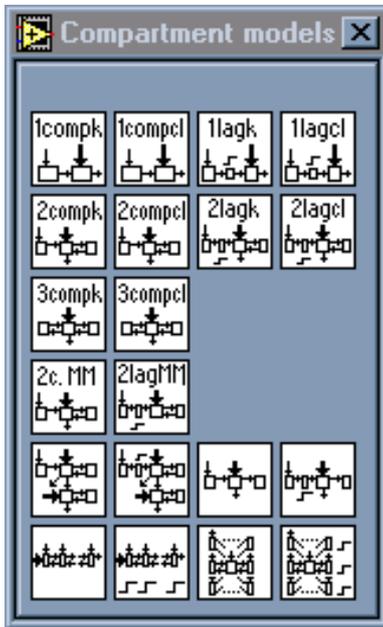


Fig. 3: :PSim: Compartment Model VIs

Fig. 4 illustrates the other library of compartment blocks. These VIs are general-purpose building blocks, aiming users with a more advanced LabVIEW knowledge.

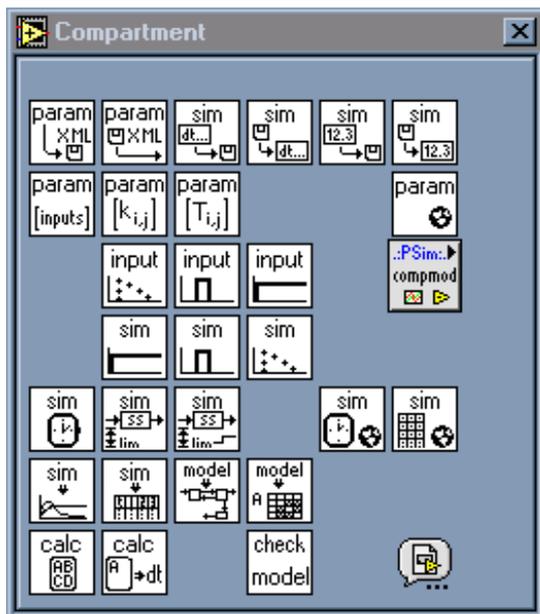


Fig. 4: :PSim: Compartment VIs for general-purpose models

### XML model description language

To facilitate data manipulation, an XML-based compartment description language was developed. XML stands for Extended Markup Language, and is widely used to create custom markup languages for various purposes. XML data files are standard text files enhanced with the markup tags to group and identify parts of the data. XML document can be effortlessly read both by humans and computers, making it possible to modify the data directly with a simple text editor. Storing data with comments increase the efficiency of data processing and retrieval.

Compartment model and simulation parameters are stored in ComPSim-XML files that are based upon the XML structure definition (ComPSim-XSD the corresponding XML Schema Document is shown in Fig. 5).

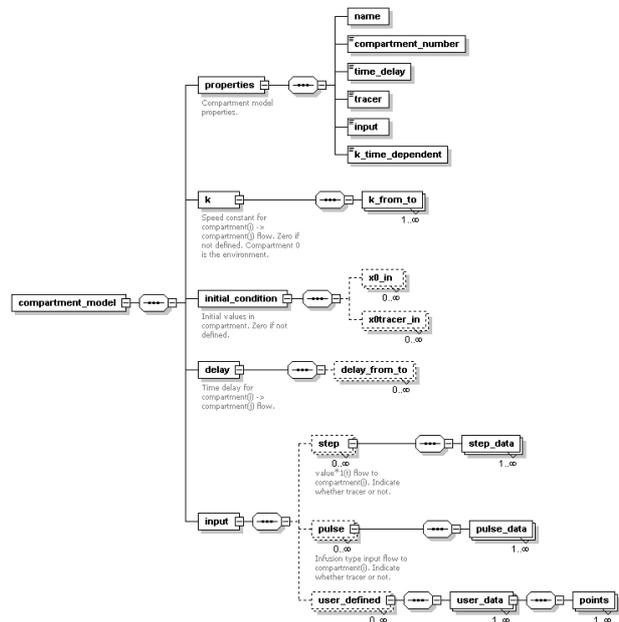


Fig. 5: ComPSim-XSD XML Schema Document

There is an XML-based general purpose modeling language under development, based on the above mentioned compartment description language.

### Source Code

So far :PSim: has been developed entirely in LabVIEW that means there are now platform dependent elements in it. There is not one DLL included, nor is there a single code interface node (CIN). (CIN calls code written in a text-based programming language, such as C, directly from a LabVIEW block diagram).

### SIMULATIONS AS LECTURE AIDS

It is not just the author's experience that most engineering subjects are taught more straightforwardly when the lecturer utilizes the fruits of computer simulation and multimedia. The best possible way to show dynamical behavior of a technical system would

be to have the system readily at hand in the classroom. However only a very small minority of real-world systems can fit in a classroom. That is where photos, schematics, and even more videos and simulation programs come up front.

As soon as the dynamic systems behave as in real life and give the appropriate response to stimuli on-line, then the style of a lecture is changed revolutionarily. Instead of presenting the theory in a rather dry fashion, the background and the inner workings of the system becomes alive.

Furthermore, simulations can be provided to aid the students' work at home, to help them deepen their understanding of the – for them sometimes really arcane – processes. In essence they can “play around” with the process, make experiments without the danger of making something irreversibly wrong. They harness the power of simulation to the effect of reducing risks, time and expenses by operating the system model.

### .:PSIM.: APPLICATIONS

The main application area of .:PSim:. is in systems and control engineering courses at BUTE. One characteristic example of the very first programs is a simulated process with a controller (Fig. 6) that groups of 2-3 students have to use to prove their ability to tune controllers according to certain criteria (stability, speed, precision). There exist a real-world version of the illustrated three-tank water-level control loop, so the simulations' results could be compared to the process' actual responses.

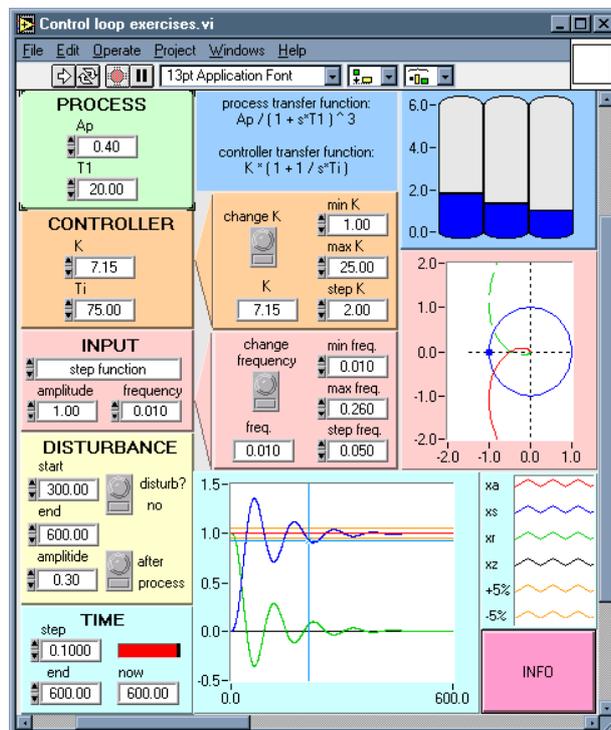


Fig. 6: .:PSim:. control engineering course example

Newer additions are .:PSim:. programs used to illustrate lectures, for example to show the connection between time and frequency domain, or to introduce and compare controller-tuning methods. Quite a large number of sample applications have been (and are continuously) developed to help students understand the theories. These samples are open to download from the department's web server with the necessary run-time application, so that they could be utilized without the LabVIEW development system. One such example aims to improve students' skill in sketching approximate Bode diagrams of transfer elements connected in series and comparing them to the exact diagram (Fig. 7).

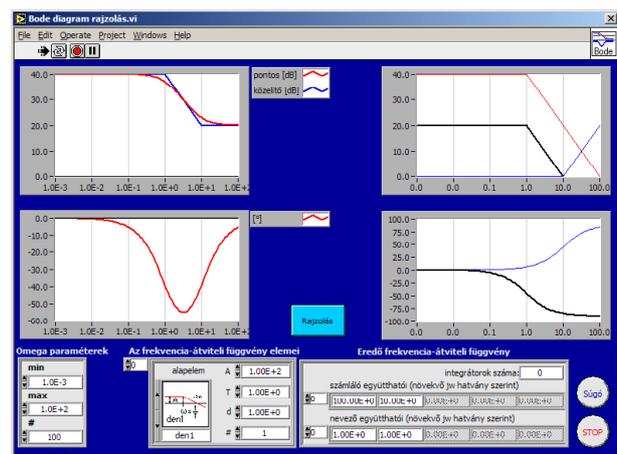


Fig. 7: Sketching an Approximate Bode Diagram

### CONCLUSION

.:PSim:. proved many times to be a powerful tool in education. .:PSim:. simulations have found quite a number of applications in systems and control engineering courses at BUTE. For one they serve as standalone practice assignments to test and enhance students' knowledge in the laboratory. They are used as an illustration prepared to help visualize real world dynamic processes in lectures. As a hands on tool .:PSim:. is used to solve simulation tasks that show up during the lecture. Furthermore students can use .:PSim:. blocks to build LabVIEW simulation programs of their own.

It is the nature of every program that its development never ends, so .:PSim:. stands before further enhancements, such as the general-purpose XML-based language mentioned above.

### REFERENCES

- Bronzino, J.D. (Editor-in-Chief). 1995. *The Biomedical Engineering Handbook*. CRC Press
- D'Argenio, D.Z., Schumitzky, A. 1997. *ADAPT II User's Guide: Pharmacokinetic/Pharmacodynamic Systems Analysis Software*. Biomedical Simulations Resource, Los Angeles
- Dorf, R.C., Bishop, R.H. 1998. *Modern Control Systems*. Addison Wesley Longman

- Kheir, N.A. (editor). 1995. *Systems Modeling and Computer Simulation*. Marcel Dekker, Inc.
- Wells, L.K.; Travis J. 1997. *LabVIEW for Everyone – Graphical Programming Made Even Easier*. Prentice Hall
- Zeigler, B.P., Praehofer, H., Kim T.G. 2000. *Theory of Modeling and Simulation*. Academic Press

## AUTHOR BIOGRAPHY



**PETRA ARADI** received her MSc and PhD in Mechanical Engineering at BUTE, in 1994 and 2000 respectively. She also obtained an MSc in Biomedical Engineering (BUTE, 2002). Since 1994 she works at the Faculty of Mechanical

Engineering of BUTE, presently as associate professor in the Department of Informatics.

Her teaching areas are systems and control engineering, as well as microcontroller applications, PLCs and Internet programming.

Her research interests cover these teaching areas, with the recent addition of co-operative mobile robotics.

# SIMULATION TOOLS IN CONTROL ENGINEERING EDUCATION

Jenő Kovács and Imre Benyó  
University of Oulu  
POB 4300, Linnanmaa, 90014 Oulun yliopisto, Finland  
Jeno.Kovacs@oulu.fi, Benyo@rit.bme.hu

György Lipovszki  
Budapest University of Technology and Economics  
Goldmann Gy. tér 3, V2 ép. 1111 Budapest, Hungary  
Lipovszki@rit.bme.hu

## KEYWORDS

Simulation, control, mathematical modelling, computer-aided education.

## ABSTRACT

The paper introduces simulation package developed for facilitating the education of discrete-time control theory at undergraduate level. The aim was not only to provide an interactive demonstration tool, but also to provide a better understanding of the applied algorithms. That aim was fulfilled by the choice of the LabVIEW programming environment, which allows the simplicity of graphical programming with the traceability of the analogue devices.

The simulation package provide basic blocks for identification and control: discrete-time filter, discrete-time state space models, identification blocks for output error method and equation error method, recursive least square estimation, Kalman-filter, state observer and general predictive control. Furthermore, the package contains a training purpose simulator demonstrating the so-called RST control structure. Beside demonstration, the simulator provides a good basis for control design.

Beyond classroom demonstration, the simulator can be an effective tool to intensify self-study and distance education. Several examples are shown to demonstrate the features of the new tool.

## INTRODUCTION

During the last decade, the development of educational and industrial software and simulation tools has been accelerated. Industrial applications focus on the replacement of expensive equipment by software tools (virtual equipment) and parallel the new technologies, e.g. Fieldbus, are strongly supported by high-tech software solutions. Also, numerous flexible software solutions for industrial human machine interface (HMI) and supervisory control and data acquisition (SCADA) have appeared in the market. The university education increasingly integrates such industry-standard programming-environment tools mainly in laboratory processes but more and more frequently also in the research and the classroom education. In education, the demonstration is the most common utilisation. Considering engineering education, demonstration involves process modelling and simulation, imitated data acquisition and process control. It requires high-

level graphical user interface providing efficient communication.

One of the most widespread industrial software used in education is the LabVIEW, a National Instrument product (National Instrument, 2003). The LabVIEW is a block-oriented graphical programming environment developed at first place for data acquisition and monitoring, but process control and modelling are also fully supported. Due to the additional toolboxes, the application area is continuously expanding.

Control engineering toolbox, TUBSIM, for mainly continuous-time modelling and control has been earlier developed in (Lipovszki and Aradi, 1995). The TUBSIM is a system simulation extension of LabVIEW interpreting an analogue computer in graphical programming environment. Besides the typical analogue computers elements, the TUBSIM Library contains different Boolean blocks, typical system engineering elements (low order transfer function elements, continuous time controllers like PI, PID, time delay) and some discrete time blocks. The TUBSIM Library is successfully used in the control engineering education at the Budapest University of Technology and Economics (Aradi, 1996).

Simulation-demonstration tools for discrete-time control engineering are presented in this paper. The tools are developed for the “Discrete-time control design” and the “Advanced control design” courses at the University of Oulu (Finland). A training purpose simulator supports the first course, providing a demonstration and control-design environment. The second is an advanced course, where one general-purpose simulator cannot cover all the topics. Therefore not complete simulators, but rather elementary blocks (e.g. state-space models, identification toolboxes) are developed and the users should build their own simulators. This way, the discussed algorithms can be better understood.

The paper first introduces the training purpose simulator – the RST simulator. Then the new blocks for advanced control engineering are described and the utilisation is demonstrated via a general predictive control example.

## RST CONTROL DESIGN

Applying the input-output, polynomial approach for system description, the training purpose simulator is based on the general presentation of a discrete-time

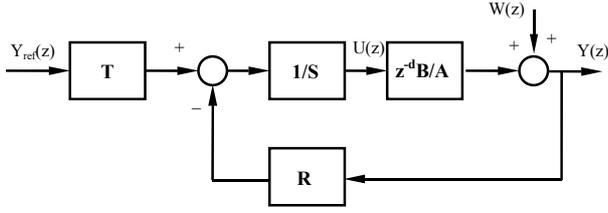


Figure 1: General two-degree-of-freedom control system.

controller: the two-degree-of-freedom controller, see in Figure 1. The process to be controlled is here defined by the pulse transfer function  $z^{-d} \frac{B}{A}$ . The controller is constructed from the three control polynomials, R, S and T; therefore it is called the RST control structure. The structure is very attractive, since most of the discrete-time controllers can be described by or transformed to the RST structure. The control design aims to define the R, S, and T polynomials to achieve certain required (dynamic and steady-state) performance for regulation and tracking, defined respectively by the pulse transfer functions

$$H_D(z^{-1}) = \frac{Y(z)}{W(z)} = \frac{A(z^{-1})S(z^{-1})}{A(z^{-1})S(z^{-1}) + z^{-d}B(z^{-1})R(z^{-1})}$$

$$H_{CL}(z^{-1}) = \frac{Y(z)}{Y_{ref}(z)} = \frac{B(z^{-1})T(z^{-1})}{A(z^{-1})S(z^{-1}) + z^{-d}B(z^{-1})R(z^{-1})}$$

The required regulation dynamics may be defined by the closed-loop characteristic polynomial,  $P_D(z^{-1})$ . Solving the Diophantine equation

$$A(z^{-1}) \cdot S(z^{-1}) + z^{-d}B(z^{-1}) \cdot R(z^{-1}) = P_D(z^{-1})$$

one can obtain the S and R polynomials. If required, additional n integrators can guarantee zero-steady state error in response to disturbance by replacing polynomial S by  $(1-z^{-1})^n S$ . Zero cancellation, which has a role in tracking, can be achieved by applying  $P_D B^+$  on the right hand side;  $B^+$  denotes the zeros to be cancelled as a factor of B,  $B = B^+ \cdot B^-$ . Further tracking requirement can be satisfied by the proper choice of polynomial T. Guaranteeing zero-steady state error in response to reference signal:

$$T(1) = \frac{P_D(1)}{B^-(1)}$$

Utilising the two-degree-of-freedom feature of the RST structure, the polynomial T may (partly) cancel the  $P_D$  dynamics, allowing the user to define different tracking dynamics via an external reference model,  $B_m/A_m$  as:

$$T(z^{-1}) = \frac{P_D(z^{-1})}{B^-(1)} \quad \text{and} \quad \frac{Y(z)}{Y_{ref}(z)} = \frac{B_m(z^{-1})B^-(z^{-1})}{A_m(z^{-1})B^-(1)}$$

Pole-placement with implicit reference model is another

alternative in (Åström and Wittenmark, 1997). In that case, the R, S and T polynomials are designed to satisfy the requirement

$$H_{CL}(z^{-1}) = \frac{Y(z)}{Y_{ref}(z)} = z^{-d} \frac{B_m(z^{-1})B^-(z^{-1})}{A_m(z^{-1})B^-(1)}$$

## RST simulator

The RST simulator was developed in LabVIEW graphical programming environment. The LabVIEW provides a block-oriented programming structure, well facilitated with additional toolboxes. The current simulator requires special continuous- and discrete-time blocks, which are available from the TUBSIM toolbox (Lipovszki and Aradi, 1995), while others are described in (Benyó *et al.* 2003).

The user interface of the simulator, shown in Figure 2, has three main areas: the *RST structure* in the upper right hand corner, the *graphical windows* below it, and the *parameter windows* on the left hand side.

The *RST structure* illustrates the control structure and provides several option for:

- choosing the type of process model (continuous- or discrete-time),
- setting the parameters of the reference signal,  $Y_{ref}$ , (amplitude, period, start time) and the disturbance signal, W, (period and amplitude of a deterministic stepwise signal, amplitude and mean value of a stochastic noise),
- applying (or not) integral action,
- choosing the type of T (polynomial or scalar),
- utilising (or not) a reference model,  $\frac{B_m}{A_m}$ .

The *graphical windows* plot the reference signal, the output signal and the sampled output (in case of continuous-time process) in the upper window, and the control input is shown in the lower one.

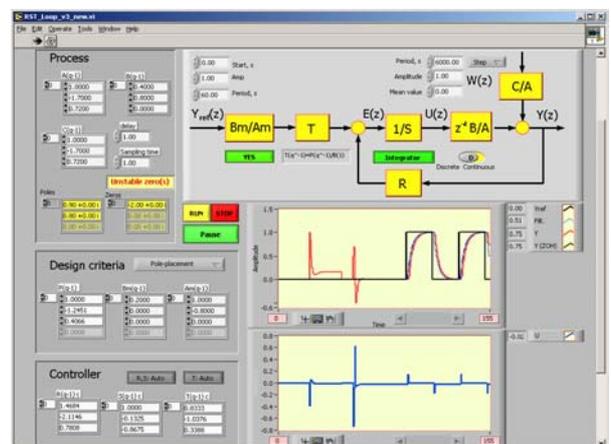


Figure 2: The user interface of the RST simulator.

The user can define the process, set the control design criteria and calculate or set manually the controller parameters in the *parameter windows*. The process parameters are the polynomials (A, B, and C) of a pulse transfer function (in discrete-time) with the time delay (d) or those of a transfer function (in continuous-time). The zeros and poles of the process are automatically calculated and warning is given when any of those is unstable. It has an importance, *e.g.*, when considering zero cancellation. After the process was defined, the control design shall be started. The user may choose from the list of offered controllers:

- pole placement controller,
- pole-zero placement controller,
- dead-beat controller, and
- minimum-variance controller.

The user sets the polynomial  $P_D$  for pole- and pole-zero placement controller (see under *Regulation criteria*), but for dead-beat and minimum variance controllers, the simulator automatically selects the  $P_D$  as  $P_D(z^{-1})=1$  and  $P_D(z^{-1})=C(z^{-1})$ , respectively.

The *tracking criteria* are determined by the choice on:

- polynomial T: if scalar then it effects only the steady-state error, if polynomial it allows to utilise the 2dof property,
- the existence of explicit reference model; if chosen, the  $B_m$  and  $A_m$  polynomials are set here.

It is important to emphasise that the choice of polynomial T and the reference model allows setting the tracking behaviour to be different from the regulation dynamics.

Based on the selected control methods, the *Controller window* automatically plots the resulted control polynomials. However, the user has the freedom to choose those manually to test any other controllers, which are not necessarily offered by the simulator. Especially, when continuous-time process is chosen, since it requires manual introduction of the control polynomials.

Several features results in a user-friendly interface, such as displaying only the necessary elements; *e.g.* the polynomials of the reference model appear only when the user selects to apply one; or any modification in process parameters or design criteria is immediately updates the R, S and T polynomials. The simulation can be paused at any time for changing signal parameters.

### Demonstration example

The RST simulator offers numerous possibilities for demonstrating or designing control structures. During an introductory course to control theory, simple example helps to understand the performance degradation caused by a stepwise output disturbance signal and how to eliminate it by applying an integrator

in the forward loop. The difference between the regulation and tracking trajectories can be easily visualised by selecting the proper controller parameters in a fast and simple manner. Later, advanced students may design their own controllers and easily test those using the simulator. Two demonstration examples are here discussed: the design steps of a pole-placement controller and the illustration of intersampling ripple phenomenon.

#### Pole-placement controller

The design steps of the pole-placement controller can be summarised the following way: a) design of the output disturbance elimination based on the required closed-loop characteristic polynomial, b) introducing additional integration action to avoid non-zero steady-state error in response to output disturbance, c) ensuring zero-steady state error in reference signal following and finally d) selecting a tracking dynamics to emphasise the two-degree-of-freedom feature of the RST control structure.

Let the process to be controlled, using  $h=1$  sec sampling time:

$$\begin{aligned} A(z^{-1}) &= 1 - 1.7z^{-1} + 0.72z^{-2}, \\ B(z^{-1}) &= 0.4 + 0.8z^{-1}, d = 1. \end{aligned}$$

The desired regulation dynamics is

$$P_D(z^{-1}) = 1 - 1.2451z^{-1} + 0.4066z^{-2}.$$

First, the simulation is run without integral action, therefore non-zero static error remains when a step output disturbance occurs. Applying the integration, the error can be eliminated. As the last step, the response to change in reference signal is tested with an explicit reference model. The responses in these three steps can be shown in the same graphical window, as seen in Figure 2, to support the full understanding.

#### Intersampling ripple

Intersampling ripple is a specific, discrete-time control phenomenon. Due to the sampling, information about the process performance can be obtained only at the sampling instants. If the continuous-time signal oscillates between the sampling instants, intersampling ripple occurs. Although this phenomenon has a remarkable importance in real-time application, it can be easily overlooked during control design when the user applies a discrete-time model for the process.

Avoiding this mistake, the RST simulator allows to define the process in continuous-time and to apply discrete-time controller. Consider the following double integrator process to be controlled,  $G(s) = 1/s^2$ . The control task is to achieve an open-loop behaviour defined by

$$H_{OL}(s) = \frac{0.2}{s(1+s)}.$$

The discrete-time transfer functions assuming ZOH and

sampling-time  $h = 1$  sec are

$$G(z^{-1}) = 0.5 \frac{z^{-1} + z^{-2}}{(1 - z^{-1})^2}$$

$$H_{OL}(z^{-1}) = 0.074 \frac{z^{-1} + 0.718z^{-2}}{(1 - z^{-1})(1 - 0.368z^{-1})}$$

The simplest open-loop controller is

$$H(z^{-1}) = \frac{H_{OL}(z^{-1})}{G(z^{-1})} = 0.148 \frac{(1 + 0.718z^{-1})(1 - z^{-1})}{(1 + z^{-1})(1 - 0.368z^{-1})}$$

In RST structure it can be applied as:

$$T(z^{-1}) = 0.148(1 + 0.718z^{-1})(1 - z^{-1})$$

$$S(z^{-1}) = (1 + z^{-1})(1 - 0.368z^{-1})$$

$$R(z^{-1}) = 0$$

The control inputs response to an impulse reference signal is

$$U(z) = H(z^{-1})Y_{ref}(z) = 0.148 \frac{(1 + 0.718z^{-1})(1 - z^{-1})}{(1 + z^{-1})(1 - 0.368z^{-1})} 1$$

The critically stable pole of the control input causes the oscillation in the control signal as therefore in the controlled output, as shown in Figure 3.

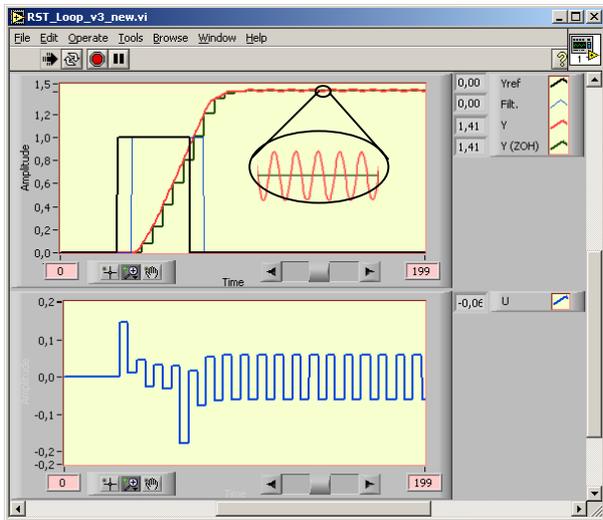


Figure 3: Intersampling ripple phenomenon. The upper graphic window also shows how the continuous-time signal oscillates while the sampled signal is constant.

## ADVANCED CONTROL DESIGN

The aim of the development of the new blocks was to facilitate the simulation of the advanced identification and control techniques. In the present state of the development, the new block library includes tools for SISO discrete time systems. The algorithms of the

implemented methods can be found in (Ikonen and Najim, 2002).

The new blocks can be divided into four subgroups according to their task: process modeling, identification tools, state observers, general predictive controller. A pop-up help window provides a brief description of each block.

### Blocks for process modelling

The new set of blocks provides several possibilities for process modelling: the input-output approach and the state-space approach. The input-output approach defines the process by its pulse transfer function. The block “ $\frac{B(q^{-1})}{A(q^{-1})}$ ” represents the pulse transfer function, defined

by the coefficients of the B and A polynomials.

The state-space subgroup consists of three different blocks. The “sys SS” block defines the process model in state-space equation. The user shall define the state-space matrices.

In control and observer design, the controllable and observable canonical forms of the state-space equation are generally applied. If the process model is only available in the form of pulse transfer function (e.g. identification in input-output approach), then the necessary state-space equation can be obtained by the “SS contr” (State-Space controllable) and the “SS obs” (State-Space observable) blocks.

### Blocks for identification

The identification subgroup contains three blocks: the recursive least square algorithm (RLS), the equation error (EE), and the output error method (OE).

The “RLS” block estimates the parameters of a given regression model using the recursive least square method. The input of the block is the regression vector and the output is the correlation vector. The covariance matrix of the parameter adaptation algorithm can be maintained by applying the offered methods: forgetting method, constant trace method or the bounded information algorithm. Since, the “RLS” block requires only the regression vector as an input, the user has the freedom to choose the model structure; consequently, the user has to create that vector.

In the “EE” and “OE” blocks, the structure of the regression model is a priori defined by the equation error method and the output error method, respectively. The parameter adaptation algorithm is in both cases the RLS algorithm. The block’s inputs are the input and output signals of the process to be identified and the order of the model. The blocks return the estimated transfer functions given by the coefficients of the numerator and denominator polynomials, the output of the regression model, and the covariance matrix.

### Blocks for State observer

Among the state observer blocks, one can choose between the Kalman-Filter (“ $\hat{X}_{KF}$ ”) and the “Fixed Gain State Observer (“ $\hat{X}_{FGF}$ ”).

The implemented Kalman-Filter estimates the state variables at the  $(k+1)^{th}$  sampling instant based on  $k^{th}$  process input and  $(k+1)^{th}$  output measurement:  $x(k+1) = f(u(k), y(k+1))$ . The outputs of the block are the state vector and the trace of the covariance matrix.

The Fixed Gain State Observer estimates the state variables in the  $(k+1)^{th}$  instant based on  $k^{th}$  process input and  $k^{th}$  output measurement. The output of the block is the  $(k+1)^{th}$  state variables vector:  $x(k+1) = f(u(k), y(k))$ .

The state-observers utilise the observable canonical form of the state-space model. The required form can be generated by the “SS obs” (State-Space Observable) block, which transforms the transfer function into observable canonical state-space form.

### Blocks for General Predictive Controller

There are two blocks relating to the GPC control algorithm. The “GPC M” calculates the gain matrices of the controller; the “GPC” block is the controller. The block realizes a GPC for SISO process. The controller is based on the state-space model of the controlled process.

The inputs of the block are the state variables, the controlled process output signal and the reference signal. The main output of the block is the control signal. (There is another signal for graphical purpose, the desired future process outputs on the prediction horizon.)

Besides the general parameters of the GPC algorithm (prediction, minimum and control horizons, weighting factors of the control error and control signal in the cost function; there is possibility to use weighting matrices for enabling weighting the terms in the appearance of time), the pulse transfer function of the desired closed-loop behavior can be defined.

In the case of online identification, it is possible to use the new identified model of the process in every time step, otherwise the time demanding computation of the gain matrices is performed only in the first time step (Adaptive button on/off).

### Demonstration examples

The outlook and the utilisation of the developed blocks are demonstrated in the following two examples. The first example is an on-line identification combined with state observation, while the second example presents a GPC control structure.

### Identification and state estimation

In this example, a simple parameter estimation and state estimation problem are solved. The “unknown” process is on-lined identified using the transfer function approach (output error method), and the identified model is transformed to state-space model that is used for the estimation of the state variable (Kalman-filter). The LabVIEW realisation, shown on Figure 4, clearly demonstrates the simplicity of programming.

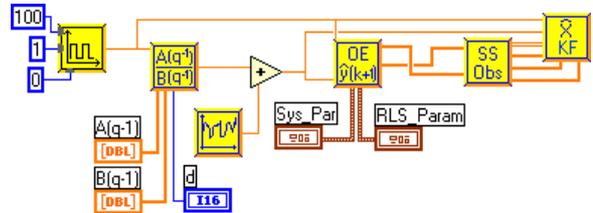


Figure 4: The program of the simulation without the time setting and graphical blocks.

The process is modelled by the output error method as

$$y(k) = \frac{B(q^{-1})}{A(q^{-1})}u(k) + e(k)$$

The pulse transfer function is changed at the 75<sup>th</sup> sampling instants as

$$\frac{B(q^{-1})}{A(q^{-1})} = \frac{bq^{-1}}{1 + aq^{-1}} = \frac{0.2q^{-1}}{1 - 0.8q^{-1}} \Rightarrow \frac{0.2q^{-1}}{1 - 0.6q^{-1}}$$

The  $e(k)$  noise is a Gaussian white noise with the variance 0.1. Within the RLS algorithm of the OE parameter estimation, the forgetting factor  $\lambda = 0.98$  was applied to maintain the covariance matrix.

The state-space model is simple since the process is a first order,

$$\begin{aligned} x(k+1) &= ax(k) + bu(k) \\ y(k) &= x(k) \end{aligned}$$

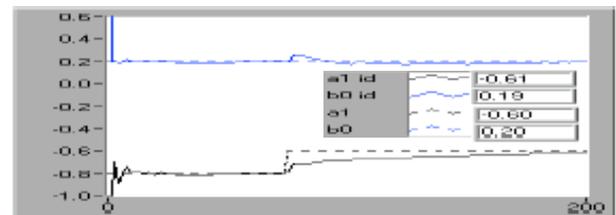


Figure 5a: LabVIEW graph: parameter estimates; true parameters (dashed lines), estimated ones (solidlines)

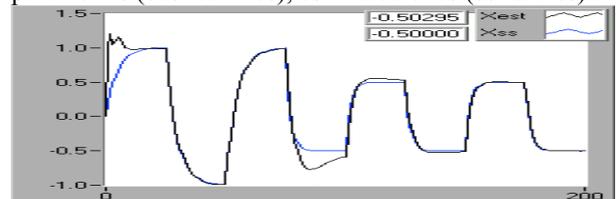


Figure 5b: LabVIEW graph: process and model state variables “Xss” and “Xest”.

The initial values of the parameters were set  $[a \ b]_{t=0} = [-1 \ 0.6]$ . Figure 5a demonstrates the parameter convergence while Figure 5b shows the process and model state variables,  $x(k)$  and  $\hat{x}(k)$ , respectively.

### General Predictive Control

This example demonstrates the application of the General Predictive Control block. The process output ( $y$ ) is controlled by the GPC to follow the reference signal ( $r$ ) with a prescribed tracking behaviour. The reference signal is a square wave signal, and the process has Gaussian white noise type measurement noise. The Fig. 8 shows the LabVIEW program of the simulation, with all of its accessories for graphical purpose. The Control Panel of the Simulation is shown on the Figure 6. On the control panel there are all of the numerical controls with which it is possible to change the parameters even in run-time.

Furthermore, all the process and control parameters are displayed on the control panel. The results of the simulation can be followed through the graphs and the numerical displays.

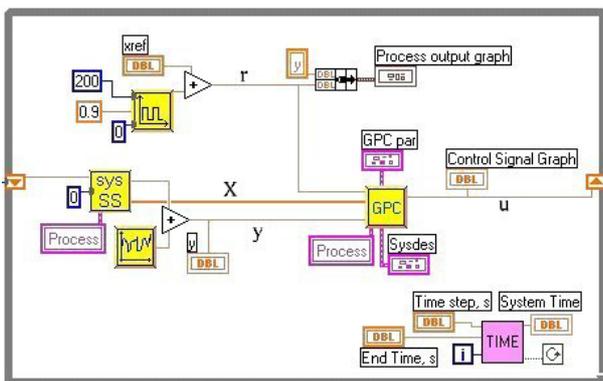


Figure 6: The program diagram of the simulation, with the same signal notation as on the previous figure.

### CONCLUSION

Computer-aided education may be an advantageous solution for increasing the efficiency of control education. The classroom education, similarly to laboratory exercises, may be further visualised by introducing target-oriented simulation/demonstration environment.

A package of advanced simulation tools is developed to serve particular courses on digital control theory and advanced control engineering. The developed simulation elements provide an easy-to-use and easy-to-understand manner to deepen the enhanced knowledge of the course.

The presented tools support the simulation of discrete-time modelling, identification, state-observer and general predictive control design. The programming environment was chosen to be the LabVIEW due to its attractiveness.

The RST simulator facilitates the discrete-time control

education, especially supporting the polynomial input-output approach for system description. The basic structure is the two-degree-of-freedom formulated by the control polynomials,  $R$ ,  $S$  and  $T$ .

Further advantage of the LabVIEW based simulator is that it can be transformed into a self-executable file and run it in any Windows based operation system. It requires only a runtime engine, which is freely available from the web site of the National Instrument.

The proposed simulator has been evaluated in practice in a course "Digital control theory" at the University of Oulu and will be used in co-operation with the Budapest University of Technology and Economics.

### REFERENCES

- Benyó, I.; Lipovszki, Gy. And Kovács, J. 2003. "Advanced Control Simulation Tools In LabVIEW Environment", *Proc. 6<sup>th</sup> IFAC Symposium on Advances in Control Education*, Finland, 275-279.
- Lipovszki, Gy. And Aradi, P. 1995. "General Purpose Block Oriented Simulation System Using LabVIEW", *NIWeek'95*, Austin, TX, USA.
- National Instruments, 2002. *LabVIEW User Manual* at <http://www.ni.com/pdf/manuals/>
- Åström, K.J. and Wittenmark, B. 1997. *Computer-Controlled Systems* (Prentice-Hall)
- Aradi, P. 1996. "Using LabVIEW in Education of Systems and Control Engineering", *NIWeek'96*, Austin, TX, USA.
- Ikonen, E. and Najim, K. 2001. *Advanced Process Identification And Control*, Dekker.



**JENŐ KOVÁCS**, born in Hungary in 1967, (M.Sc. 1991 Budapest, Hungary, Ph.D. 1998 Oulu, Finland) is a senior assistant at the Systems Engineering Laboratory, University of Oulu, Finland. His research interests include adaptive control, constrained control, advanced modelling and their application to energy systems and power plant control problems.



**IMRE BENYÓ** is born in Budapest, Hungary in 1975. He was graduated at the Technical University of Budapest, as mechanical engineer. He is researching at the System Engineering Laboratory,

University of Oulu, Finland. His research area covers the predictive control, system identification problems, and its applications in the power plant control.



**GYÖRGY LIPOVSZKI** was born in Miskolc, Hungary and went to the Budapest University of Technology and Economics, where he studied electronics and graduated

in 1975. He is now an Associate Professor at the Department of Department of Production Informatics Management and Control and his research field is development of different simulation frame systems.

# Simulation Tool for Cooperative Mobile Robots

Petra Aradi  
Department of Informatics  
Faculty of Mechanical Engineering

Budapest University of Technology  
and Economics  
Műgyetem rkp. 3., Budapest  
H-1111, Hungary  
e-mail: petra@rit.bme.hu

Miklós Zs. Soós  
Department of Informatics  
Faculty of Mechanical Engineering

Budapest University of Technology  
and Economics  
Műgyetem rkp. 3., Budapest  
H-1111, Hungary  
e-mail: soos@rit.bme.hu

Edit Stevensné-Száday  
Department of Building Services  
Engineering  
Faculty of Mechanical Engineering  
Budapest University of Technology  
and Economics  
Műgyetem rkp. 3., Budapest  
H-1111, Hungary  
e-mail: szaday@rit.bme.hu

## KEYWORDS

mobile robots, cooperation, simulation

## ABSTRACT

This paper reports on the first attempts to create a simulation program for cooperative mobile robots. The organization of the robot family is borrowed – at least to some degree – from nature. A bee-hive was chosen as the organizational basis of the robot family. The queen of the hive is represented by a PC, the workers are mobile robots. Mobile robots can only communicate with the queen via an RF link. The task for the robots is to patrol an area and check the state of lights (operational, emergency, or darkness). The position of the robots are given by ceiling-mounted cameras directly to the queen. The queen can modify the route of the robots with directions via RF. The simulation system is developed in LabVIEW. Based on the results of simulation experiments robots are being constructed and used in a test environment similar to the simulated one.

## INTRODUCTION

Intelligent electronic systems penetrate everyday life as it stands. The advent of this era began, when transistors first appeared in the 1950s. The miniaturization of transistor-based systems, especially of digital computing devices opened up a totally new world in electronics. Microprocessors, personal computers and microcontrollers are common nowadays not just in industrial applications, but in customer electronic products, too. Self-contained devices with onboard processors and controllers are to be found almost everywhere: in cars, cameras and mobile phones, just to mention a few everyday gadgets. Household appliances, like washing machines, refrigerators, and vacuum cleaners are also beginning to be equipped with such a brain. An even faster growing area in applying intelligent electronics is the toy-market.

The word robot comes from Karel Capek and is now applied to machines that accomplish certain tasks without getting tired or bored, always with the same precision and quality. Industrial robots are widely used

for assembly, welding, painting, transporting heavy objects and other tough tasks, often in environments not suitable for humans. Industrial robots are characteristically fixed in arrangement; they can not change their location, just move one or more “limbs”. Mobile robots on the other hand are autonomous machines, and are able to do locomotion.

## Mobile Robots

A mobile robot could very well be defined as a mobile machine that interacts with its environment through sensors, and attempts to achieve some objective. The objective could be relatively simple e.g. following a line drawn on the floor, or it could be quite difficult e.g. operating as a member of a team of robots playing soccer. Mobile robots are to be found in toy boxes of today’s children in an exponentially increasing number, they even seem to be a must. Robot pets, like Sony’s Aibo dog (Fig. 1) are equipped with a special purpose computer, that makes the robot understand and fulfill commands, process sensory data and handle accordingly.



Fig. 1. Sony’s Aibo

## Robotic Kits from Toy Manufacturers

Popular construction toys like LEGO or fischertechnik have robotic kits (Figures 2 and 3). These kits inherit the building-block approach from their predecessors and contain at least the few most important elements for robot construction. Programming the robots in both toys is done – among other freely available tools – with a block- and function-oriented graphical language.

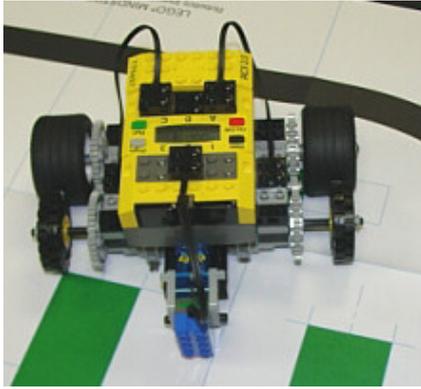


Fig. 2. A LEGO robot



Fig. 3. fischertechnik robots

## MOBILE ROBOTICS IS INTRODUCED IN THE DEPARTMENT OF INFORMATICS, BUTE

Staff and students of the Department of Informatics (Faculty of Mechanical Engineering, Budapest University of Technology and Economics) started designing and building mobile robots from LEGO and fischertechnik kits in the end of 2003. These kits were chosen because they offer a cheap and flexible alternative to robotics kits and ready-made mobile robots.

### Stand-alone Robots vs. Robot-families

Robot-building started quite naturally with the sample designs from the manuals, however individual designs appeared very soon afterwards. Stand-alone mobile robots were the first construction tasks, nevertheless such self-contained robots require a very complicated brain and they are prone to errors more, than a team of similar, but a bit dumber robots organized in a “family”. The organizational structure for the cooperative robot family is borrowed from nature; ants and bees were chosen as the basis, although the analogy is far from thorough. The Robot Queen represents the central intelligence, being the organizer in the family, and coordinating the workers. Workers can communicate only with the Queen as of now, however communication among workers is also under consideration.

## Let Us Simulate

An MSc thesis dealing with the development of a robot-family has been written and successfully defended at the Department of Informatics in June 2004. This thesis, and all the preparations turned the attention to simulate robot-families.

From the first development efforts it very soon occurred, that a simulation system for such robot families helps a great deal to create the hierarchical, organizational and control structure of the family. As it is very well known the power of simulation is in reducing risks, time and expenses by operating the system model instead of the actual system. The above-mentioned advantages of simulation are especially emphasized, when the system model is implemented as a computer program.

That is why a simulation system for studying the behavior of robot-families was envisioned, and is now under development and testing.

## SIMULATING A ROBOT-FAMILY

### The First Project

First of all a suitable assignment had to be defined with the details clearly worked out. The first project is set to deal with a building services application. Mobile robots move around in a building to check the operation of certain devices, like light, heating and air conditioning. Robots check for example, whether lights are working properly or lights are working unnecessarily (no one is in the room). As this task is a rather complex one, only the operation of lights is examined first.

There is an area specified (Fig. 4), where certain parts have to be lighted always, such as corridors leading from offices to staircases and lifts. Lights can have three states: normal operation, emergency and no lights at all. Robots have to patrol the area in regular time intervals and report the state of lights. Robots have to communicate their position to each other, so that they can cover the whole area optimally.

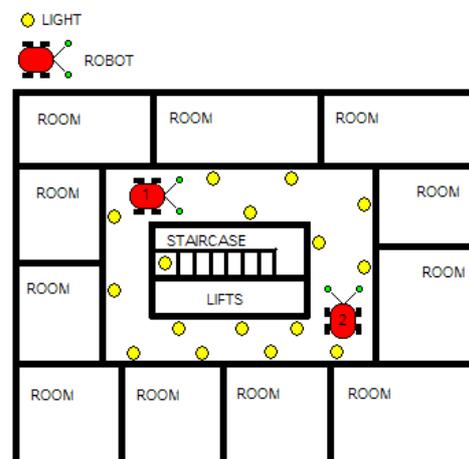


Fig. 4: Environment Layout

Each robot is equipped with an RF transmitter, but they can only communicate with their superior, the Queen. The position of each robot is detected with the help of ceiling-mounted cameras, just like in robotic soccer, however robots can detect their position from special markings on the ground, too and then transmit it to the Queen. The information from the cameras goes also into the Queen. The Queen in this case is PC equipped with the appropriate RF devices to communicate with the robots. In the simulated environment the Queen is a different program task running on the same PC as the simulation. Cameras are substituted by the continuously computed trajectory of the robots projected on the layout. In certain positions of the layout markings are placed and when a robot moves over such a marking, its detection is simulated.

The speed of the robots can either be zero or a slow fixed speed. Acceleration to and deceleration from this speed is almost instantaneous, that makes simulation a lot more easier. A robot's speed is also zero, when it has to align its orientation according to the commands from the Queen.

The lights can be switched by a preprogrammed algorithm, or by the user. As of now, there are no external obstacles allowed in the area.

### Implementation

National Instruments' LabVIEW was chosen as a development tool. LabVIEW stands for Laboratory Virtual Instrumentation Engineering Workbench, and is available in various computer platforms, such as Linux, MacOS, Windows, Solaris, HP-Unix. LabVIEW applications are called virtual instruments or VIs for short. VIs have a user interface as shown in Fig. 5 (front panel in LabVIEW terminology) and a block diagram (Fig. 6), where the actual program is built with LabVIEW's graphical components.

The ease of use and programming, together with LabVIEW's excellent connection to the outside world through data-acquisition, I/O and network protocols makes it the ideal tool for scientists and engineers.

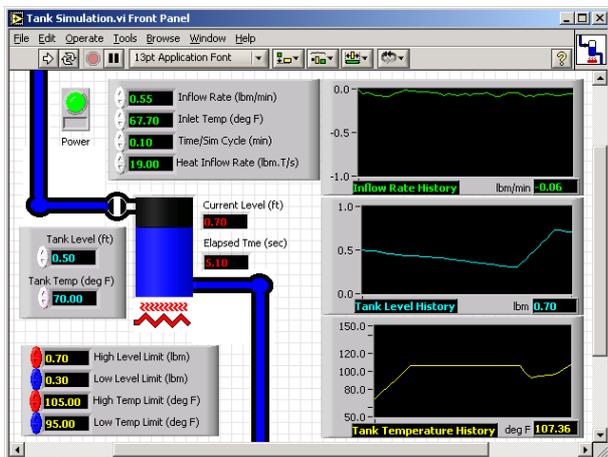


Fig. 5: Front Panel in LabVIEW

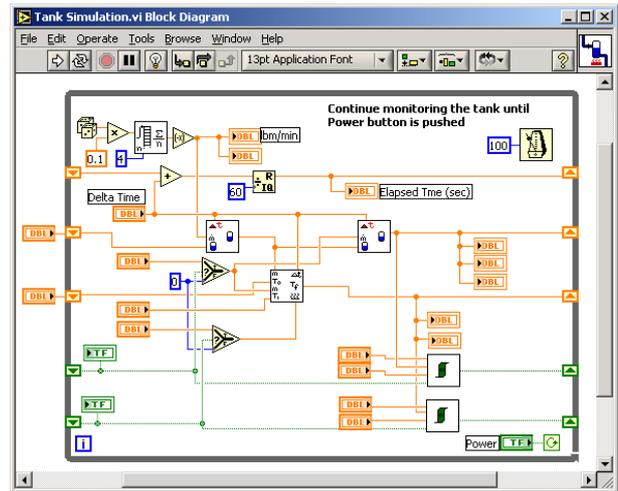


Fig. 6: Block Diagram in LabVIEW

The above mentioned properties of LabVIEW make it an ideal environment for rapid application development, that is one of the major reasons for the authors' choice.

Fig. 7 shows the main screen of the mobile robot simulator program.

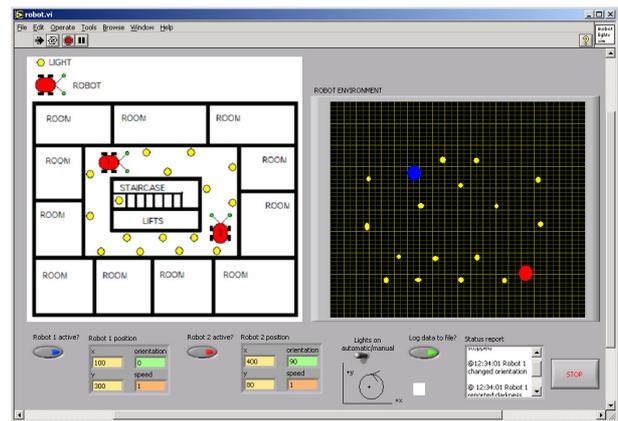


Fig. 7: Mobile Robot Simulator Front Panel

The mobile robot simulation program has the following functionality:

- maximum two robots can be simulated simultaneously
- the operation of lights can be set to a predefined time-sequence or
- lights can be operated manually
- the ceiling-mounted cameras' information is simulated by the calculated position of robots
- robots' reports on the state of lights is written to the screen and optionally to a log file

### FURTHER PROJECT IDEAS

To check the operation of heaters and air conditioners, temperature and humidity sensors have to be mounted on the robots. Further on an intelligent control can be developed, which uses the number of people in the certain room, not just temperature and humidity values. To accomplish this task either a passive door-mounted

people-counter has to be used or wall-mounted cameras are necessary or even robots can count people.

Besides it is important to develop the necessary communication among robots, something similar to the queen-robot RF links.

Such intelligent robotic wardens are a good alternative in older building, where there is no building intelligence installed. To build and intelligent building costs less, when new premises are build, however upgrading an existing structure could cause problems, not just the financial art, but the noise and inconvenience of construction work, too.

## CONCLUSION

The very early development stage of a cooperative mobile robot family was presented. During the work quite a number of aspects came up. These aspects are important for both the actual construction and programming of the robots, as well as the development of the simulation system.

The results show, that when the actual robot control and communication is developed, previous simulation experiments provide considerable help.

Thanks has to be expressed to FÖGÁZ Rt. for their support to bring mobile robotics within reach.

## REFERENCES

- Dorf, R.C., Bishop, R.H. 1998. *Modern Control Systems*. Addison Wesley Longman
- Ferrari, M., Ferrari, G. and Hempel, R. (technical editor). 2002. *Building Robots with LEGO Mindstorms*. Syngress Publishing, ISBN 1-928994-67-9
- Kheir, N.A. (editor). 1995. *Systems Modeling and Computer Simulation*. Marcel Dekker, Inc.
- McComb, G. 2001. *The Robot Builder's Bonanza*. 2nd ed. McGraw-Hill, ISBN 0-07-136296-7
- Nehmzow, U. 2003. *Mobile Robotics: A practical Introduction*. 2nd ed. Springer, ISBN 1852337264
- Niku, S.B. 2001 *Introduction to Robotics. Analysis, Systems, Applications*. Prentice Hall, ISBN 0-13-061309-6
- Szabó, R. 2001. *A mobil robotok szimulációja (Simulation of Mobile Robots)* ELTE Eötvös Kiadó, Budapest, Hungary, ISBN 963 463 476 1
- Wells, L.K.; Travis J. 1997. *LabVIEW for Everyone – Graphical Programming Made Even Easier*. Prentice Hall
- Zeigler, B.P., Praehofer, H., Kim T.G. 2000. *Theory of Modeling and Simulation*. Academic Press

## AUTHOR BIOGRAPHY



**PETRA ARADI** received her MSc and PhD in Mechanical Engineering at BUTE, in 1994 and 2000 respectively. She also obtained an MSc in Biomedical Engineering (BUTE, 2002). Since 1994 she works at the Faculty of Mechanical Engineering of BUTE, presently as associate professor in the Department of Informatics.

Her teaching areas are systems and control engineering, as well as microcontroller applications, PLCs and Internet programming.

Her research interests cover these teaching areas, with the recent addition of cooperative mobile robotics.



**MIKLÓS ZS. SOÓS** received his BSc in Mechanical Engineering at János Bolyai Military Technocal College (BJKMF), in 1998, then an MSc also in Mechanical Engineering at BUTE, in 2004.

He starts his PhD studies in academic year 2004/2005 in the Department of Informatics, Faculty of Mechanical Engineering, Budapest University of Technology and Economics. His main research interest is cooperative semi-autonomous mobile robotics.



**EDIT STEVENÉ-SZÁDAY** received her MSc in Mechanical Engineering at BUTE, in 1994. She has been working at the Faculty of Mechanical Engineering of BUTE since 1994, recently as an assistant researcher. Her main research interest is

Displacement ventilation and simulation in which she is completing her dissertation. Her teaching areas include ventilation and air-conditioning systems.

# THE SIM-SERV ASSOCIATION: SERVICES FOR USERS AND SUPPLIERS OF SIMULATION IN PRODUCTION AND LOGISTICS

Johannes Krauth, Sim-Serv Quality and Services Manager  
Adolf-Reichwein-Str. 32, 28329 Bremen, Germany  
Email: Johannes.Krauth@sim-serv.com

**Abstract:** The Sim-Serv Association, which was founded with financial support by the EU, offers a range of services to users and suppliers of simulation. The core services for potential users are its website, the help desk and individual consultation by a team of neutral experts. For researchers and suppliers, there are platforms to present themselves, to publish results and news on achievements, to join or initiate working groups and find partners for joint projects. The Sim-Serv association is open to new members at any time. The paper explains its services and how to take advantage of them.

## INTRODUCTION

Sim-Serv has been set up as a “Virtual Institute” in the year 2001. It received financial support from the EU Fifth Framework Programme for Research and Development. It is one of 17 Virtual Institutes which focus on different areas of technology. Their mission is to stimulate and co-ordinate application oriented research and development in dedicated areas, and to ensure a smooth transition of results into applications in industry and society. Each Virtual Institute (VI) is financed by the EU for a limited period, but is meant to operate beyond this period, hence each VI will establish a self-financing organisation which will carry on the activities at the end of the EU funded project. To this end, the Sim-Serv Association was founded in spring 2004. It is a non-profit organisation where every organisation involved with simulation (researchers, developers, vendors, users) can become a member.

Sim-Serv ([www.sim-serv.com](http://www.sim-serv.com)) focuses on product- and production-oriented simulation. It provides practical support to researchers and developers, to those offering tools and services on a commercial basis as well as to industrial users. The following sections of this paper provide a rough survey of the status of simulation in Europe, explain the objectives of Sim-Serv, its organisation, and the services it is offering to its members and to potential users in industry.

## SIMULATION IN EUROPE: THE CURRENT SITUATION

It is commonly accepted that simulation – in spite of its obvious power and benefits – is not widely used in industry, clearly not as wide as it should be. The

estimated potential for savings and improvement in European industry, which could be achieved by proper use of simulation, is enormous. And those companies who used it express a high degree of satisfaction. In the nineties, simulation was considered a rapidly growing market, and the American Integrated Manufacturing Technology Initiative ([www.imti21.org](http://www.imti21.org)) ranked simulation as one of the four most important technologies for future manufacturing. It states that ‘no other approach offers more potential for improving products, perfecting processes, reducing design-to-manufacture times, and reducing product realisation costs’.

In reality however, the quantitative growth of simulation was rather slow – the growth of its use in European industry as well as the growth of European simulation suppliers.

A study of the use of system simulation in UK manufacturing industry /1/ showed a 92% satisfaction rate amongst users, yet a penetration into industry of less than 10%. That pattern is reported as common across Europe /2/. A more recent study shows that not much has changed in the last decade /3/.

We see three major reasons for this slow take-up: Sorted by priority, they are:

1. Simulation and its benefits are still insufficiently known in industry, especially among those who make the decisions whether to use it or not /4/.
2. Simulation is difficult to justify from an economic point of view. Simulation is often used to prove the viability of a system design. But what is the added value of such a proof in terms of money? Maybe the model helps improve the design, but maybe the improvement is also possible without simulation? It is very difficult to make a convincing financial case for simulation: The extra expenses of using it can be estimated, but not the savings or benefits achieved.
3. Simulation appears as an extra effort, compared to the “conventional” way of working. It is not fully integrated in current planning and management procedures. Problems that should indeed be solved using simulation used to be solved without simulation in the past. Obviously, people who work on such problems and do not (yet) use simulation need a strong motivation to start using it.

On the other hand, the supply side in Europe consists of numerous small or even micro-enterprises, offering in many cases highly specialised tools and solutions. Many

of them are recently born spring-offs from research institutes. Their products are often of top quality, but for such small companies it is not easy to get Europe-wide visibility and find customers from a broad range of industry sectors. It is especially difficult because the market is dominated by a number of powerful globally operating suppliers of general simulation packages with professional marketing and sales activities. Hence, in spite of technical brilliance, European products and service suppliers find it difficult to survive on the European market.

To summarise: Excellent technology is available, but not applied to the extent it should. Both sides – suppliers as well as users – are suffering from this relatively low level of application. And many opportunities for further development are lost because the communication of both sides is insufficient.

## THE OBJECTIVES OF SIM-SERV:

### Promote the Application of Simulation in European Industry

The main objective of Sim-Serv is to stimulate a wider use of simulation technology in European industry and thus help European companies (especially small and medium-sized enterprises) meet the challenges of global competition. What is required to achieve this wider use? If the three barriers stated above are correct, then it is necessary

1. to spread information about simulation to industry managers, and provide detailed information on the benefits achieved by other users.
2. to gather data about cost and benefits of simulation application in industry, in various sectors and for a range of different problems.
3. to make simulation a standard tool for daily work of engineers and managers, and to integrate it into existing methodologies and tool sets. Until this has been achieved, simulation services of proven quality should be easily available at reasonable cost.

Therefore, Sim-Serv's main activity is dissemination: providing a central entry point for those seeking information, guidance and support. Newcomers find general information, case studies about successful applications (with an emphasis on the business dimension), links to experts, suppliers and tools as well as a help desk and a group of neutral, vendor-independent experts ready to answer any questions regarding the use of simulation. Details of the service offered to industrial enterprises are explained below.

### Stimulate and Co-ordinate Research and Development in the Area of Simulation

The second objective of Sim-Serv is to strengthen the development of simulation technology in Europe. This implies surveys of the state of the art, an analysis of needs and gaps, and the initiation of R&D activities addressing the identified gaps. The main mechanisms of Sim-Serv to meet this objective are the Working Groups which are introduced below.

### Support European Simulation Suppliers

Methods, techniques and tools developed in Europe (often funded by European tax payers' money) should be economically exploited in Europe. This means: a strong basis of simulation professionals is necessary: developers and vendors of commercial tools and solutions as well as professional service providers.

The above mentioned dissemination activities also support European simulation suppliers in that they help expand the market for their tools and services. Sim-Serv is particularly keen on spreading information about innovative techniques and tools, new application areas etc. Through its wide dissemination and marketing activities, Sim-Serv facilitates access to a European market even for small or micro providers. Sim-Serv also assists these small suppliers in developing suitable presentation material.

On the long run Sim-Serv aims at the development of widely accepted standards and quality criteria as well as a suite of tools, which comply the standards and criteria and are offered under a common brand.

The services and dissemination options available for suppliers are explained below.

## SIM-SERV'S ORGANISATION

The basic idea of Sim-Serv is to act as a neutral "mediator" facilitating a smooth matching and interaction of demand and supply. Wherever possible and reasonable, this interaction should use electronic media, however it is understood that face to face meetings cannot always be replaced by virtual interaction. Hence local presence is as essential as European wide recognition.

To get started, Sim-Serv was initially funded by the European Union. In order to continue the activities after the end of the start-up period in October 2004, the not for profit **Sim-Serv Association** was founded. This association is open for all organisations involved in developing, selling or using simulation technology in Europe.

The Sim-Serv Association runs a **Website** and a **Help Desk** which provides industry with general information and consulting regarding the benefits of simulation and possibilities to apply this technology.

The Sim-Serv Association and its central Help Desk are locally supported by a network of **Local Contact**

**Points.** Local Contact Points provide information and services in local languages and are available for face to face meetings whenever the need arises.

The **Sim-Serv Suppliers Group** consists of currently more than 60 members, the number is steadily growing. They represent a good mix of complementary skills and cover the majority of EU member states plus some Central European countries. There is a good balance of academic and commercial partners in the group.

The major role of suppliers is to provide input to the association's web site and to deliver customised solutions on commercial terms, whenever Sim-Serv received a request from an industrial user.

All suppliers are presented on Sim-Serv's web site. The presented material can be easily edited by the suppliers themselves.

The Suppliers Group is open to new members at any time. It is indeed one of the aims of Sim-Serv to expand this group substantially.

In addition, **Working Groups** (WGs) are being set up dealing with various technical or commercial issues of common interest. WG members may be suppliers or users of simulation tools and services, members or non-members of the Sim-Serv network. Currently, the following Working Groups are operating:

- Modular Design of Simulation Tools
- Open Digital Factory
- Simulation Accuracy for Plastics and Rubber Production
- Simulation Assisted Automation Testing
- Quantitative Benefits of Simulation
- Business and Enterprise Modelling
- Simulation of Traffic and Transportation Systems
- Human-Centred Modelling and Simulation
- Road Map of Simulation in Process Industries
- Road Map of Simulation in Manufacturing and Logistics

Sim-Serv also stimulates co-operation and joint developments of suppliers. Sim-Serv supports its members by bringing together partners with similar or complementary aims and skills, and helping them form consortia. Sim-Serv supports the development of research projects and the application for research funding.

Finally, Sim-Serv provides an overview of the state of the art in simulation, it identifies trends and unsolved

problems, and thus provides guidance for research and - last not least – research policy. Sim-Serv will play an active role also in future EU research programmes.

Working Groups generated elsewhere are also invited to use Sim-Serv's facilities and make themselves known via Sim-Serv.

## **SIM-SERV'S SERVICES TO INDUSTRY**

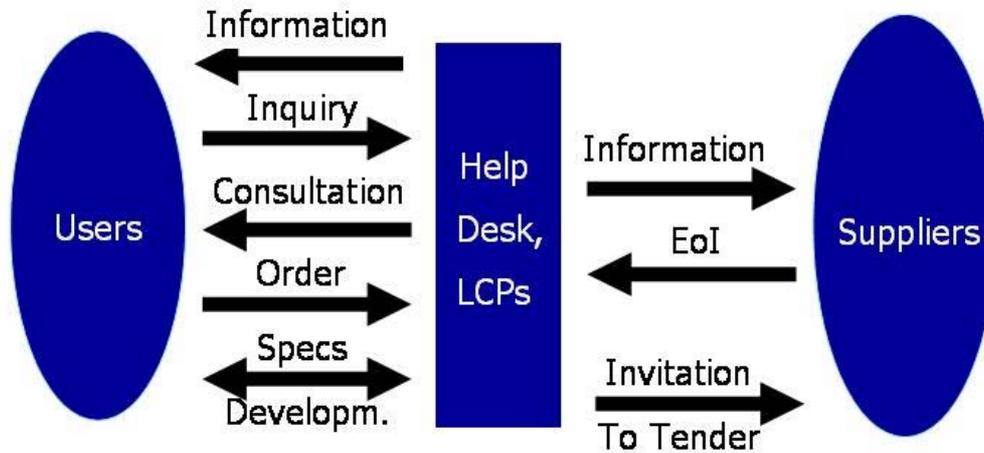
The central service of Sim-Serv is the web site. It contains general information about simulation, a database of technical and scientific information, news and information on relevant events, case studies and success stories, a list of suppliers and presentations of simulation tools.

A relatively new service is the self evaluation tool, which is freely available on our web site. A simple questionnaire of 10-12 questions help an industrial user define very roughly his situation and aims, and to check if simulation should be applied in this particular situation. 2-3 days after submitting the questionnaire, he will receive a report and a recommendation written by one of our experts.

Besides, Sim-Serv offers the following services to potential simulation users in industry:

- the help desk answers specific questions and offers a first and rough evaluation of problems
- independent technical consultation supports the user in analysing his problem and checks if simulation should be applied and how,
- a supplier-independent functional specification of the application /solution is developed on request by technical experts
- the best suited supplier(s) of the specified solution are found (see below for details)
- project management support, quality assurance and an evaluation of the solution after its implementation are offered

Figures 1 and 2 show how these services guide a novel user from first information and contact through to a co-operation with a member of the Sim-Serv Suppliers Group.



*Figure 1: The interaction of Sim-Serv, Users and Suppliers (part 1)*

## TEST CASES: SOLUTIONS DELIVERED TO INDUSTRY

In a number of test cases, the Sim-Serv approach proved feasible and beneficial to both customers and suppliers. Test cases are industrial applications of simulation where Sim-Serv assisted the customer analyse the problem, checked applicability of simulation and searched for suitable suppliers. Here are some examples of test cases:

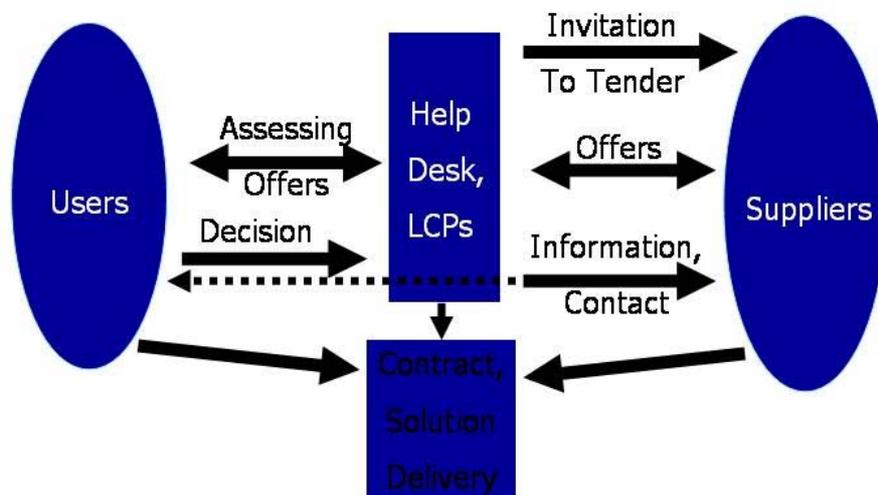
A medium sized UK based manufacturer of aluminium parts was looking for tools to support production scheduling. They had already made a pre-decision for a particular tool, but did not know how to make best use of it. Sim-Serv established a contact to Riga Technical University who developed a simulation model of the plant and used it for testing out several scheduling strategies and predicting the effects of various configurations of the tool. As a result, the company managed to reduce stocks of raw material by 50% /5,6/.

A Finnish manufacturer of rubber and plastic parts had to reduce product development time. With the support of Sim-Serv they tested a simulation tool and decided to introduce it. Production cost was reduced by 30% on average /7/.

A French manufacturer of laminates was offered a new technology for producing multi-ply laminates. They were looking for a proof of technical and economic feasibility of the proposed process. Sim-Serv brought them in touch with two simulation service suppliers: one of them developed a model of the entire process which allowed estimating production output and cost. The second made a very detailed model of the most critical part of the process in order to prove technical feasibility /8/.

An application in a German manufacturer who had to design a new assembly line is described in /9/.

/10/ describes a simulation study carried out by an Italian Sim-Serv member in a UK based manufacturing enterprise. This paper as well as /9/ is particularly remarkable in that both describe in detail the procedures applied by Sim-Serv to match demand and supply. Both cases gave rise to modifications of Sim-Serv's procedures, and both led to an ongoing co-operation of supplier and user. Both user companies were sceptical before the project, and now want to continue using the model for production planning and scheduling. This proves that Sim-Serv' approach is well suited to introduce simulation in enterprises which never used this technology before.



*Figure 2: The interaction of Sim-Serv, Users and Suppliers (part 2)*

In /11/, the introduction of a simulation based scheduling tool in a German SME is described. This case as well as /9/, /6/ and /7/ prove that simulation is applicable also in SMEs.

## SERVICES TO RESEARCHERS AND SIMULATION PROFESSIONALS

To researchers and simulation professional (suppliers), Sim-Serv offers essentially two advantages:

- the chance to present themselves to potential users in industry and thus to find additional partners / customers
- the chance to network and co-operate with other simulationists, to co-ordinate and join forces.

More specifically, the following is offered:

- space on our web site to present themselves, their expertise, their successful projects, and their tools
- guidance and support for the preparation of this material
- support for its translation to other European languages
- news and information about relevant events on the web site
- a data base containing up to date technical information

- working groups as a possibility to co-operate with other suppliers
- the chance to contribute to joint (funded) research and development projects
- professional, European- wide dissemination activities to attract potential users to our web site, and to acquire commercial or research projects
- offerings for additional commercial projects acquired by the Core Team

Sim-Serv helps its members find partners/customers, and it searches customers itself, e.g. by means of the self evaluation tool. Whenever a customer approaches Sim-Serv and asks for some simulation service, Sim-Serv offers the services described above: neutral advice, supplier independent development of a functional specification, search for suited supplier(s). The search for suppliers is done by the following “internal bidding procedure”:

A functional specification is developed by Sim-Serv in close co-operation with the customer. This specification and an Invitation to Tender are then circulated to the Suppliers Group or to a subset of members pre-selected jointly by Sim-Serv and the customer.

The members who are interested in the offered projects submit their offers, which contains details of the solution they offer, the price and the earliest possible delivery date. Based on criteria defined by the customer, Sim-Serv evaluates these offers and presents them to the customer who makes a final selection.

In reality, some iteration may be needed, e.g. suppliers may ask for more information before they submit an offer, or the customer asks for modifications of the offers. In general our experience shows that this procedure is considered effective and fair by both sides.

## SIM-SERV'S FUTURE

Sim-Serv as a funded project terminates end of October 2004. From then on, the Sim-Serv Association continues providing the services and operating the web site.

Sim-Serv closely co-operates with other virtual institutes, mainly with the virtual institute for advanced manufacturing technologies ADMAN ([www.max-serv.com](http://www.max-serv.com)). These two institutes seem to complement each other in a most natural way.

Sim-Serv is aware of the existence of numerous organisations active in the simulation field. The intention is by no means to compete with them. We rather intend to complement the more science-oriented organisations such as the national and international simulation societies or EuroSim, and to support the commercial organisations in order to promote our common goal:

*To improve the general knowledge about simulation and its benefits, particularly in industry, to stimulate and facilitate a wider take-up, and to create an environment for fruitful and exciting further developments of simulation technologies.*

All researchers, commercial suppliers and users of simulation technology are invited to join Sim-Serv, use our services and become members of the Association in order to contribute to the shaping of the future of simulation in Europe.

For more information, please contact:

Dr. Johannes Krauth

Sim-Serv Services and Quality Manager

Adolf-Reichwein-Str. 32, D-28329 Bremen

phone +49.421-437 3676

email: [Johannes.Krauth@sim-serv.com](mailto:Johannes.Krauth@sim-serv.com)

[www.sim-serv.com](http://www.sim-serv.com)

## BIOGRAPHY

Dr.-tech. Dipl.-Math. Johannes Krauth, initiator and Services and Quality Manager of Sim-Serv, has more than 25 years of experience in simulation software development (continuous and discrete event simulation) and in manufacturing applications. He worked in four research institutes in Germany, at the Hungarian Academy of Sciences in Budapest, and at the University of Patras in Greece. For several years, he ran a small consultancy business. His research

interests include conditions for simulation success and failure, and simulation quality criteria.

## REFERENCES

- /1/ Hollocks, B.W.: A Well Kept Secret? Simulation in Manufacturing Industry Reviewed. OR Insight, vol 5, no 4, October 1992, pp12-17.
- /2/ Kerckhoffs E.J.H., Vangheluwe, H.L. and Vansteenkiste, G.C., Report of ESPRIT Basic Working Group 8467 SiE: Simulation in Europe. Brussels 1994.
- /3/ Abdel-Malek L, Wolf C, Johnson F and Spencer III T (1999) OR Practice: "Survey Results and Reflections of Practising INFORMS Members", Journal of the Operational Research Society, 50, 10, 994-1003, October
- /4/ Wisniewski M, Jones C, Kristensen K, Madsen H and Ostergaard P.: Does Anyone Use the Techniques We Teach?, OR Insight, 7., April-June 1994, pp 2-7.
- /5/ Merkuryeva, G.: Decortpart: Manufacturing Planning and Capacity Optimisation.  
[www.sim-serv.com/success\\_stories.php](http://www.sim-serv.com/success_stories.php) (2003)
- /6/ Merkurieva, G., Shires, N., Morrisson, R., de Reuver, M.: Simulation Based Scheduling for Batch Anodising Processes. [www.sim-serv.com/pdf/whitepapers/whitepaper\\_30.pdf](http://www.sim-serv.com/pdf/whitepapers/whitepaper_30.pdf) (2004)
- /7/ Ture, T. (2003) Production cost of rubber and plastics parts reduced by 30%. [www.sim-serv.com/success\\_stories.php](http://www.sim-serv.com/success_stories.php)
- /8/ Mallinson, R.: Flexible Manufacture of Thermoplastic Composite Multiply Laminates. [www.sim-serv.com/success\\_stories.php](http://www.sim-serv.com/success_stories.php) (2004)
- /9/ Krauth, J., Krug, W., Pullwitt, S.: Layout Optimisation and Scheduling Support – A Sim-Serv Case Study. Paper accepted for the Industrial Simulation Conference. Malaga, June 7-9, 2004
- /10/ Aldini, L., Maccioni, R., Munster, K.: The FGWilson Test Case of Sim-Serv. Paper accepted for 45<sup>th</sup> Conference on simulation and Modelling SIMS 2004, Copenhagen, September 23- 24, 2004. See also [http://www.sim-serv.com/pdf/whitepapers/whitepaper\\_52.pdf](http://www.sim-serv.com/pdf/whitepapers/whitepaper_52.pdf).
- /11/ Druyen, J., Noche, B.: Vergangenheit und Zukunft – Der Mittelstand rüstet auf. Paper accepted for the conference: Experiences from the Future: Simulation in Production and Logistics. Berlin, October 4 and 5, 2004.