Jin-Woo Park
Tag-Gon Kim
Yun-Bae Kim (Eds.)

Communications in Computer and Information Science 5

AsiaSim 2007

Asia Simulation Conference 2007
Seoul, Korea, October 2007
Proceedings

Springer

Communications
in Computer and Information Science 5

Jin-Woo Park
Tag-Gon Kim
Yun-Bae Kim (Eds.)

AsiaSim 2007

Asia Simulation Conference 2007
Seoul, Korea, October 10-12, 2007
Proceedings

Springer

Volume Editors

Jin-Woo Park
Seoul National University
599 Gwanagno, Gwanak-gu, Seoul, 151-742, Korea
E-mail: autofact@snu.ac.kr

Tag-Gon Kim
KAIST
373-1 Guseong-dong, Yuseong-gu, Daejeon, 305-701, Korea
E-mail: tkim@ee.kaist.ac.kr

Yun-Bae Kim
SungKyunKwan University
300 Chunchun-dong, Jangan-gu, Suwon, Kyunggi-do, 440-746, Korea
E-mail: kimyb@skku.edu

Library of Congress Control Number: 200794566

CR Subject Classification (1998): I.6

ISSN 1865-0929
ISBN-10 3-540-77599-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-77599-7 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12213696 06/3180 5 4 3 2 1 0

Preface

On behalf of the Conference Organization Committee, I would like to welcome you all to the 2007 Asia Simulation Conference, or AsiaSim 2007.

The Asia Simulation Conference is an annual conference organized by the Chinese Association for System Simulation, the Japan Society for Simulation Technology, and the Korea Society for Simulation. Since the first gathering of the Asian simulation community in Seoul in the year 2001, the Asia Simulation Conference has provided a forum for scientists, academics, and professionals from all around Asia to present their latest and exciting research findings in various fields of modeling, simulation, and their applications.

This year, we received about 120 papers and only 42 papers were selected for the final program after a very stringent refereeing process. As such, this significantly strict selection percentage produced the very high-quality papers in the fields of modeling and simulation methodology, e-business, supply chain management, general and engineering applications, manufacturing, medical, military, networks, transportation, traffic and virtual reality.

The success of AsiaSim 2007 was mainly the result of many individual contributions, and I would like to thank all authors, reviewers, Program Committee members, session Chairs, and other participants. Especially, my special thanks go to the Program Chair, Yun Bae Kim, and all members of the Organizing Committee for their excellent effort and sacrifice in a very short and concentrated period of time. Thanks are also due to our sponsors for their financial, technical, and publication support.

The Hoam Faculty House of Seoul National University provided excellent meeting facilities and a comfortable setting for the exchange of ideas with fellow researchers. I invite all of you to enjoy the conference proceedings.

November 2007 Jin Woo Park

Organization

Committee Chairs

Honorary Chair	Chan-Mo Park Pohang University of Science and Technology, Korea Seong-Joo Park Korea Advanced Institute of Science and Technology, Korea Doo-Kwon Baik Korea University, Korea
General Chair	Jin-Woo Park Seoul National University, Korea
General Co-chairs	Faul Fishwick University of Florida, USA Cellier Francois University of Arizona, USA Toshiharu Kagawa Tokyo Institute of Technology, Japan Myeong Hee Kim Ewha Womans University, Korea Cellier Francois Korea Advanced Institute of Science and Technology, Korea Axel Lehmann Universität der Bundeswehr Muenchen, Germany Bohu Li Beijing University of Aeronautics and Astronautics, China Bernard P. Zeigler University of Arizona, USA
Local Organizing Committee Chair	Sung-Do Chi Hankuk Aviation University, Korea
International Program Committee Chair	Yun Bae Kim Sungkyunkwan University, Korea
Publication Committee Chair	Tae-Ho Cho Sungkyunkwan University, Korea

Local Organizing Committee

Woo-Jin Chang: Seoul National University, Korea
Jong-Sik Lee: Inha University, Korea
Kang-Sun Lee: MyongJi University, Korea
Soo-Hyun Park: KookMin University, Korea
Dong-Won Seo: KyungHee University, Korea

International Program Committee

Jung Gil Choi: Korea Institute of Industrial Technology, Korea
Norman Foo: University of New South Wales, Australia
Sumit Ghosh: The University of Texas at Tyler, USA
David Goldsman: Georgia Institute of Technology, USA
Young Shin Han: University of Sungkyul, Korea
Alemdar Hasanoglu: University of Kocaeli, Turkey
Fernando J. Barros: University of Coimbra, Portugal
V. Kachitvichyanukul: Asian Institute of Technology, Thailand
Yukio Kagawa: Nihon University, Japan
Kee Bum Kang: Naval Postgraduate School, USA
Ki-hyung Kim: Ajou University, Korea
Haklin Kimm: ESU of Pennsylvania, USA
Jong-Sik Lee: Inha University, Korea
Hessam S. Sarjoughian: University of Arizona, USA
Ryo Sato: University of Tsukuba, Japan
Mi Ae Shon: Sungkyunkwan University, Korea
Shingo Takahashi: Waseda University, Japan
Shinsuke Tamura: Fukui University, Japan
Xiao Tianyuan: Tsinghua University, China
Peng Xiaoyuan: Beijing University, China
Guang Leng Xiong: Tsinghua University, China
Mohd Nazmi Yosoff: International Islamic University, Malaysia

Publication Committee

Sung-Do Chi: Hankuk Aviation University, Korea
Woo-Jin Chang: Seoul National University, Korea
Hyung Jong Kim: Seoul Women's University, Korea
Yun Bae Kim: Sungkyunkwan University, Korea
Jong-Sik Lee: Inha University, Korea
Kang-Sun Lee: MyongJi University, Korea
Soo-Hyun Park: KookMin University, Korea
Dong-Won Seo: KyungHee University. Korea
Hee Seok Seo: Korea University of Technology and Education, Korea
Mi Ra Yi: Mokpo National Maritime University, Korea

Table of Contents

Agent Based Simulation

Aero Space

System Dynamics

Numerical Simulation - II

Web Based Simulation

General Engineering

Logistics Simulation

Simulation and AI

Simulation Tools

Statistical Methods

Health Care/Education

Methodology

Event Graph Models for Generic Manufacturing Systems with Push and Pull Policies

Truong Hong Trinh[1] and Voratas Kachitvichyanukul[2]

[1] College of Economics, University of Danang
Danang, Vietnam
[2] Industrial Engineering & Management
Asian Institute of Technology
P.O. Box 4, Klong Luang, Pathumthani 12120, Thailand
hongtrinh@vnn.vn, voratas@ait.ac.th

Abstract. The study presents two simulation models for generic manufacturing systems operated under MRP-push policy and Kanban-pull policy. The models are designed as long run planning tools for improving manufacturing performance. A simulation language for discrete-event simulation based on event graph is used to construct the models for pull, push systems [5]. The models in this paper are totally data driven. Input data includes machines, products, processes, bill of materials, and demand data. An iterative heuristic algorithm is employed to determine the initial model parameters: the number of Kanban for pull systems, and safety stock levels for push systems. Simulation experiments are carried out to demonstrate the effectiveness of the heuristic and the simulation models. The experiments are conducted in an environment involving changes in two operational factors: demand and processing time. The experimental results indicate that for the same throughput, the pull system outperforms the push system in terms of lead time and work in process (WIP) inventory in such environment.

Keywords: Kanban, Pull system, Push system, performance, Simulation, SIGMA.

1 Introduction

Manufacturing industry has faced increasing competition in the global market where competitive strategy requires the company to maintain high productivity and flexibility to adapt to market requirements. Many researchers have developed simulation models for investigating operational factors on performance of pull, push systems (see [1], [2], [3], [4], and [7]). Most of these works are focused on a specific situation and modifications are often required to adapt the model to fit the new situations.

The paper presents a generic simulation tool for the evaluation of the performance of the pull, push systems for manufacturing systems operated in a job shop environment with bill of material consideration. The models are developed to be totally data-driven and the user can customize the model by simply changing the input data.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 1–11, 2007.

2 Methodology

In simulation it is impossible to plug a set of possible decision variables into a single closed form formula. In addition, the objective function can not be evaluated exactly due to randomness in the output. One way to deal with this problem is to replicate the simulation n times at a set of input factors of interest, and use the average value of output across these replications as an estimate of the objective function at that point.

The following heuristic procedure is used in the simulation to find out a set of input factors that can satisfy the production requirements. .

(1) Estimate the upper bound for each value of input factor. This upper bound can be estimated in deterministic environment so that desired total output is achieved.

(2) From BOM structure, items (jobs) are classified in levels. Choosing a job as a candidate using a top-down rule, job with higher level will be given higher priority over job with lower level.

(3) For each input value, experiment with n replications is made to estimate average output value. The hypothesis test is applied to determine whether the average output reaches the desired output value or not. The smallest input value that yields the desired output is kept for later use.

(4) Repeat for another job with the same procedure until all jobs have been considered.

The pull system controls part flows in the shop. According to JIT philosophy, the pull system should synchronize demand with production rate. WIP level is also number of Kanbans that is considered as input factors for the system. The push system controls part flow by controlling the released orders. Order release rate affects the WIP level and the time jobs spent in the system.

Following the principle of Factory Physics, the WIP level is a convenient control for the system and within the normal operating conditions, the throughput is increasing with work in process hence the Little's formula can be used to guide the initial search of parameters. Consequently, a point can be found that the model reaches the desired total output at the favorable values of input factors. The heuristics may not be optimal, but it provides an algorithmic strategy to reduce system WIP that terminates in finitely many steps.

3 Development of Simulation Models

3.1 The Conceptual Model

The models developed in this research are generic, i.e., they can be used for varieties of products and processes. The input data are divided into two parts: structural and operational. The structural input data include machines, products, job routing, usage quantities (BOM), and processing time. The operational input data include monthly demand (firmed, forecasted), initial stock (Kanbans, safety stock), and order lot size. The input data is read from disk files at the beginning of the experiments. The structural data is used to configure the manufacturing system where the operational data is used to set the initial values for operation conditions. Here the initial stocks are

synonymous to the number of Kanbans in the pull system and safety stock in the push system.

The output statistics generated from the model are total throughput and throughput by product type, average time in system, average work in process.

3.2 JIT Simulation Model

In a pull system, production at each station is triggered in response to an actual demand arrival at the end of the lines that associated with final products. Customer demand and order arrival schedules to be released from MPS. The information about order releases at the end of line is transmitted to the rest of the line via part specific Kanban. Production activities of the final assembly line are connected in a chain manner to the preceding processes or subcontractors of the entire processes.

The simulation model is divided into five sections: 1) Data Input, 2) Machine Capacity Planning, 3) Order Releasing, 4) Pull Control System, and 5) Simulation Experiments. The JIT simulation event graphs model is presented as shown in Figure 1. In the simulation model, the resident entities are machines, stores, and buffers associated with jobs and machine groups. The transient entities are jobs, Kanbans flowing through the system. By ignoring the details and taking a high-level view of the event graph. The effort can be focused on Kanban system in order to explore the most distinctive feature of Kanban system. The pull control system includes three basic activity cycles relating to Kanban information and material flows. A rectangular cycle, a triangular cycle, and a compound activity cycle are considered as follows:

The rectangular cycle (STORE, PKPOS, BUFFER, WKPOS and back to STORE) is designed to control the Kanban flows (Production Kanban and Withdrawal Kanban). The process will start when the jobs are released at the STORE vertex where the right quantities and type of Kanban are exchanged. The pull control mechanism schedules its predecessor through PKPOS, BUFFER, and WKPOS vertexes. The cycle stops when the tasks on the job routing are completed.

The first triangular cycle (BUFFER, WKPOS, STORE and back to BUFFER) is designed to control material flow between workstations. This activity cycle plays the roles for both information and material flow. While the compound activity cycle (JOINQ, and the cycle of BUFFER, START, FINISH, NEXTQ and back to BUFFER) is designed for only material flow within particular workstation, this cycle shows busy or idle status for each machine group. If no machines are idle, then the job joins in queue. Also, first job in the queue is released every time a job of the same type is finished. Machines continuously assemble components to produce parts until available components in BUFFER are depleted.

The rectangular cycle is viewed as the activity of transient entity that moves between resident entities (from machine to machine) while the triangular cycles are the activities of a resident entity (machine group) processing successive transient entities (Jobs).

3.3 MRP Simulation Model

In MRP system, the key inputs include master production schedule (MPS), bill of material (BOM), and inventory records. These data are constructed in data files, and

are often updated during operations. Using this information, the MRP system identifies actions that operations must take to stay on schedule, such as releasing planned orders, and order quantities.

MRP explosion is used to convert the requirements of various final products into specific requirements of the subassemblies, components and raw materials needed by the final products. Moreover, in order to control part flows in the system, feeder lines are designed to identify component types and quantities on each operation of final products.

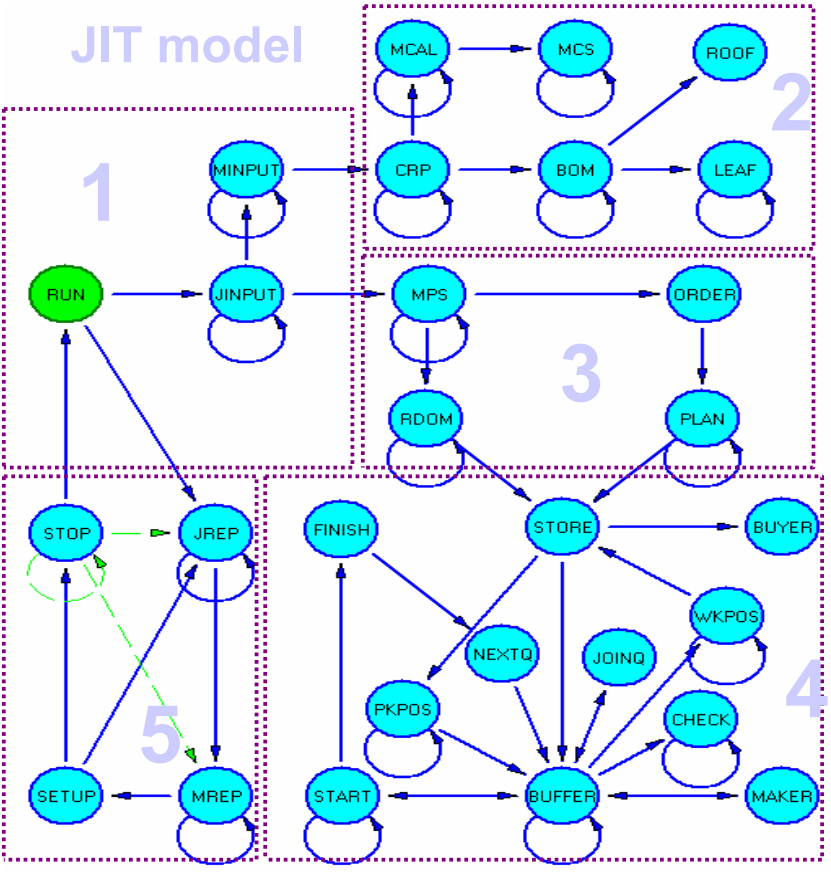


Fig. 1. JIT simulation model in Sigma

When an order is released, MRP explosion updates information about order (job type, lot size), and computes material requirements corresponding to an order lot. Raw material authorized to be moved to the buffer immediately, the process starts as soon as any machines are available. Completed parts are moved to buffer for the next operation. The inventory activities that occur between workstations also happened in the same manner.

MRP system also determines the timing and size of order quantities to suppliers. A lot-sizing rule must be assigned to raw material types in advance. The choice of lot sizing rule proposed to the system is periodic fixed order quantities to smooth the production.

The MRP simulation model includes 1) Data Input, 2) Machine Capacity Planning, 3) Lot Sizing, 4) Order Releasing, 5) the Push Control System, and 6) Simulation Experiment. An event graph model is shown in the Figure 2.

The model can be briefly described by taking a high-level view of the event graph. The effort can be focused on MRP push control system. The control mechanism of jobs is based on model of network of queue that have two basic activity cycles, one is used for controlling movements between resident entities (machines, stores, buffers), and the other is for controlling transient entities (jobs, Kanbans, raw material).

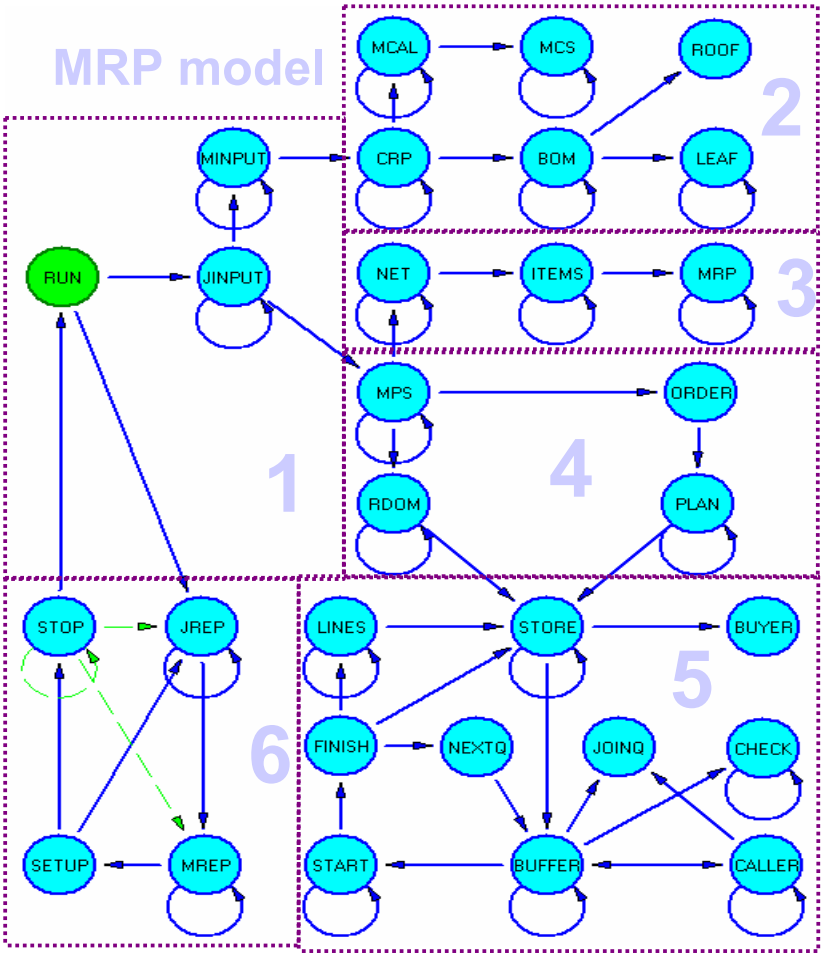


Fig. 2. MRP simulation model in Sigma

The activities on the FINISH, LINES, BUFFER nodes is designed to control part flows between workstations, while the activities on JOINQ, BUFFER, START, FINISH, NEXTQ nodes present the status of machines (idle or busy), and changes between machine groups.

4 Experiment Design

4.1 Experimental Factors

Order release mechanism is based on planned demand that comes from firmed orders and forecasted orders. For the firmed orders, the release is planned as a deterministic process, and rate of order release is uniform to smooth production. For the forecasted orders, the timing and quantity of orders is a stochastic process. Based on historical data, forecasted proportion can be estimated as a ratio of planned demand. In the study, forecasted order release is assumed with a predetermined rate, but probability of job release is estimated on scale of each forecasted job orders.

Processing time is assumed to be normally random variable with known mean, and standard deviation of 10% mean.

4.2 Cases in Simulation Experiments

Simulation experiments are performed for three cases:

Case 1: “No variability”. Simulation models are evaluated under deterministic environment where input and output processes are deterministic over time.

Case 2: “Processing time variability”. Simulation models are evaluated under processing time variability. Processing times are assumed to be normally random variable with known mean, and standard deviation of 10% of the mean.

Case 3: “Processing time variability and Uncertainties in demand”. Simulation models are evaluated under processing time variability and uncertainties in demand where the standard deviation of demand is proposed as a fraction of monthly demand from MPS.

Each experiment includes 10 replications (runs) of length 1 day long. The independence of replications is accomplished by using different random seeds.

5 Analysis and Results

5.1 Determining Values of Input Factors

The initial number of Kanban in JIT model, and initial safety stock level in MRP model are determined by applying the heuristic procedure described earlier. A Kanban set for each job includes 10 combinations, in which each combination requires 1 run for deterministic process (case1), and 10 runs for stochastic processes (case 2 and 3). Upper bounds are predetermined for each case. For instance, upper bounds of case 1 are (10,10,30), and that of case 2 and 3 is (15,10,30). In the MRP model, safety stocks are required for final products (job A, B, and C) and raw materials (job D, E, and F). For raw material, safety stock can remain at a high level

over some periods. In the simulation model, this level can be estimated from maximum value of shortage over time, and assumed to be 10 for all raw materials. Set of Safety Stock is proposed for simulation experiment where upper bounds are predetermined. More specific, upper bounds (20,15,10) are assigned for all cases.

Since the length of each run is 1 day long, the desired total output is 420 per day. Average total output is estimated from the experiments. A test with 5% significant level is performed to check whether the average at certain combination achieves the desired total output or not, a test with 5% significant level is performed. The table 1 below shows results obtained by using the heuristic procedure for the number of Kanbans, and safety stock estimated.

Table 1. Initial Kanban and Safety Stock

JIT model	Number of Kanbans (A, B, C)			Output of jobs (A, B, C)			Total output
Case 1	5	7	23	240.0	120.0	60.0	420.0
Case 2	10	8	24	240.0	119.9	59.8	419.7
Case 3	11	10	28	239.7	119.9	59.8	419.4
MRP model	Safety stocks (A, B, C)			Output of jobs (A, B, C)			Total output
Case 1	15	9	1	240.0	120.0	60.0	420.0
Case 2	16	10	2	239.9	119.7	60.0	419.6
Case 3	17	13	4	237.6	121.2	60.0	418.8

5.2 Analyzing Steady State Throughput

Welch's procedure [6] is employed to estimate a warm-up period. An experiment for case 3 is carried out by making 10 replications over 24 hours, and the throughput is observed every hour. The mean throughput is averaged value of 10 replications. The following graphs show mean throughput over 24 hour runs of JIT, MRP models.

The graphs of steady state throughput are given in Figure 3. A warm-up period of 4 hours is estimated for JIT, MRP models. Steady state mean hourly throughputs that estimate from average throughput of the last 20 hours are 17.095 for two models.

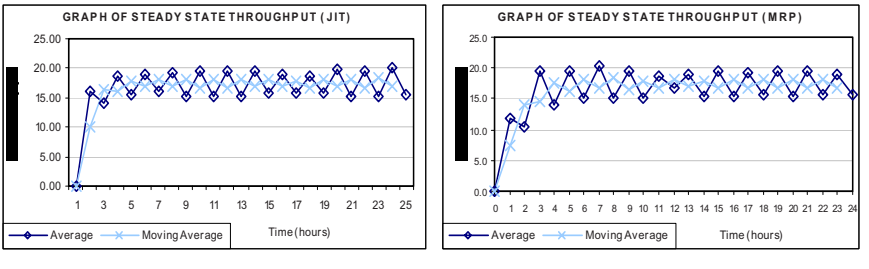


Fig. 3. Steady State Throughput for Systems under JIT and MRP

5.3 Evaluating Performance Measures

To make a fair comparison, models are conducted in the same environment, and with the same values of input factors. For illustrative purpose, Case 3 with processing time variability and demand uncertainty, is chosen for comparison of performance measures between models.

The visual performance measures of an experiment with random seed of 12345 are plotted on the same graphs over 1 day. Figures 4a, 4b, 4c, and 4d show side by side the performance indicators for the JIT and the MRP cases. The four performance indicators are: the total production output of the systems.

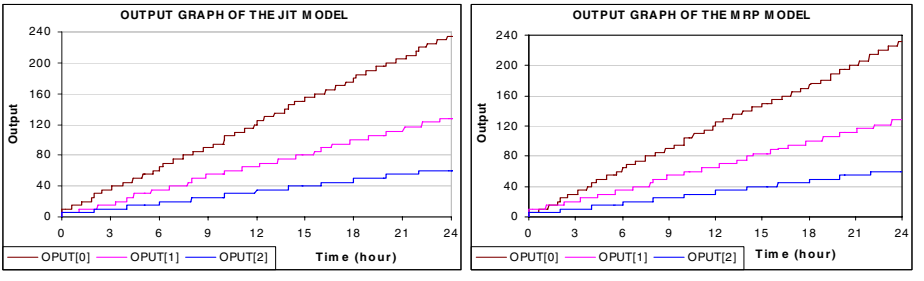


Fig. 4a. Graphs of Total output for JIT and MRP models

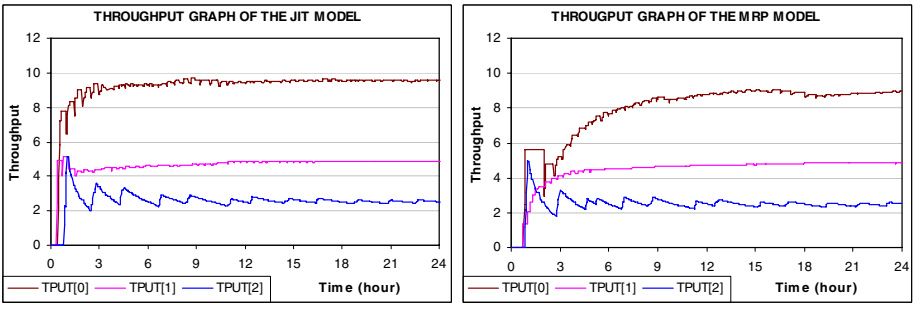


Fig. 4b. Graphs of throughput for JIT and MRP models

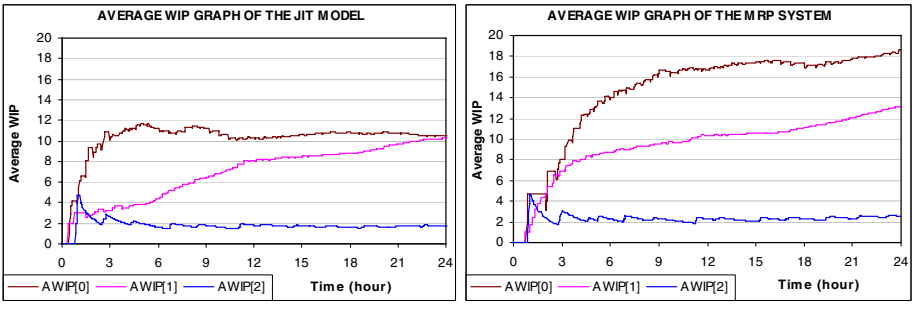


Fig. 4c. Graphs of average WIP for JIT and MRP models

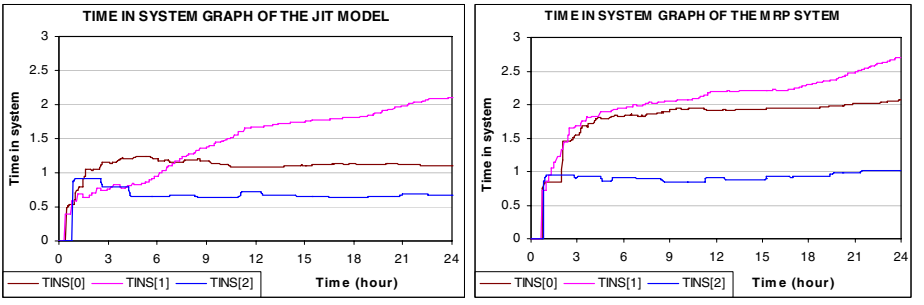


Fig. 4d. Graphs of time in system for JIT and MRP models

Experiment with 10 replications of length 1 day long, and a warm-up period of 4 hours is employed to estimate throughput, time in system, and work in process. Experiment results are summarized in the tables 2 and 3 below.

Table 2. Simulation Outputs for JIT Case

JIT MODEL	Average output	Throughput rate	Time-in-system	Average WIP
Job 0 (A)	240.0	9.839	1.285	12.689
Job 1 (B)	124.0	4.903	1.871	9.237
Job 2 (C)	60.0	2.503	0.689	1.727

Table 3. Simulation Outputs for MRP Case

MRP MODEL	Average output	Throughput rate	Time-in-system	Average WIP
Job 0 (A)	240.0	9.738	2.107	20.536
Job 1 (B)	121.9	4.942	2.380	11.799
Job 2 (C)	60.0	2.507	0.841	2.110

The hypothesis test assumes 5% significant level, and values of t-test are determined from level of significance and degree of freedom. The mean and standard deviation of performance measures are summarized in tables 4-5 below.

Table 4. Throughput Statistics for JIT Case

Throughput	Mean throughput		Standard deviation (STD)		Test statistic	Evaluation Results
	(JIT)	(MRP)	(JIT)	(MRP)		
Job 0 (A)	9.839	9.738	0.252	0.260	0.884	No difference
Job 1 (B)	4.903	4.942	0.155	0.122	-0.635	No difference
Job 2 (C)	2.503	2.507	0.003	0.004	-2.008	No difference

Table 5. Throughput Statistics for MRP Case

Time in system	Mean time in system		Standard deviation (STD)		Test statistic	Evaluation Results
	(JIT)	(MRP)	(JIT)	(MRP)		
Job 0 (A)	1.285	2.107	0.265	0.153	-8.493	Difference
Job 1 (B)	1.871	2.380	0.493	0.426	-2.473	Difference
Job 2 (C)	0.689	0.841	0.030	0.128	-3.650	Difference

By comparing values of t-test with test statistics, evaluations can be concluded from the comparisons. There are no significant differences on mean throughput between two models. However, average time in system of MRP model is greater than that of the JIT model.

6 Summary and Conclusion

The success of the manufacturing system depends on both the design and operational factors associated with the system. The design factors that were studied include the number of machines for each group, values of input factor. While operational factors like processing time and uncertainties in demand were investigated in the study.

In applying simulation to manufacturing system, the times required to initially build the model and late modify the model are quite significant. A generic model where changes can be done via changing of data can provide significant time saving.

The differences between the pull and the push system can be utilized as an advantage to build a manufacturing system that encompasses the positive attributes of the different mechanisms. Integration is to compensate for the weakness on both sides. The recommend is that using MRP for planning and JIT for the execution combines two powerful tools into an efficient manufacturing system.

Some general findings from the study are summarized as follows:

The analysis also indicated that there are interactions between the number of Kanbans and average WIP level in the pull system. Once the number of Kanban increases, the average WIP also increases. The push system has a similar relationship between safety stock and lead-time.

In the same experiment environment, the average time in system of the push system is higher than that of the pull system. This also causes the higher WIP level in the push system. Another result is given that average machine utilization of the pull system is higher than that of the push. This concludes that the pull system is better at controlling activities on shop floor than the push system.

In highly stochastic environment, the system requires larger values of input factors to meet desired total output. It was observed that these values were assigned for parts differently in control systems. In the pull system, part with lower level in BOM structure that serves for many higher levels requires larger number of Kanbans. Meanwhile, in the push system, part with higher level that is processed through many stages (longer lead-time) requires larger safety stock.

References

1. Andijani, A.: Trade-off between maximizing throughput rate and minimizing system time in Kanban systems. International Journal of Operations and Production Management 17, 429–445 (1997)
2. Beamon, B.M., Bermudo, J.M.: A hybrid push/pull control algorithm for multi-stage, multi-line production systems. Production planning and control 11, 349–356 (2000)
3. Rees, L.P., Huang, P.Y., Taylor III, B.W.: A comparative of an MRP lot for lot system and a Kanban system for a multistage production operation. International Journal of Production Research 27, 1427–1443 (1989)
4. Sarker, B.R., Fitzsimmons, J.A.: The performance of push and pull systems: a simulation and comparative study. International Journal of Production Research 27, 1715–1731 (1989)
5. Schruben, L.W.: Graphical simulation modeling and analysis. The Sciencific Press, Massachusetts (1995)
6. Welch, P.D.: On the problem of the initial transient in steady-state simulation. IBM Watson research center, New York (1981)
7. Yan, H.: The optimal number of Kanbans in a manufacturing system with general machine breakdowns and stochastic demands. International Journal of Operations and Production Management 15, 89–103 (1995)

A Model and Analysis of the Bullwhip Effect Using a SCOR-Based Framework

Haejoong Kim[1], Jongkyung Park[1], Tai-Woo Chang[2], Hanil Jeong[3], Ki Tae Kim[4], and Jinwoo Park[1]

[1] Department of Industrial Engineering, Seoul National University, Seoul, 151-744, Korea
haejoong.kim@gmail.com, pjkyoung@empal.com, autofact@snu.ac.kr
[2] Department of Industrial Engineering, Kyonggi University, Suwon, 443-760, Korea
keenbee@kgu.ac.kr
[3] Department of IT Business Engineering, Daejeon University, Daejeon, 300-716, Korea
hijeong@dju.ac.kr
[4] Department of International Trade & Logistics, Youngsan University, Yangsan, 626-790, Korea
jeha@ysu.ac.kr

Abstract. Increasing dynamics are making efficient management of the supply chain difficult. The bullwhip effect is known to be a major source of this inefficiency so there is a considerable amount of literature analyzing its causes. However, the reported outcomes are not comparable because they consider different business environments and make different assumptions regarding operative processes. We propose a SCOR (Supply-Chain Operations Reference-model)-based framework for modeling and analyzing the bullwhip effect so that practitioners can evaluate the performance of their supply chain, and establish a strategy to improve its performance.

Keywords: supply chain, bullwhip effect, SCOR, simulation.

1 Introduction

Good supply chains must respond rapidly, effectively, and efficiently to ever-increasing competition in the global market. However, it is not easy to achieve good performance. One of the main causes of this difficulty is the "bullwhip effect".

The bullwhip effect is defined as the amplification of demand variation in a supply chain from downstream to upstream, and causes inventory excesses or shortages that lead to increased cost to the company and the entire chain.

Most research has focused on the following areas: 1) providing empirical evidence to support the existence of the bullwhip effect, 2) analytically demonstrating the existence of the bullwhip effect, 3) identifying the possible causes of the bullwhip effect, and 4) developing strategies to reduce the impact of the bullwhip effect.

Since Forrester described the supply chain dynamics [1], research has focused on identifying causes of the bullwhip effect and providing measures for reducing its impact [2-7]. Through statistical modeling with time series or simulation models, the

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 12–20, 2007.

causes of the bullwhip effect were verified. For example, the well-known 'Beer Game' is widely used and to verify the bullwhip effect and analyze its causes.

However, these results are still based on a simple supply chain structure, and the studies did not investigate the interaction among the causes by fully considering one or two factors individually. Moreover, the studies consider different business environments, and make different assumptions regarding operative processes and the supply chain cost structure. Therefore, different model parameters are selected, so the outcomes are hardly comparable [8]. While these studies do help practitioners understand the basic phenomenon, they do not provide good guidelines to assist practicing managers in minimizing the impact of demand uncertainty on the performance of their supply chains [9].

In order to evolve an efficient and effective supply chain, it is necessary to construct a comprehensive supply chain model by which performance can be assessed. In this paper, we propose a framework, based on SCOR model, for modeling and analyzing the bullwhip effect. We then try to verify our framework with a simple case.

The organization of this paper is as follows: Section 2 reviews the previous research on the bullwhip effect, and supply chain metrics and measurement. In section 3, we propose a framework for modeling and analyzing the bullwhip effect through the SCOR-based supply chain model. In Section 4, the simulation and analysis based on the proposed SCOR-based supply chain model is described. Finally, conclusions are presented in Section 5.

2 Literature Review

Previous research on supply chain performance has occurred in the following areas: 1) identifying and analyzing the possible causes of the bullwhip effect, 2) developing metrics to assess supply chain performance.

2.1 The Causes of the Bullwhip Effect

The extensive literature analyzing the causes of the effect [1-11] is summarized in Fig. 1. Most research has been conducted using an analytical model (time series), or a simulation model (Beer Game). Both model-types led to similar conclusions about causes of the bullwhip effect. For example, larger order lead times lead to larger variances of orders

Although many empirical findings verify the existence of the bullwhip effect, more and more data from economists show that in many industries the variance of production is smaller than that of sales. Furthermore, the prevalence of production smoothing has increased over time [12, 13]. Li *et al*. demonstrated the existence of the 'anti-bullwhip effect', which states that the outgoing orders in a supply chain have a smaller variance than the incoming orders. They also demonstrated 'lead time paradox', meaning that information transformation at higher stages decreases as lead time at a lower stage is increased [14]. Baganha and Cohen constructed an analytical model to demonstrate that inventories can sometimes have a stabilizing effect on supply chains, and that the bullwhip effect is not always present throughout the supply chain [15].

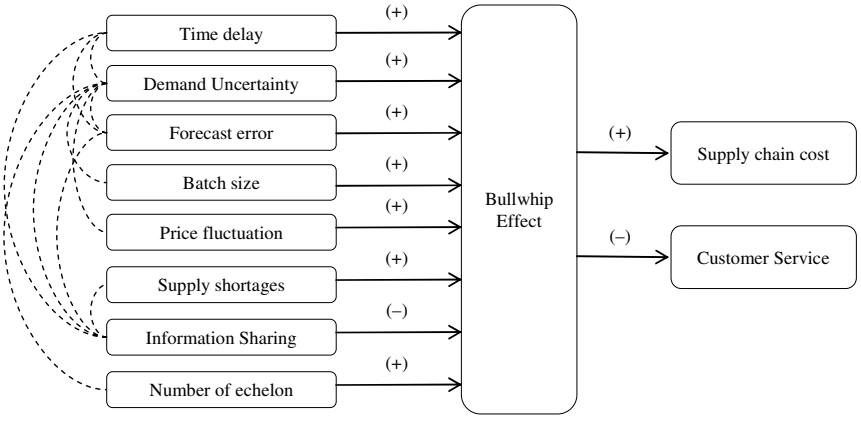


Fig. 1. The causes of bullwhip effect and the results

2.2 Metrics and Measur.ement in Supply Chain Performance

While many firms realize the potential of supply chain management, they often lack the insight to develop of effective performance measurement and metrics needed to achieve a fully integrated supply chain [16]. In order to enhance a supply chain, researchers and practitioners have paid much attention to organizational performance measurement and metrics [17, 18].

Sanchez et al. explored the relationship between the dimensions of supply chain flexibility and firm performance in a sample of automotive suppliers. This research shows the positive relationship between superior flexibility and firm performance [19]. Milgate developed a three-dimensional view of supply chain complexity: uncertainty, technological intricacy, and organizational systems. Of these three, only uncertainty was found to be significantly related to both delivery speed and reliability [20].

While many analytical and numerical models have been proposed to handle operational and design issues, formal models for strategic planning are scarce. To establish these formal models, the Supply Chain Council (SCC) developed the SCOR model a strategic planning tool used to simplify supply chain management, including the standard metrics used to measure process performance [21].

3 SCOR-Based Framework for Supply Chain Simulation

As mentioned above, researchers who try to improve supply chain performance consider the various environments in which the supply chain works, and the assumptions regarding the operative processes and supply chain cost structure. Thus, a framework for modeling and analyzing the bullwhip effect is needed to allow practitioners to evaluate the performance of the supply chain objectively, and to provide a strategy for improving its performance. Fig. 2 shows our proposed framework.

By summing up the previous research based on the standard processes and metrics of the SCOR model, we present the SCOR-based supply chain model including the structures/processes, the environmental/operational parameters, and the performance measures. Specific parameters are set to reflect a specific case. Environmental

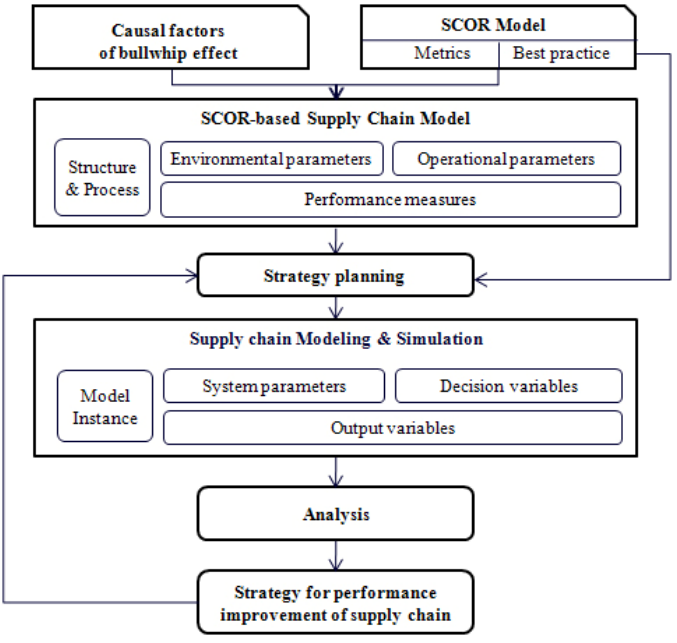


Fig. 2. SCOR-based framework for supply chain performance improvement

parameters are related to the external processes, and operational parameters are related to the internal processes. Environmental and operational parameters are converted into system parameters which cannot be changed by cooperative strategy, and input (decision) variables which can be controlled by the individual member's decision. Results of the simulation are analyzed, and the strategy for improving the performance of the supply chain is established by a planning process which refers to SCOR best practices. Details are presented in the following sub-sections.

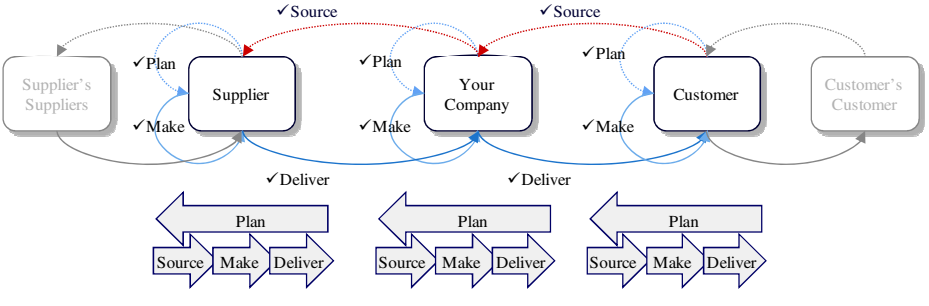


Fig. 3. Basic structure & process of supply chain

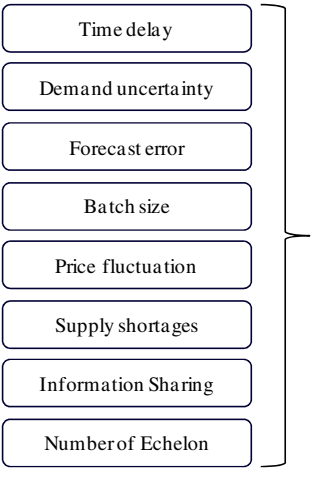
3.1 Structures and Processes of the Supply Chain

The structures and processes of supply chains are described using SCOR process building blocks as shown in Fig. 3. While the SCOR model contains the five primary

management processes of PLAN, SOURCE, MAKE, DELIVER and RETURN, we exclude the RETURN process because reverse logistics are not considered in our model. By linking the external suppliers or customers, and specifying their processes, it is possible to describe the complex supply chain easily.

3.2 Environmental and Operational Parameters and Performance Measures

From previous research on the bullwhip effect, we sum up the parameters related to its causes in Fig. 4.



Cost	•Holding cost
	•Backorder cost
	•Ordering cost
	•Transportation cost
Lead time	•Information lead time
	•Production lead time
	•Transportation lead time
Demand patterns	•Constant/Step function
	•AR(p) /ARMA(p,q) / ARIMA(q, d, q)
	•Seasonality/ Trend
Forecasting methods	•Moving average
	•Exponential smoothing
	•Regression analysis
Inventory control policy	•Base stock model
(Replenishment policy)	•(Q, r) Model
	•(s, S) Model
Batch size	•Fixed size
(Order batching)	•Variable size
Information sharing	•VMI
	•POS
Production capacity	•Unlimited
	•Limited
	•Defect ratio
Price fluctuation	•Constant
	•Variable
Supply chain structure	• Serial 1:1 (2echelon or more)
	• Divergent 1:N (2 echelon)
	• Converse N:1 (2 echelon)

Fig. 4. Causal factors for bullwhip effect and its parameters

These parameters are classified into environmental and operational parameters reflecting the SCOR metrics in Table 1. In the SCOR model, there are five performance attributes: Reliability, Responsiveness, Flexibility, Costs, and Assets. While assets are related to the cash cycle time, fixed assets, and return on working capital, we do not consider them, and thus exclude this attribute.

Metrics in the SCOR model do not necessarily relate to a SCOR Level 1 process, we try to link the performance attributes and process. This will help to analyze and evaluate them against other supply chains with competing strategies. Just as you would describe a physical object like a piece of lumber using standard characteristics (e.g., height, width, and depth), the description of a supply chain requires standard

Table 1. Environmental & operational parameters and performance measures

			Parameters of bullwhip effect	Process
Environmental Parameters			Information sharing	Plan
			Demand pattern	Plan
			Supply chain structure	Plan
Operational Parameters	Reliability	Forecast accuracy	Information lead time	Plan
			Forecasting methods	Plan
		% of Orders produced in Full	On-time production	Make
		% of Orders Delivered in Full	On-time delivery	Deliver
	Respon-siveness	Plan cycle time	Planning lead time	Plan
		Source cycle time	Order lead time	Source
		Make cycle time	Production lead time	Make
		Deliver cycle time	Transportation lead time	Deliver
	Flexibility	Upside Source Flexibility		Source
		Upside Source Adaptability		Source
		Downside Source Adaptability	Safety stock	Source
		Upside Make Flexibility	Overtime	Make
		Upside Make Adaptability	Time needed to increase additional capacity	Make
		Downside Make Adaptability	Time needed to decrease supplemental capac-ity	Make
		Upside Deliver Flexibility	Time needed to increase additional delivery capacity	Deliver
		Upside Deliver Adaptability	Time needed to increase additional capacity	Deliver
		Downside Deliver Adaptability	Time needed to decrease supplemental capac-ity	Deliver
	Cost	Cost to Plan	Information cost	Plan
		Cost to Source	Ordering carrying cost	Source
			Inventory cost	Source
		Cost to Make	Production cost	Make
		Cost to Deliver	Transportation cost	Deliver
			Backorder cost	Deliver
Performance Measures			Bullwhip Effect	
			Service level	
			Total supply chain cost	

characteristics. Without these characteristics it is extremely difficult to compare an organization that chooses to be the low-cost provider to an organization that chooses to compete on reliability and performance.

4 Simulation Modeling and Experiments

4.1 Setting Parameters

To verify our framework, we consider the case of a Beer Game model with a four-echelon supply chain consisting of retailer(R)-wholesaler(W)-distributor(D)-manufacturer(M). Suppose that the managers of the supply chain want to improve their performance by reducing the production lead time at the factory or in transportation from one location to another. Due to resource and investments, the lead time at

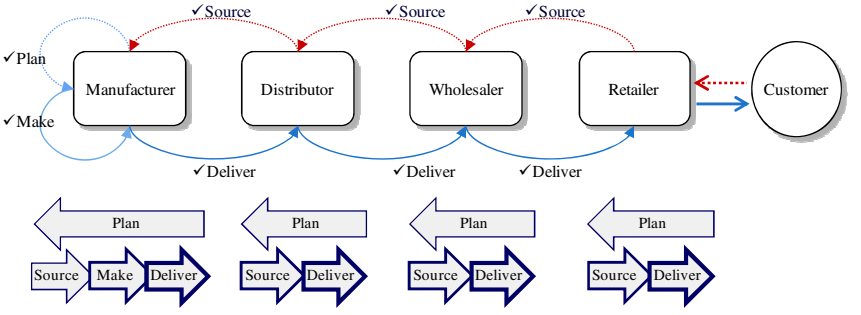


Fig. 5. Structure & processes of Beer Game model with a four-echelon

one or two points can be reduced to 1 or 2 days. First, we construct the structure and process of the model as in Fig. 5, and then parameters are set according to Table 2.

Table 2. Parameter setting

		Level
System parameters	Demand pattern	Step function (10 → 20 units/period)
	Information sharing	None
	Planning lead time	1 period
	Replenishment rule	Base stock model
	Inventory cost	1 $ /unit period
	Backorder cost	2 $ /unit period
	Forecasting Method	Moving average
Decision variables	Lead time	Current states, Alternatives (See Table 3)
Output variables	Total supply chain cost	Inventory cost + Backorder cost

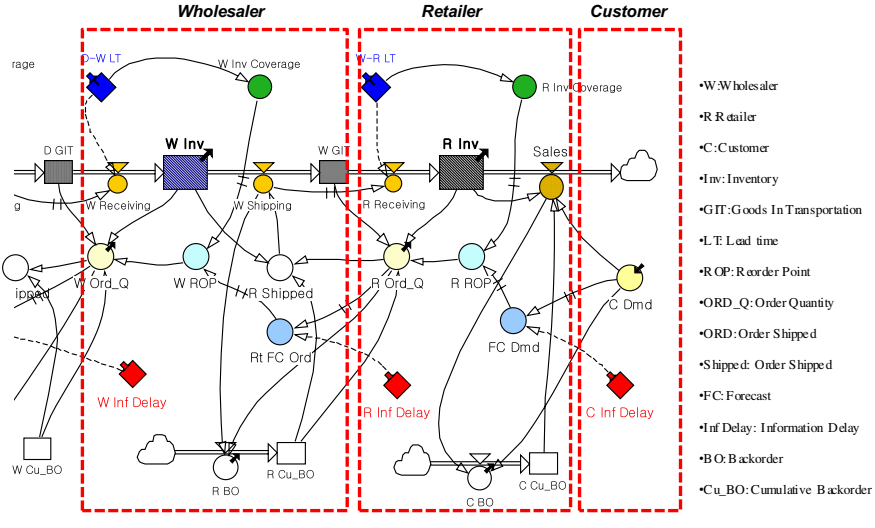


Fig. 6. System dynamics model

4.2 Simulation Modeling Using System Dynamics and Results

Fig. 6 shows a piece of the whole simulation model which was built using Powersim™ software. Fig. 7 shows the order rate and inventory changes, which are very similar to those in previous research.

Table 3 shows the simulation results for the current and alternate models. The results indicate that alternative B6 is the best for our case.

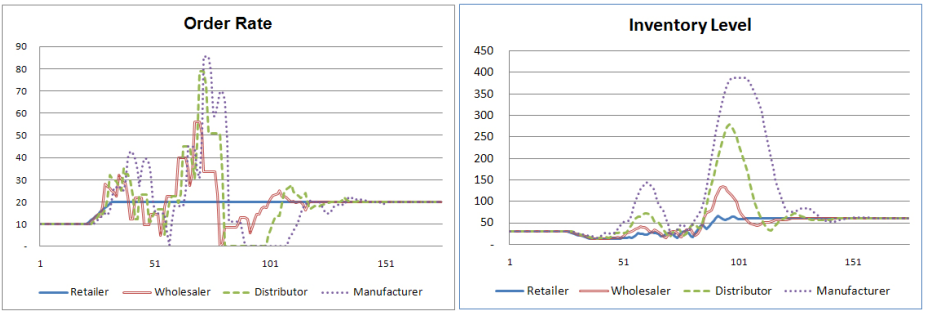


Fig. 7. Order rate and inventory change in supply chain

Table 3. Simulation results

	Lead time (period)				Simulation Result			
	M	*M-D*	*D-W*	*W-R*	Order Variance*	Inventory cost	Backorder cost	Total cost
Current	4	4	4	4	12.35	39,277	29,024	68,301
A1	4	4	4	1	10.21	36,112	27,751	63,863
A2	4	4	1	4	8.28	31,393	23,987	55,380
A3	4	1	4	4	5.60	28,764	19,623	48,387
A4	1	4	4	4	5.53	29,759	23,303	53,062
B1	4	4	2	2	7.90	33,779	21,619	55,397
B2	4	2	4	2	6.05	29,425	22,810	52,236
B3	2	4	4	2	6.22	30,609	24,496	55,105
B4	4	2	2	4	5.78	28,281	21,254	49,535
B5	2	4	2	4	4.91	29,260	17,511	46,771
B6	2	2	4	4	4.79	29,984	14,675	44,658

* Order variance=Var(*R*)/Var(*M*).

5 Conclusions

In this paper we proposed a framework for the modeling and analysis of the bullwhip effect utilizing the SCOR model, a widely accepted industry standard. We conducted simulation modeling for a Beer Game, analyzed the results and then investigated the most effective strategy to improve its performance. Despite the apparent simplicity of the model, it does provide us with a good framework for improving system performance.

To improve the supply chain efficiency in a more realistic situation, we have to consider more environmental and operational parameters such as the customer order

decoupling point and the supply chain relationship. It is also necessary to describe the strategy planning procedure and statistical analysis in more detail.

References

1. Forrester, J.W.: Industrial dynamics. MIT Press, Cambridge (1961)
2. Towill, D.R.: Supply chain dynamics. International Journal of Computer Integrated Manufacturing 4, 197–208 (1991)
3. Wikner, J., Towill, D.R., Naim, M.: Smoothing supply chain dynamics. International Journal of Computer Integrated Manufacturing 22, 231–248 (1991)
4. Lee, H.L., Padmanabhan, V., Whang, S.: Information distortion in a supply chain: the bullwhip effect. Management Science 43, 546–558 (1997)
5. Sterman, J.D.: Modelling managerial behavior: Misperception of feedback in a dynamic decision making experiment. Management Science 35, 321–339 (1989)
6. Lee, H.L., So, K., Tang, C.: The value of information sharing in a two-level supply chain. Management Science 46, 626–643 (2000)
7. Dejonckheere, J., Disney, S.M., Lambrecht, M.R., Towill, D.R.: Measuring and avoiding the bullwhip effect: a control theoretic approach. European Journal of Operational Research 147, 55–74 (2003)
8. Miragliotta, G.: Layers and mechanisms: A new taxonomy for the bullwhip effect. International Journal of Production Economics 104, 365–381 (2006)
9. Zhao, X., Xie, F., Zhang, W.F.: The impact of information sharing co-ordination on supply chain performance. Supply Chain Management: An International Journal 7, 24–40 (2002)
10. Gavirneni, S., Kapuschinski, R., Tayur, S.: Value of information in capacitated supply chains. Management Science 45, 16–24 (1999)
11. Chen, F., Drezner, Z., Ryan, J.K., Simchi-Levi, D.: Quantifying the Bullwhip Effect in a simple supply chain: The impact of forecasting, lead times and information. Management Science 46, 436–443 (2000)
12. Krane, S.D., Braun, S.N.: Production Smoothing Evidence from Physical-Product Data. Journal of Political Economy 99, 558–581 (1991)
13. Gorman, M.F., Brannon, J.I.: Seasonality and the production-smoothing model. International Journal of Production Economics 65, 173–178 (2000)
14. Li, G., Wang, S., Yan, H., Yu, G.: Information transformation in a supply chain: a simulation study. Computers & Operations Research 32, 707–725 (2005)
15. Baganha, M.P., Cohen, M.A.: The stabilizing effect of inventory in supply chains. Operations Research 46, 72–83 (1998)
16. Gunasekaran, A., Patel, C., Tirtiroglu, E.: Performance measures and metrics in a supply chain environment. International Journal of Operations & Production Management 21, 71–87 (2001)
17. Otto, A., Kotzab, H.: Does supply chain management really pay? Six perspectives to measure the performance of managing a supply chain. European Journal of Operational Research 144, 306–320 (2003)
18. Chan, T.S., Qi, H.J., Chan, H.K., Lau, C.W., Ralph, W.L.: A conceptual model of performance measurement for supply chains. Management Decision 41, 635–642 (2003)
19. Sanchez, A.M., Perez, M.P.: Supply chain flexibility and firm performance. International Journal of Operations & Production Management 25, 681–700 (2005)
20. Milgate, M.: Supply chain complexity and delivery performance: an international exploratory study. Supply Chain Management: An international Journal 6, 106–118 (2001)
21. Supply Chain Council, SCOR Model v.8.0 (2006)

Dynamic Control of a Simulator Servo System Using CMAC Controller RBF Identifier-Based

Wang Weihong[1] and Yu Lei[2]

[1] Automatic Control Department Beijing University of Aeronautics and Astronautics, 37# Xueyuan Rd, Haidian District, Beijing China
wwh2005@buaa.edu.cn
[2] Automation Institute Qingdao University of Science and Technology, Qingdao, Shandong, China
frank_yu@126.com

Abstract. Neural control technique is well-suited to the problem of controlling dynamic system. A kind of CMAC controller for flight simulator servo system RBF identifier-based is presented. The controller consists of the RBF identifier, the single neuron PID controller and the CMAC feedforward controller. The RBF neural network is used to identify the model of the plant and adjust the single neuron PID controller's parameter. The suitable parameter of the controller is given as fast as possible by used searching. This method could shorten transient response time clearly. The simulation experiment has proved the method can improve control precision and the response rapidness.

Keywords: CMAC, flight simulator, RBF, servo system.

1 Introduction

Multi-layer feed-forward neural network is the network model which is studied and used most widely in the control field, it use the error back-propagation(BP) method to adjust weights. In theory feed-forward BP networks with more than three layers can approach to any nonlinear functions, so we can use it to identify dynamic model of plant and structure controller. But BP networks are global approximation networks, it must readjust all weights at every train, so converges slowly, and the optimization algorithm quadrics gradient-descent method based is easy to get in local minimum. So BP networks can't satisfy the real-time demand of online identification and control. Radial basis function (RBF) neural network is a kind of local approximation networks, whose architecture is similar to BP networks. Differently BP networks can have several hidden layers, while RBF networks have only one hidden layer, and their activation functions of hidden layers are different.

Cerebellar model articulation controller (CMAC) is a kind of local approximation networks with efficient address mapping. Its learning can't get in local minimum, and its least mean squares (LMS) algorithm is simple and quick, so it can approach nonlinear time-varying map in servo system quickly, and be suitable for real-time control fields[1][2].

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 21–28, 2007.

Three-shaft flight simulator is a complex electromechanical servo system with nonlinear taches, it requires high performance of quick trace to dynamic signal. Because of above characters of CMAC network and RBF networks, we use them for system control, perhaps can obtain high control performance. So we combine RBF identifier with compound CMAC controller, and present a new servo control method.

2 Description of the Plant

Flight simulator servo system is composed of control computer, preamplifier, PWM, DC- power supply, DC-motor, velocity sensor, position sensor and so on. Preamplifier, PWM, DC- power supply can be regarded as amplifier , and main parts are DC-motor, shaft, load. The model of plant is shown as fig. 1.

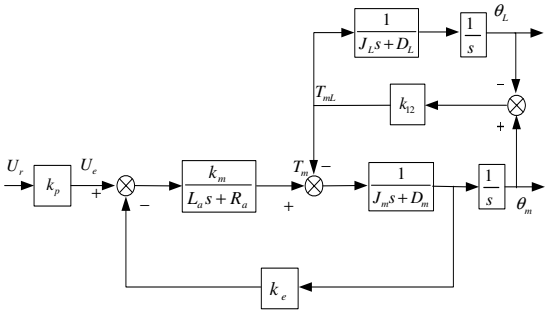


Fig. 1. Model of servo system

In fig.1, U_r is control voltage, U_a, L_a, R_a are motor's armature voltage, inductance and resistance, J_m is motor's moment of inertia, J_L is load's moment of inertia, D_m , D_L denote viscous damping coefficient of motor and load respectively, k_p is the gain of amplifier, k_m is motor's electromagnetic moment coefficient, k_e is motor's back electromotive force coefficient, k_{12} is rigidity coefficient of shaft, θ_m is corner of motor's shaft.

We choose amplifier, PWM, motor and load as generalized plant in design.

3 Controller Design

The architecture of servo system is shown as fig.2. Controller includes RBF identifier, single neuron PID controller and CMAC feed-forward controller. The RBF network is used to identify the Jacobian parameter $\partial y / \partial u$ and the parameter is used to regulate the Kp and Kd.. The control signal input to the plant is the sum of the terms from the CMAC module and the single neuron PID feedback controller. At the beginning, the

PID controller provides all of the initial servo system control signal, after each control cycle, the CMAC weights are updated. As training progresses, the CMAC output signal improves, and the resulting control error decreases, diminishing the control effort required by the PID controller and increasing that of the CMAC. The CMAC therefore progressively takes over control from the single neuron PID. Training is performed on-line in real time.

3.1 RBF Identifier

The RBF identifier in fig.2 chooses three-layer RBF network with 3-6-1 architecture, that is, three input nodes, six hidden nodes, one output node. Computation process of the network is as follows:

Input vector is $X=[x_1,x_2,x_3]^T$, where $x_1=u\_1$, $x_2=u\_2$, $x_3=r\_1$. Suppose that u, r are the control output and command at k time, u_1, u_2 are control output at k-1 time and k-2 time respectively, r_1 is command at k-1 time.

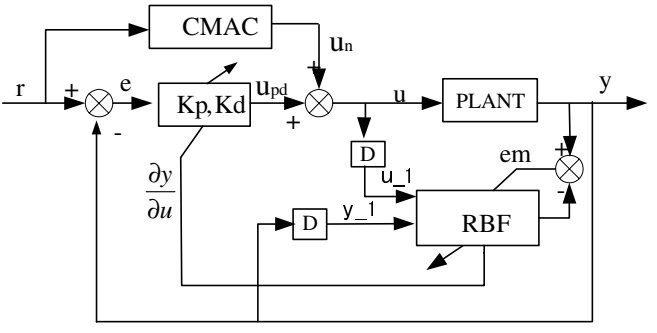


Fig. 2. Structure diagram of neural network feed-forward control system

Hidden layer vector is $H=[h_1,h_2,h_3,h_4,h_5,h_6]$,where, $h_j=\exp\left[-\frac{\|X-C_j\|}{2b_j^2}\right]$

Central vector is $C_j=[C_{j1},C_{j2},C_{j3}]^T$;

Basis width vector is $B=[b_1,b_2,b_3,b_4,b_5,b_6]^T$, and $b_j>0(j=1,2,3,4,5,6)$;

Weight vector of network is $WC=[wc_1,wc_2,wc_3,wc_4,wc_5,wc_6]^T$;

Then, the output of RBFNC is $U_m=H\bullet WC=wc_1h_1+wc_2h_2+wc_3h_3+wc_4h_4+wc_5h_5+wc_6h_6$

The control performance function is chosen as $J_C=\frac{1}{2}(r(t+1)-y(t+1))^2$

Then by the gradient-descent method, we can get the adjusting algorithm of output weight, central vector and basis width vector of hidden layer as follows:

Let $e_c=r(t+1)-y(t+1)$, then.

$$wc_j(t+1)=wc_j(t)+\eta_w e_c \cdot yu \cdot h_j+\alpha_w[wc_j(t)-wc_j(t-1)] \tag{1}$$

$$\delta_{bj} = e_c \cdot yu \cdot wc_j \cdot h_j \cdot \frac{\left\| X - C_j \right\|^2}{b_j^3} \tag{2}$$

$$b_j(t+1)=b_j(t)+\eta_b \delta_{bj}+\alpha_b[b_j(t)-b_j(t-1)] \tag{3}$$

$$\Delta C_{ji} = e_c \cdot yu \cdot wc_j \cdot \frac{X_i - C_{ji}}{b_j^2} \tag{4}$$

$$C_{ji}(t+1)=C_{ji}(t)+\eta_c \Delta C_{ji}+\alpha_c[C_{ji}(t)-C_{ji}(t-1)] \tag{5}$$

Where, j=1,2,3,4,5,6, i=1,2,3, η_w, η_b, η_c are learning rate of output weight, basis width and center of hidden layer respectively; $\alpha_w, \alpha_b, \alpha_c$ are their factors of momentum respectively; yu is the Jacobian parameter of plant. The system carries identifying parameter to single neuron controller, and adjusts Kp, Kd. This can get the suitable value as quickly as possible, make system stable, and shorten the transient state time.

3.2 Single Neuron Controller

Single neuron is a multi-input and single-output process unit with ability of self-learning and self-adaptation, its architecture is shown as fig.3, its input-output characteristic is denoted as formula (6).

$$y = K \sum_{i=1}^{n} w_i x_i - \theta \tag{6}$$

In formula (6), K is coefficient of proportionality; W_i is the weight corresponding with x_i, we can compute it by certain learning algorithm.

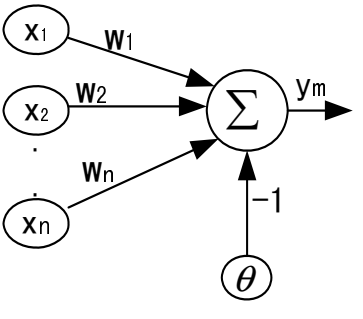


Fig. 3. Architecture of single neuron

Single neuron has small computational cost, so it has special application value in some situation. For example it can be used as identifier in real-time control system, or as controller to satisfy real-time and quick performance requirement[4].

3.3 CMAC Controller

CMAC controller uses three-layer neural network with input layer, hidden layer and output layer. We optimize weights by training, and introduce single neuron to judge control performance of CMAC, this can strengthen system stability and attenuate disturbance. In learning process, CMAC's control output u_c becomes the whole output of the system gradually by learning control voltage U incressantly[3].

The controller design process can be summarized as following steps:

(1) Choose command signal U as the input of CMAC, then quantify it.

Divide the input signal in input layer, and make each input correspond with a unique quantification unit. Middle layer is composed of many judgment intervals. For any input, only a few intervals' outputs are nonzero. Generalization parameter C stipulates the region size which influences network output.

(2) Adopt hashing mapping which uses remainder-divided method, this can project input swatch to the addresses of concept memory. It is divided by a number, and then we can obtain a remainder as actual memory address.

The "conflict" problem caused by hashing mapping can be ignored in algorithm design, if parameter is chosen appropriately. This is because the conflict can be regarded as a stochastic disturbance and its influence can be reduced gradually by iterative process, when the conflict is not fierce. So it can't influence output. In fact Albus the proposer of CMAC also proved that the probability of conflict is nearly impossible, so long as the number of memory cells is one hundred times as great as generalized constant C.

(3) In the primary step of CMAC control, command signal provide the next step's desired state, CMAC network generates corresponding output u_c(k), output of routine controller is u_{pd}(k), the whole control signal of system is sum of above two signal. When weight training begins, weight is adjusted according to the error between u_c and u until it satisfy precision requirement.

When CMAC is used as controller, it adopts the algorithm as follows:

$$y(k) = rin(k+1) \tag{7}$$

$$u(k) = u_{pd}(k) + u_c(k) \tag{8}$$

$$\varepsilon_k = u(k) - u_c(k) \tag{9}$$

$$w_j(k+1) = w_j(k) + \eta \cdot \frac{\varepsilon_k}{c} \tag{10}$$

When the system begins working, it sets weight vector W to be zero vector, now u_c =0, u= u_{pd} , system is controlled by routine controller, that is, error vector $e_i = x_i - x$ is very great, and inputted to routine controller. After computing, it

outputs control signal u_{pd}. Along with continuous learning of network, output u_c of CMAC whose input is x_d will approach to the requiring u which can make the system trace desired state. ξ is acceptable learning error. When $|u-u_c| \le \xi$, weights are changeless; When $|u-u_c| \ge \xi$, weights are adjusted according to formula (10).

3.4 Study by Simulation

Formula (11) is the model of some servo system plant. We simulate this model, the input is sinusoidal signal, its amplitude is 1.0 degree, its frequency is 4Hz. Simulation result is shown as fig.4, in which the dotted line is command signal, dashdotted line is system's respond.

$$G(s) = 1100/(s^2 + 80s) \tag{11}$$

Fig.5 is amplificatory figure of respond curve, Fig.6 is output of controller. Comparatively Fig.7 and Fig.8 is the curve when system has no RBF identifier, Fig.7 is the amplificatory figure of initiatory transition process tracing sinusoidal signal, andFig.8 is output curve of controller.

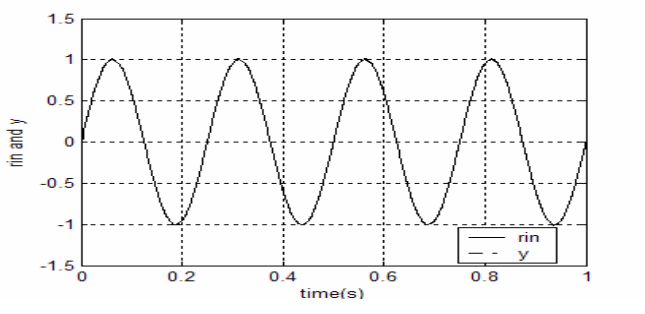


Fig. 4. Simulation curve tracking sinusoidal signal

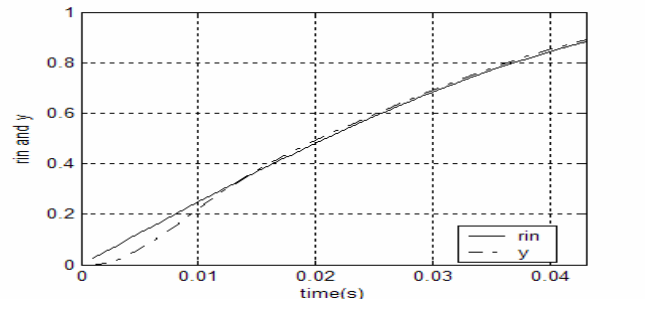


Fig. 5. Amplificatory figure of output respond (with RBF identifier)

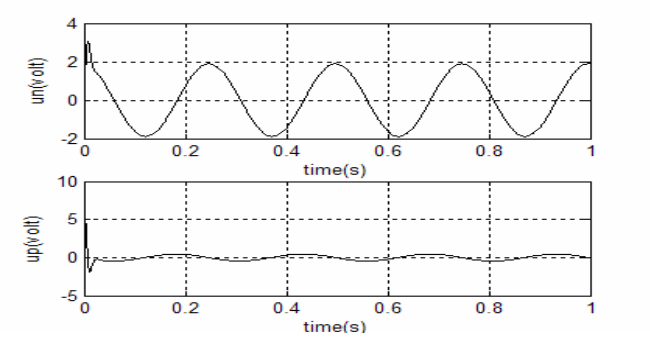


Fig. 6. Control signal (with RBF identifier)

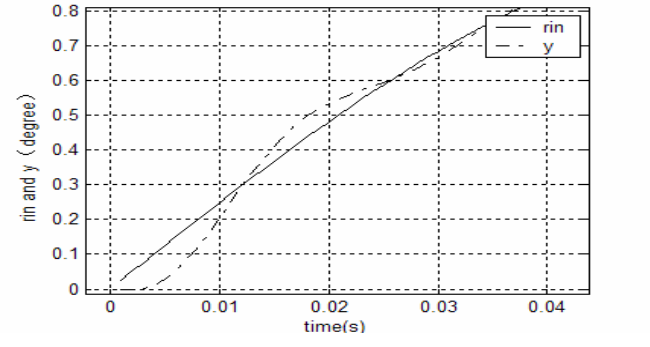


Fig. 7. Amplificatory figure of output respond (without RBF identifier)

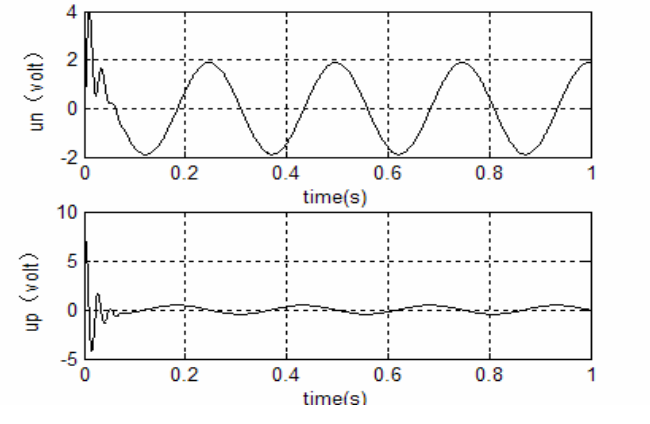


Fig. 8. Control signal (without RBF identifier)

Comparing control signal and real-time respond curve of these two cases, the system with RBF identifier reaches stability after only one surge, while the system without RBF identifier surges many times, and its transient respond time is long.

4 Conclusion

The paper proposes a kind of CMAC controller for servo system RBF neural network identifier-based. Identify Jacobian parameter of the plant by RBF identifier, and carry it to single neuron controller, then adjusts Kp, Kd. This can get the proper value as quickly as possible, make system stable, and shorten the transient state time. It realizes the control rapidness. The simulation results show that controller is effective, and adjustment to transient respond time is obvious by the increase of frequency.

References

1. Mese, E.: A rotor position estimator for switched reluctance motors using CMAC[J]. Energy Conversion and Management 44, 1229–1245 (2003)
2. Gang, S., Zengqiang, C., Zhuzhi, Y.: The theory and application of CMAC[J]. Yi Qi Yi Biao Xue Bao 24(S4), 269–273 (2003.8)
3. Rudenko, O.G., Bessonov, A.A.: CMAC neural network and its use in problems of identification and control of nonlinear dynamic objects[J]. Cybernetics and Systems Analysis 41(5), 647–5816 (2005)
4. Xie, Z., Liang, D., Cheng, Z., Wen, H.: Research on Control System of Linear PM Brushless DC Motor Based on Single Neuron PID[C], Electrical Machines and Systems. In: ICEMS 2005, Proceedings of the Eighth International Conference on, September 27-29, 2005, vol. 1, pp. 385–387 (2005)

A Strict LMI Condition for ESPR Property of Continuous-Time Descriptor Systems

Jian Liung Chen[1], Wei-Der Chang[2], Chia-Hung Lin[1],
Meei-Ling Hung[3], and Yen-Lon Chen[2]

[1] Department of Electrical Engineering, Kao Yuan University,
Kaohsiung County 821, Taiwan, R.O.C.
{clchen,eecl53}@cc.kyu.edu.tw
[2] Department of Computer and Communication, Shu-Te University,
Kaohsiung County 824, Taiwan, R.O.C.
wdchang@mail.stu.edu.tw,
skuddc@hotmail.com
[3] Department of Electrical Engineering, Far East University,
Tainan County 744, Taiwan, R.O.C.
meeiling@cc.feu.edu.tw

Abstract. In this paper, a new strict linear matrix inequality (LMI) condition for extended strictly positive real (ESPR) characterization of continuous-time descriptor systems is addressed. Compared with existing nonstrict LMI conditions, the proposed conditions are expressed in terms of definite LMIs with no equality constraint, which are more tractable and reliable in numerical computation. A numerical example is demonstrated to verify the theoretical results obtained in this paper.

Keywords: Continuous-time descriptor systems, Extended strict positive realness, Strict LMI.

1 Introduction

The descriptor model is a natural mathematical representation for many practical systems because it provides a description of the dynamic as well as the algebraic relationships between the chosen descriptor variables simultaneously. Due to its more direct and general description than the state-space representation of dynamic systems, such a model has been employed in different areas of research, e.g. robotics, chemical process control, power systems, and highly interconnected large-scale systems [1].

The (strict) positive realness of transfer matrix is an essential property in network and circuit theory. Due to contributions of Yakubovich and Kalman, the Kalman-Yakubovich-Popov lemma (also called the positive real lemma) has many important researches in control system theory, e.g., stability analysis [2], absolute stability [3], adaptive control [4], optimal control [4] and robust control [4], etc. Besides, Haddad and Bernstein [5] propose algebraic conditions to guarantee the property of extended

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 29–35, 2007.

strict positive realness (ESPR), a stronger sense of the strict positive realness (SPR), of a transfer matrix and study the connection with robust stability.

Viewing the importance of (extended) SPR and the generality of descriptor system models, investigation of the property for descriptor systems and its control problem becomes an essential and attractive topic. In this paper, we propose a now strict LMI condition, which definite LMI with no equality constraint in it, for the descriptor ESPR property. Compared with existing nonstrict LMI condition [6], the proposed condition is expressed in terms of definite LMI with no equality constraint, which are more tractable and reliable in numerical computation. Finally, a numerical example is demonstrated to show the result.

Notations employed in the paper are explained as follows: Let M be a matrix of complex numbers with proper dimension, M^T and M^* stand for the transpose and the hermitian of M, individually; $M > 0$ (or $M < 0$) means that M is positive (or negative) definite; while, by $M \geq 0$ (or $M \leq 0$), M is positive (or negative) semi-definite; and $M > N$ (or $M \geq N$) is equivalent to $M - N > 0$ (or $M - N \geq 0$). The identity matrix with dimension $r \times r$ is denoted by I_r and we simply use I to indicate any identity matrix with proper dimension. In the sequel, $\Re^n$ denotes the n dimensional Euclidean space, $\Re^{n\times m}$ denotes the set of all $n \times m$ real matrices, and $\mathrm{Re}[s]$ denote the real part of s.

2 Preliminaries

We consider the following descriptor system

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bw(t) \\ z(t) &= Cx(t) + Dw(t) \end{aligned} \tag{1}$$

where $x \in \Re^n$ is the descriptor variable, $w \in \Re^{n_w}$ is the exogenous input, and $z \in \Re^{n_z}$ is the external output. The matrix E possibly being singular, i.e. $rank(E) = r \leq n$, and other matrices are constant of appropriate dimensions. For brevity, we will use (E, A, B, C, D) to denote all the considered descriptor systems. When only the behavior of the descriptor variable in an unforced system (1) is concerned, the simple notation (E, A) will be employed. Some important features to the study of descriptor systems are recalled below.

Definition 1. [7] A pair (E, A) is called regular if $\det(sE - A)$ is not identically zero, called impulse-free, i.e. $(sE - A)^{-1}$ is proper, if degree of $\det(sE - A)$ is equal to $rank(E)$, and called stable if all the roots of $\det(sE - A) = 0$ lie in the open left-half plane. Furthermore, we call the pair (E, A) admissible if it is regular, impulse-free, and stable.

If the descriptor system (1) is regular, the transfer matrix from w to z is well defined by

$$T_{zw} = C(sE - A)^{-1}B + D$$

and the existence and uniqueness of solution to pair (E, A) are guaranteed for any specified initial condition. It is prove in [7] that the system (1) is regular and impulse-free, then there exist two nonsingular matrix Ψ and Φ such that

$$\breve{E} = \Psi E \Phi = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \text{ and } \breve{A} = \Psi A \Phi = \begin{bmatrix} \breve{A}_r & 0 \\ 0 & I_{n-r} \end{bmatrix}$$

Therefore, the system (1) is converted into the Weierstrass normal form

$$(\breve{E}, \breve{A}, \breve{B}, \breve{C}, \breve{D})$$

where

$$\breve{B} = \Psi B = \begin{bmatrix} \breve{B}_1 \\ \breve{B}_2 \end{bmatrix}, \quad \breve{C} = C\Phi = \begin{bmatrix} \breve{C}_1 & \breve{C}_2 \end{bmatrix}$$

are partitioned accordingly and $\breve{D} = D$. The transfer matrix from w to z can be rewritten as

$$T_{zw} = \breve{C}_1 \left(sI_r - \breve{A}_r \right)^{-1} \breve{B}_1 + \breve{D} - \breve{C}_2 \breve{B}_2 .$$

3 Main Results

In this section, the strict LMI condition to check the admissibility and the ESPR property of linear time-invariant descriptor systems is proposed. At first, we give the definitions about extended strictly positive realness of the rational transfer matrix.

Definition 2. [5,6,8] Let $H(s)$ be a square real rational transfer matrix in s. $H(s)$ is said to be extended strictly positive real (ESPR) if it is analytic in $\mathrm{Re}[s] \geq 0$ and $H(j\omega) + H^*(j\omega) > 0$ for all $\omega \in [0, \infty]$.

Next, we recall a generalization of the well-known ESPR lemma for system (1). This lemma provides a necessary and sufficient LMI condition for the admissibility of (E, A) pair with T_{zw} being ESPR.

Lemma 2. [6,8] Consider system (1) and suppose that $D + D^T > 0$. Then the following statements are equivalent.

(I) (E, A) is admissible and T_{zw} is ESPR.
(II) There exists a nonsingular matrix Q such that

$$\begin{bmatrix} AQ + (AQ)^T & B - (CQ)^T \\ B^T - CQ & -(D + D^T) \end{bmatrix} < 0$$
$$EQ = (EQ)^T \geq 0. \tag{2}$$

The LMI conditions in Lemma 1 is nonstrict, since it contains equality constraint $E^T P = P^T E \geq 0$ or $EQ = (EQ)^T \geq 0$. As mention in [9], these equality constraints may cause a big trouble in checking the conditions numerically. To overcome this drawback, we propose strict LMI condition for ESPR property of system (1) which statements as follows.

Theorem 1. Consider system (1) and suppose that $D + D^T > 0$. Then the following statements are equivalent.

(I) (E, A) is admissible and T_{zw} is ESPR.

(II) There exist a positive definite matrix P and a matrix S such that

$$\begin{bmatrix} A(PE^T + VS) + (PE^T + VS)^T A^T & B - (PE^T + VS)^T C^T \\ B^T - C(PE^T + VS) & -(D + D^T) \end{bmatrix} < 0, \tag{3}$$

where the matrix V which is of full column rank and composed of based of $Null\ E$.

Proof: [(II)⇒(I)] By setting $Q := (PE^T + VS)$ in (3), we obtain the first inequality of (2). Note that $EQ = E(PE^T + VS) = EPE^T = (EQ)^T \geq 0$, this implies that Q satisfies the second inequality of (2) with the equality constraint. By Lemma 1, we conclude (E, A) is admissible and T_{zw} is ESPR.

[(I)⇒(II)] By taking the singular value decomposition (SVD) on E, there exist two real unitary matrices M and N such that

$$\overline{E} := MEN = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}$$

where $\Sigma_r \in \Re^{r \times r}$ is diagonal with the singular values of E. Then, by defining $\overline{x} := N^{-1} x$, system (1) is converted into an SVD coordinate system

$$(\overline{E}, \overline{A}, \overline{B}, \overline{C}, \overline{D}) := (MEN, MAN, MB, CN, D)$$

Note that, from the block structure of $\overline{E}$, the matrix $\overline{V} = \begin{bmatrix} 0 & \overline{V}_2^T \end{bmatrix}^T$, where $\overline{V}_2 \in \Re^{(n-r) \times (n-r)}$ is nonsingular matrix, is composed of base of $Null\ \overline{E}$. Therefore, the matrix V can be represented as

$$V := N\overline{V}.$$

If the statement (I) holds then it is equivalent to $(\overline{E}, \overline{A})$ is admissible and $\overline{C}(s\overline{E} - \overline{A})\overline{B} + \overline{D}$ is ESPR. Therefore, by Lemma 1, there exists a nonsingular matrix $\overline{Q}$ satisfying

$$\begin{bmatrix} \overline{A}\overline{Q} + (\overline{A}\overline{Q})^T & \overline{B} - (\overline{C}\overline{Q})^T \\ \overline{B}^T - \overline{C}\overline{Q} & -(\overline{D} + \overline{D}^T) \end{bmatrix} < 0 \tag{4}$$

$$\overline{E}\overline{Q} = (\overline{E}\overline{Q})^T \geq 0 \tag{5}$$

Partitioning $\overline{Q}$ in accordance with the block structure of $\overline{E}$ as

$$\overline{Q} = \begin{bmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{bmatrix}.$$

From (5) and $\overline{Q}$ is nonsingular, we can conclude that $Q_2 = 0$ and $\Sigma_r Q_1 = (\Sigma_r Q_1)^T > 0$. Let

$$\overline{P} = \begin{bmatrix} Q_1\Sigma_r^{-1} & (Q_3\Sigma_r^{-1})^T \\ Q_3\Sigma_r^{-1} & \alpha I \end{bmatrix}$$

and $\overline{S} = \overline{V}_2^{-1} Q_4 [0 \quad I_{n-r}]$, it is easy to show $\overline{P}\overline{E}^T + \overline{V}\overline{S} = \overline{Q}$. Note that we can get $\overline{P} > 0$ by choose $\alpha > 0$ such that $\alpha I - Q_3(Q_1\Sigma_r)^{-1}Q_3 > 0$. Hence (4) can be rewritten as

$$\begin{bmatrix} \overline{A}(\overline{P}\overline{E}^T + \overline{V}\overline{S}) + (\overline{P}\overline{E}^T + \overline{V}\overline{S})^T \overline{A} & \overline{B} - (\overline{P}\overline{E}^T + \overline{V}\overline{S})^T \overline{C}^T \\ \overline{B}^T - \overline{C}(\overline{P}\overline{E}^T + \overline{V}\overline{S}) & -(\overline{D} + \overline{D}^T) \end{bmatrix} < 0$$

$$\Leftrightarrow \begin{bmatrix} MAN(\overline{P}N^T E^T M^T + N^{-1}\overline{V}\overline{S}M^{-T}M^T) + \\ \quad (\overline{P}N^T E^T M^T + N^{-1}\overline{V}\overline{S}M^{-T}M^T)N^T A^T M^T \\ (MB)^T - CN(\overline{P}N^T E^T M^T + N^{-1}\overline{V}\overline{S}M^{-T}M^T) \end{bmatrix}$$

$$\begin{matrix} MB - (\overline{P}N^T E^T M^T + N^{-1}\overline{V}\overline{S}M^{-T}M^T)^T (CN)^T \\ -(D + D^T) \end{matrix} \Bigg] < 0$$

Let $P := N\overline{P}N^T > 0$ and $S := \overline{S}M^{-T}$, the above inequality can represented as

$$\begin{bmatrix} M & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A(PE^T + VS) + (PE^T + VS)^T A^T & B - (PE^T + VS)^T C^T \\ B^T - C(PE^T + VS) & -(D + D^T) \end{bmatrix} \begin{bmatrix} M^T & 0 \\ 0 & I \end{bmatrix} < 0,$$

which is equivalent to (3).

4 Numerical Example

Example 1. Consider the following descriptor system

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \dot{x}(t) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix} x(t) + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} w(t)$$

$$y(t) = [1 \quad 0 \quad 0] x(t) + w(t)$$

where $D + D^T > 0$. It is desired to know whether the pair (E, A) is admissible and the transfer matrix T_{zw} is ESPR. The MATLAB LMITOOLBOX [10] is used to check the feasibility of LMI (3) derived in Theorem 1. In this example, the matrix V can be chose as

$$V = [0 \quad 0 \quad 1]^T.$$

After several iterations, the LMITOOL solver indicates that the LMI is feasible and the solver returns

$$P = \begin{bmatrix} 3.9731 & 0 & 0 \\ 0 & 3.4009 & 0 \\ 0 & 0 & 6.5879 \end{bmatrix}; \; S = \begin{bmatrix} 0 & 0 & 0.7647 \end{bmatrix}.$$

This means that (E, A) is admissible and the associated transfer function

$$\frac{s+1}{s+2} \tag{6}$$

is ESPR as well. By the Definition 1 and the matrices E and A are diagonal, it is easy to check the pair (E, A) is admissible. To verify T_{zw} is ESPR, we calculate the value of sum of (6) and its Hermitian for ω ranging from 0 to 100. Fig. 1 displays the whole outcome of the calculation. It is clear to see that the transfer function is ESPR.

5 Conclusion

We have presented the strict LMI condition for extended strictly positive real lemma that can be used to check the admissibility and the ESPR property of linear time-invariant descriptor systems. The condition is much more tractable and reliable in numerical computations than existing condition which contain equality constraints.

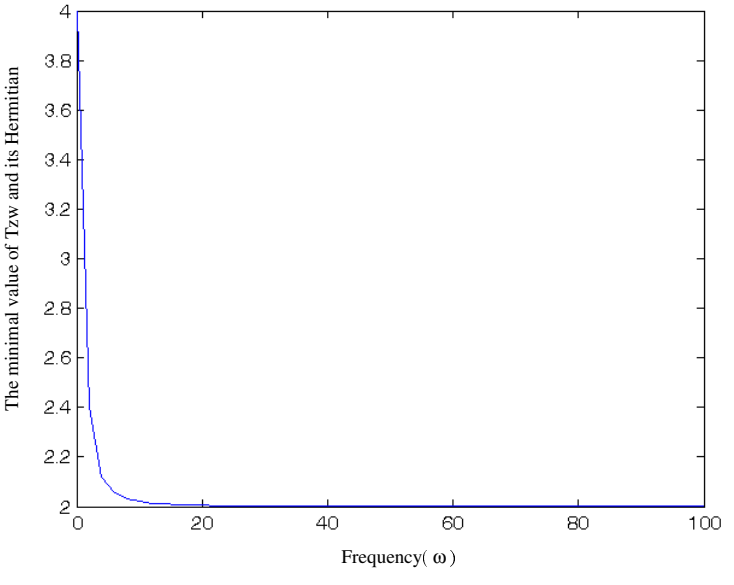


Fig. 1. The value of $T_{zw}(j\omega) + T_{zw}^{*}(j\omega)$

References

1. Lewis, F.L., Mertzios, V.G.: Recent Advances in Singular Systems. Circuits, Systems, and Signal Processing 8 (1989)
2. Popov, V.M.: Hyperstability of Control Systems. Springer, New York (1973)
3. Xiao, C., Hill, D.J.: Concepts of Strict Positive Realness and The Absolute Stability Problem of Continuous-time Systems. Automatica 34, 1071–1082 (1998)
4. Barabanov, N.E., Gelig, A.K., Leonov, G.A., Likhtarnikov, A.L., Matveev, A.S., Smirnova, V.B., Fradkov, A.L.: The Frequency Theorem (Kalman-Yakubovich Lemma) in Control Theory. Automat. Remote Control 57, 1377–1407 (1996)
5. Haddad, W.M., Bernstein, D.S.: Explicit Construction of Quadratic Lyapunov Functions for The Small Gain, Positivity, Circle, and Popov Theorems and Their Application to Robust Stability. Part I: Continuous-time Theory. Int. J. Robust and Nonlinear Control 3, 313–339 (1993)
6. Zhang, L., Lam, J., Xu, S.: On Positive Realness of Descriptor Systems. IEEE Trans. Circuits Syst. I 49, 401–407 (2002)
7. Dai, L.: Singular Control Systems-Lecture Notes in Control and Information Sciences. Springer, Berlin (1989)
8. Wang, H.S., Yung, C.F., Chang, F.R.: The Positive Real Control Problem and The Generalized Algebraic Riccati Equation for Descriptor Systems. Journal of the Chinese Institute of Engineerings 24, 203–220 (2001)
9. Uezato, E., Ikeda, M.: Strict LMI Conditions for Stability, Robust Stabilization, and H_∞ Control of Descriptor Systems. In: Proc. of the 38th CDC, pp. 4092–4097 (1999
10. Gahinet, P., Nemirovski, A., Laub, A., Chilali, M.: The LMI Control Toolbox. The MathWorks, Inc. (1995)

Design of Fuzzy Sliding-Mode Controller for Chaos Synchronization

Chao-Lin Kuo[1], Cheng-Shion Shieh[1], Chia-Hung Lin[2], and Shun-Peng Shih[3]

[1] No. 49, Jung-Hua Road, Hsin-Shih Township, Tainan County, 744, Taiwan, R.O.C.
Department of Electrical Engineering, Far-East University
clkuo@cc.fue.edu.tw
[2] Lu-Chu Hsiang, Kaohsiung County 821, Taiwan, R.O.C.
Department of Electrical Engineering, Kao-Yuan University
[3] Yen Chau, Kaohsiung County 824, Taiwan, R.O.C.
Department of Computer and Communication, Shu-Te University

Abstract. This paper presents a fuzzy controller to solve a master-slave chaos synchronization problem. At first, the method of traditional sliding mode control is considered, which utilizes the discontinuous sign function to make the system state reaching a sliding surface. Next, fuzzy rules are determined according to the Lyapunov theorem, and the fuzzy controller is designed for chaos synchronization. Finally, an example of chaos synchronization for an uncertain Duffing-Holmes system is presented to illustrate the validity and feasibility of the proposed controller.

1 Introduction

A chaos synchronization problem means making both chaotic oscillators behave exactly the same. Generally two chaotic systems in synchronization are called a drive system and a response system, respectively. Chaos synchronization can be applied in many areas such as in chemical reactions, power converters, signal process, communication, and biological systems [1, 3, 4, 8, 12]. There are many methods for synchronization of a chaotic system such as adaptive control method [5, 6, 14], back-stepping control method [7], H^{∞} control method [10], sliding mode (variable structure) control method [2, 13, 16], and fuzzy control method [15].

Zadeh [17, 18] initiated a fuzzy set theory. The fuzzy logic control schemes have been widely developed for almost 40 years, and have been successfully applied to many applications [13, 15]. We can easily apply the fuzzy logic control to control an ill-modeled system by experiments of skilled operators. Although there have been some successful applications of fuzzy logic control, it still has some drawbacks in the design procedure. For example, the fuzzy control rules are often experience-oriented and suitable membership functions should be given by time-consuming trial-and-error procedures. Besides, the dynamic behavior of control system cannot be specified precisely. In recent years, some chaos synchronization based on fuzzy system has been proposed [15].

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 36–45, 2007.

In this paper, we are devoted to the research of design fuzzy controller for synchronizing the state trajectories of two Duffing-Holmes systems with differential initial conditions, system uncertainties and external disturbances. This paper is organized as follows: Section 2 described the dynamics of mater-slave chaos synchronization system. In section 3 described the design approaches of fuzzy controller. Numerical simulations that confirm the validity and feasibility of proposed method are shown in Section 4. Finally, conclusions are given in Section 5.

2 System Description and Problem Formulation

Consider the following two n-dimensional chaotic systems,

$$\begin{cases} \dot{x}_i = x_{i+1}, & 1 \le i \le n-1 \\ \dot{x}_n = f(x,t) \end{cases}, \quad x \in R^n \tag{1}$$

$$\begin{cases} \dot{y}_i = y_{i+1}, & 1 \le i \le n-1 \\ \dot{y}_n = f(y,t) + \Delta f(y) + d(t) + u \end{cases}, \quad y \in R^n \tag{2}$$

where $u \in R$ is a control input, f is a given nonlinear function of x and t, $\Delta f(y)$ is an uncertain term representing the unmodeled dynamics or structural variation of the system (2) and $d(t)$ is the disturbance of system (2). In general, the uncertain term $\Delta f(y)$ and disturbance term are assumed bounded, i.e.

$$\left|\Delta f(y) < \alpha\right| \text{ and } \left|d(t) < \beta\right|, \tag{3}$$

where α and β are positive.

It is assumed that $f(x,t)$, $f(y,t)$ and $\Delta(y,t)$ satisfy all the necessary conditions, such that system (1) and (2) have a unique solution in the time interval $[t_0, +\infty)$, $t_0 > 0$, for any given initial condition $x_0 = x(t_0)$ and $y_0 = y(t_0)$. The dynamics of system (1) display a chaotic motion without control input $(u = 0)$.

The control problem considered in this paper is that for different initial conditions of systems (1) and (2), the two coupled system, i.e. the master system (1) and the slave system (2), to be synchronized by designing an appropriate control $u(t)$ which is attached to the slave system (2) such that

$$\lim_{t \to \infty} \left\| x(t) - y(t) \right\| \to 0, \tag{4}$$

where $\|\cdot\|$ is the Euclidean norm of a vector.

3 Fuzzy Controller for Design Methodology

In this section, we want to address the sliding-mode control and design procedures of a fuzzy controller for chaos synchronization.

3.1 Sliding-Mode Control

Let the error state be $e_i = y_i - x_i$, $i = 1,2,\ldots,n$, and $g(e,t) = f(e+x,t) - f(x,t)$, the error dynamic equations are

$$\begin{aligned} \dot{e}_i &= e_{i+1} \quad ; 1 \le i \le n-1 \\ \dot{e}_n &= g(e,t) + \Delta f(e+x) + d(t) + u. \end{aligned} \tag{5}$$

Using the concept of extended systems, the standardized state space equations of the error states can be obtained as

$$\begin{aligned} \dot{e}_i &= e_{i+1} \quad ; 1 \le i \le n-1 \\ \dot{e}_n &= g(e,t) + \Delta f(e+x) + d(t) + u = e_{n+1} \\ \dot{e}_{n+1} &= \frac{d}{dt}(g(e,t) + \Delta f(e+x) + d(t)) + \dot{u} \end{aligned} \tag{6}$$

System (6) is of the controllable canonical form. In such a case, there are no internal dynamics [11]. Based on the control law proposed by [2], the sliding surface can be defined as

$$s = e_{n+1} - e_{n+1}(0) + \int_0^t \sum_{j=1}^{n+1} c_j e_j \, dt = 0 \tag{7}$$

where $e_{n+1}(0)$ denotes the initial state of e_{n+1}. Eq. (7) can also be formulated as

$$\dot{e}_{n+1} = -\sum_{j=1}^{n+1} c_j e_j \tag{8}$$

with the initial condition $e_{n+1}(0) = e_{0(n+1)}$, and the sliding mode dynamics can be described by the following system of equations:

$$\begin{aligned} \dot{e}_i &= e_{i+1} \quad ; 1 \le i \le n \\ \dot{e}_{n+1} &= -\sum_{j=1}^{n+1} c_j e_j \end{aligned} \tag{9}$$

or in a matrix equation form as

$$\dot{e}_i = \begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ -c_1 & -c_2 & \cdots & \cdots & \cdots & -c_{n+1} \end{bmatrix} e_i = A_i e_i \quad ; 1 \le i \le n+1 \tag{10}$$

with the initial states being $e_i(0) = [e_{0(1)} \; e_{0(2)} \; \cdots e_{0(n+1)}]^T$. The design parameters c_j can be determined by choosing the eigenvalues of A_i such that the corresponding characteristic polynomial

$$P(e) = \dot{e}_{n+1} + \sum_{j=1}^{n+1} c_j e_j \tag{11}$$

is Hurwitz. These eigenvalues are also relative to the speed of system response.

3.2 Fuzzy Controller Designs

The fuzzy logic control arose from the desire to describe complex control with linguistic descriptions. The fuzzy logic control is easy to understand and simple to implement, because fuzzy logic emulates human control. The fuzzy controller and expert systems have been successfully applied in many complex industrial processes. In this section, we utilize the fuzzy logic and fuzzy propositions to design the controller. The block diagram of fuzzy controller with chaos system is illustrated in Fig. 1. We have the error state from the master and slave systems. The fuzzy controller can determine $u(t)$ by the error and system states.

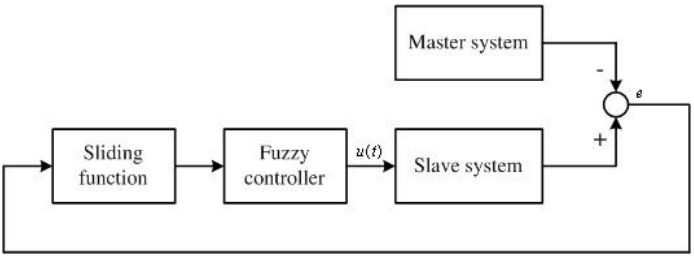


Fig. 1. Diagram of the fuzzy controller

The fuzzy controller has a two-input and a single-output. Input variables are the normalized sliding function (7) and the derivative of sliding function $\dot{s}$. The overall control output is chosen as

$$u_f(t) = F(s, \dot{s}) . \tag{12}$$

where $F(s, \dot{s})$ denotes the functional characteristics of the fuzzy linguistic decision schemes. The membership function of the input linguistic variables s and $\dot{s}$, and the membership functions of the output linguistic variable $u_f(t)$ are shown in Fig. 2, respectively. They are decomposed into seven fuzzy partitions expressed as negative big (NB), negative small (NS), zero (ZE), positive small (PS), and positive big (PB). The fuzzy rule table is designed in Table 1.

The reaching law can be chosen as

$$\dot{s} = -k_f F(s, \dot{s}) . \tag{13}$$

where k_f is a positive constant value. From Eqs. (7) and (13), we can obtain

$$\dot{s} = \dot{e}_{n+1} + \sum_{k=1}^{n+1} c_k e_k = -k_f F(s,\dot{s}) . \tag{14}$$

The differential equation of control input $u(t)$ is

$$\dot{u} = -\frac{d}{dt}\left[g(e,t) + \Delta f(y) + d(t)\right] - \sum_{k=1}^{n+1} c_k e_k - k_f F(s,\dot{s}) . \tag{15}$$

In the real world, the external disturbance $d(t)$ and the system uncertainty $\Delta f(y)$ are unknown. So, the implemented control input is described as

$$\dot{u} = -\frac{d}{dt} g(e,t) - \sum_{k=1}^{n+1} c_k e_k - k_f F(s,\dot{s}) . \tag{16}$$

After integration of Eq. (16), we can obtain the control input of the slave system

$$u = \int_0^t \left[-\frac{d}{dt} g(e,t) - \sum_{k=1}^{n+1} c_k e_k - k_f F(s,\dot{s}) \right] dt . \tag{17}$$

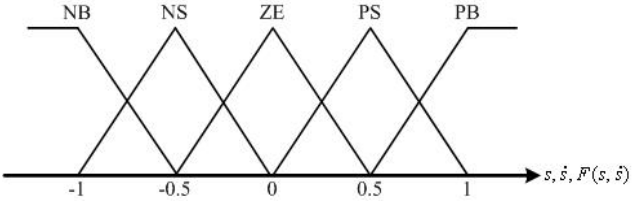


Fig. 2. The membership functions of input variables and output variable

Table 1. Rules-table of fuzzy controller

$F(s,\dot{s})$		s				
		NB	NS	ZE	PS	PB
$\dot{s}$	NB	NB	NB	NB	NS	ZE
	NS	NB	NB	NS	ZE	PS
	ZE	NB	NS	ZE	PS	PB
	PS	NS	ZE	PS	PB	PB
	PB	ZE	PS	PB	PB	PB

Theorem 1. Consider the master-slave system (1) and (2), the two systems are synchronized by the controller $u(t)$ (17) for the slave system. Then the error state trajectory converges to the sliding surface $s(t) = 0$.

Proof. We define a Lyapunov function as

$$V = \tfrac{1}{2} s^2 . \tag{18}$$

Taking the time derivative of Eq. (16), we have

$$\dot{V} = s\dot{s} = s\left\{ \frac{d}{dt}\left[\Delta f(y) + d(t)\right] - k_f F(s, \dot{s}) \right\}. \tag{19}$$

Let $\left| \frac{d}{dt}\left[\Delta f(y) + d(t)\right] \right|$ is bounded, and $\left| \frac{d}{dt}\left[\Delta f(y) + d(t)\right] \right| \le k_f$. There, we have

$$\dot{V} \le -k_f |s| . \tag{20}$$

The reaching condition $s\dot{s} < 0$ is maintained and $\lim_{t\to\infty} s(t) \to 0$. This completes the proof.

4 Numerical Example

In this section, simulation results are presented to demonstrate the effectiveness of the proposed fuzzy sliding-mode controller for chaos synchronization problem. Consider two coupled Duffing-Holmes systems as follows

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -p_1 x_1 - p_2 x_2 - {x_1}^3 + q\cos(\omega t) , \end{aligned} \tag{21}$$

$$\begin{aligned} \dot{y}_1 &= y_2 \\ \dot{y}_2 &= -p_1 y_1 - p_2 y_2 - {y_1}^3 + \Delta f(y) + d(t) + q\cos(\omega t) + u(t) . \end{aligned} \tag{22}$$

The second equation of the slave system (22) is perturbed by an uncertainty term $\Delta f(y)$ and interfered with a disturbance $d(t)$ and the control input $u(t)$ is attached to the slave system. Let us define the synchronization errors between the master system and slaver system as $e_1 = y_1 - x_1$ and $e_2 = y_2 - x_2$. Subtracting (22) from (21), we have the synchronization error dynamics as

$$\begin{aligned} \dot{e}_1 &= e_2 \\ \dot{e}_2 &= -p_1 e_1 - p_2 e_2 - {y_1}^3 + {x_1}^3 + \Delta f(e + x) + d(t) + u(t) . \end{aligned} \tag{23}$$

Then, the standardized state space equations can be described as

$$\begin{aligned} \dot{e}_1 &= e_2 \\ \dot{e}_2 &= e_3 \\ \dot{e}_3 &= -p_1 e_2 - p_2 e_3 - 3{y_1}^2 y_2 + 3{x_1}^2 x_2 + \frac{d}{dt}\left(\Delta f(e + x) + d(t)\right) + \dot{u}(t) \end{aligned} \tag{24}$$

Let the sliding surface be defined as

$$s = e_3 - e_3(0) + \int_0^t (c_3 e_3 + c_2 e_2 + c_1 e_1) dt \,. \tag{25}$$

The eigenvalues corresponding to the sliding surface can be decided by $[c_3 \quad c_2 \quad c_1]$ and these eigenvalues dominate the converging rate of the error dynamics. They can arbitrarily be assigned. Choose the reaching law as in Eq. (17). The control input is determined as

$$u = \int_0^t \{p_1 e_2 + p_2 e_3 + 3{y_1}^2 y_2 - 3{x_1}^2 x_2 - (c_3 e_3 + c_2 e_2 + c_1 e_1) - k_f F(s, \dot{s})\} dt \,, \tag{26}$$

with the initial condition $u(0) = 0$.

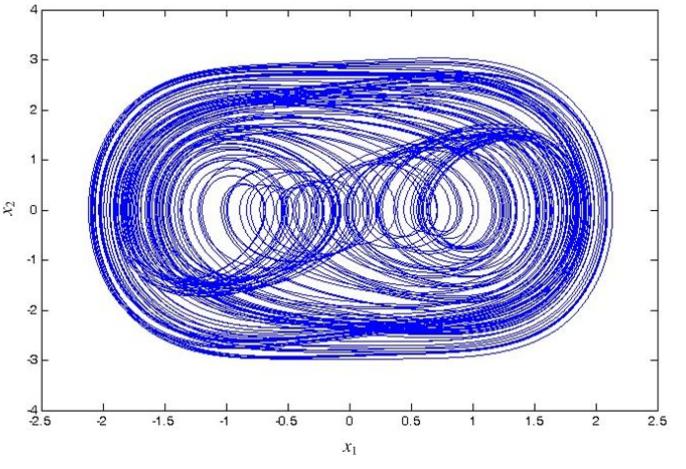


Fig. 3. Hyperchaotic behavior of the Duffing-Holmes oscillator in $x_1 - x_2$ plane

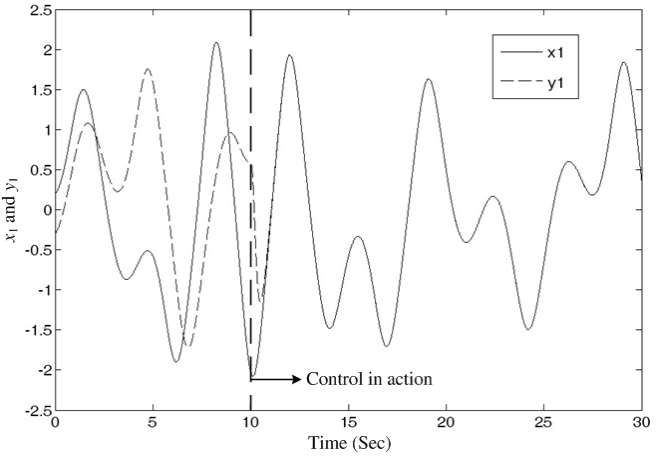


Fig. 4. The time response of x_1 and y_1

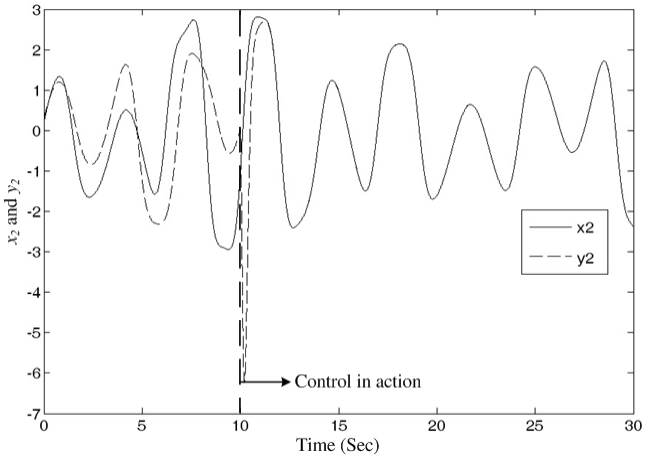


Fig. 5. The time response of x_2 and y_2

For the master-slave synchronization control systems (21) and (22), the parameters are $p_1 = -1.1$, $p_2 = 0.4$, $q = 2.1$ and $\omega = 1.8$, the master system (21) displaces chaotic behavior [9]. The master system (21) will exhibit a hyperchaotic behavior, as shown in Fig. 3, where the attractor is shown on the $x_1 - x_2$ plane by giving initial conditions $x_1(0) = 0.1$ and $x_2(0) = 0.1$. It is supposed that the disturbance $d(t) = 0.2\cos(\pi t)$ and the uncertainty term $\Delta f(y) = -0.05 y_1$ in the slaver system (22). The eigenvalues corresponding to the sliding surface are chosen as $[c_1 \quad c_2 \quad c_3] = [600 \quad 200 \quad 20]$, and coefficient of the sliding-mode controller is chosen as $k_f = 50$. The simulation step size was $0.001\,\text{sec}$.

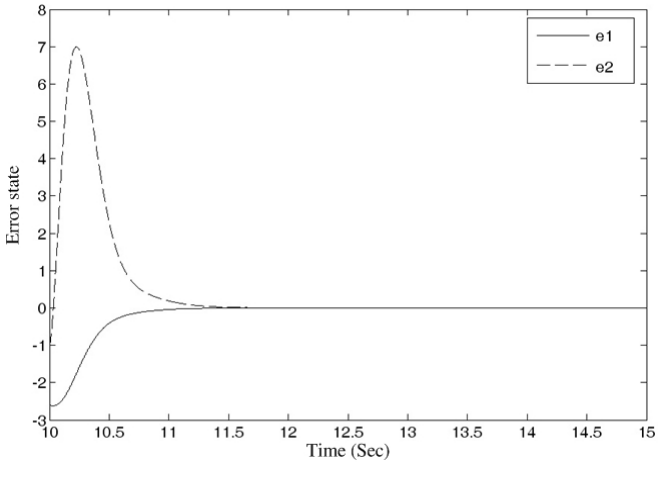


Fig. 6. The error state responses

The simulation results with initial conditions $x_1(0) = 0.2$, $x_2(0) = 0.2$, $y_1(0) = -0.3$, and $y_2(0) = 0.3$ are shown in Fig. 4 and Fig. 5. Those show that the slave and the master systems can reach synchronization when control operation in the slave system at $t = 10$ secs. Fig. 6 shows the error state responses. The control input of slave system was demonstrated in Fig. 7.

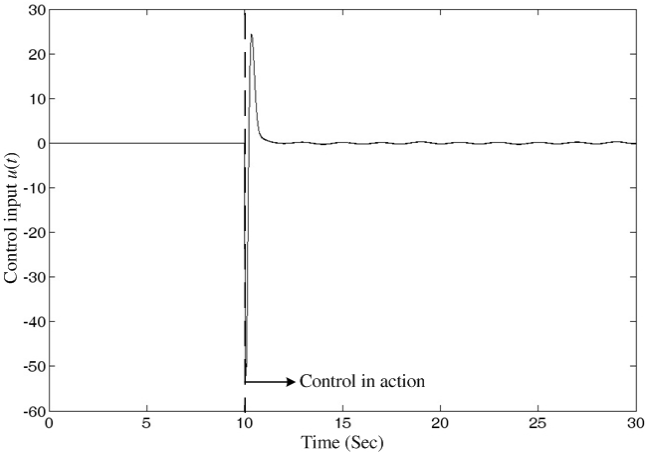


Fig. 7. The control input of slave system

5 Conclusions

In this paper, fuzzy sliding-mode control for chaos synchronization has been proposed. The fuzzy controller based on the sliding-mode and Lyapunov stability theory, which is designed for the regulation of the error state vector to a desired point in the state space. Numerical simulation results demonstrate that the proposed method can be successfully applied to synchronization problems of a Duffing-Holmes system. The derived controllers are robust so that the closed-loop system is stable in the presence of uncertainties and disturbance. The chattering phenomenon of conventional switching type sliding controls does not occur in this study.

Acknowledgement

This work is supported by the National Science Council, Taiwan, Republic of China, under the project NSC96-2218-E-269 -001.

References

1. Carroll, T.L., Pecora, L.M.: Synchronizing chaotic system. IEEE Trans. Circ. Syst. I: Fundam. Theory Appl. 38, 453–456 (1991)
2. Chen, C.L., Lin, W.Y.: Sliding mode control for non-linear system with global invariance. Proc. Inst. Mech. Engrs. 211, 75–82 (1997)

3. Chen, G., Fong, G.: Chaotification via arbitrarily small feedback controls: theory, method, and applications. Int. J. Bifur. Chaos 10, 549–570 (2000)
4. Chen, G., Dong, X.: From chaos to order: methodologies, perspectives and applications. World Scientific, Singapore (1998)
5. Liao, T.-L.: Adaptive synchronization of two Lorenz system. Chaos Solitons & Fractals 9, 1555–1561 (1998)
6. Lian, K.-Y., Liu, P., Chiang, T.-S., Chiu, C.-S.: Adaptive synchronization design for chaotic systems via a scalar driving signal. IEEE Trans. Circuit Syst. I 49, 17–27 (2002)
7. Lu, J., Zhang, S.: Controlling Chen's chaotic attractor using backstepping design based on parameters identification. Phys Lett. A 286, 145–149 (2001)
8. Lü, J., Chen, S.: Chaotic time series analysis and its application. Wuhan University Press, China (2002)
9. Nayfeh, A.H.: Applied nonlinear dynamics. Wiley, New York (1995)
10. Slotine, J.E., Li, W.: Applied nonlinear control. Prentice-Hall, Englewood Cliffs (1991)
11. Suykens, J.A.K., Curran, P.F., Vandewalle, J.: Robust nonlinear synchronization of chaotic Lur'e system. IEEE Trans. Circuit Syst. I 44, 891–904 (1997)
12. Tamasevicius: Reproducible analogue circuit for chaotic synchronization. Electron. Lett. 33, 1105–1106 (1997)
13. Tanaka, K., Ikeda, T., Wang, H.O.: A unified approach to controlling chaos via LMI-based fuzzy control system design. IEEE Trans. Circ. Syst. I 45, 1021–1040 (1998)
14. Yau, H.-T.: Design of Adaptive Sliding Mode Controller for Chaos Synchronization with Uncertaintie. Chaos Solitons & Fractals 22, 341–347 (2004)
15. Yau, H.-T., Kuo, C.-L., Yan, J.-J.: Fuzzy Sliding Mode Control for A Class of Chaos Synchronization with Uncertainties. International Journal of Nonlinear Sciences and Numerical Simulation 7(3), 333–338 (2006)
16. Yin, X., Ren, Y., Shan, X.: Synchronization of discrete spatiotemporal chaos by busing variable structure control. Chaos Solitons & Fractals 14, 1077–1082 (2002)
17. Zadeh, L.A.: Outline of a New Approach to Analysis of Complex System and Decision Process. IEEE Transactions on Systems, Man, and Cybernetics 3(1), 28–44 (1973)
18. Zadeh, L.A.: Fuzzy Logic. IEEE Computer 21(4), 83–93 (1988)

Simulation-Based Procedure for Bottleneck Identification

Chompoonoot Kasemset and Voratas Kachitvichyanukul

Industrial Engineering & Management
School of Engineering and Technology
Asian Institute of Technology
P.O. Box 4, Klong Luang, Pathumthani 12120, Thailand
Tel: (66 2) 0-2524-5675; Fax: (66 2) 0-2524-5697
{st104182,voratas}@ait.ac.th

Abstract. This study presents a simulation-based procedure to identify bottleneck station(s) related to Theory of Constraints (TOC). Bottleneck identification starts by running simulation model of existing system to collect data of the utilization of each machine/process and the time between arrivals and departures of each machine/process. Three factors, the machine/process utilization, the process utilization factor (ρ) and the product bottleneck rate (Rb) are used to identify potential bottleneck candidates. The real bottleneck must come from the process that has high value of two factors; the utilization of machine/process and the utilization factor and low value for the bottleneck rate. The simulation is used again to evaluate the solution by increasing capacity of a bottleneck candidate. If no improvement in throughput is observed, then the station is not a bottleneck and the procedure is run again using the other bottleneck candidate. Numerical examples are given to illustrate the proposed procedure in both case of single bottleneck and multiple bottlenecks.

Keywords: Theory of Constraints (TOC), Simulation, Bottleneck Identification, Single Bottleneck, Multiple Bottlenecks.

1 Introduction

This paper presents a simulation-based procedure to identify bottleneck in the context of the Theory of Constraints (TOC) concept. In this study, the simulation-based tool for TOC implementation proposed in [1] is used (See also [2]). By applying this simulation tool, the data such as machines, jobs, and processes are read into the physical flow model to identify bottleneck in the first step in TOC implementation. Data such as machine utilization and the time between arrivals and departures of parts are collected for used as indicators to the true bottleneck. The performance of the system can be measured by using throughput data obtained from simulation tool and compared with the target demand by using the mean confident interval comparison provided by ARENA. The procedure stops when the target demand is met. Finally, the real bottleneck can be identified.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 46–55, 2007.

The practices of Theory of Constraints (TOC) have been extensively developed and provide a total solution to managing a factory to optimize on time delivery, inventory and operation cost [3]. In fact, every process is a chain of operation, and the strength of the chain is the same as its weakest link, called "Constraint".

The key TOC concept is "The system output rate is limited by the slowest rate of any machine" [4]. In TOC, there are only two types of machines; a bottleneck machine or CCR (Capacity constraint resource) and a non-bottleneck machine or non-CCR. Bottleneck Resource or CCR is the resource with capacity equal or less than the demand. There are fewer resources of this type in the factory. The other type is the non-bottleneck machine or non-CCR. It is the resource with capacity greater than the demand. The main point of this procedure is to balance flow of product by making the output rate at the bottleneck equal to or a little bit more than the market demand. The five steps to implement the TOC are given below:

Step 1: Identify the system constraint.
Step 2: Exploit the system constraint.
Step 3: Subordinate the system constraint.
Step 4: Elevate the system constraint.
Step 5: Repeat the process from the first stefgfhggp if the constraint is violated.

The procedure can be summarized in a diagram as given in Figure 1[5].

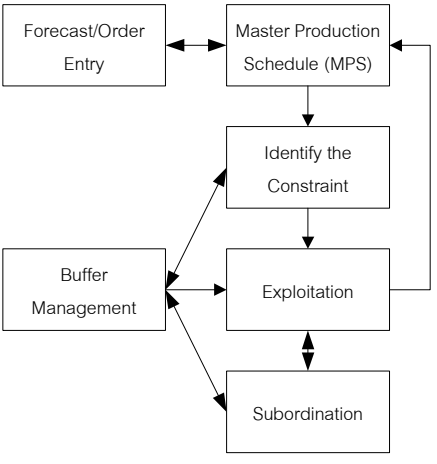


Fig. 1. Steps for implementing TOC, [5]

The most important step is to identify the system constraint, the first step of TOC implementation. The bottleneck identification is the key to TOC implementation because the bottleneck is the one machine that can limit the overall system performance. If the true system bottleneck can be identified, the following steps to improve system performance can definitely work. However, to identify the real system bottleneck is not as simple as it sounds. Theoretically, the selection is made based on the economic scarcity of the resource, i.e., the constraint resource must be operated at close to 100% utilization [3]. Moreover, machines that have the highest

utilization and the longest queue are candidate bottleneck machines. On the other hand, under job shop environment, it is difficult to identify real system bottleneck because different products have their own routes and usually share machines. In the case that the bottleneck machine is located behind another machine that operates at a slow rate, it may be difficult to identify the bottleneck simply from the utilization or the number in queue.

Consider the following flow shop as shown in Figure 2, process A has the lowest processing rate. Too much WIP occur in front of it and if the utilization is measured, it is likely to be high. In this case, A is designated to be the bottleneck.

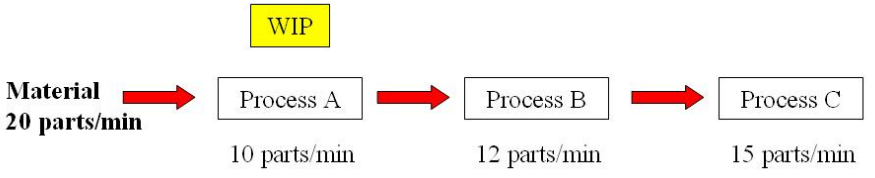


Fig. 2. Bottleneck Identification Case 1

Compare with Figure 3, if only the WIP and the utilization are used, process A will clearly be identified as the bottleneck. This conclusion may be misleading.

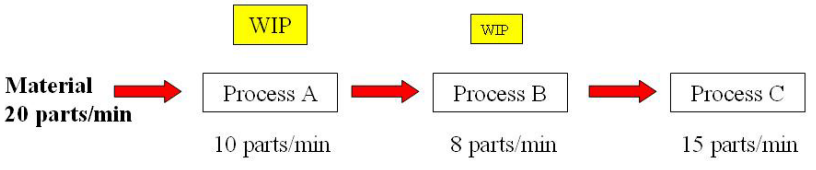


Fig. 3. Bottleneck Identification Case 2

The real bottleneck station should be the station that has high utilization and the lowest throughput rate, so in Figure 3, Process B should be identified as the system constraint. These examples are only for flow shop case. If Job-shop is considered, there are many other relating factors such as product mixed ratio, product flow paths, etc.

Given that it is not so easy to identify the real system bottleneck, this study proposes the simple procedure to handle this step.

2 The Proposed Procedure

In this study, three factors are introduced for bottleneck identification, Utilization of machine/ process, Utilization factor (Based on queuing theory) and Bottleneck Rate (Based on Factory Physics [7]). By applying simulation model, the utilization of each machine/ process, the time between arrivals and departures of parts can be collected and analyzed to identify potential bottleneck machine / process.

1. Utilization of machine/ process: These data can be directly collect from simulation model. Processes or machines having high utilization are selected to be bottleneck candidates.

2. Utilization factor: Based on queuing theory, a single-server system with finite queue of no more than K units, M/M/1/K. The utilization factor (ρ) is calculated from $\rho = \lambda / \mu$.When λ is the average number of units added to the time buffer per unit time and μ is the average service rate unit per unit time[6]. In this study, λ and μ are arrival and departure rate for each process/ machine. If processes or machines have high utilization factor, they are identified to be bottleneck candidates.

3. Bottleneck Rate: From Basic Factory Dynamic [7], Bottleneck Rate (Rb) is the rate (parts per unit time or jobs per unit time) of the workstation having highest long-term utilization. This concept can be used to find additional bottleneck candidates. The output rate from each station comes from:

$$\text{Output Rate} = \frac{\text{Number of part processed}}{\text{Operation time}}$$

In each product route, the output rate of every station is calculated. Rb is the minimum value of output rate among all stations. The station with Rb is identified to be bottleneck candidates.

3 How to Identified Real Bottleneck?

Among all bottleneck candidates, the station that meets three following criteria;

1. High value of the machine/process utilization
2. High value of the process utilization factor (ρ)
3. Low value of the product bottleneck rate (Rb)

It is identified to be real bottleneck system. To confirm the effectiveness of the solution from this approach, the simulation tool from [1] is applied to test the effectiveness of this approach.

In the next section, the illustrative example shows how this approach works in a test problem.

4 Illustrative Example

4.1 Case 1: Single Bottleneck

An example job shop comprises of seven different machines. Five different types of products move through the job shop. Each product-type requires a number of operations to be performed at specified kinds of machines in specified sequence. All operation times and set up times are exponentially distributed. (The detail of problem can be found in [1])

The example job shop will be used to demonstrate how to identify the bottleneck process by using the information obtained from simulation model to effectively and efficiently meet the target demand at 3833 parts per month while operating the job shop using the TOC policy.

The data collected from simulation tool are shown and analyzed in Table 1.

From Table 1, A, B, D and G are identified to be bottleneck candidates because they meet one from three criteria. In this problem, only G meets all three criteria, so G is identified to be the system bottleneck.

This proposed procedure identifies G to be system bottleneck. After G is identified, the simulation should be rerun to test whether the system throughput can be improved by adding one G machine and compare with based situation. The result from simulation shows that adding G machine can improve system performance from based situation (under the 75% confident interval level) (see in Figure 4).

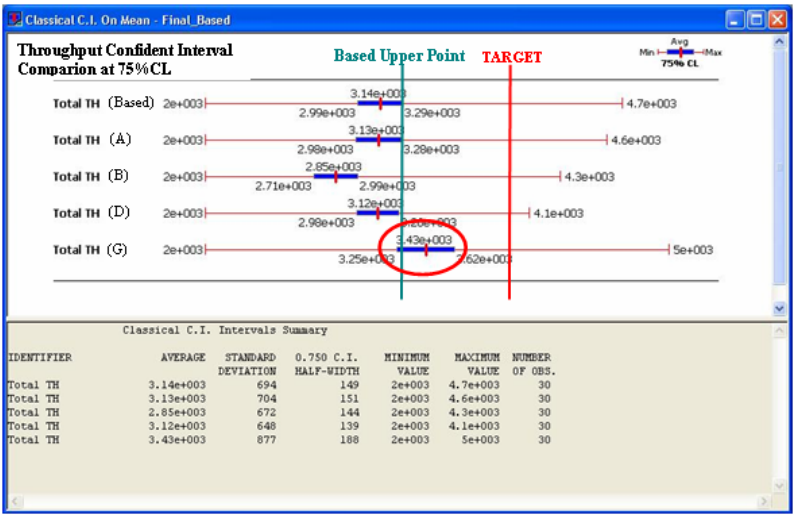


Fig. 4. Mean Interval Comparison for bottleneck identification Case 1

Table 1. Data collected from simulation tool and analyzed data (Case 1)

Machine Name	**A**	**B**	C	**D**	E	F	**G**
% Utilization	**78.26**	61.63	43.72	57.36	22.18	58.76	**75.06**
Time between arrival (min)	572.4	857.5	1181.8	1150	2108.8	942.6	616.3
Time between leaving (min)	634.5	900.4	1185	1187.1	2073.6	982.4	657
λ (part)	0.00175	0.00117	0.00085	0.00087	0.00047	0.00106	0.00162
μ (part)	0.00158	0.00111	0.00084	0.00084	0.00048	0.00102	0.00152
Utilization Factor, ρ	**1.109**	1.05	1.002	1.032	0.983	1.042	**1.066**
R_b (part/min)		**0.123 (P.1, P.2)**		**0.123 (P.3, P.5)**			**0.123 (P.3, P.5) and 0.119 (P.4)**

Table 2. Data collected from simulation tool and analyzed data (Case 2A)

Machine Name	**A**	**B**	C	D	**E**	**F**
% Utilization	**81.5**	67.98	47.93	74.2	**85.43**	59.25
Time between arrival (min)	16.8	16.7	21.3	15.4	13.4	21.5
Time between leaving (min)	18.7	17	21.5	15.8	15.4	21.9
λ (part)	0.0595	0.0599	0.047	0.065	0.0748	0.0466
μ (part)	0.0534	0.0589	0.0566	0.0633	0.065	0.0456
Utilization Factor, ρ	**1.114**	1.017	1.008	1.027	**1.151**	1.02
R_b (part/min)		**0.062 (P.2)**			**0.048 (P.3)**	**0.050 (P.1)**

Note: 1.) Bold letter mean that one of three criteria is met in order to identify bottleneck candidate.

2.) Example for Rb 0.123 (P.1, P.2) means for product 1 and product 2 bottleneck rate is 0.123 at B machine

4.2 Case 2: Multiple Bottleneck

Three different types of jobs move through the production shop. Three job-types are designated as type 1, type 2, and Type 3. Each job-type requires that operation be performed at specified kinds of machines in a specified sequence. All operation times are exponentially distributed. Job arrives at the shop with exponential inter-arrival times.

The goal is to improve the system in order to meet the external demand, target throughput ~ 200 parts per week.

The data collected from simulation tool are shown and analyzed in Table 2.

From Table 2, A, B, E and F are identified to be bottleneck candidates because they meet one from three criteria. In this problem, only E meets all three criteria, so E is identified to be the system bottleneck.

This proposed procedure identifies E to be system bottleneck. Same as in Case 1, the simulation should be run to check whether the system performance can be improved by increasing E when compared with based situation. In this case, adding only E machine cannot improve system throughput (under the 75% confident interval level) (See Figure 5). It means that there may be more than one bottleneck in this system.

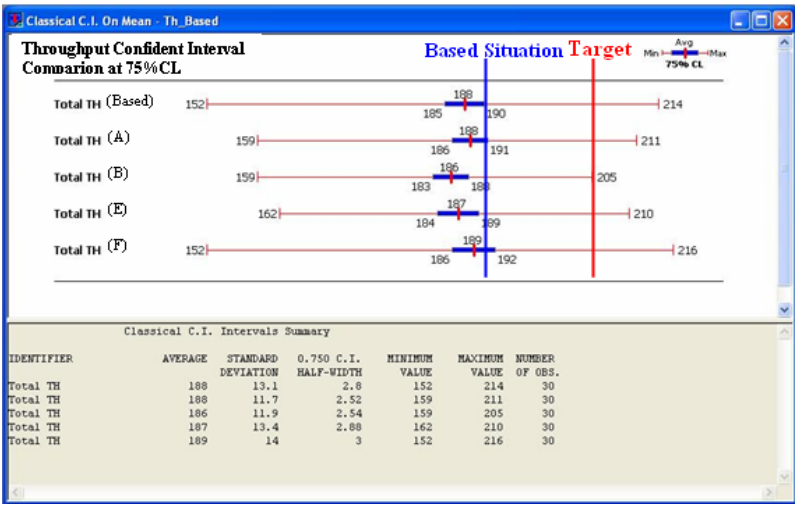


Fig. 5. Mean Interval Comparison for bottleneck identification Case 2

Thus, the proposed procedure should be rework again and applying same simulation model but adding one of machine E and collect data from simulation run.

From Table 3, A, B, E and F are identified to be bottleneck candidates same as before, but in this time only A that meets all three criteria, so A is identified to be the system bottleneck with E.

After the proposed procedure is reconsidered again, A and E are identified to be system bottleneck. The simulation should be rerun to test whether adding A and E machines can improve the system throughput. The result from simulation shows that adding A and E machine can improve system performance from based situation (see in Figure 6).

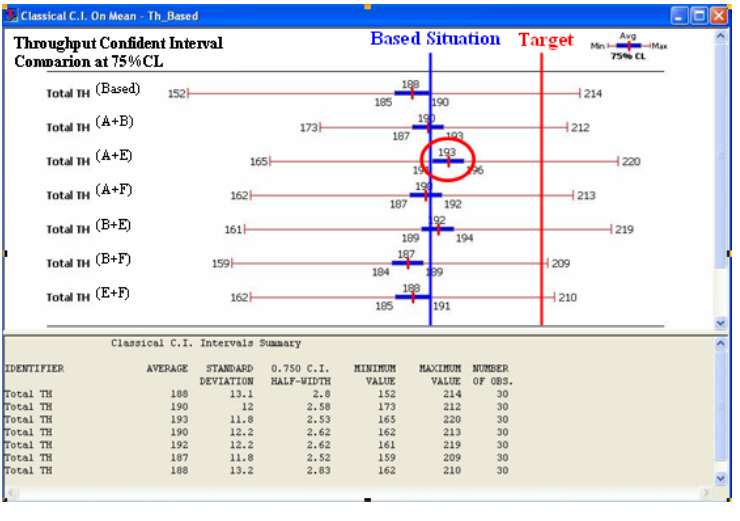


Fig. 6. Mean Interval Comparison for multiple bottleneck identification

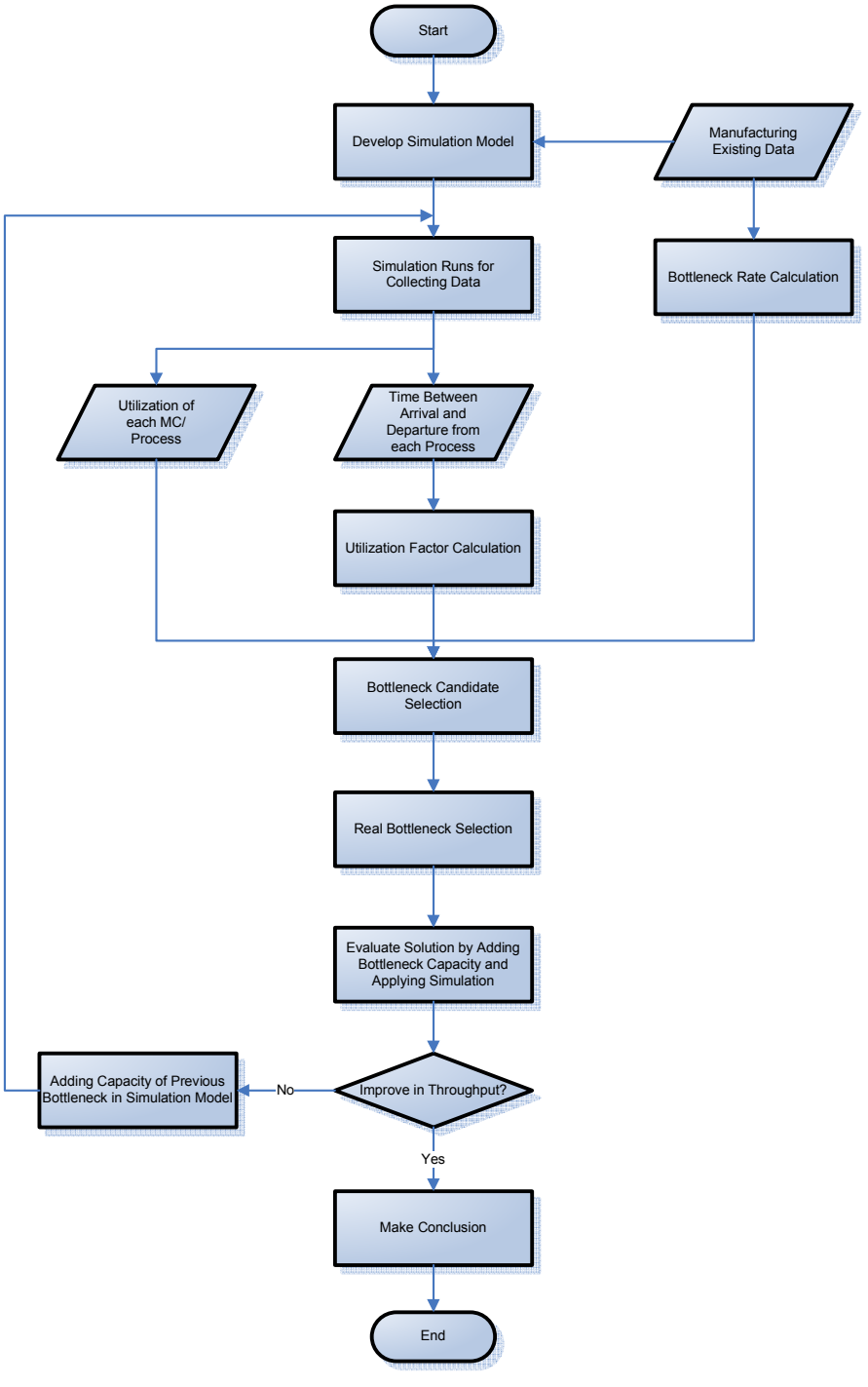


Fig. 7. Proposed Procedure Flow Chart

Table 3. Data collected from simulation tool and analyzed data (Case 2B)

Machine Name	A	B	C	D	E	F
% Utilization	**82.33**	67.05	48.21	76.25	76.23	58.87
Time between arrival (min)	16.6	16.8	21	15.1	13.3	21.3
Time between leaving (min)	18.4	17.3	21.1	15.9	15.1	21.7
λ (part)	0.06	0.0596	0.0477	0.0661	0.0754	0.047
μ (part)	0.0544	0.0577	0.0475	0.0629	0.0661	0.0462
Utilization Factor, ρ	**1.103**	1.032	1.004	1.051	**1.141**	1.019
R_b (part/min)	**0.051 (P.3)**	**0.062 (P.2)**				**0.050 (P.1)**

Note: 1.) Bold letter mean that one of three criteria is met in order to identify bottleneck candidate.

2.) Example for Rb 0.050 (P.1) means for product 1 bottleneck rate is 0.050 at machine F

5 Summary

This paper presents an approach to use simulation as a decision support tool to identify system bottleneck related to the TOC policy. In the proposed procedure, three factors are considered, Utilization of machines, Utilization Factor and Bottleneck Rate, to be the keys to identified system bottleneck. The real bottlenecks are the process or machine that high value of two factors; the utilization of machine/process and the utilization factor and low value for the bottleneck rate. The simulation is applied to verify the effective of the solution. If there is an improvement in system throughput compared with based situation when the bottleneck's capacity is increased, the conclusion of system bottleneck can be made. On the other hand, if result from simulation does not indicate that there is an improvement in system throughput, the procedure should be rerun again by applying simulation model adding capacity of previous identified bottleneck to find another bottleneck station. The proposed procedure conclusion chart is shown in Figure 7.

Reference

1. Kasemset, C.: Simulation-based Tool for Theory of Constraints (TOC) Implementation, Master's Thesis, Asian Institute of Technology, Bangkok, Thailand (2005)
2. Kasemset, C., Kachitvichyanukul, V.: Simulation-based Tool for Implementing Theory of Constraints. In: The 6th APIEMS Conference, Manila, Philippines (December 2005)
3. Jose' Luis Pe'rez, R.: TOC for World Class Global Supply Chain Management. Computers Industrial Engineering 33(1-2), 289–293 (1997)

4. Goldratt, E.M., CoX, J.: The Goal: A Process of Ongoing Improvement, Revised Edition. North River Press (1986)
5. Stein, R.E.: Re-Engineering the Manufacturing System, 2nd edn. Marcel Dekker, New York (2003)
6. Radovislsky, Z.D.: A Quantitative Approach to Estimate the Size of the Time Buffer in the Theory of Constraints. International Journal Production Economics 55, 113–119 (1998)
7. Hopp, W.J., Spearman, M.L.: Factory Physic, 2nd edn. McGraw-Hill, New York (1999)

Self-adaptive Fault-Tolerance of HLA-Based Simulations in the Grid Environment

Jijie Huang[1], Xudong Chai[2], Lin Zhang[1], and Bo Hu Li[1]

[1] Beijing University of Aeronautics and Astronautics, Beijing 100083, China
chxhjjco-p@163.com,
[2] Beijing Simulation Center, Beijing 100854, China
xdchai@263.net

Abstract. The objects of a HLA-based simulation can access model services to update their attributes. However, the grid server may be overloaded and refuse the model service to handle objects accesses. Because these objects have been accessed this model service during last simulation loop and their medium state are stored in this server, this may terminate the simulation. A fault-tolerance mechanism must be introduced into simulations. But the traditional fault-tolerance methods cannot meet the above needs because the transmission latency between a federate and the RTI in grid environment varies from several hundred milliseconds to several seconds. By adding model service URLs to the OMT and expanding the HLA services and model services with some interfaces, this paper proposes a self-adaptive fault-tolerance mechanism of simulations according to the characteristics of federates accessing model services. Benchmark experiments indicate that the expanded HLA/RTI can make simulations self-adaptively run in the grid environment.

Keywords: Grid, HLA/RTI, fault-tolerance, self-adaptive.

1 Introduction

In a large scale distributed simulation, simulation federates running at different locations are liable to termination because of the failure of computing resources or the net congestion at a grid server [1]. The fault-tolerance of distributed simulations has been a hotspot in current M&S community and many fault-tolerance methods have been proposed, such as the federate migration method [2][3] and the cloning federate method [4] etc. To migrate or clone a federate, the RTI (Run Time Infrastructure) needs to transmit a large number of information among federates in the same federation. However, when a HLA-based simulation runs in the grid environment, the transmission latency between a federate and the RTI may exceed one second [5] and simulations are required to decrease the number of data exchange times in order to increase the TAG (TimeAdvanceGrant) rate. Thereby, the above two fault tolerance methods are not perfect when they are used in grid environment.

In fact, HLA-based simulations in grid environment are activities based on the access to grid services. Each federate can call model service interfaces to update

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 56–65, 2006.

object instance attributes and interact with other federates by the LRC (local RTI component) which calls the CRC (core RTI component) services interfaces. So the Quality of Service (QoS) of these grid services should be improved to ensure a HLA-based simulation success.

This paper firstly analyzes the characteristic of federates accessing model services and the requirements of the self-adaptive ability of a model service. Then, by cooperating with the Service-enabled RTI (SE-RTI) and expanding the OMT and interfaces of HLA, the self-adaptive ability of a model service is transformed into the self-adaptive fault-tolerance ability of the simulation.

2 The Characteristics of Federates Accessing Model Services in the Grid Environment

When service-oriented technologies are used in a simulation in the grid environment, object instances of a federation can access model services to update their attributes. From the point of view of a model service, the access of an object instance forms a job that has a prominent feature, namely its running lifetime equals to the lifetime of the corresponding object instance. During this lifetime, the object instance can randomly and concurrently access the model service. The model resource can simultaneously provide services to many object instances, this process can be illustrated as Figure 1. It needs to record the access middle state of an object instance so that it can calculate attributes for this object instance next time. If an object instance fails to access a model service, this may result in occurring an error or termination of the simulation.

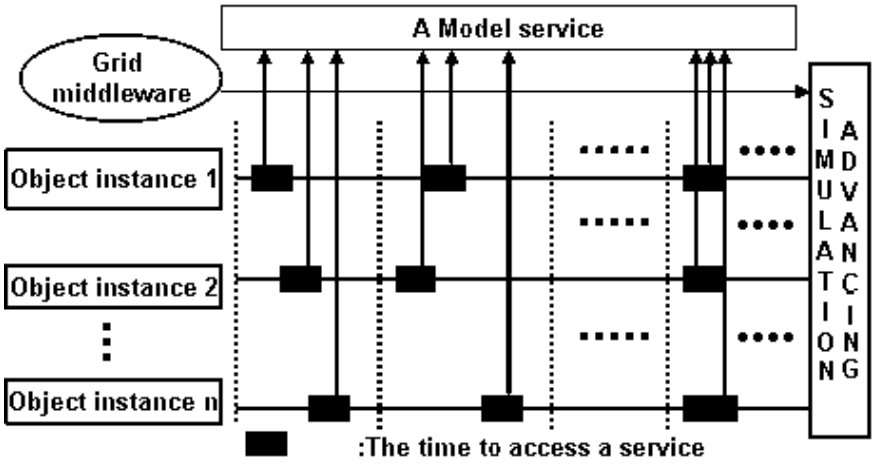


Fig. 1. Object instances randomly access a model service

An object instance of a federation is dispatched to access a model service by a grid middleware. Because a simulation advance period of a federation is

very long corresponding to the time spent in calculating attribute values, the grid middleware may dispatch others object instances to access this model service when it find the grid server is idle. However, these object instances maybe concurrently access the model service in some future time as shown in Figure 1, the grid server will be overloaded and refuse the model service to handle objects accesses. Because these objects have accessed this model service during last simulation loop and their middle state date are stored in this server, this may terminate the simulation.

A grid middleware can adopt many kinds of dispatching tactic, such as the resource Reservation tactic [6][7], Computational Economy tactic [8][9] and the Priority tactic etc. The resource reservation and the computational Economy tactic demand a job to give the resource amount needed by it. A job formed by an object instance in a federation is difficult to tell the grid middleware this value because the federation can not estimate how long time it will run. The transmission latency in grid, which spends almost all the time of a simulation loop, varies from several hundred milliseconds to several seconds and is changeable during simulation period. A federation usually needs to run thousands of loops, this will be a very long period. However, the grid resource actually spends a short time in servicing this federation. Apparently, the resource reservation tactic and the computational Economy tactic waste a large amount of grid resources.

On the other hand, the resource reservation tactic admits a grid middleware to dispatch an access of an object instance only if it can satisfy the resource amount needed by the object instance. Otherwise the middleware can not service the object instance and the simulation can not run immediately.

The Priority tactic firstly calculates the priority grade of a job according to some parameters, such as the user priority or the job type etc. Then the grid middleware queues all jobs According to their priority grade value and handles them in turn. A job formed by an object instance in a federation usually has a very long period. If the grid middleware handles this job firstly, other jobs must wait a long time to be dispatched by the grid middleware. Thereby, the grid middleware usually adds this job to the tail of its jobs list so that the simulation can not run immediately yet.

So the dispatching tactics of a grid middleware are not suitable for the jobs formed by the accesses of object instances. Each model services needs to dispatch the accesses of object instances if it perceives the grid environment not suitable for it to provide services.

3 The Self-adaptive Model Service

In order to ensure the service-oriented simulation in the grid environment to successfully access a model service, a self-adaptive mechanism [10][11] should be introduced to the model service so that it can sense the environment changes and take self-protection measures. The self-adaptive model service is apperceives the environment changes by communicating with a supervisor agent that is a daemon process deployed on a grid server. The supervisor agent inspects the hardware

usage of the server, such as CPU and memory, etc. When the supervisor agent apperceives that the hardware resource utilization ratio reaches a critical point, it finds out the reason from the following aspects: the grid service container; other processes on the grid server. If the overuse of resources is caused by the WS container, the supervisor will find a self-adaptive model service deployed in the container as the main reason and inform it to adjust its service tactics. If it is other processes that result in the overuse, the supervisor will inform all the self-adaptive model services deployed in the WS container to adjust their service tactics.

When a self-adaptive model service is informed by a supervisor agent, it will know that it is concurrently being accessed by too many jobs and running under a critical point of utilization ratio. In order to finish a job access in time on this situation, the self-adaptive model service migrates the access to a coordinate self-adaptive model service (CAMS). It repeats this process until it is informed by a supervisor agent that the grid server has enough spare resources to allow it to directly handle accesses of jobs.

The self-adaptive model service has saved several CAMS URLs in its private data section. Before migrating an access, it needs to select the best CAMS URLs according to the computing ability of each grid server on which a CAMS runs. It also needs to allocates a handle for each job so that a job can access its service interfaces by this handle. Each interface of the self-adaptive service must save job access's internal states after the jobs finish accessing it. This method comes from the WSRF service tactic that separates a grid service into two parts: they are the web service and the resource document. When a self-adaptive model service decides to migrate an access of a job to the best CAMS, it fetches a new handle from the CAMS and records the relationship between the new and the old handle. In order to let the CAMS finish the access of the job, the self-adaptive model service needs to copy the internal state information of the job to the CAMS by calling one service interface of it.

4 The Self-adaptive Ability of Simulations

An object instance of a HLA-based simulation in grid environment corresponds with a job when it continually calls a self-adaptive model service to calculate its attributes. When an access of an object instance has been migrated by the self-adaptive model service, the object instance needs to sense this migration in time and turn its next access to the new self-adaptive model service so that the simulation ca self-adaptively run. For that, the self-adaptive model service should notify the object instance asynchronously.

Both OGSA and WSRF provide a grid service with asynchronous notification ability, they allow clients to subscribe their service data and publish a message to the subscribers when one of the service data is changed. An Object instance subscribes a service data by joining a topic in which there are many other subscribers. If a self-adaptive model service publishes the migrating message of an object instance to a topic, other subscribers can also receive the message. It is

obvious that the network resource is wasted and the communication efficiency is played down. In order to notify a single object instance, therefore, the self-adaptive model service must recur to the federate callback mechanism between the LRC and the CRC.

When an object instance initializes an access to a self-adaptive model service, it should transmit its simulation information to the service, which includes the federate handle and the federation name etc. To describe the information, a tuple M (u,f,h,o) is used (where: u is the URL of the CRC; f is the name of the federation; h is the handle of the federate; o is the handle of the object). The self-adaptive model service returns an access.handle for the object instance, so an one-to-one relationship between the tuple M and the access handle is formed.

During the initializing process of a federate, the LRC embedded in the federate sends all used self-adaptive model service URLs to the CRC. The CRC handles the many-to-many relationship between federates and all used self-adaptive model services in a federation by adopting subscription mechanism. For that, the SE-RTI expands the OMT by adding model service URLs to it and correspondingly adds a model management section to the FED file. The structure of model management section likes the spaces section:

(Models
(Model name 1 (ModelURL CAMS URL 1) ... (ModelURL CAMS URLn))
...
(Model name N (ModelURL CAMS URL1) ... (ModelURL CAMS URL m))).

Thus the CRC allocates a handle for each model used by a federation. After a federate of the federation gets a model handle by the model subscription function, an object instance of the federate can be identified uniquely when it accesses a self-adaptive model service or the CRC component.

4.1 The Self-adaptive Simulation Process

Before beginning a self-adaptive simulation, a supervisor daemon process and many self-adaptive model services have been started in a grid server. All the URLs of self-adaptive model services that are used by the federation can be written to the FED file. For example, a federation uses the movement model and three self-adaptive movement model service URLs have been found by simulation programmers, the model management section of its FED file can be written in the following formats:

(Models (model movement
(modelURL HTTP://192.168.6.168:8080/axis/services/MS1)
(modelURL HTTP://192.168.5.171:8080/axis/services/MS2)
(modelURL HTTP://192.168.4.16:8080/axis/services/MS3)))

During the initialization of the simulation, the SE-RTI modifies three operations of each federate:

1. after creating a federation, the CRC analyses the contents of the model section in the FED file and allocates a handle for each model name. At the same time, the CRC records the given URL of each CAMS and notifies every CAMS of the federation information and other CAMS URLs;

2. after joining a federation, a federate needs to subscribe its used models from the CRC so that the CRC can recode the used model information for each federate;

3. after creating an object instance, a federate firstly designates a self-adaptive model service URL as the access destination for the object instance, then it calls the JoinModelManage interface added for the RTI ambassador in order to transmit the contents of the tuple M to the self-adaptive model service and get an access handle from it.

The simulation advances as if it is running under local net condition. Once a self-adaptive model service finds that it is the result of an object instance access from this federation that the system resource utilization ratio has reached a critical point, it migrates the object instance access to new CAMS. The simulation can also sense the migration and modify the object instance access to the new CAMS when the object instance calculates its attributes next time. The self-adaptive process of the simulation is shown in Figure 2.

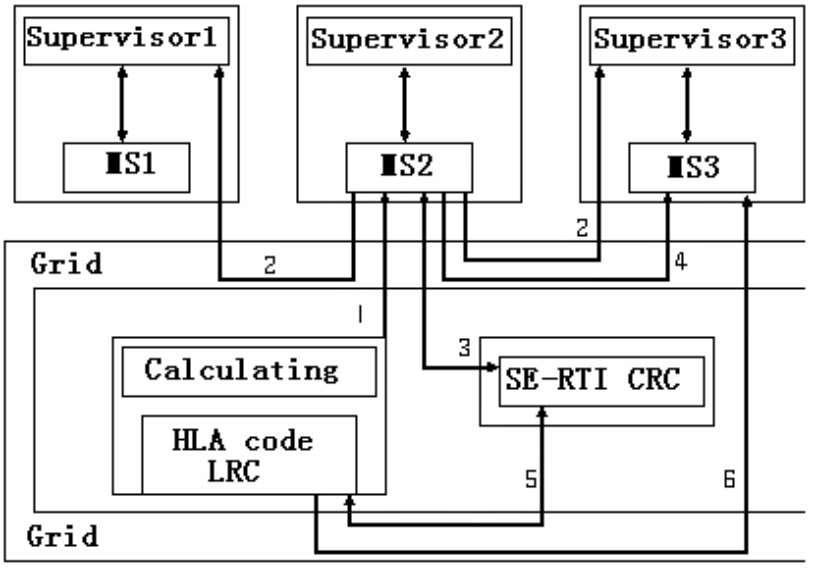


Fig. 2. The self-adaptive process of a simulation

1. The self-adaptive model serviceMS2senses that the system resource utilization ratio has reached a critical point by a supervisor agent;

2. MS2 calculates the evaluation value of environment by accessing other two supervisors and select MS3 as the better CAMS before migrating an access of one object instance;

3. MS2 finds the CRC According to the tuple M and informs the CRC of the object instance handle and MS3 URL;

4. MS2 gets a new access handle from MS3 and copies the internal state information of the object instance to MS3. MS3 executes the access from MS2

and returns the result to MS2 that sends the result to the object instance. This process continues until the object instance also senses the migration of its access.

5. The CRC informs the corresponding federate that needs to migrate the access of the object instance to MS3 by the callback link between the CRC and the LRC embedded to the federate;

6. The object instance senses that MS2 has migrated its access and turns to access MS3 from now on.

At last, the CRC informs MS1 and MS2 and MS3 that they need to clear the internal state information about the federation before destroying the federation execution.

5 The Added Interfaces for a Self-adaptive Simulation

In order to realize the communication between the SE-RTI and the self-adaptive model service, some interfaces need to be added for them. The added interfaces of HLA services are listed bellow:

1. SubscribeModelClass: It is added to the RTI ambassador in order that a federate can inform the CRC of its used model and get a handle from the CRC.

2. JoinModelService: It is added to the RTI ambassador in order that an object instance of a federate can transmit the contents of its tuple M to a self-adaptive model service before updating its attributes.

3. NotifyModelChange: It is added to the federate ambassador. The CRC uses it to inform a federate that it's an object instance access that is migrated by a self-adaptive model service to a new CAMS and the object instance needs to access the CAMS when this instance updates its attributes.

4. RequestFederateMigrate: It is added to the CRC in order that a self-adaptive model service can inform the CRC of an object instance access migration.

The added interfaces of an model service are listed bellow:

1. FederationRegiste: It is called by the CRC when the CRC finishes creating a federation, which makes the self-adaptive model service be able to create a data section to recode the internal stat of object instances in this federation.

2. FederationLogout: It is called by the CRC before the CRC destroying a federation, which makes the self-adaptive model service be able to clear internal information about this federation..

3. AddModelURL: It is called by the CRC when the CRC finishes creating a federation, which makes the self-adaptive model service be able to record the URLs of CAMS in FED file;

4. FederateRequestJoin: It is called by the LRC in order that the LRC can transmit the content of a federate tuple M to it and can get an access handle from it;

5. ModelNeedMove: It is called by a supervisor agent to inform that it should migrate a job access;

6. MoveToNewModel: It is called by the old self-adaptive model service to copy its internal state of an object instance to the new one when the old one is ready to be migrated.

The called procedure of above interfaces is shown in Figure 3.

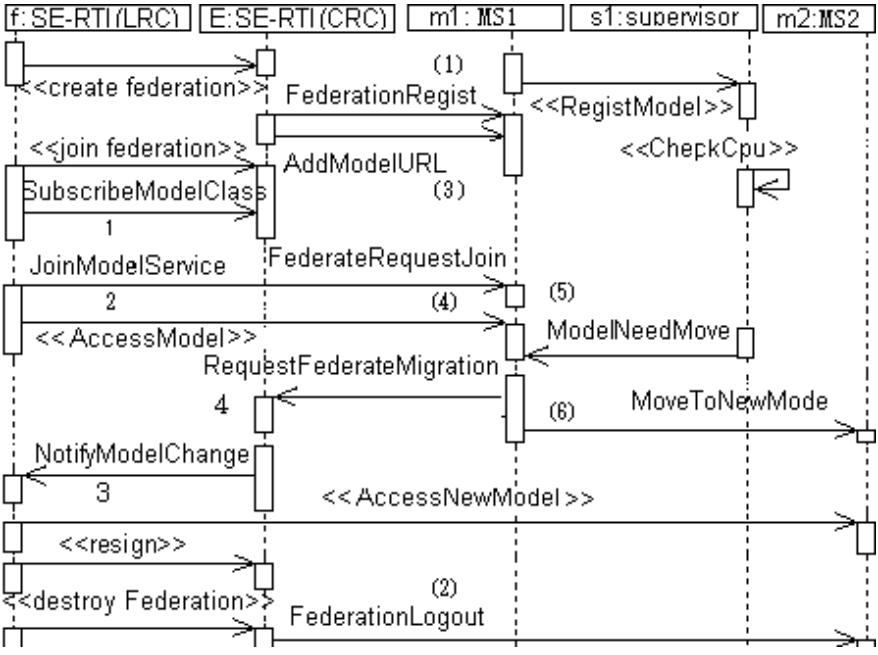


Fig. 3. A simulation calls the added interfaces

6 Experimental Results of the Self-adaptive Simulation

Federates in experiments have run in one computer, which is located in Tsinghua University (TsingHua). The CRC grid service of SE-RTI is located in Beijing University of Aeronautics and Astronautics (BUAA). This is shown in the above Figure 2. There are six federates in the federation. Each federate has a tank object instance. Firstly the movement model is designed as a normal service that has no self-adaptive abilities. The Six tanks calculate their attributes by accessing the MS2 movement service. When the simulation advances to the 50th loop, the concurrently access situation is happened because the movement service needs to calculate one million times of multiplications, which is a added condition in order to inspire concurrently access. This situation is shown as Figure 4A. It is obvious that the simulation is paused.

Then, the movement model is designed as a self-adaptive model service. when the simulation advances to the 50th loop, the access of the first tank object instance makes the supervisor agent find that the machine is in a critical state. This is shown in Figure 4B. The last line of Figure 4B illustrates that MS2 has been informed by the supervisor agent that the number one interface (get-newposition) of it needs to be migrated when it is accessed next time. When

Fig. 4. A:The concurrently access to a non-self-adaptive model service; B:The concurrent access to a self-adaptive model service; C:The CRC transmits the migration notification to an object instance

the second tank object instance begins to access the getnewposition interface, the MS2 decides to migrate the access to MS3, which is shown in Figure 4C. Apparently, the Qos of MS2 is improved.

The last line of Figure 4C illustrates that the CRC has sent a migrating callback information to the receiver service that is running on the same local network with the second federate. After finishing the fifth loop of the simulation, the second tank object instance senses its access has been migrated to MS3 and begins to access MS3. The same process is done by the third and the fourth tank object instance. On the other hand, the fifth and the sixth tank object instance begin to access MS1. So the simulation can adaptively run and has a fault-tolerance mechanism.

The transmission latency between a TsingHua federate and the CRC grid server is 0.2677 second. The fault-tolerance methods based on redundancy and replication technology increase the communication times and result in remarkable decrease of simulation TAG rate.

7 Conclusions

Object instances can randomly and concurrently access a model service. This can result in that the grid server will be overloaded and refuse the model service to handle objects accesses. So a fault-tolerance mechanism must be introduced into HLA-based simulations in grid environment. improving the QoS of these grid services is a better way than traditional fault-tolerance methods. In order to transform the self-adaptive ability of a model service into the self-adaptive fault-tolerance ability of the simulation, the OMT needs to be expanded by adding

model service URL to it. Some interfaces needs to be added to HLA services and model services. The RTI of HLA also needs to be rebuilt to a Service-enabled RTI.

References

1. Möller, B., Löfstrand, B., Karlsson, M.: Developing Fault Tolerant Federations Using HLA Evolved[C]. In: Proceeding of 2005 Spring Simulation Interoperability Workshop, paper no. 05S-SIW-048, San Diego, California, USA (April 2005)
2. Zajac, K., Bubak, M., Malawski, M., et al.: Towards a Grid Management System for HLA-based Interactive Simulations[A]. In: Proceedings of the Seventh IEEE International Symposium on Distributed Simulation and Real-Time Applications (DS-RT 2003)[C]. Delft (2003)
3. Yuan, Z., Cai, W., Low, W.Y.H., Turner, S.J.: Federate Migration in HLA-Based Simulation[C]. In: Proceedings of the 2004 International Conference on Computational Science, Krakw, Poland, pp. 855–864 (2004)
4. Chen, D., Turner, S.J., Gan, B.P., Cai, W., Wei, J.: Management of Simulation Cloning for HLA-based Distributed Simulations[C]. In: European Simulation Interoperability Workshop 2004 (Edinburgh, UK), 04E-SIW-010 (2004)
5. Xie, Y., Teo, Y.M., Cai, W., Turner, S.J.: Service. Service Provisioning for HLA-based Distributed Simulation on the Grid[C]. Provisioning for HLA-based Distributed Simulation on the Grid. In: Proceedings of the 19th ACM/IEEE/SCS Workshop on Principles of Advanced and Distributed Simulation, pp. 282–291. IEEE Computer Society Press, Monterey, USA (2005)
6. Vazhkudai, S.: Enabling the co-allocation of grid data transfers[C]. In: Proceedings of the 4th International Workshop on Grid Computing, Phoenix, AZ, USA, pp. 44–51. IEEE Computer Society Press, Los Alamitos (2003)
7. In, J.-U., Avery, P., Cavanaugh, R., et al.: Policy based scheduling for simple quality of service in grid computing[C]. In: Proceedings of the 18th International Parallel and Distributed Processing Symposium, Santa Fe, USA, pp. 315–324. IEEE Computer Society Press, Los Alamitos (2004)
8. Wolski, R., Plank, J., Brevik, J., et al.: Analyzing market-based resource allocation strategies for the computational grid[J]. International Journal of High-performance Computing Applications 15(3), 258–281 (2001)
9. Zhang, H.-J., Li, Q.-H., Ruan, Y.-L.: Resource co-allocation via agent-based coalition formation in computational grids[C]. In: Proceedings of the 2003 International Conference on Machine Learning and Cybernetics, Xi'an, China, pp. 1936–1940. IEEE Computer Society Press, Los Alamitos (2003)
10. Kephart, J., Chess, D.: The Vision of Autonomic Computing[R].IEEE Computer Society pp. 41–59 (2003)
11. Jann, J., Browning, L.A., Burugula, R.S.: Dynamic Reconfiguration: Basic Building Blocks for Autonomic Computing on IBM pSeries Servers[J]. IBM Systems Journal 42(1), 29–37 (2003)

Constant-Time Record Management in a Java Embedded Small Device

Haklin Kimm

Computer Science Department
East Stroudsburg University of Pennsylvania
East Stroudsburg, PA 18301
haklkimm@esu.edu

Abstract. In this paper, we present an approach that can provide faster and constant-time record operations with a smaller memory usage while consumes less battery power, using the Record Management System (RMS) of Java Mobile Information Device Profile (MIDP) in small embedded devices. The comparable program segments are presented and analyzed in order to show the efficient use of Java MIDP RMS with or without Record-Enumeration interface. The proposed approach in this paper supports the faster access of record traversal that can retrieve and update any records in the MIDP RMS file in constant time, while the previous approach takes O (log n) time.

Keywords: Java MIDP, Record Management File System, Embedded Devices.

1 Introduction

Java 2 Micro Edition is becoming the becoming the important platform when running applications on resource-limited embedded devices such as handheld devices, cell phones, TV set-top boxes, Personal Digitals Assistants (PDAs) and in-vehicle systems. Since many of these systems are memory constrained, it is extremely important to keep the memory footprint of Java applications under control [1-4]. Java 2 Micro Edition (J2ME) is a Java toolkit to provide an easy development of applications running on small embedded devices and gadgets. These small devices are not considered as general-purpose computing units and may be part of an appliance, and they tend to have personalized nature as seen in wireless mobile devices such as cell phones and PDAs [11, 15, 18] whose uses are seen in health care management as well [19].

J2ME is equipped with a modular and scalable architecture to support the various types of embedded devices. J2ME can provide the services for Connected Device Configuration (CDC) and Connection Limited Device Configuration (CLDC), which depend on the type of Java Virtual Machines (JVM): C Virtual Machine (CVM) or K Virtual Machine (KVM). The Configuration of CDC or CLDC is a specification to a particular device category that shows device constraints: memory, power supply and networking. J2ME architecture is briefly explained in the followings.

K Virtual Machine (KVM) is designed for resource-constrained embedded computing devices with as little as a few kilobytes total memory, including actual

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 66–74, 2007.

virtual machine and java class libraries themselves [9, 13]. CDC provides basic services for the devices that equip the followings [10, 12-14]: 32-bit processor devices, 75 Mhz or better, devices with less than 2 Mbytes, fixed or battery powered, high bandwidth and high fidelity, full Java class library, and C-virtual Machine (CVM).

In the meantime, CLDC is being used for the smaller devices that requires less than CDC. CLDC provides basic services for the devices that require the followings: 16- or 32-bit processor devices, 16 Mhz or better, devices with less than 512 Kbytes, mainly battery powered, often used in Wireless networking, limited Java class library, and K-virtual Machine (KVM).

J2ME configurations do not provide enough functionality for most application development. Profile is a specification as configuration is; however, profile defines the capabilities for a kind of devices such as PDAs or cell phones. Mobile Information Device Profile (MIDP) is a profile that provides the API for user interfaces, persistent storage, and networking for wireless devices such as cell phones and pagers

The rest of this paper is organized into four sections. In section 2, the persistent storage of MIDP RMS is described with its relevant Java classes and interfaces including Record Enumeration. In section 3, previous approach using MIDP RMS in O (log n) time is discussed. In section 4, the constant time operations of MIDP RMS are presented. In section 5, we conclude with some remarks.

2 MIDP Record Management System (RMS)

Memory in handheld devices is being used for implementing applications and storing persistent data to retrieve them later. The package javax.microedition.rms is used for handling data storage where small memory is available in the cell phones and pagers. The Java RMS package for wireless devices consists of two classes, three interfaces and five exceptions [10, 11].

Record store names are case sensitive and can be up to 32 characters in Java RMS. Users can create and/or open a record store file in the following:

```
rs=RecordStore.openRecordStore("data", true);
rs=RecordStore.openRecordStore("data", false);
```

The first line with *true* creates the file *data* in MIDlets, and the second line with *false* opens the existing file *data.* MIDlets are referred as applications that only use the APIs defined by the MIDP and CLDC. MIDlets that share codes or data are packaged together as MIDlet suites [11]. The openRecordStore throws RecordStoreNotFoundException and/or RecordStoreException, if a file is not accessible.

A record store file consists of a directory header and data blocks that are structured with a linked list. The header keeps a link to the first data block and another link to the first block of free spaces. The header also contains the information about a file: the number of records, version number, last modified time in millisecond, and next recordID. They work like a typical file control block in file systems.

The recordID is used as a primary key (integer) to read or write a record from and/or to a RMS file. The first record of the RMS file is with the value of recordID (= 1), and any additional records are assigned with a new recordID that is one larger than

the current recordID. In RMS each record should be read as a whole in bytes after finding the corresponding record with a recordID:

byte[] line = rs.getRecord(recordID);

A new record can be added into the RMS file *data* with a recordID as shown in the following statements:

int recordID = rs.getNextRecordID();
rs.addRecord(line, 0, line.length);

The addRecord() throws RecordStoreFullException if there is not enough storage available. In the same way as addRecord(), a record can be removed from a RMS file by using a recordID:

rs.deleteRecord(recordID);

The deleted recordIDs from deleteRecord (recordID) can not be used again. However, the data blocks of deleted records are marked free and are not completely removed yet from the RMS file. We can say that a record deletion in RMS does not mean to allow the size of the RMS data file smaller immediately. The deleted record is just marked free for future use, and the deleted recordID cannot be used with setRecord() yet. The efficient use of memory cannot be achieved in this way since users cannot access the deleted area to add a new record. As shown in the following statement, we tried to use the deleted recordID again but happened to see a runtime error since we were trying to use the just-deleted recordID again.

rs.setRecord(just_deleleted_ID, data, 0, data.length);

In order to facilitate a faster record traversal in a RMS file, the interface RecordEnumeration is being used with a RecordStore object. With the use of RecordEnumeration, the recordIDs (not the records) of a RMS file are logically maintained as the sequence of list elements.

public RecordEnumeration enumerarateRecords (RecordFilter filter, RecordComparator comparator, Boolean keepUpdated);

Both arguments of RecordFilter and RecordComparator from the above example are used to filter and sort the record in RecordEnumeration objects. If the first two arguments of enumerateRecords are null, then the records in the RMS file are stored without any specific order. To store or retrieve the records in a specific order it is imperative to define a filter and/or a comparator. The last argument, Boolean keepUpdated, determines whether the enumeration tracks the changes of the RMS record store or not. If the keepUpdated is true, the enumeration that updates itself as records is added to or deleted from a record store file. The enumerateRecords throws RecordStoreNotOpenException, which is a subclass of RecordStoreException, if the file is not open.

3 Previous Approach

A basic problem of using persistent RMS storage in small embedded devices is adding or deleting variable sized records into or from a list without the prior information. One

simple approach for resolving this problem is the use of a linked list, which can grow easily to any size and supports sequential traversal but cannot support random access [9]. Since the RMS file is structured in a singly linked list, the time to search a record in the RMS file requires O (n) time. In order to improve the retrieval time of a persistent RMS file, the use of RecordEnumeration interface is recommended in MIDP RMS. The RecordEnumeration restructures the RMS file to a doubly linked list from a singly linked list, by using recordIDs only. As we know, the doubly linked list provides the easier forward and backward movements in a list than the singly list does since the doubly linked list connects records with the two pointers: forward and backward. Each node of the RecordEnumeration doubly linked list consists of an integer field and two pointer fields, which require 6 to 12 bytes depending on the size of the pointer field [7, 8]. The RMS file with the RecordEnumeration interface would require a less average lookup time than the RMS file without the RecordEnumeration would; however, in the worst case, it still needs O (n) time. We know that the RMS RecordEnumeration structure is similar to a file system with a single-level structure where all files are contained in the same directory.

Our previous method retrieves a record in the RMS file in O (log n) time. The previous method uses the same or smaller amount of memory than the Record-Enumeration interface requires. A separate index file consists of a field (column) to store integer recordIDs and is used to search the corresponding record.. Then the records in the RMS file do not need to be equipped with a RecordEnumeration interface. In the previous method, every read or write operation begins with the index file and implements the record operation on the RMS file after. The traversal operation of the index file can be done with the binary search method, which implements in O (log n) time. The recordID found during the binary search method is used to retrieve the corresponding record from the RMS file in O (1) time. Thus, the overall time to retrieve a record improves from O (n) time to O (log n) time.

We show an example to use recordIDs with a byte array. When using a small mobile cell phone or PDA, users usually do not store too many number of data elements. For the brevity of our example, we assume that the number of data elements is not larger than 256. Here we use the byte array that can store recordIDs. Each byte can have at most 256 different integer values: 0-255. In the following program segments, user data are typed in from the RIM Blackberry screen and saved in the RMS file, say dataK.

```
byte[] key = new byte[256];
RecordStore rs = RecordStore.openRecordStore("dataK", true);
while (record_is_in) {
   String typeIn;
      typein = mainScreen.getString();
      byte[] bname = typeIn.getBytes();
      int j = rs.getNextRecordID();
      rs.addRecord(bname, 0, bname.length);
```

```
        key[i] = (byte) j;
        i++
    }
```

Users can add their own information from the screen, and the typed-in information can be fetched with mainScreen.getString() method. However, the event handling method to save the user input into a RMS file is defined and implemented with the CommandListner interface. The string value received from the user requires to be changed into bytes in order to be stored into the RMS file. With the implementation of typeIn.getBytes() and addRecord() methods, the typed-in data are added into the RMS file as a new record and the corresponding recordID together. Lastly, we need to store the integer recordID into the byte array key[]. We need to convert the integer value into the byte value because the recordID is an integer value that uses 4 bytes in MIDP RMS. The following example shows the safe conversion from integer to byte, in which the leftmost 3 bytes of the integer values (129, 127) are truncated.

```
key [i] = (byte) j;
```

Integer	***Byte***
00000000 00000000 00000000 10000001 (129)	*10000001*
00000000 00000000 00000000 01111111 (127)	*01111111*

There is a minor problem in converting from an integer to a byte for storing and retrieving index values in the byte array. As shown in [8], a byte is defined as a 2's complement number so that the range of integer values is -128 through 127, not 0 through 255. As shown on the above, 129 will be -127 when retrieving it from the byte array, key []. This requires extra efforts when retrieving the records from the RMS file.

Byte	***Integer***
10000001	*11111111 11111111 11111111 10000001*
01111111	*00000000 00000000 00000000 01111111*

As we see from the above conversion, the positive integer 129 turns out to be a negative integer number when converting back from a byte to an integer. To avoid this problem, we need to have the following statement to replace all 1's with all 0's from the leftmost three bytes in the integer value:

```
id = (int) (key[i] & 255);
```

```
   11111111 11111111 11111111 10000001
& 00000000 00000000 00000000 11111111 (255)
  00000000 00000000 00000000 10000001 (129)
```

After applying the above approach using an extra byte array instead of using RecordEnumeration interface, there needs to be different approaches for modifying and/or deleting a record in the RMS file. When deleting a record, we like to mark the corresponding index in the byte array key[i], where i is the corresponding index. We use the following statement to complete the deletion:

```
key [i] = 0;
```

The value of key[i] is marked for deletion by replacing its content with zero, which does not create any conflict since no recordIDs in the MIDP RMS can have the value of zero. The marked recordID for deletion can be reused again anytime to store a new record. Whenever trying to use the deleted recordID and storage again in Java MIDP RMS with or without RecordEnumeration interface, you will run into a runtime error message. Even though the deleted space is available in the current RMS file, you cannot reuse that space because of the deletion of the recordID.

4 Constant Time Operations

Here we present an approach that can achieve O(1) time to operate records with the RMS files. As mentioned above, we are still using an indexed file approach, however, with the modified method. A byte array is used to represent a recordID from 0 to 255, but 0 and 255 will be used only to provide addition and deletion operations. The key 0 field (byte) will be used as storage to keep track of changes such as add or delete operation. Whenever to add or delete a record, the topmost field, say key[0], is used to point to a record that has been marked for deletion for the first time. The record marked for deletion will keep the value (255) if there are no more records for deletion.

We present an example how to operate records in constant time. Since a small index file is used to store recordIDs, every record read/write operation begins with the index search and followed by record operations.

4.1 Delete a Reczord with an Index

The following steps are implemented for deleting records in constant time. The following step does not create any conflict since no recordIDs in the MIDP RMS do use the value of -1 (255);

Table 1. Deletion of records

0	~~00~~->04
1	01
2	~~02~~->ff
3	03
4	~~04~~->~~ff~~->05
5	~~05~~->~~ff~~->02
6	06
…	…

Step 1: Mark the corresponding index with a record to delete in the byte array key[del_idx] with -1:

```
key[del_idx] = -1;
```

Step 2:
if (key[0]) == 0
{
key[0]= del_idx;
key[key[0]] = del_idx;
key[del_idx] = -1;
}

The marked records are permanently deleted later when exiting the program. Assume that records are deleted such as 4->5->2; these changes are shown in Table 1.

- delete record 4
  - key[0] = 4; key[4] = -1(ff);
- delete record 5
  - key[4] = 5; key[5] = -1;
- delete record 2;
  - key[5] = 2; key [2] = -1;

Table 2. Addition of records

0	~~04~~->~~05~~->~~02~~->00
1	01
2	~~02~~->ff
3	03
4	~~05~~->04
5	~~02~~->05
6	06
…	…

4.2 Reuse of Marked Records

The following steps are implemented for adding records in constant time.

Step 1:
Use the key[0] to store the address of the first deleted record; marked_addr represents the recordID to be used for resetting;
marked_addr = key[0];

Step 2:
key[0] is restored with the address of the next deleted record;
key[0] = key[key[0]], where key[0] <> -1;

Assume that three records are to be added; these changes are also shown in Table 2.

•first addition
  •marked_addr= key[0];
  •key[0] = key[key[4]]; //key[0] = 5;
•second addition
  •marked_addr= key[0];
  •key[0] = key[key[5]]; //key[0] = 2;
•third addition;
  •marked_addr= key[0];
  •if (key[key[0]] == ff) key[0] = 0;

5 Conclusions

There will be more personal information devices applying the Java technology as the market for mobile devices and phones is continuing to increase at a rapid rate [3, 6]. In this work, we present that constant-time operations are available to Java Information Device Profile (MIDP) Record Management System (RMS) by applying indexed file approach.

Since the RMS RecordEnumeration creates a doubly linked list based on recordIDs, each node in the list consists of a recordID as on integer and two pointers as short or integer which are depending on the memory address to access. If we need to create a doubly linked list to store 256 recordIDs, then the list needs 2048 = 256*8 (1 integer for a recordID and 2 short integers for pointers) or 3072 = 256*12 (3 integers) bytes depending on the pointer type. In the previous approach using a byte array requires 256 bytes at most to store 256 recordIDs, with providing extra steps to convert integer to byte when to store a recordID into the byte array and vice versa. Therefore, the use of a byte array to store recordIDs in bytes spends 8 or 12 times less memory than the use of the interface RecordEnumeration does.

The previous approach [16] using a byte array(s) to store recordIDs provides random access and binary search that takes O (log n) time in its worst case. However, this time we can achieve those operations in constant time by applying a modified scheme for deletion, and especially addition by reusing the space marked for deletion efficiently.

References

1. Clausen, L.R., Schultz, U.P, Consel, C., Muller, G.: Java Bytecode Compression for Low-End Embedded Systems. ACM Transactions on Programming Languages and Systems 22(3), 471–489 (2000)
2. Chen, G., Kandemir, M., Vijaykrishnan, N., Irwin, M., Mathiske, M.: Heap Compression for Memory-Constrained Java Environments. In: ACM OOPSLA Conference of 2003, pp. 282–301 (October 2003)
3. Chen, G., Kandemir, M., Vijaykrishnan, N., Irwin, M., Wolf, W.: Energy Saving Through Compression in Embedded Java Environments. In: ACM CODES Conference of 2003, pp. 163–168 (May 2003)
4. Rollins, S., Almeroth, K., Milojicic, D., Nagaraja, K.: Power-Aware Data Management for Small Devices. In: ACM WoWMoM Conference 2002, pp. 80–87 (September 2002)

5. Riggs, R., Taivalsaari, A., VandenBrink, M.: Programming Wireless Devices with the Java 2 Platform. Addision Wesley, Reading (2001)
6. Java2 Platform micro edition (J2ME) technology for creating mobile device (white paper), http://java.sun.com/products.cldc/wp/KVMwp.pdf
7. Aho, A., Ullman, J.: Foundations of Computer Science. W.H. Freeman, New York (1992)
8. Arnold, K., Gosling, J.: The Java Programming Language: 2nd edn. Sun Microsystems (1998)
9. Brodnik, A., Carlsson, S., Demaine, D., Munro, J., Sedgewick, R.: Resizable Arrays in Optimal Time and Spaces. In: Proceedings of the 6 th International Workshop on AlgorIthms and Data Structures (1999)
10. Feng, Y., Zhu, J.: Wireless Java Programming with J2M2, Sams Publishing (2001)
11. Giguere, E.: Java 2 Micro Edition. John Wiley & Sons, Chichester (2000)
12. General BlackBerry Programming Techniques for RIM Developers, http://developer.rim.net/ knowledge/ faq/index.shtml
13. Giguere, E.: Record Management System Basics, http://wireless.java.com/
14. Mahmoud, Q.: MIDP: Database Programming using RMS: a Persistent Storage for MIDlets, http://wireless.java.com/
15. Hemphill, D.: Tutorial: J2ME Design and Development Considerations. In: OOPSLA 2001 (October 2001)
16. Kimm, H., Shin, S., Shim, C.: Two Approaches to Improve Java MIDP Record management System in Wireless Devices. In: Proceedings of 2005 IEEE International Conference on Electro Information Technology (May 2005)
17. Silberschatz, A., Gagne, G., Galvin, P.B.: Operating System Concepts, 7th edn. John Wiley & Sons, Chichester (2005)
18. Brakmo, L.S., Wallach, D.A., Viredaz, M.A.: A Technique for Reducing Energy Consumption in Handheld Devices. In: MobiSys Conference, Boston, MA, June 6-9,
19. Lapinsky, S.E., et al.: andhel computers in critical care. Journal of Critical Care 5(4) (2001)

Crisis Management Simulation: Spread of Diseases in National University of Singapore

Gary S.H. Tan and R. Lau

School of Computing
National University of Singapore
Computing 1, Law Link, Singapore 117590
gtan@comp.nus.edu.sg

Abstract. In an attempt to study the spread of diseases in the National University of Singapore (NUS), a simulation model unique to the NUS environment was developed to evaluate the effectiveness of different intervention policies during the event of a disease outbreak. The system allows the user to introduce a user-defined disease into the NUS population to study its impact. The effectiveness of various intervention policies on selected diseases are evaluated and presented in the paper.

1 Introduction

A crisis can be defined as a crucial time or state of affairs in which a decisive change is impending. In modern day context, a crisis could refer to a variety of events, including terrorist attacks, chemical spills and global pandemics. In crisis management, there are three main steps – identifying the crisis, planning a response to the crisis, and resolving the crisis.

In this paper, we investigate the crisis of a disease outbreak due to the constant threat of an epidemic. Throughout history, there have been records of countless epidemics, resulting in massive loss of lives. In other cases, outbreaks of epidemics have caused intensive economic losses [7]. The severe impacts of epidemics make it extremely beneficial for us to be able to understand, anticipate and control the behavior of epidemics. Knowing the behaviors of epidemics allows us to plan responses before any outbreak, and allow for rapid and suitable decisions in times of crisis.

In an attempt to control and prevent the occurrence of epidemics, scientists have been trying to predict and mimic the behaviors of epidemics through the use of various tools. Compartmental models using differential equations remain as important tools for epidemic simulation.

This paper describes a model for a small population, or more specifically, the population at the National University of Singapore (NUS) campus. Unlike traditional models of epidemiology, this model will take into consideration factors such as space and motion. The spread of a disease will then be simulated based on this model. From the simulation results, we are able to analyze the impact of various diseases assuming that there is an outbreak in NUS. Using this model, we could then test the effectiveness of different intervention policies in their ability to resolve the outbreak.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 75–84, 2007.

2 Background and Related Work

2.1 Definition of Epidemiology, Basic Reproductive Ratio, etc

Epidemiology is defined as the branch of medical science that deals with the incidence, distribution, and control of disease in a population. Epidemics can occur as a result of a very infectious disease. When the epidemic spreads across a large region, such as a continent or worldwide, it is described as pandemic. For the purpose of our study, disease shall refer to an impairment of the normal state of the living animal or plant body or one of its parts.

The Basic Reproductive Ratio, R_0, is commonly referred to as the epidemiological threshold. It is defined as the average number of secondary cases caused by an infectious individual in a totally susceptible population during its entire infectious period [2]. Mathematically, it can be represented in the equation below.

R_0 = (mean contact rate) * (transmission probability) * (infectious duration)

If R_0 is less than 1, then the epidemic is unlikely to persist. If R_0 is greater than 1, then the epidemic is likely to spread. The Basic Reproductive Ratio presents three categories of measures to control epidemics - reducing transmission probability, decreasing mean contact rate and reducing the length of infectious period. Interventions usually aim at affecting these three factors to control and stop an epidemic.

The transmission of a disease requires the presence of a susceptible individual, an infected individual and an effective contact between them. Upon an effective contact, there is a risk that the disease will actually be transmitted. While it is unethical to conduct experiments to induce infections and actually compute the transmission probability, a good estimate of this value is the Secondary Attack Rate [3]. The Secondary Attack Rate can be defined mathematically as:

Secondary Attack Rate = (# secondary cases) / (# susceptibles)

Figure 1 shows a typical disease life cycle.

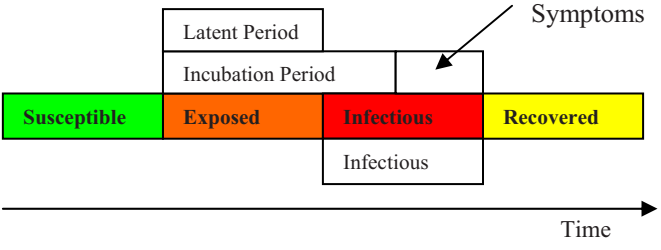


Fig. 1. Disease Life Cycle

2.2 Epidemic Models

Many epidemic simulations are now based on the compartmental model, dividing people into different categories based on the state of their health. Having established this, the next step is to model how the infection actually spreads. The consensus regarding this is much less unified, with various approaches being discussed.

Compartmental Models
Traditional epidemic simulation depended on compartmental models to represent the population .The compartmentalizing of the population gives rise to its name. A variety of compartmental models exist, and the choice of model depends on the nature of the disease [4].

Differential Equations
The earliest form of epidemic simulation relied on differential equations to model the flow of the population from one phase to another. This is a mathematical approach to model the progress and impact of infectious disease. Consider the basic SIR model: a set of assumptions, such as fixed population size, zero incubation period, completely homogeneous and well-mixed population are used [5].

Lattice Models
One motivation of using lattice models is the failure of the differential equations approach to capture space and motion, producing unrealistic results when simulating for small populations. Agents in lattice models are placed on a regular grid that is usually two-dimensional. In lattice models, a well-mixed population is not assumed. Agents can interact only with their neighbours and their resultant state is a function of its neighbours. Movement to adjacent cell can also be modelled. This ability to represent space and motion greatly overcomes the failure of the differential equations approach in attempting to model small populations.

3 The NUS Epidemic Model

The lattice model approach was adopted for our NUS epidemic model. Lattice models focus on the spatial aspects and the physical location of each agent. Data for this aspect is available and we could try to mimic the behaviour and movement of each agent during the course of the day.

3.1 Agents and Their World

Agents are living beings that reside in the world. They can refer to any objects that we are interested in studying about, such as birds, mosquitoes and humans. There may be diseases such as dengue fever that depends on vectors like mosquitoes. However, for the purpose of this study, agents refer exclusively to humans.

Agents are divided into two categories, students and staff. Staff comprises faculty members, research staff, administrative and professional staff, as well as general staff. A student is required to take cross faculty modules as part of academic requirements and hence is likely to move to other zones for lectures and tutorials. Also, within their home zones, they are likely to be very mobile, moving from one venue to another to attend lessons. A member of the staff is less likely to be as mobile at their home zone, travelling less compared to a student.

The world that Agents reside in is the NUS campus. High-density areas consisting of buildings are identified. Special attention is given to the halls of residence. This is because agents staying on campus residences continue to have activities after office hours. To the agents, the world is made up of many cells, and each agent is located in one of the cells. Agent movements are restricted to the cells that are defined in the world.

The spread of disease can only be achieved when an infected agent is located in the same cell with other susceptible agents. When there is a contact, there is a chance that the disease will be transmitted. When an agent is infected, they will progress through the four biological stages of disease infection.

3.2 Model Assumptions

There are certain assumptions that were made to simplify the model. While these assumptions limit the realism of the model, it is important to note that all the significant factors have been taken into consideration. Some of the assumptions are:

1. NUS is a closed population. This means that there is no incoming population or outgoing population.
2. There can be no outside source of infection. Agents do not contract the disease when they are not on campus. The only way an Agent can get infected with the disease is to interact with another infected Agent. As a result, the entire outbreak is due to a single "patient zero" in NUS.
3. A five-day working week is assumed. Working hours take place from 8:00am to 6:00pm. Hall activities end at 12:00am. On Saturdays and Sundays, it is assumed that no activities take place in NUS.

4 Implementation

A number of software packages and library add-ons are available in the realm of agent-based simulation and were considered for this simulation, e.g. Repast, Swarm and Mason. Generally, they provide a conceptual framework as well as software libraries. These libraries are used together with a programming language, usually Java, when implementing a model. Knowledge of the programming language is expected, resulting in a steeper learning curve.

In [1], Jackson also did a review of the available software platforms for Agent Based Models (ABMs). After careful consideration, the simulation was implemented in NetLogo, a cross-platform multi-agent programmable modeling environment.

NetLogo [6] is a high-level platform that uses its own programming language. This is very different from other platforms, which provides libraries to supplement programming languages already in use, like Java, to aid the creation of ABMs. Despite being a simpler language compared to the other platforms, it contains sufficient control and structuring capabilities to implement complex models. The availability of high-level structures and primitives simplifies the task of programming.

Speed-wise, NetLogo out-performs some of the other platforms. Built-in graphical interfaces are provided to implement visual output. Tools for statistical output and real-time graphing for data are available. If the need arises, Java extensions could be created and used in the models created.

4.1 Graphical User Interface

The user interface is divided into three parts, the map display on the left, the input parameters on the right and the output at the bottom (figure 2).

Users can select a default disease scenario from the drop down box at the top right hand side. Custom input parameters can also be specified using the slider bars on the right. At the bottom, there are four charts that plot in real time, the number of susceptible, exposed, infected and recovered Agents.

Animated display on the left shows the real time location of agents as they move about on NUS. Green dots represent susceptible agents. (orange - exposed, red - infected, yellow - recovered).

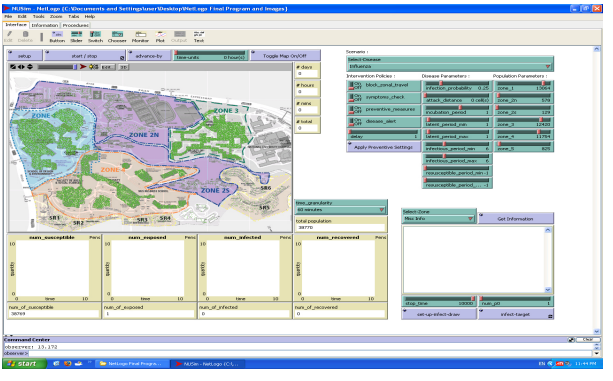


Fig. 2. Graphical User Interface

Users can choose from a list of pre-defined diseases, or input a custom disease. To characterize a disease, a set of parameters is needed. Each of the parameter is explained below:

Infection Probability

When an infected Agent comes into contact with a susceptible agent, the infection probability is one of the factors that influence the likelihood of infection. The higher the infection probability, the more likely it is that an infected agent will pass the disease to a susceptible Agent. The infection probability is based on Secondary Attack Rate that is calculated from available data.

Attack Distance

Different types of diseases have different mechanisms by which they spread. For example, Influenza and SARS spread mainly via close contacts. Hence, it is unlikely that an agent can infect an agent in another cell. A value of 0 indicates that infection can only occur within the same cell.

Incubation Period

After a number of days, an infected agent starts exhibiting symptoms. A value of 0 indicates that the agent is symptomatic only when he/she is infectious, while a value of 1 indicates that the agent is symptomatic one day after he/she is infectious.

Latent / Infectious / Resusceptible Period

The duration in days that an agent spends in each phase (S, E, I or R) can be specified. A range of values can be given as input or a fixed value by setting the minimum and maximum to the same value. If the disease profile does not contain a particular phase, e.g. the Exposed phase, both maximum and minimum values to 0.

4.2 Initialization

The world is set to a default size of 600 by 450 cells. During the SARS outbreak in 2003, a geographical zonal plan was drawn up, dividing the NUS campus into six different zones. To set up the world, this NUS zonal map is used to divide the world. JPEG files marking out the six zones were imported and then used to mark out the zones. Subsequently, the high-density areas and residences were also identified by importing two other JPEG files.

It is important to note that each agent is not a cell. The cell is merely a "container", or a square of physical space that the agent resides in. The agents are created based on students and staff statistics that were obtained from the NUS Registrar's Office and Office of Human Resources.

5 Experiments and Results

As the scope of epidemiology is very large, we have chosen to focus on three specific diseases (Influenza, SARS, and Avian Flu) and three different intervention policies in the event of an outbreak. The effectiveness of each of the intervention policy is observed for each of the different disease, and the aim is to find out which intervention policy is the most effective.

Disease profiles and characteristics were obtained from reliable sources such as World Health Organization. However, information on diseases such as incubation period is often incomplete or does not contain a fixed value. At times, available data could, at best, give an estimate of what are the likely characteristics of diseases. For our experiments, we have used the set of parameters for the three diseases based on available data as shown in table 1.

Table 1. Disease Parameters Used For Experiments

Parameter	Influenza	SARS	Avian Flu
Incubation Period	1 day	0 day	1 day
Latent Period	1 day	4 days	6 days
Infectious Period	6 days	14 days	14 days
Infection Probability	0.25	0.31	0.25
Attack Distance	0 cell	0 cell	0 cell

As mentioned earlier, the Basic Reproductive Ratio presents three categories of intervention policies to control epidemics - reducing transmission probability, decreasing mean contact rate and reducing the length of infectious period. While there is not much that the authorities can do to reduce the length of infectious period, we can try to reduce the impact of the other two factors.

Default Scenario

The default scenario is used as a control scenario by which the subsequent experiments are compared to. For the default scenario, each of the three diseases is introduced into the population, but none of the intervention policies are put into place. This is to investigate the impact of the outbreak if no intervention policy is implemented at all, and the agents continue going about their daily routine.

Restricting Zonal Travel

Restricting Zonal Travel attempts to reduce the mean contact rate of agents by restricting agents movements to their home zone only. Although agents are not allowed to travel to other zones, they can still move about within their own zone. During working hours, this restriction on zonal travel is implemented. However, agents staying in campus halls or residences still get to meet with people from other zones after working hours when they are back to their living areas.

Preventive Measures

Preventive Measures aims to reduce the likelihood of the transmission of infection from one agent to another. This form of intervention policy cannot be enforced and is usually advisory, such as reminding people to maintain personal hygiene via frequent washing of hands. To simplify the study, the overall transmission probability is reduced by 20%, and we investigate if the relative gains from implementing such a policy is more than, less than or equal to the reduction in transmission probability.

Symptoms Checking

The goal of Symptoms Checking is to reduce the mean contact rate of infected agents. If infected agents do not come into contact with susceptible agents, there is no way to propagate the infection. At the start of each working day, every agent is checked to see whether they are symptomatic. This intervention policy assumes that all symptomatic agents are isolated and removed from the population. They only rejoin the population when they have recovered for their illness.

Disease Alert

Disease Alert is not considered an intervention policy. It refers to the awareness of the agents that there is a possible risk of getting infected with the disease. As with the SARS outbreak, people are aware of the risk of the disease and as a result, their behavior deviated from their norm. They are less likely to travel and when they are symptomatic, they are more likely to seek medical attention and stay at home. When the accumulated total of the infected and recovered agents reaches 0.1% of the population, Disease Alert is activated after a delay of one day.

For our experiments, we used an AMD Sempron Processor 3000+ 1.8GHz desktop PC with 512MB of Ram, installed with Microsoft Windows XP and NetLogo 3.1.3. At the start of the experiment, a single agent is infected with the disease we are interested in studying. Disease Alert and the relevant intervention policy is activated one day after the cumulative total of the infected and recovered population exceeds 0.1% of the total population.

In total, four sets of experiments were conducted. The results and discussions on each set of experiment are presented below (figures 3-5). (black line: default scenario, green short dash: restricting zonal travel, red long dash: preventive measures, blue long-short dash : symptoms checking). Figure 6 gives the overall results.

5.1 Discussion

For all three experiments, a similar shape was observed for the susceptible population and infected population. Initially, the drop in susceptible was very gentle. However, as the infection picked up its pace, the drop becomes more and more drastic before it eventually flattens out towards the end. A jagged curve was observed for Influenza's infected population, and is probably due to its short latent period, which could lead to short bursts in infections. Unlike Influenza, SARS and Avian Influenza both have much

longer latent period. In terms of the speed of outbreak, Influenza is much more deadly, taking only 2500 time units for the outbreak to end. SARS, in comparison, spread the slowest, taking up to three times the amount of time. In terms of the total number of infections, Avian Influenza infected the largest percentage of the population at 85%. In contrast, SARS once again was least dangerous, infecting 52% of the population.

An interesting point to note is that on paper, SARS seems like an extremely dangerous disease. Among the three, it has the highest infection probability. Also, it has a

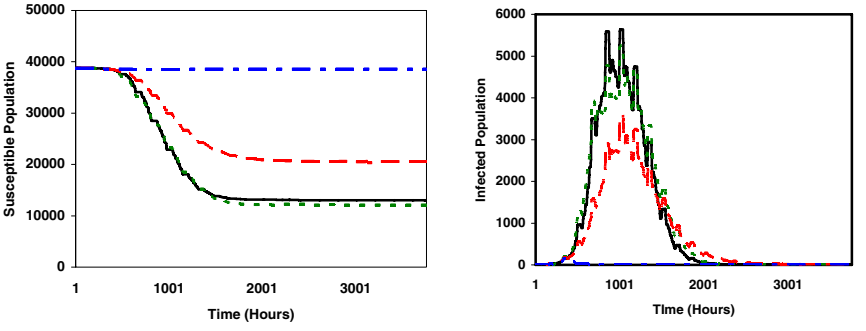


Fig. 3. Influenza Results

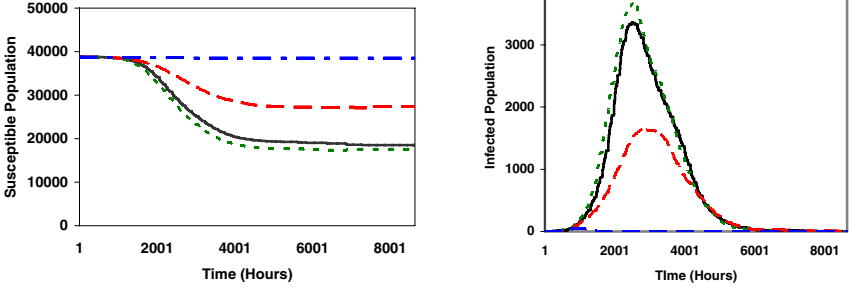


Fig. 4. SARS Results

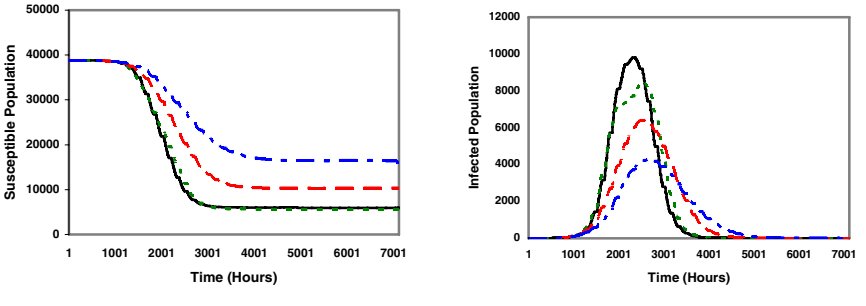


Fig. 5. Avian Influenza Results

relatively short latent period, and a long infectious period. Yet, results indicate that it is the least deadly. Its speed of spread and impact did not reach the level of both Influenza and Avian Influenza. This could because SARS, unlike Influenza and Avian Influenza, is only contagious when the infected individuals are symptomatic. Influenza and Avian Influenza have a "safe window" to transmit to susceptible individuals without showing any symptoms. Symptomatic individuals have a chance of withdrawing from the population via staying at home and thus, help to slow the outbreak.

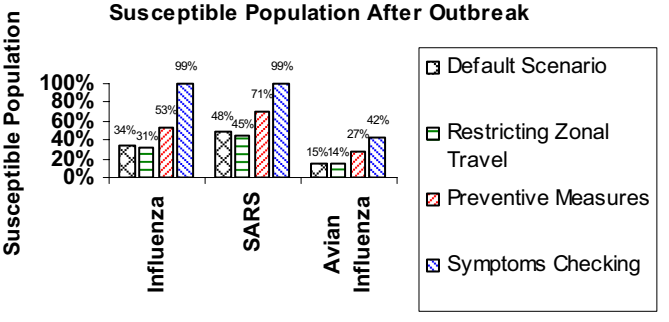


Fig. 6. Overall Results

Restricting Zonal Travel

For all three experiments, the implementation of zonal restriction did not seem to have a significant effect on both the eventual susceptible population as well as the speed of the outbreak. This is surprising, as the restriction of zonal movement should have slowed or contained the outbreak within specific zones.

There are two possible explanations to the experiments result. Firstly, the implementation of this zonal policy could have been too late to contain the outbreak. If there are already infected agents in each zone, the implementation of such a policy would do nothing to affect the spread of the disease within each zone. Secondly, the restriction of zonal movement is only implemented during working hours. After office hours, agents are allowed to return to their campus hostels or residences. This gives them the opportunity to interact with people from other zones, creating chances of spreading the infection between zones.

Preventive Measures

From the results observed, Preventive Measures could potentially be very effective in outbreak control. By decreasing the overall transmission probability by 20%, the susceptible population at the end of the outbreak experienced varying degrees of increase. The susceptible population increased by 56%, 48% and 80% respectively.

Given that 20% is a very conservative estimate, the actual improvements due to the active adoption of Preventive Measures by individuals could deny the infection from spreading throughout the entire population. This could explain why Preventive Measures is often advocated in times of epidemic, such as the SARS outbreak in 2003.

Symptoms Checking

Symptoms Checking is the most effective intervention policy. It managed to stop both Influenza and SARS before the outbreak spread out of control. Even when it did not

stop the outbreak of Avian Influenza, it was able to slow it down and diminish its severity. The ability of Symptoms Checking in controlling SARS is consistent with our theoretical understanding. Because SARS is contagious only when symptoms are exhibited, Symptoms Checking is able to remove infectious agents efficiently.

It is interesting to note that while Symptoms Checking was effective for Influenza, it was not effective for Avian Influenza. Firstly, Influenza has a short latent period of 1 day, compared to Avian Influenza's 6 days. Also, its infectious period was 6 days compared to 14 days of Avian Influenza. As Symptoms Checking removes all symptomatic agents from campus in the morning, we can possibly eliminate infectious period as the reason. This leaves us with the short latent period of Influenza.

One possible explanation is that the short latent period makes the turnover rate of Influenza Agents extremely high. Agents took only one day to go from susceptible, to exposed, to infectious state. The effect of Symptoms Checking, coupled with weekends where Agents do not need to study or work, quickly exhausted the number of exposed and infectious Agents.

Unlike Influenza, Avian Influenza has a much longer latent period, and the impact of weekends is much reduced. Exposed Agents are not eliminated as quickly as it was the case for Influenza. The ability to spread the disease while not showing any symptoms renders Symptoms Checking ineffective. Thus, if the latent period of Influenza had been longer, it would not have been contained by Symptoms Checking.

6 Conclusion

An epidemic model specific to NUS was described and presented in this paper. A total of three sets of experiments were conducted, each investigating the effectiveness of an intervention policy with respect to a particular disease. From the experiment results, it was observed that Symptoms Checking is the most effective in minimizing the effect of all three diseases.

The use of different types of intervention policies is only one specific aspect of epidemiology that is studied. Other important areas include the speed of intervention policies, impact of population density and effect of contact patterns.

References

1. Jackson, S.K., Lytinen, S.L., Railsback, S.F.: Agent-based Simulation Platforms: Review and Development Recommendations. Simulation 82(9), 609–623 (2006)
2. Keeling, M.J.: The Mathematics of Diseases. Plus Magazine. Issue 14 (2001)
3. Shih, C.F.: Modelling Epidemic Spread Using Cellular Automata. Department of Computer Science and Software Engineering, The University of Western Australia (2002)
4. Trottier, H., Philippe, P.: Deterministic Modeling Of Infectious Diseases: Theory And Methods. The Internet Journal of Infectious Diseases 1(2) (2001)
5. Weisstein, E.W., Weisstein, T.: Kermack-McKendrick Model, MathWorld (2004), http://mathworld.wolfram.com/Kermack-McKendrickModel.html
6. Wilensky, U.: NetLogo, The Center For Connected Learning (1999), http://ccl.northwestern.edu/netlogo/
7. World Health Organization, http://www.who.int

An Early Warning System for Loan Risk Assessment Based on Rare Event Simulation

Hong Zhou, Yue Qiu, and Yueqin Wu

School of Economics & Management, Beihang University
Beijing 100083, P.R. China
h_zhou@buaa.edu.cn

Abstract. System simulation is one of important tool for risk assessment. In this paper, a new method is presented to deal with credit risk assessment problems for commercial banks based on rare event simulation. The failure probability of repaying loans of listed company is taken as the criterion to measure the level of credit risk. The rare-event concept is adopted to construct the model of credit risk identification in commercial banks, and cross-entropy scheme is designed to implement the rare event simulation, based on which the loss probability can be assessed. Numerical experiments have shown that the method has a strong capability to identify the credit risk for commercial banks and offers a good tool for early warning.

Keywords: rare event, simulation, credit risk, cross-entropy.

1 Introduction

Many countries have attached great importance to bank loan risk management since the Asia Financial Crisis. Credit risk is currently one of the biggest challenges for the Chinese domestic commercial banks. And credit risk identification is of special importance to these domestic financial institutions. After China's entry into the World Trade Organization (WTO), the domestic commercial banks have been confronting with both opportunities and challenges of the international financial market and credit risk management has become a more and more imperative task. Undoubtedly, early warning management is a very important link in the risk management of commercial bank loans. China has been concerned with the risk involved with bank loans and has urged the application of modern information technology to address this problem. Thus, a financial early warning system is currently in demand.

Due to the increase of the overall risk level and the structural changes of the global financial market, overseas academic and financial institutions have been putting much emphasis on credit risk management, and many theoretical and application achievements have been made in the field of credit risk measurement over the last 20 years [7]. The traditional methods used to evaluate credit risk, such as discriminate analysis and logistic, are all based on statistical analysis. Domestic researchers have also made great progress based on the overseas credit risk management theories.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 85–94, 2007.

However, due to the lagged credit risk management of Chinese commercial banks, the traditional credit risk measurement methods could not provide satisfactory results with limited and incomplete historical data. In recent years, the statistical methods are replaced with soft computing techniques because soft computing techniques have been proven to be effective and efficient in a number of fields. Therefore, neural network models are developed and applied quickly to credit risk evaluation in recent years because of their excellent performances of treating non-linear data with learning capability [15], [16], and [22]. However, the shortcoming of neural networks is also significant due to a "black box" syndrome and the difficulty in dealing with qualitative information, which limited its applications in practice[1],[21]. Besides, a relatively slow convergence speed is also the disadvantage of NNs model.

In this paper, a credit risk identification model based on rare event simulation is proposed. The failure probability of repaying loans of listed company is taken as the criterion to measure the level of credit risk. The failure event of repaying loans is treated as rare event due to the relatively low probability, and the probability is assessed through the rare event simulation method. The cross-entropy (CE) method is a new generic approach to rare event simulation [17]. The CE method was motivated by an adaptive algorithm for estimating probabilities of rare events in complex stochastic networks by Rubinstein in 1997, which involves variance minimization [18]. It was soon realized that a simple cross-entropy modification could be used for not only estimating probabilities of rare events but solving difficult COPs as well [19], [20]. Cross entropy method has been widely used recently [2], [4], [9]. Cross entropy has practical applications in early warning systems both in theory and technology. The proposed model has the characteristics of self-adaptation and self-learning, and is suitable for the domestic commercial banks due to their historical data are limited and incomplete. Cross entropy method is attempted to apply to commercial banks credit risk assessment. Moreover, we wish to expand the application areas of cross-entropy method and propose a commercial bank's credit risk warning method in a new way.

The rest of this paper is organized as follows: Section 2 describes methodology; Section 3 explains the data source and the structures of model based on rare event simulation; an experiment analysis is shown in section 4 and finally some concluding remarks are drawn from section 5.

2 Cross-Entropy Methods

Rare event simulation has attracted extensive attentions since the concept was first proposed, but there haven't been many research results yet [8], [10]. Performances of modern systems, such as coherent reliability systems, inventory systems, insurance risk, storage systems, computer networks and telecommunications networks, are sometimes characterized by probabilities of rare events and usually studied by simulation [3], [5], [11], and [13]. Straightforward simulation for rare events requires a large number of trials and hard to implement because the occurrence of rare events are extremely little in a standard simulation [12], [14], hence new methods are needed to be investigated and developed. IS is based on the idea to make the occurrence of rare events more frequent, or in other words, to speed up the simulation. Technically, IS

aims to select a probability distribution that minimizes the variance of the IS estimate. The efficiency of IS depends on obtaining a good change of measure. Large deviation results have proven their applicability for small systems with medium complexity [6].

Let $x = (X_1, \ldots, X_n)$ be a random vector taking values in some space Ω and $f(x)$ be a probability density function on Ω. Let s be some real-valued function on Ω. Suppose we are interested in the probability that $s(x)$ is greater than or equal to some real number r and this probability can be expressed as $l = P\{s(x) \geq r\} = EI_{(s(x)\geq r)}$. If this probability is very small, say smaller than 10^{-5}, we call $\{s(x) \geq r\}$ a *rare event*.

A straightforward way to estimate $l = P\{s(x) \geq r\} = EI_{(s(x)\geq r)}$ is to use crude Monte-Carlo simulation: draw a random sample $X_1, \ldots, X_N$ from $f(x)$, then

$$\hat{l} = \frac{1}{N}\sum_{i=1}^{N} I_{\{s(X_i)\}} = \frac{1}{N}\sum_{i=1}^{N} I_{\{s(X_i)\geq r\}} . \tag{1}$$

is an unbiased estimator of l. However this poses serious problems when $\{s(x) \geq r\}$ is a rare event. In that case a large simulation effort is required in order to estimate l accurately, i.e., with small relative error or a narrow confidence interval.

An alternative is based on importance sampling: take a random sample $X_1, \ldots, X_N$ from an *importance sampling density* $g(x)$ on Ω, then

$$l = \int_{\{s(x)\geq r\}} f(x)dx = \int_{\{s(x)\geq r\}} L(x)g(x)dx . \tag{2}$$

where $L(x) = \frac{f(x)}{g(x)}$ is called the *likelihood ratio*. It is well known that the best way to estimate l is to use the change of measure with density

$$g^*(x) = \frac{I_{(s(x)\geq r)} f(x)}{r} . \tag{3}$$

The estimator (3) has zero variance, and we need to produce only N = 1 sample. The obvious difficulty is of course that $g^*(x)$ depends on the unknown parameter, so it cannot be computed in practice. In estimate rare event probability: draw a random sample $X_1, \ldots, X_N$ from $g(x)$, and then get estimator according to (2)

$$\hat{l}_{IS} = \sum_{i=1}^{R} I_{\{s(X_i)\geq r\}} L(X_i) . \tag{4}$$

We will call the estimator in (4) the *importance sampling estimator*.

To obtain the optimal – or at least a good – change of measure from the original sampling distribution $f(x)$ to a new distribution $g(x)$, is the main challenge with respect to making importance sampling efficient and robust. From equation (2), the only restriction to the probability density $g(x)$ is that $g(x) > 0$ for all samples x. This is a necessary condition which serves as a guideline for choosing $g(x)$. The efficiency of the importance sampling estimate in (4) is dependent on an appropriate choice of the new distribution. An unfortunate choice of $g(x)$ may cause the variance of (4) to be larger than the variance of the estimates obtained by direct simulation. For some $g(x)$, the variance may be infinite, and, hence, it is crucial to find a good $g(x)$. Moreover, it is often convenient to choose a $g(x)$ in the family of densities $\{f(\cdot;v)\}$. The idea now is to choose the parameter vector, called the reference parameter (sometimes called tilting parameter) v such that the distance between the densities $g_{opt}(x)$ and $f(\cdot;v)$ is minimal. A particular convenient measure of distance between two densities g and h is the Kullback-Leibler distance, which is also termed the cross-entropy between g and h. The Kullback-Leibler distance is defined as:

$$\begin{aligned} & D(g,h) \\ &= E_g \ln \frac{g(x)}{h(x)} \\ &= \int g(x) \ln g(x) dx - \int g(x) \ln h(x) dx \end{aligned} \tag{5}$$

Minimizing the Kullback-Leibler distance between $g_{opt}(x)$ in (10) and $f(\cdot;v)$ is equivalent to choosing v such that $-\int g_{opt}(x) \ln f(x;v) dx$ is minimized, which is equivalent to solving the maximization problem

$$\max_v \int g_{opt}(x) \ln f(x;v) dx \tag{6}$$

Substituting $g^*(x)$ from (**3**) into (**6**) we obtain the maximization program

$$\max_v \int \frac{I_{\{s(x)\geq r\}} f(x;U)}{l} \ln f(x;v) dx \tag{7}$$

This is equivalent to the program

$$\max_v D(v) = \max_v E_U I_{\{s(X)\geq r\}} \ln f(x;v) \tag{8}$$

Using again importance sampling, with a change of measure $f(\cdot;w)$ we can rewrite (8) as

$$\max_{v} D(v) = \max_{v} E_w I_{\{s(X)\geq r\}} W(x;U,w) \ln f(x;v) \tag{9}$$

Where

$$W(x;u,w) = \frac{f(x;U)}{f(x;w)}$$

is the likelihood ratio, at x, between $f(\cdot;U)$ and $f(\cdot;w)$. The optimal solution of (9) can be written as

$$v^* = \arg\max_{v} E_w I_{\{s(X)\geq r\}} W(x;U,w) \ln f(x;v) \tag{10}$$

We may estimate v^* by solving the following stochastic program

$$\begin{aligned} &\max_{v} \hat{D}(v) \\ &= \max_{v} \frac{1}{N} \sum_{i=1}^{N} I_{\{s(X_i)\geq r\}} W(X_i;U,w) \ln f(X_i;v) \end{aligned} \tag{11}$$

Where $X_1,\ldots,X_N$ is a random sample from $f(x;w)$. In typical applications the function in (11) is convex and differentiable with respect to v, thus, the solution of (11) may be readily obtained by solving (with respect to v) the following system of equations:

$$\frac{1}{N} \sum_{i=1}^{N} I_{\{s(X_i)\geq r\}} W(X_i;U,w) \nabla \ln f(X_i;v) = 0$$

The advantage of this approach is that the solution of (**11**) can often be calculated analytically.

3 Structure of the Commercial Bank Loan Risk Early Warning System

The data of samples used in this study are selected from the listed corporations of China. Those corporations can be divided into two categories: one category consists of a set of ST (special treated) corporations and the other one consists of non-ST corporations. According to the industry classification in China stock market, 10 industries observations are selected in 2005, including 10 ST and 10 non-ST corporations. Dates are all from the www.stockstar.com and www.sse.org.cn websites.

The financial ratios include four classifications: earning capability ratio, running capability ratio, short term repayment capability ratio and long term repayment capability ratio. Additionally, each classification has its own set of detail indexes. The original data is disposed as follows: the maximum value within the industry is seen as the standard value and each index value divided by the standard; each classification is assigned to the average weighted. The formula for the input data is: $\sum weight \times (value / s\tan dard)$ (Table 1).

Table 1. Training Samples

No.	Input Vector			
1	0.28	0.20	0.09	0.21
2	0.30	0.19	0.09	0.20
3	0.28	0.21	0.10	0.18
4	0.31	0.18	0.08	0.19
5	0.29	0.18	0.10	0.21
6	0.31	0.17	0.08	0.18
7	0.30	0.20	0.09	0.22
8	0.28	0.18	0.12	0.19
9	0.29	0.18	0.11	0.22
10	0.28	0.20	0.11	0.18
11	0.27	0.19	0.09	0.20
12	0.31	0.19	0.12	0.17
13	0.28	0.20	0.09	0.17
14	0.30	0.20	0.08	0.18
15	0.25	0.12	0.04	0.08
16	0.24	0.10	0.07	0.09
17	0.24	0.08	0.06	0.12
18	0.26	0.13	0.06	0.13
19	0.23	0.11	0.05	0.10
20	0.19	0.09	0.07	0.09

It is equivalent to a rare event from the commercial banks default incident itself; and it is to predict rare event probability from the default probability of loans. Moreover, the cross-entropy method is very good in the field of forecast the rare events probability. So, applying cross-entropy method to credit risk early warning is feasible. In general, the steps are as follows:

1. Choose the synthetic score as the estimate function: $s(x) = -\sum weight \times index$ and get the critical value r. Take the 0.28% of the ranking as the initial critical value r based on Standard & Poor's.
2. Using the critical value r to confirm default probability, moreover, estimate the critical value of default probability.
3. The default probability of test samples is determined and gets the conclusion of whether the default event will occur.

The above rationale results in the following algorithm:

step 1. Pre-sample and estimate the initial value of estimate function r ;

step 2. Compute the default probability of ST $\hat{l}$ and non-ST corporations $\bar{l}$:

step2.1. Define $v_0 = u$, $\rho = 0.1$;

step2.2. Generate a sample $X_1, \ldots, X_N$ from importance sampling probability density function $f(.; v_{t-1})$ and compute the sample $(1-\rho)$-quantile r_t of the performances, provided $r_t < r$,otherwise set $r_t = r$.

step2.3. Use the same sample $X_1, \ldots, X_N$ to solve the stochastic program (11). Denote the solution by v_t .

step2.4. If $r_t < r$, set $t = t+1$ and reiterate from step 2.2. Else proceed with step 2.5.

step2.5. Estimate the rare-event probability using the importance sampling estimator (4).

step 3. Estimate the critical value of default probability;

step 4. Compute default probability of testing sample.

It is important to note that the above algorithm is useful only in the case where the probability of the "target event" $\{s(x) \geq r\}$ is not too small, say $l \geq 10^{-5}$ .For default probability of non-ST corporations; however, the above algorithm is difficult to carry out. Namely, due to the rareness of the events $\{s(X_i) \geq r\}$, most of the indicator random variables $I_{\{s(X_i) \geq r\}}$, $i = 1, \ldots, N$ will be zero, for moderate N . A *multi-level* algorithm can be used to overcome this difficulty. The idea is to construct a sequence of reference parameters $\{v_t, t \geq 0\}$ and a sequence of levels $\{r_t, t \geq 1\}$ .That is, v_t and r_t satisfies $p_{v_{t-1}}(s(x) \geq r_t) \geq \rho$.

4 Experiment Results

A pair of testing samples is entered into the model, and the rare event probability is calculated. The critical value of rare event occur is $r = 0.074$, the critical value of default probability for Non-ST corporation is $\bar{l} = p(s(x) > r) = 3.4379 \times 10^{-8}$, the critical value of default probability for ST corporation is $\hat{l} = p(s(x) > r) = 0.0435$. This means that when the sample is considered Non-ST then the output is between $\bar{l}$ and $\hat{l}$, but when the output is bigger than $\hat{l}$ the sample is considered ST (Table 2).

The testing results show the actual output is consistent with the desired output. It is very appropriate to apply this early warning model to analyze commercial bank loan crisis.

Using the estimated optimal parameter vector, the final step with N=1000 gave now an estimate of 3.4379×10^{-8} with a 90%H.W. of 9.8737×10^{-9} . The simulation time was only 4 seconds, using a Matlab implementation on a Pentium III 500 MHz processor. In

Table 2. Testing Sample and Default Probability

sample	Earning capability	Running capability	Short term repayment capability	Long term repayment capability	Default probability
Non-ST	0.26	0.23	0.11	0.18	8.6482e-008
ST	0.22	0.12	0.10	0.21	0.1520

contrast, the CPU time required for the CMC method with N=1000000 samples is approximately 9000 seconds and gave an estimate of 6.6667×10^{-7} with an 90%H.W. 5.3344×10^{-7} . We see that with a minimal amount of work we have reduced our simulation effort (CPU time) by roughly a factor of 2250. In the following table "90% H.W." denotes the half-width of a 90% confidence interval and N is the sample size (Table 3).

Table 3. Simulation Efficiency Comparison

method	Default probability	N	simulation time	90%H.W.
MC	6.6667×10^{-7}	1000000	9000seconds	5.3344×10^{-7}
CE	3.4379×10^{-8}	1000	4seconds	9.8737×10^{-9}

The above results indicate high efficiency of the CE method for estimating rare events probabilities, where the naive Monte Carlo method fails. For events that are not very rare, the CE method may still help in terms of providing estimates with smaller variance.

5 Conclusion

The research of credit risk is an important problem faced by the banking. This paper investigated the modeling and performances of rare event simulation method on a problem of credit risk evaluation, based on a set of financial data selected from China listed corporations. And put forward a risk predicting solution based on Cross-Entropy. The credit loan is the main income of a bank, which is also the most risky product. With the implementation of the system, credit risk management will be improved and it will make an important role in the risk control domains. The model constructed by the Cross-Entropy metrod is compatible to the enterprises in China and can reduce the credit risk of the state-owned commercial bank greatly. The CE method was motivated by an adaptive algorithm for estimating probabilities of rare events in complex stochastic networks (Rubinstein, 1997); there is a broad space for development both in theory and in application. I hope this approach can promote the research of our country credit risk evaluation, and also it raises ability and levels of the risk management of our country financial industry.

Acknowledgment

The authors would express sincere appreciation to the support from the National Natural Science Foundation of China (Project no. 70531010, 70521001), and Program for New Century Excellent Talents in University of China (Project no. NCET-04-0175).

References

1. Shapiro, A.F.: The merging of neural networks, fuzzy logic, and genetic algorithms. Insurance:Mathematics and Economics 31, 115–131 (2002)
2. Alon, G., Kroese, D.P., Raviv, T., Rubinstein, R.Y.: Application of the Cross-Entropy Method to the Buffer Allocation Problem in a Simulation-Based Environment. Annals of Operations Research 134, 19–67 (2005)
3. Al-Qaq, W.A., Devetsiklotis, M., Townsend, J.K.: Stochastic Gradient Optimization of Importance Sampling for the Efficient Simulation of Digital Communication Systems. IEEE Trans. Communications 43, 2975–2985 (1995)
4. Chepuri, K., Homem-de-Mello, T.: Solving the Vehicle Routing Problem with Stochastic Demands using the Cross-Entropy Method. Annals of Operations Research 134, 53–181 (2005)
5. Devetsikiotis, M., Townsend, J.K.: An Algorithmic Approach to the Optimization of Importance Sampling Parameters in Digital Communication System Simulation. IEEE Trans. Communications 41, 1464–1473 (1993)
6. Siegmund, D.: Importance Sampling in the Monte Carlo Study of Sequential Tests. The Annals of Statistics 4, 673–684 (1976)
7. Altman, E.I., Saunders, A.: Credit Risk Measurement: Development over Last 20 Years. Journal of Banking and Finance 21, 11–12, 1721–1742 (1998)
8. Glynn, P.W.: Efficiency Improvement Techniques. Annals of Operation Research 53, 175–197 (1994)
9. Hui, K.-P., Bean, N., Kraetzl, M., Kroese, D.P.: The Cross-Entropy Method for Network Reliability Estimation. Annals of Operations Research 134, 101–118 (2005)
10. Bucklew, J.A.: Introduction to Rare Event Simulation [M]. Springer, Heidelberg (2004)
11. Townsend, J.K., Haraszti, Z., Freebersyser, J.A., Devetsikiotis, M.: Simulation of rare events in communications networks. IEEE Communications Magazine, 36–41 (1998)
12. Sadowsky, J.S.: Large Deviations Theory and Efficient Simulation of Excessivebacklogs in a GI/GI/m Queue. IEEE Transactions on Automatic Control, 1383–1394 (1991)
13. Villén-Altamirano, M., Martínez-Marrón, A., Gamo, J., Fernández-Cuesta, F.: Enhancement of the Accelerated Simulation Method RESTART by Considering Multiple Thresholds. In: Proceedings 14th International Teletraffic Conference, pp. 797–810 (1994)
14. Hsieh, M.: Adaptive Importance Sampling for Rare Event Simulation of Queuing Networks Ph. D. thesis, Stanford University (December 1997)
15. Odom, M.D., Sharda, R.A.: Neural Network Model for Bankruptcy Prediction[A]. In: Proceedings of the IEEE International Joint Conference on Neural Networks[C], vol. 2, pp. 163–168 (1990)
16. Jackson, P., Perraudin, W.: Regulatory implications of credit risk modeling. Journal of Banking and Finance 24, 1–14 (2000)
17. de Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A Tutorial on the Cross-Entropy Method[J]. Annals of Operations Research, 19–67 (2005)

18. Rubinstein, R.Y.: Optimization of Computer Simulation Models with Rare Events. European Journal of Operations Research 99, 89–112 (1997)
19. Rubinstein, R.Y.: The Simulated Entropy Method for Combinatorial and Continuous Optimization. Methodology and Computing in Applied Probability 2, 127–190 (1999)
20. Rubinstein, R.Y.: Combinatorial Optimization, Cross-Entropy, Ants and Rare Events. In: Uryasev, S., Pardalos, P.M. (eds.) Stochastic Optimization: Algorithms and Applications, pp. 304–358. Kluwer, Dordrecht (2001)
21. Piramuthn, S.: Financial credit risk evaluation with neural and neurofuzzy systems. European Journal of Operational Research 12, 310–321 (1999)
22. Tam, K.Y., Kiang, M.Y.: Managerial applications of neural networks: The case of bank failure predictions[J]. Management Science 38, 926–947 (1992)

Forecasting Model for IPTV Service in Korea Using Bootstrap Ridge Regression Analysis

Byoung Chul Lee, Seho Kee, Jae Bum Kim, and Yun Bae Kim*

Dept. of Systems Management Engineering, Syungkyunkwan University, 300, Cheoncheon-dong, Jangan-gu, Suwon, Gyeonggi-do, Korea
kimyb@skku.edu
http://iesys.skku.ac.kr/~simul

Abstract. The telecom firms in Korea are taking new step to prepare for the next generation of convergence services, IPTV. In this paper we described our analysis on the effective method for demand forecasting about IPTV broadcasting. We have tried according to 3 types of scenarios based on some aspects of IPTV potential market and made a comparison among the results. The forecasting method used in this paper is the multi generation substitution model with bootstrap ridge regression analysis.

1 Introduction

The advent of Internet Protocol Television (IPTV), which is a method of distributing television contents over IP [1], will change individual daily life dramatically by offering a glittering array of new services not available under the traditional broadcasting technologies.

IPTV is a packet-based real-time broadcasting system used to deliver various video information and on-demand digital contents over the Web and is regarded as the new generation broadcasting system.

In Korea, many telecom firms recognize and expect the great potential effect of IPTV. However, there exist skeptical predictions because there are so many uncertain factors with respect to the purchasing behavior of customers under the present telecom markets' environment. It makes marketers cannot assert that IPTV will result in substantial revenue from this new business.

So, it is essential to forecast the demand of IPTV by analyzing the market more systematically and considering more varied and robust scenarios which might happen in these markets.

This paper describes the analysis of effective methods for demand forecasting which we used, respectively, according to 3 types of scenarios based on some aspects of IPTV's potential market. We performed comparisons with the results of each of the methods with the others.

The findings of the analysis in this paper can help firms become ready to deploy IPTV, making their strategic decision with more certainly. The results of this paper might be applied with fundamental data in order to establish the communication broadcasting convergence policy of the government.

* Corresponding author.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 95–102, 2007.

2 Forecasting Potential IPTV Market

2.1 Forecasting Method for Potential IPTV Broadcasting Market

In the potential market, there are only a few of the existing services that may be substituted with IPTV. In the sense of understanding the development of broadcasting technology, CATV may be called the first generation, the next following generation is satellite TV, and IPTV as third generation technology. CATV has variety of channels that traditional ground wave TV does not have. Satellite TV has nationwide broadcasting network. However, both services are local broadcasting and single direction services. The IPTV appearance may change a total paradigm. Service providers are actively conducting trials of IPTV based services.

So, a multi-generation diffusion model is applicable for the broadcasting markets that each service is divided in each generation while considering technology and service differences.

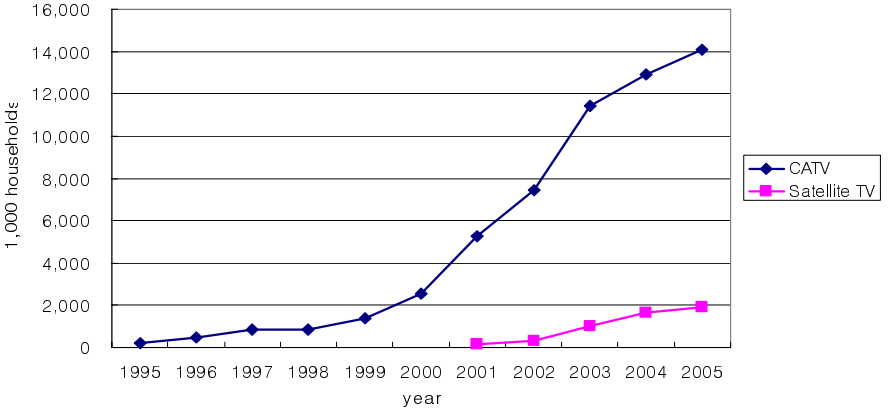


Fig. 1. The number of households that are Cable TV or Satellite TV subscribers

The Bass model has a customer's behavioral rationale. This rationale is consistent with social science studies and is based on the simple premise about the hazard-rate function (the conditional probability that an adoption will occur at time *t* given that an adoption has not yet occurred). The Norton and Bass multi-generation model is the extension of the classical Bass model to a multi-generation situation. In their model, the generations remain in the market and remain competing for market share. After the second generation is introduced in the market, the share of the first generation keeps growing by the first generation's potential market. The diffusion process of the first generation to the second generation is taking place, simultaneously [2].

The generalized Norton and Bass model for multi-generation can be expressed as follows:

$$S_1(t) = m_1 F_1(t)[1 - F_2(t - \tau_2)] \tag{1}$$

$$S_2(t) = F_2(t - \tau_2)[m_2 + m_1 F_1(t)][1 - F_3(t - \tau_3)] \qquad (t < \tau_2) \tag{2}$$

$$S_3(t) = F_3(t - \tau_3)[m_3 + F_2(t - \tau_2)][m_2 + m_1 F_1(t)] \qquad (t < \tau_3) \tag{3}$$

$$F_i(t) = \frac{[1 - e^{-(p_i + q_i)t}]}{[1 + (q_i / p_i)e^{-(p_i + q_i)t}]} \qquad (t < \tau_i) \tag{4}$$

$S_i(t)$: The number of adopters for the i th generation at time t

$F_i(t)$: The cumulative proportion of the adopters for the i th generation at time t

m : The potential maximum number of adopters

τ_i : The time of the i th generation introduced

2.2 Ridge Regression Analysis with Bootstrap Method

The coefficients of the multi-generation Bass model can be estimated usually by the polynomial regression analysis if the model is allowed to be dealt with the discrete form of the polynomial equations. However, multicollinearity exists because the regressor variables of the model are truly not independent.

Serious multicollinearity prohibits precise statistical inference, i.e., estimation of regression coefficient or perhaps prediction is rendered poor due to the condition. Thus, any attempt to highlight the individual roles of the variables is badly clouded [3].

Consider the regression from of our model

$$y = X\beta + u, \tag{5}$$

where y is a $T \times 1$ vector of the dependent variable, X is a $T \times p$ matrix of p regressors whose first column consists of all ones. The matrix X has rank p and u is a $T \times 1$ vector of independent and identically distributed (iid) errors with elements u, and an unknown, possibly nonnormal true distribution function F with mean zero and variance σ^2.

The ordinary least squares (OLS) estimator, b, is a special case of the ridge estimator, $b_\lambda = (X'X + \lambda I_p)^{-1}$, when the biasing parameter λ equals 0. The main motivation in using ridge regression is to deal with multicollinearity by using $\lambda > 0$, such that the mean squared error (MSE) is reduced. One recognizes a trade-off between variance and bias and injects small bias in exchange for a major reduction in variance. Among the various stochastic choices of λ a popular choice is due to Hoerl, Kennard, and Baldwin[4] : $\lambda_{hkb} = ps^2 / b'b$, where s2 is defined below. The λ_{hkb} depends on the y data through s2 and the denominator. The ridge residuals are defined as $\tilde{e} = y - Xb_\lambda$ whose mean $\bar{e}$ is not necessarily zero. Denote by $e = \tilde{e} - \bar{e}$ the recentered ridge residuals. The covariance matrix of b_λ is approximated by

$$\text{cov}(b_\lambda) \approx s^2 (X'X + \lambda I_p)^{-1} X'X (X'X + \lambda I_p)^{-1}, \tag{6}$$

$$s^2 = (e'e)/(T-p). \tag{7}$$

An empirical cumulative distribution function (CDF) of recentered ridge residuals puts probability mass *l/T* at each e_t and is denoted here by F_e. Now the basic bootstrap idea is to use F_e with mean zero and variance σ_e^2 as a feasible, approximate, nonparametric estimate of the CDF of the true unknown errors denoted by F_u. Let *J* be a suitably large number (= 1000, say). We draw .*J* sets of bootstrap samples of size *T*, with elements denoted by e_{*jt} (*j = 1, 2, . . . ,.J and t = 1,2, . . . , T*) from F_e using random sampling with replacement. This generates *J* sets of $T \times 1$ vectors denoted by e_{*j} having elements e_{*jt} (*t = 1,2, . . . , T*).

Hence the pseudo y data are obtained by

$$y_{*j} = Xb_\lambda + e_{*j} \tag{8}$$

for *j = 1, 2, . . . , .J* , yielding a large number *J* of regression problems to be used for bootstrap inference described below. Note that the bootstrap size *J* has been chosen to be a large number, consistent with the defining probability $P(e_{*jt} = e_t) = 1/T$.

Hence their variance is

$$\sigma_e^2 = E(e_{*jt}^2) = \sum_{t=1}^{T} (1/T)\,\mathrm{var}(e_t) = (1/T)\sum_{t=1}^{T} e_t^2 = s^2(T-p)/T \cdot \tag{9}$$

If each residual is 'scaled up' by multiplication by $\sqrt{[T/(T-p)]}$, the variance of the resealed e_{*jt} should be equal to s^2. Applying ridge regression to the pseudo *y* data leads to

$$b_{*j} = (X'X + \lambda I_p)^{-1} X' y_{*j}, \tag{10}$$

$$\mathrm{cov}(b_{*j}) \approx s^2 (X'X + \lambda I_p)^{-1} X'X (X'X + \lambda I_p)^{-1}, \tag{11}$$

for *j = 1,2, . . . ,J*. Although we have not included the third subscript λ in b_{*j}, we are bootstrapping the ridge estimator here.

In this paper, we estimate the regressor coefficient of our forecasting model by using such the bootstrap ridge regression analysis. In addition, we suppose that the saturation point of the subscript in each service is inferred from the number of the entire subscribers in the high speed internet market.

3 Designing Scenarios

Our work is composed of three of the scenarios (an optimistic, a realistic, and a conservative view) in order to observe various demand forecasting results in a future situation. The main factors of the scenario can be classified with the service quality side and the environmental side. The service quality side is the variety of contents, rate level, and broadband capacity. The other side is the thrust of governmental IPTV policy and society diffusion of IPTV.

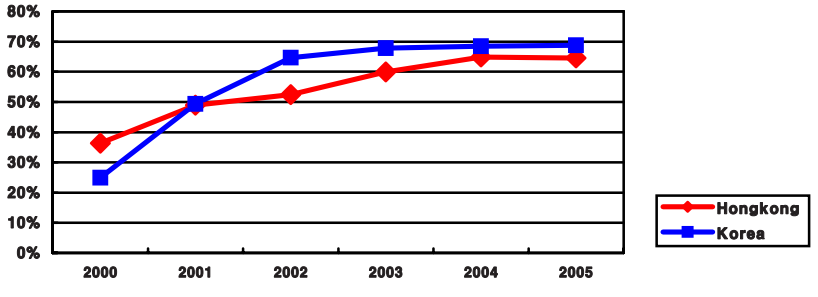
The broadband-powered offerings promise bigger possibilities for end users than traditional solutions. The primary plus of these offerings is that the service makes it

easier and more cost-effective to deliver a wide range of contents. Thus, IPTV will be activated, if it is generalized as the next generation broadband including using as delivery platforms, VDSL and FTTH. From the contents side, the offerings must diversify IPTV contents by repeatable ground wave broadcasting. The conservative viewpoint for IPTV service in Korea is contrary to optimistic conditions. Finally, the realistic view can be interpreted as taking a neutral attitude on the market status of IPTV.

Table 1. Factors of the scenarios

Factor		Scenarios in Korea		
		Optimistic	Realistic	Conservative
Service quality	Contents	Retransmitting broadcasts & Various Contents	Retransmitting broadcasts & general contents	Prohibition of retransmitting broadcasts & Poor Contents
	Charges	Free of a set-top box	Proper Charge for a set-top box	A high charge for a set-top box
	Broadband capacity	More than 6Mbps	6Mbps	Less than 6Mbps
IPTV environment	Government	Permit to 2007	Permit to 2007	Permit to 2007
	Society	The rate of joining to Paid TV: 85%		

In connection with our method, we wish to address some assumptions. These assumptions were applied to the Hong Kong subscriber data with reference data for the parameter presumption of the model.



We used this data because there is no past data about a Korea domestic IPTV market. It is highly probable that the broadcasting and telecommunication market of Hong Kong and Korea are structurally analogous to each other.

4 Forecasting Results

The demand forecasting results of these methods from three interpretations of IPTV service, in addition to being divided into three branches of market status in the future are now presented. We observe 4 million to 5 million subscribers (households) for the

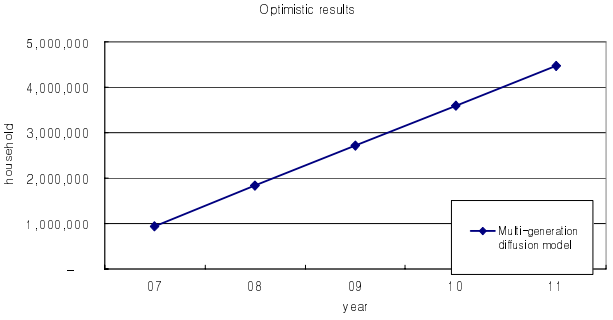


Fig. 2. The results in the optimistic view

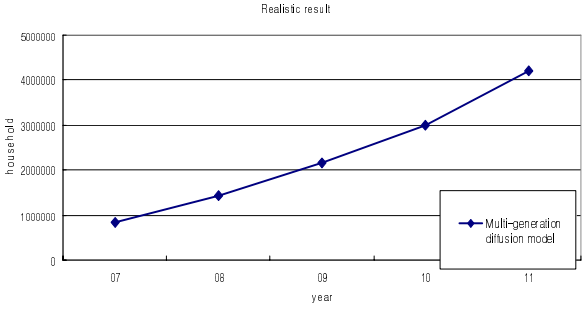


Fig. 3. The results in the realistic view

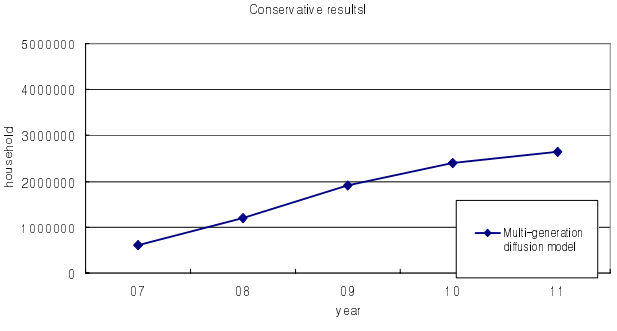


Fig. 4. The results in the conservative view

result of the optimistic view and 2 million to 4million subscribers (households) for the result of the conservative view until the year 2011.

With regard to forecasting methods, the results of the multi-generation diffusion model are the highest for the others' view. Both the optimistic and realistic views appear to have a gap of 1 million subscribers between the forecast methods. However, there is a gap of 2 million subscribers between the maximum and minimum values.

5 Discussion and Conclusion

Digital convergence brings together the separate worlds of audio, video, data, and voice communication services into one unified service. Even today, we can access the same services and contents (e-mail, music television) using different terminals over different types of networks. In a convergent world, IPTV, the convergence services of television and the Internet, is rapidly penetrating into the public around the world. Thus, consumers will be in control of their entertainment and media content. As a result, IPTV is a common theme among telecommunications and cable television suppliers. These companies think that by offering a portfolio of bundled voice, video, and data services, they will be able to gain customers and increase their market share.

In this study, we applied forecast method to the market for IPTV subscribers in Korea through the broadcasting market in the future. The forecast procedure was composed of three scenarios (an optimistic, a realistic, and a conservative view) to observe various demand forecasting results in the future. ***As a result, the IPTV potential market was estimated at about 4~5 million subscribers (households) by 2011 in Korea.***

For further study, there are at least a few remaining problems. This study begins with limited factual data (i.e. contents charge and highly selected content or program features). The issue of economies of scale and the competitions among IPTV providers need further study.

References

1. Shin, D.H.: Potential user factors driving adoption of IPTV: What are customers expecting from IPTV, Technological Forecasting & Social Change (in press)
2. Kim, Y.B., Seo, S.Y., Lee, Y.T.: A Substitution and Diffusion Model with Exogenous Impact: Forecasting of IMT-2000 subscribers in Korea, Vehicular Technology Conference – Fall, pp. 948–952 (1999)
3. Myers, R.H.: Classical and modern regression with applications. PWS Publishers (1986)
4. Hoerl, A.E., Kernnard, R.W., Baldwin, K.F.: Ridge regression: some simulations. Communications in Statistics A4 (1975)
5. Shin, Y.H., Park, Y.J., Choi, M.K.: Study on the IPTV Demand Forecasting Using Markov Chain and Logistic Model. In: 2005 KMIS International Conference, pp. 638–640 (2005)
6. Joo, C.M.: A Study of Characteristics Integrating Broadcasting and Telecommunication of IPTV and Introduction Plans Thereof. The Korean Journal of Cyber Communication Society 15, 161–197 (2005)
7. Kang, J.W., Lee, S.W.: A Convergence-Adoption Model for IPTV Use: Functional Similarity & Media Substitution. The Korean Journal of Journalism & communication studies 50(2), 5–33 (2006)
8. Kim, T.G.: South Korea Turns on Internet-Based TV, Korean Times, 10-09-2006
9. Kim, Y.B., Seo, S.Y., Lee, Y.T.: A Substitution and Diffusion Model with Exogenous Impact: Forecasting of IMT-2000 subscribers in Korea, Vehicular Technology Conference – Fall, pp. 948–952 (1999)
10. Hwang, S.Y.: KT aims to be a leader in TV service on internet, The Korea Herald, 11-27-2006

11. Dalkey, N., Helmer, O.: An experimental application of the Delphi method to the use of experts. Management Science 9(3), 458–467 (1963)
12. Okoli, C., Pawlowski, S.D.: The Delphi method as a research tool: an example, design considerations and applications. Information & Management 42, 15–29 (2004)
13. Banuls, V.A., Salmeron, J.L.: A Scenario-Based Assessment Model—SBAM, Technological Forecasting & Social Change (in press)
14. Norton, J.A., Bass, F.M.: A Diffusion Theory Model of Adoption and Substitution for Successive Generations of High-Technology Products. Management Science 33(9), 215–227 (1987)
15. Meade, N., Islam, T.: Forecasting with growth curves: An empirical comparison. Elsevier Science B.V., Amsterdam (1995)
16. Norton, J.A., Bass: A Diffusion Theory Model of Adoption and Substitution for Successive Generations of High-Technology Products. Management Science 33(9) (1987)
17. Mahajan, V., Sharma, S.: A Simple Algebraic Estimation Procedure for Innovation Diffusion Models of New Product Acceptance. Technological Forecasting and Social Change 30, 331–345 (1986)
18. Linstone, H.A., Turoff, M.: The Delphi Method: Technology and Application. Addison-Wesley, Reading (1975)

Simulation and Evaluation on the Performance of the Proposed Constellation of Global Navigation Satellite System

Guangming Liu*, Ying Liao, Yuanlan Wen, Jun Zhu, and Xiangjun Feng

College of Astronautics and Material Engineering, National Univ. of Defense Technology, Changsha 410073, China
mingandfang79@yahoo.com.cn

Abstract. The appropriated deployment of the proposed constellation of global navigation satellite system (GNSS) will offer improvements in accuracy and reliability compared with GPS. The availability, accuracy and reliability evaluation for the new constellation versus GPS were presented using continuous coverage, navigation accuracy and position dilution of precision (PDOP). Simulation results show that the navigation performance of the proposed constellation is improved evidently, and they prove that the designed constellation concept is feasible and rational.

Keywords: new constellation of GNSS, GPS, position dilution of precision (PDOP), continuous coverage, navigation accuracy.

1 Introduction

The development of global navigation satellite system (GNSS) and wireless communication systems have made instantaneous positioning, navigation, and communication available at all times almost everywhere on Earth. Now there are two global navigation satellite systems, GPS in U.S.A. and GLONASS in Russia. But they provide low accuracy positioning service for civil user in peacetime. So it is necessary to build a global navigation satellite system on Chinese own account.

The GALILEO System will be an independent, European-controlled, satellite-based global navigation system and will provide a number of guaranteed services to users equipped with GALILEO-compatible receivers [1]. The Japanese Quasi-Zenith Satellite System (QZSS) is under investigation through a cooperative effort between the government and the private sector. QNZZ is considered a multi-function satellite system, as it is able to provide communication, broadcasting, and positioning services for mobile users in a specified region with a high elevation angle [2].

Based on the navigation performance of global navigation satellite system, the paper designs a global navigation satellite constellation which will provide continuous, 3-D,

* Corresponding author.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 103–111, 2007.

and high accuracy navigation service. The predominance of the designed global navigation satellite system is proved by simulating and analyzing continuous coverage, constellation availability and navigation accuracy compared with the exiting GPS constellation. The paper will then continue by defining some important performance measures for a GNSS.

2 Performance Measures for GNSS

The performance measures for GNSS are foundations of constellation design and navigation performance evaluation, and they mainly include continuous coverage, space formation, and constellation availability etc.

(1) Continuous Coverage

The GNSS should provide more than 4 navigation signals continuously for the coverage areas. If the higher demand of the positioning accuracy and reliability is put forward by users, or receiver autonomous integrity monitoring (RAIM) is considered, the capability of the GNSS providing more than 6 navigation signals continuously should be achieved [3]. Furthermore, the user receiver elevation cut-off angle should be taken into account for the continuous coverage of the GNSS.

(2) Space Formation

The geometrical intensity of the designed navigation constellation and the space distribution of the satellites in some observation time are usually evaluated by geometric dilution of precision (GDOP). GDOP is the satellite geometry dependent quantity that maps the user equivalent range error (UERE) into the user accuracy. GDOP can also be decomposed into parts, namely position dilution of precision (PODP), horizontal dilution of precision (HDOP), vertical dilution of precision (VODP), and time dilution of precision (TDOP) respectively. Dilution of precision (DOP) is good measures of system availability as they represent the geometric strength of the solution [4]. DOP values can also be used to represent system accuracy when it is assumed that all range measurements have the same UERE. In many navigation situations, the user is interested in using a GNSS to obtain a horizontal position. For this reason, PDOP will be used as a measure of availability and accuracy.

Users' positioning accuracy is directly decided by UERE and PDOP. UERE represents the physical performance of system signal, and PDOP represents the geometrical intensity of satellites' space distribution. If UERE is a certain value, smaller is the value of PDOP, higher is the user's positioning accuracy. PDOP is a key index of performance evaluation, which includes the characteristic performance of navigation constellation's indexes of performance evaluation else. So, PDOP can demarcate constellation coverage area and analyze constellation availability.

(3) Constellation Availability

Currently, the navigation constellation availability can be represented by DOP or positioning accuracy. The constellation availability analysis usually adopt reference grids method which counting all of single grid availability for reference grids by certain longitude-latitude interval.

The navigation constellation availability can be defined as the time percent of the provided PDOP less than a certain given value in the coverage area. So, the PDOP availability of the navigation constellation can be computed using the following equation.

$$A_j = \frac{N_j^n}{N_j} \times 100\% \tag{1}$$

where A_j is the constellation availability for $PDOP \le n$ of grid point j on the coverage area, N_j is the time statistic for $PDOP$ of grid point j in an observation period, N_j^n is the time statistic for $PDOP \le n$ of grid point j in an observation period, and n is the threshold value for PDOP of the constellation availability.

3 Simulations Scenarios

3.1 Constellation Design

Nowadays, the circular orbit satellite is adopted well by the GNSS in use, such as GPS in U.S.A. and GLONASS in Russia. Because the circular orbit satellite can ensure the altitude and the coverage area maintenance a certain value, and the circular orbit satellite can ensure the strength of the signal received by users maintenance approximately equality.

Consequently, the designed constellation of global navigation satellite system (CGNSS) consisted of 36 medium earth orbiting (MEO) satellites in 6 orbital planes with 60 degrees inclination. For the simulation, 6 satellites are assumed to be equally spaced in each plane (60 degrees between satellites). The ascending nodes of the six orbital planes of the simulated constellation are arbitrarily assigned right ascensions of 0, 60, 120, 180, 240 and 300 degrees, respectively.

The GPS constellation used is the 'real' GPS constellation as it existed at the beginning of GPS almanac from GPS Week 1089 (November 19, 2000) [5]. This is dome by using the actual GPS almanac from that week. The constellation consists of 27 satellites unequally space in six orbital planes. The orbital parameters for each satellite of CGNSS and GPS are shown in Table 1.

Table 1. The orbital parameters for CGNSS and GPS options

Constellation Option	CGNSS	GPS
Element Type	MEO	MEO
Number of Satellites	36	27
Number of Orbital Planes	6 (plane spacing: 60°)	6 (plane spacing: 60°)
Number of Satellites in an Orbital Plane	6 (phase spacing: 60°)	4 or 5 (phase spacing is unequally)
Satellite Altitude	20000 km	20230 km
Orbit Inclination	60°	55°
Orbit Period	655 min	718 min
Orbit Eccentricity	0.001	0.001

3.2 Observation Scenarios

The simulations are conducted to evaluate the navigation performance of CGNSS and GPS, respectively, furthermore, to compare CGNSS with GPS. In the simulations described in this section, it is assumed that only the two-way range observations are used. Users on the earth surface are assumed to be stationary and are estimating their three-dimensional position vector $(x, y, z)^T$. The users are both attempting to compute their instantaneous positions, when sufficient observations are available, and are accumulating observations over time. Each observation contributes one row to the design matrix, A, which has the form using the following equation [6].

$$A_i = \left[\frac{x_s - x_r}{\rho}, \frac{y_s - y_r}{\rho}, \frac{z_s - z_r}{\rho} \right] \tag{2}$$

where the s and r subscripted coordinates correspond to the satellite and receiver, respectively, and ρ is the range between the satellite and receiver.

The theoretical navigation performance of a GNSS can be evaluated based on the satellite-user geometry using assumptions about the quality of the observations. Simulated users are positioned at 10° intervals in latitude and longitude around the Earth. Each user attempts to observe all passing satellites every 1 min. The user receiver elevation cut-off angle is employed 15°.

4 Simulations Results

4.1 Continuous Coverage

Figure 1 shows the statistic levels of the global continuous coverage on CGNSS, in which the maximum and the minimum of the statistic is 12 and 6 in a sidereal day, respectively, and the curves of the statistic go without wild fluctuation. So, the designed constellation of CGNSS can provide 6 navigation signals continuously for the global users completely.

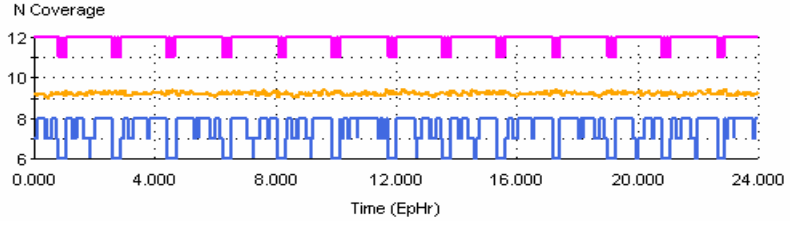


Fig. 1. The statistic of the global continuous coverage on CGNSS. The abscissa of the coordinate represents the simulation time from the start, and the unit of the simulation time is hour. The three curves in this figure represent the maximum, the average, and the minimum the statistic levels from the top down.

Figure 2 shows the statistic of the global continuous coverage on CGNSS by geographic latitude, in which the statistic increases from the earth's equator to North Pole and South Pole. The minimum statistic is 9, and the maximum statistic is 11, so the CGNSS can provide navigation signals continuously for the polar region users completely.

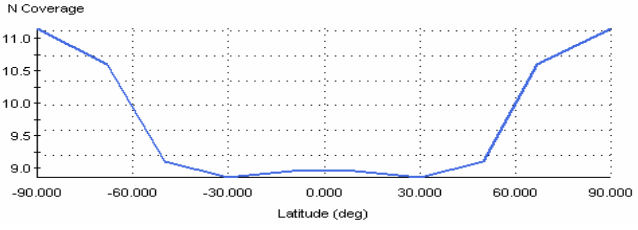


Fig. 2. The statistic levels of the global continuous coverage on CGNSS by geographic latitude

Figure 3 shows the statistic levels of the global continuous coverage on GPS, in which the maximum and the minimum of the statistic is 9 and 2, respectively, and the average is 6 in a sidereal day. So, the designed constellation of GPS can't provide 6 navigation signals continuously for the global users completely. The coverage gap with 6 occurs in 65% area, and the coverage gap with 4 occurs in 8.5% area. The existence of the coverage gap will reduce the accuracy of positioning continuously, and it is one of reasons that GPS need to be further improved.

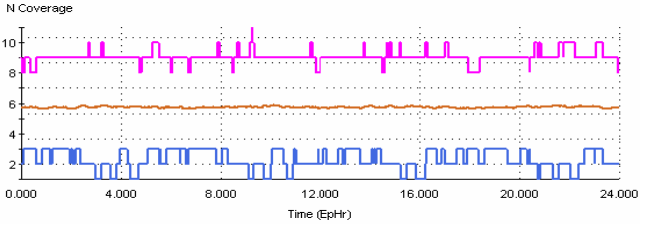


Fig. 3. The statistic levels of the global continuous coverage on GPS

4.2 Constellation Availability

Figure 4 shows the statistic satisfaction with PDOP in the global reference grids on CGNSS, and the threshold of the satisfaction is $PDOP \leq 4$. The constellation availability on CGNSS is 86.65% by simulation.

Figure 5 shows the statistic PDOP on CGNSS by geographic latitude, in which the statistic PDOP increases speedily form the middle-latitude area to North Pole and South Pole, and the maximum PDOP reach 140 indeed. When $PDOP \geq 10$, the observation equation gets ill-conditioning, and the user's positioning accuracy is hard to achieve. So, the geometrical space configuration on CGNSS is not good, and it needs to be improved.

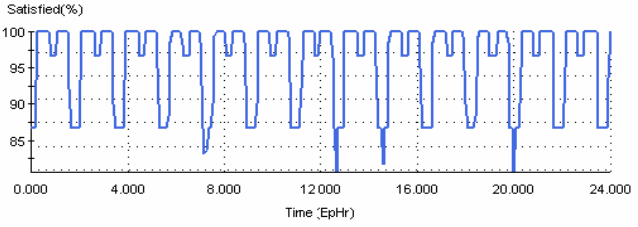


Fig. 4. The statistic satisfaction with PDOP in the global reference grids on CGNSS

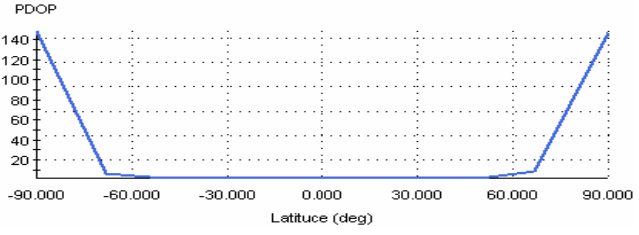


Fig. 5. The statistic PDOP on CGNSS by geographic latitude

Figure 6 shows the statistic PDOP on GPS by geographic latitude. In this figure, comparatively, the curve of statistic PDOP goes without wild fluctuation, and the maximum PDOP is 7. The constellation availability with $PDOP \leq 4$ on GPS is only 38.63%.

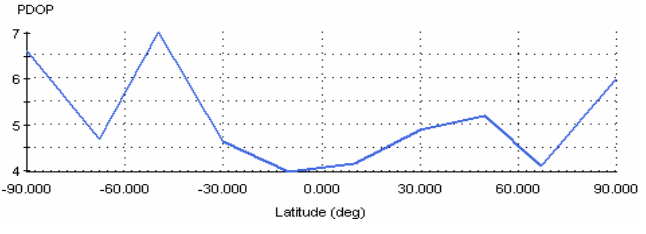


Fig. 6. The statistic PDOP on GPS by geographic latitude

Analyzing the space configuration of GPS, the space distribution of CGNSS is improved. The satellites in the second, third, forth, fifth and sixth planes are advanced by 30, 60, 90, 120, 150 degrees with respect to the satellites of the first plane.

Figure 7 shows the statistic PDOP on the improved CGNSS by geographic latitude, in which the curve of statistic PDOP goes stable, and the maximum PDOP is only 2.6. The constellation availability with $PDOP \leq 4$ on improved CGNSS is 100%, so the geometrical intensity of the space configuration on improved CGNSS is heightened evidently.

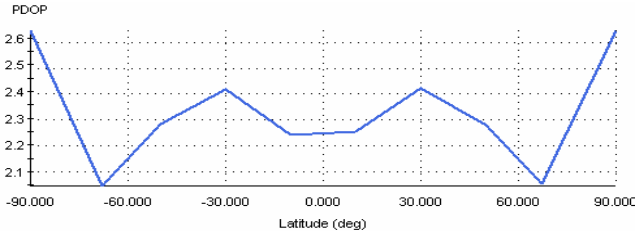


Fig. 7. The statistic PDOP on improved CGNSS by geographic latitude

4.3 Navigation Accuracy

The simulations adopt instantaneous positioning with broadcast ephemeris using C/A pseudo-ranges with the equivalent range error 13.9 meters.

Figure 8 shows the statistic navigation accuracy on the improved CGNSS by time. In this figure, the curve of statistic navigation accuracy goes stable, and the average and the minimum navigation accuracy is 12 meters and 7 meters respectively in a sidereal day.

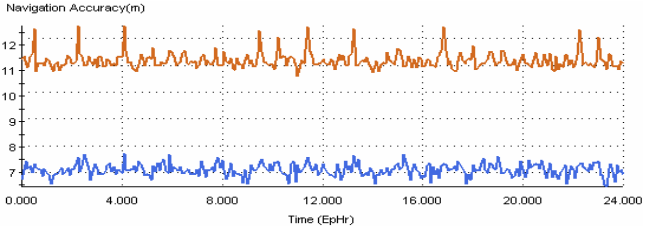


Fig. 8. The statistic navigation accuracy on the improved CGNSS. The two curves in this figure represent the average and the minimum of the statistic navigation accuracy from the top down.

Figure 9 shows the statistic navigation accuracy on the improved CGNSS by geographic latitude, in which the curve of statistic navigation accuracy goes stable, and the maximum navigation accuracy is only 13 meters. The constellation availability with the navigation accuracy less than 15 meters on improved CGNSS is 100% . So, the improved CGNSS can provide high accuracy navigation service for the global users.

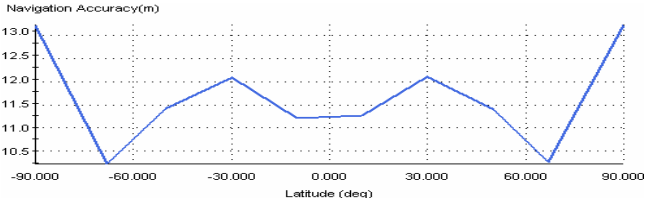


Fig. 9. The statistic navigation accuracy on the improved CGNSS by geographic latitude

Figure 10 shows the statistic navigation accuracy on the GPS, in which the curve of statistic navigation accuracy goes with some fluctuation compared with the improved CGNSS. The statistic navigation accuracy is lower, and the average navigation accuracy is about 50 meters in a sidereal day.

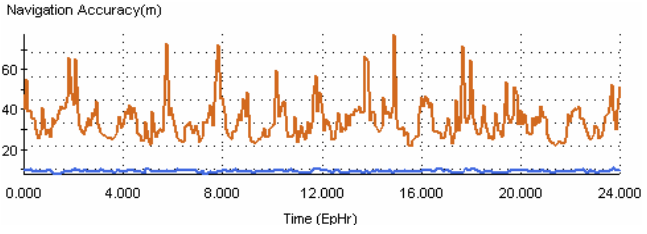


Fig. 10. The statistic navigation accuracy on the GPS. The two curves in this figure represent the average and the minimum of the statistic navigation accuracy from the top down.

Figure 11 shows the statistic navigation accuracy on the GPS by geographic latitude, in which the statistic navigation accuracy on low-latitude is 22 meters, and the statistic navigation accuracy on high-latitude is 33 meters, moreover, the statistic navigation accuracy on middle-latitude is 40 meters. The constellation availability with the navigation accuracy less than 15 meters on GPS is only 51.14% . So, compared with GPS, the improved CGNSS can provide high accuracy navigation service for the global users.

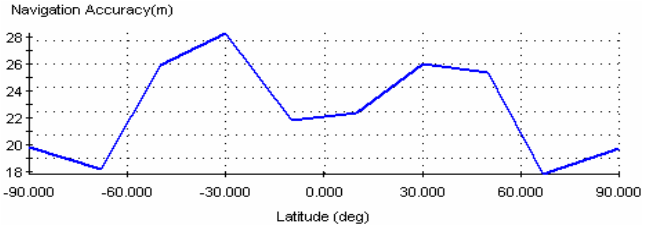


Fig. 11. The statistic navigation accuracy on GPS by geographic latitude

5 Conclusions

The design demand of the proposed constellation of global navigation satellite system is to provide more than 6 navigation signals continuously and to achieve the good space configuration and high navigation accuracy using fewer satellites compared with GPS. The simulations show that the performance indexes on the CGNSS are satisfied, and the designed constellation can provide high precision and successional navigation service for global users.

References

[1] O'Keefe, K., Lachapelle, G., Skone, S.: Assessing the Global Availability and Reliability of the Mars Network, a Proposed Global Navigation Satellite System for Mars. Canadian Aeronautics and Space Journal 51(1), 23–34 (2005)
[2] Falin, W.U., Kubo, N., Yasuda, A.: Performance Evaluation of GPS Augmentation Using Quasi-Zenith Satellite System. IEEE Transactions on Aerospace and Electronic Systems 40(4), 1249–1261 (2004)
[3] Yuling, D.U., Zhang, X., Zhang, J.: Research on New Constellation of Regional Satellite Navigation System Covering China. Computer Simulation 24(2), 49–52 (2007) (in Chinese)
[4] Parkinson, B.W., Spilker, J.J.: GPS: Theory and Applications Volume I. Progress in Astronautic and Aeronautics 163 (1996)
[5] Shuai, P., Guangji, Q.U., Chen, Z.: Studies on the design and analysis of regional navigation constellations. Chin. J. Space Sci. 26(4), 268–276 (2006)
[6] Chavez, F.R., Lovell, T.A.: Relative-Orbit Element Estimation for Satellite Navigation and Guidance. In: AIAA/AAS Astrodynamics Specialist Conference and Exhibit, Rhode Island (2004)
[7] Liu, G., Wen, Y., Liao, Y.: Research on Dynamic Simulation of Satellite Based on Multi Software Platform. Journal of System Simulation 19(2), 308–311 (2007)

Re-entry Tracking Control Via State-Dependent Riccati Equations

Chen Gao[1], Huifeng Li[2], and Ping Zhang[1]

[1] Automatic Science and Electrical Engineering School, Beihang University, Beijing 100083
[2] School of Astronautics Beihang University Beijing 100083
{Chen GAO,Huifeng LI,Ping ZHANG gaochen1981}@gmail.com

Abstract. Sate-Dependent Riccati Equation (SDRE) method is applied here to track optimal entry trajectory which can be generated online by On Board Three Dimension Trajectory Generation. In this paper, the design of a longitudinal entry tracking control law based on SDRE was proposed in detail. Since its control value and weigh metric depend on the current state which can be a good and precise compensator to disturbance. Besides, from the design procedure, SDRE method is very easy to implement whether on the mathematical deduction or the online calculation. Numerical simulation for SDRE tracking law was carried out, and compared to traditional tracking method to demonstrate the simplicity, robustness, and doable performance of SDRE controller.

Keywords: Re-entry tracking control, SDRE, Trajectory tracking.

1 Introduction

Re-entry tracking control of Reusable Launch Vehicle (RLV) is a very challenging task since the vehicle dynamics changes dramatically as the altitude and Mach number vary from atmospheric entry at hypersonic speed to subsonic approach and landing[1].

With respect to tracking control, traditional method is gain schedule, which is the wide-used control method applied in present Reusable Launch Vehicle such as space shuttle in US. However, its draw back is obvious due to its time-consuming, laborious and less robust. In this situation, various advanced control method is attempted in this field. Jeff Caplin chose dynamic inversion which usually depends on the knowledge of the system dynamics and parameters. Since perturbation during the entry flight is often unpredictable and inevitable, dynamic inversion has to combine with other methods. [1] Yuri Shetessel adopted sliding mode control to deal with this problem because of its good robustness property. However, "perfect" control requires high actuator power which may not be available, and chattering of control input which may be harmful for the system. [2]Besides, adaptive control is very efficient to deal with parameter uncertainties, but for unmodeled dynamics such as disturbance, it will not be a good compensator. [3]

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 112–118, 2007.

Recently a methodology named State Depended Riccati Equation (SDRE) has emerged as an effective method for nonlinear system which expands tradeoff flexibility design as linear H2 to nonlinear field. [4] SDRE has the same structure as the linear Quadratic regulator (LQR), but the coefficient matrices are the function of current state. Thus controller can be designed for specific state which is the robust character of this method. Furthermore, since it is LQR based design, it may inherit the tradeoff character between state error and control input, which is not readily available in other nonlinear control methods. In addition, because the penalty metrics Q and R can be chosen as state-depended, this gives the designer more intuitive control than traditional tradeoff in LQR method. [3]

In this paper, SDRE was adopted as tracking control strategies to track the real-time optimal reference trajectory. An overview of SDRE is presented in detail in section 2. In section 3, the dynamic equations of re-entry vehicle were presented. Then SDRE technique was applied in tracking control was introduced in detail and in section 4 simulation results are analyzed. Finally, conclusion follows in section 5.

2 The SDRE Method

Consider the general infinite-horizon of autonomous nonlinear systems affine in control, given the nonlinear differential equation:

$$\dot{x} = f(x) + g(x)u \tag{1}$$

The performance index i.e. the cost function can be written:

$$J = \frac{1}{2}\int_0^\infty (x^T Q(x)x + u^T R(x)u)dt \tag{2}$$

Where, $x \in R^n, f \in R^n, B \in R^{n\times m}, u \in R^m, Q \in R^{n\times n}, R \in R^{m\times m}$ and the assumption is $f(0) = 0, g(x) \neq 0$. By mimic the design proceeding of LQR formulation, SDRE method can be obtained:[5][6]

1) Use direct parameterization to bring the nonlinear dynamics to the state-dependent form:

$$\dot{x} = A(x)x + B(x)u \tag{3}$$

2) Follow the LQR design proceeding, the state-depended feedback controller can be obtained:

$$u(x) = -K(x)x = -R^{-1}(x)B^T(x)P(x)x \tag{4}$$

The symmetric positive-definite P(x) can be obtained via state depended Riccati Equation:

$$\begin{aligned} &A^T(x)P(x) + P(x)A(x) + Q(x) \\ &- P(x)B(x)R^{-1}(x)B^T(x)P(x) = 0 \end{aligned} \tag{5}$$

From the above-mentioned mathematical deduction, we have to notice that linear-like form is not like the traditional linearization, but just like an algebraic formulation which can avoid inevitable errors in linearization. Furthermore, it is obvious that for each A(x) choosing, there will be different P(x) and controller u. theoretically, there exists only one P(x) yields to the performance index, which is regarded as optimal value. Other P(x) and control value are "suboptimal". Usually, performance of P(x) depends on the A(x) chosen, and it is difficult to choose A(x) yield to optimal control, which seems to be a potential drawback of SDRE method. However, compared to benefits such as flexible design in several investigations, the limitation is rather insignificant. The only criterion of choosing [A(x),B(x)] is to guarantee controllability for all state x in the domain of interest.

By choose suitable state-depended weight metric Q(x) and R(x), the designer can directly control the performance. Since the weight metric can vary as function of state, different modes of behavior can be imposed in different regions of the state-space.[3]

3 Longitudinal Tracking Control Design

In this section the longitudinal dynamics of Re-entry is chosen as the mathematical model since it is adopted by On Board Three Dimension Trajectory Generation, [7] which is our reference trajectory. The Longitudinal state variables are chosen as $\{r\,\gamma, s_{togo}, \int s_{togo}\}$, control variable are chosen as $\{\sigma, \alpha\}$and longitudinal dynamic is shown as follow: [7]

$$\begin{cases} \dot{r} = \dfrac{\sin\gamma}{D} \\ \dot{\gamma} = \dfrac{1}{V^2 D}[L\cos\sigma + (V^2 - \dfrac{1}{r})(\dfrac{\cos\gamma}{r})] \\ \dot{s} = -\dfrac{\cos\gamma}{rD} \\ \int \dot{s} de = s \end{cases} \tag{6}$$

Where r is the radial distance from the center of the earth to the vehicle, γ is the flight path angle, s_{togo} is the range from the vehicle to the tangency of the heading alignment cone (HAC) near an end of the runway, and $\int s_{togo}$ is the integral value of s_{togo} with respect to the energy e.

According to the SDRE design procedure, first of all, we will rewrite the longitudinal dynamics differential equation as linear-like structure via direct parameterization:

$$\begin{bmatrix} \dot{r} \\ \dot{\gamma} \\ \dot{s} \\ \int \dot{s} \end{bmatrix} = \begin{bmatrix} 0 & \dfrac{\sin\gamma}{D\gamma} & 0 & 0 \\ \dfrac{L(\cos\sigma+\sigma)}{V^2 D\cdot r} & \dfrac{-\cos\gamma}{DV^2 r^2\gamma} & 0 & 0 \\ \dfrac{-(\cos\gamma+\gamma)}{r^2 D} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} r \\ \gamma \\ s \\ \int s \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \dfrac{-L}{V^2 D} & \dfrac{-\cos\gamma}{D\cdot r\cdot\alpha} \\ 0 & \dfrac{\gamma}{D\cdot r\cdot\alpha} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \sigma \\ \alpha \end{bmatrix} \tag{7}$$

In order to guarantee the valid solution of SDRE, controllability of the pair {A(x), B(x)} should be inspected:

The controllability matrix is shown as follow:

$$W_{ctrb}(x) = [B \mid AB] = \begin{bmatrix} 0 & 0 & \dfrac{-L\sin\gamma}{V^2 D^2\gamma} & \dfrac{-\cos\gamma\sin\gamma}{D^2 r\gamma\alpha} \\ \dfrac{-L}{V^2 D} & \dfrac{-\cos\gamma}{D\cdot r\cdot\alpha} & \dfrac{-L^2(\cos\sigma+\sigma)}{V^4 D^2 r} & \dfrac{(\cos\gamma)^2}{D^2 V^2 r^2\alpha} \\ 0 & \dfrac{\gamma}{D\cdot r\cdot\alpha} & 0 & 0 \\ 0 & 0 & 0 & \dfrac{\gamma}{D\cdot r\cdot\alpha} \end{bmatrix}$$

It is full ranks for all state x in the domain of interest, as the result, it is doable to choose A(x), B(x) as this form.

Then follow the steps given by SDRE in section 2, the nonlinear feedback controller equation is obtained finally:

$$u = -R^{-1}B(x)^T P(x)[r_{c-}r_s, \gamma_c - \gamma_s, s_c - s_s, \int s_c - \int s_s] \tag{8}$$

Where P(x) is obtained by State-Depended Riccati Equation:

$$\begin{aligned} &A^T(x)P(x) + P(x)A(x) + Q(x) - \\ &P(x)B(x)R^{-1}(x)B^T(x)P(x) = 0 \end{aligned} \tag{9}$$

The subscript s and c are real and control value respectively.

4 Simulation Result and Analysis

The reference tracking trajectory comes from On Board Three Dimension Trajectory Generation, [7] in which X 33 aerodynamic model was chosen. After a real time optimal re-entry trajectory is generated by [7], SDRE based tracking controller will control the aircraft to track it. During the simulation, we adopt SDRE and traditional gain-schedule method together under the same disturbance to compare their robustness and tracking precision.

Entry condition is shown below:

$$h = 121518m, V = 7622m/s,$$

$$\gamma = -1.4379^\circ, \alpha = 45^\circ, \sigma = 0$$

The disturbance:

CL/CL*=-0.1,CD/CD*=0.1.

Where CL and CD are lift coefficient and drag coefficient respectively and CL* and CD* are normal lift coefficient and drag coefficient respectively. In other word, during the simulation, the values of aerodynamic coefficients such as lift coefficient CL, drag coefficient CD used in the simulations are perturbed by constant percentages up to $\pm 10\%$ relative to their nominal value, which is used to inspect the robustness of the flight control system.

Under this situation, the following figures give the tracking performance via simulation between Traditional and SDRE method.

The state tracking performance:

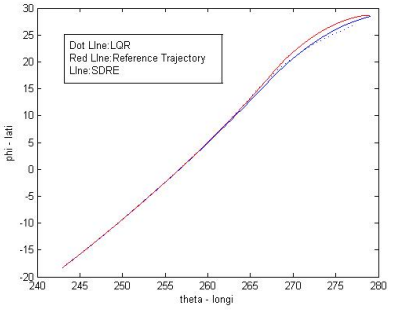


Fig. 1. Longitude VS Latitude

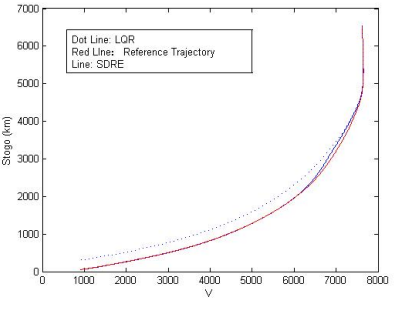


Fig. 2. Longitude VS Latitude

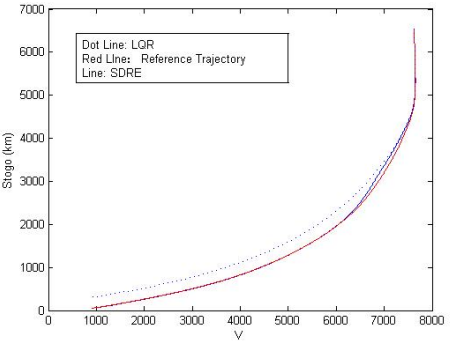


Fig. 3. S_togo VS Velocity

The control value:

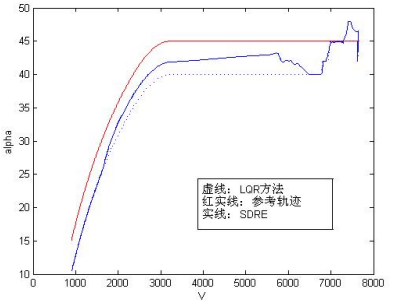


Fig. 4. Attack angle

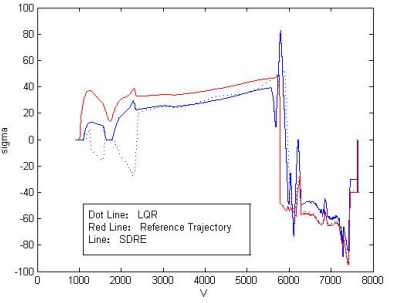


Fig. 5. Bank Angle

From these curves, it is obvious to find that under certain perturbation SDRE achieves more accurate tracking result for the Re-entry phase of the reusable launch vehicle and being more robust to external disturbances and plant uncertainties than traditional gain schedule method.

With respect to control value, red line is the recommended control value which is regarded as the performance index. It is obvious that SDRE control value gives more perfect tracking than traditional method.

In addition, the following table shows the final tracking error under different disturbance while the tracking figure of them will no be shown here due to limited article length.

Table 1. Comparison of TAEM Condition Errors with Aerodynamic Dispersions

Dispersions	Method	Range_err	Altitude_err
CL/CL*=-10% CD/CD*=+10%	SDRE	0.345km	-0.257km
	Conventional	3.781km	-1.247km
CL/CL*=+20% CD/CD*=-20%	SDRE	0.423km	-0.354km
	Conventional	10.700km	-11.905km
CL/CL*=-30% CD/CD*=+30%	SDRE	0.545km	-0.357km
	Conventional	13.781km	-16.247km

From this table, we can easily find out that SDRE can high tracking precision and a good compensator to disturbance.

5 Conclusions

This paper introduces a novel method to deal with entry tracking control method. Since its control value and weigh metric depend on the current state which can be a

good and precise compensator to disturbance. Besides, from the design procedure, SDRE method is very easy to implement whether on the mathematical deduction or the online calculation. Finally, the performance such as simplicity, robustness, and flexibility are demonstrated from the simulation result.

References

1. Shtessel, Y., Krupp, D.: Sliding Mode Control of Reusable Launch Vehicle in Launch and Re-entry mode. [C] IEEE Transaction on Control Systems Technology (1999)
2. Caplin, J.: Descent-phase Control of a Reusable Launch Vehicle. [C] IEEE Transaction on Control Systems Technology (2002)
3. Evrin, B.E.: Analysis and Implementation of State-Dependent Riccati Equation Controlled System [D], pp. 36–38. University of Illinois at Urbana-Champaign (2001)
4. Erdem, E.B., Alleyne, A.G.: Design of a Class of Nonlinear Controllers via State Dependent Riccati Equations[C]. IEEE Transaction on Control Systems Technology (2004)
5. Cloutier, J.R., Stansbery, D.T.: Nonlinear Hybrid Bank-to-Trun/ Skid-to-Turn. Missile Autopilot Design[C] AIAA Guidance, Navigation and Control Conference and Exhibit (2001)
6. Hammett, K.D., Hall, C.D., Ridgely, D.B.: Controllability issues in nonlinear the State-Dependent Riccati Equation control. Journal of Guidance Control & Dynamics 21(5), 767–773 (1998)
7. Shen, Z., Lu, P.: On-Board Generation of Three-Dimensional Constrained Entry Trajectories. [J] Journal of Guidance Control and Dynamics 26(1), 111–121 (2003)

A Data Envelopment Analysis Model for Selecting Material Handling System Designs

Fuh-Hwa Franklin Liu[1,2] and Wan-Ting Kuo[2]

[1,2] Department of Industrial Engineering and Management, National Chiao Tung University, Taiwan (R.O.C.)
fliu@mail.nctu.edu.tw
[2] Department of Industrial Engineering and Management, National Chiao Tung University, Taiwan (R.O.C.)
fruitli0809.iem92@nctu.edu.tw

Abstract. The material handling system under design is an unmanned job shop with an automated guided vehicle that transport loads within the processing machines. The engineering task is to select the design alternatives that are the combinations of the four design factors: the ratio of production time to transportation time, mean job arrival rate to the system, input/output buffer capacities at each processing machine, and the vehicle control strategies. Each of the design alternatives is simulated to collect the upper and lower bounds of the five performance indices. We develop a Data Envelopment Analysis (DEA) model to assess the 180 designs with imprecise data of the five indices. The three-ways factorial experiment analysis for the assessment results indicates the buffer capacity and the interaction of job arrival rate and buffer capacity affect the performance significantly.

1 Introduction

The material handling system under design is an unmanned job shop with an automated guided vehicle that transport loads within the processing machines. The four design factors are the following: the ratio of processing time to minimum transportation time (P/T ratio), mean job arrival rate γ, the buffer spaces for the incoming and outgoing jobs at each processing machine, C, and vehicle control strategies, S.

With the combination of these four factors, 180 possible design alternatives are experimented on through simulations. Each design alternative is simulated to collect the upper and lower bounds of the five performance indices in 30 repeated simulation runs, which are as follows:

PI_1: the average riding time (seconds) of one job on the vehicle per trip
PI_2: the frequency of machine blockage occurring
PI_3: the average duration (seconds) of each machine blockage event
PI_4: the shop throughput
PI_5: the average machine utilization

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 119–132, 2007.

Various related literature states that the average value of each index is used for output analysis. However, it would not preserve the dynamic properties of the system. In this present endeavor, we develop a data envelopment analysis (DEA) model to assess the performance of the 180 design alternatives with multiple indices that are interval data.

The three-way factorial experiment analysis for the assessment results indicates the buffer capacity factor and the interaction of job arrival rate and buffer capacity, which affects the performance significantly. As such, the Fisher Least Significant Difference (LSD) Test shows that designs with a single buffer space are not suggested. This paper thus introduces an unprecedented approach for simulation output analysis to select material handling system designs.

Section 2 of this paper contains the literature review of simulation output analysis on material handling system design and DEA models. Section 3 describes the material handling system under design. The collected simulation output of a design alternative is also depicted. In the fourth section, we introduce the assessment DEA model that we developed. Likewise, we depict the assessment results for the 180 design alternatives, and show the statistical hypothesis tests for the assessment results. Finally, the paper's conclusions are listed in the last section.

2 Literature Review

The three categories literatures related to this research are reviewed here.

2.1 Simulation Outputs Analysis of Material Handling Systems

Liu and Hung (2001) proposed three deadlock-free control strategies for a single multi-load automated guided vehicle. The results of performance indices were expressed through its mean values. The superior strategy can be found by comparing the values of the performance indices. If the performance value of the strategy outperforms the others, then it can be interpreted to mean that the strategy dominates the other strategies. In addition, each strategy was discussed individually in each condition. Based on the simulation experiments, it seemed that we could not find the standard to evaluate the strategies.

Lin, Wang, and Yang (2005) used the least significant difference (LSD) method to compare the strategies under each condition. The paired test analysis was used to rank the significant difference of the strategies. The means of each of the performance values were collected in repeated runs. This would provide some guidance for decision makers in the selection of preferable strategies on the basis of the different operation, environment, and performance measures. The preferable strategies can be observed by comparing the performance value of each condition.

2.2 Original DEA Models

The DEA model was first introduced by Charnes, Cooper, and Rhodes (1978). It is generally referred to as a CCR model. This model evaluates the relative efficiency of

decision-making units (DMUs) based on linear programming while at the same time considering the multiple inputs and outputs. DEA deals with the ratio of weighted sum of outputs to weighted sum of inputs. The optimal objective value (θ^*) is called the *ratio efficiency* of the DMU. The excess in inputs and shortfalls in outputs are revealed by the optimal solution called *slacks*. A DMU with the full ratio efficiency, θ^*=1, and with no *slacks* in any optimal solution is called *CCR-efficient*. The DMUs on the DEA frontier are relatively more efficient than other DMUs with maximum output levels given input levels, or with minimum input levels given output levels. As such, the efficient DMUs are the references of inefficient DMUs.

Tone (2001) proposed a slacks-based measure (SBM) of efficiency based on input excesses and output shortfalls. Each DMU can be improved and become more efficient by deleting the input excess and augmenting the output shortfalls. The SBM model can be interpreted as a product of input and output inefficiencies. The objective function is defined as the ratio of the percentage of sum of input excess to the percentage of sum of output shortfalls. It can eliminate the problems of unit variance and become monotone decreasing in each slack input and output.

2.3 DEA Models with Imprecise Data

In the work of Jablonsky et al. (2004), the CCR DEA model with interval data is formulated as [P1]. The n DMUs are evaluated by m inputs and s outputs indices. For each DMU_j, it transforms the amount of resource i within the interval $x_{ij} \in [x_{ij}^L, x_{ij}^U]$, i=1,…, m, and obtains the output r within the interval, $y_{rj} \in [y_{rj}^L, y_{rj}^U]$, r=1,…, s. The decision variables v_i and u_r denote the weights assigned to the lower bound values of input index i and output index r, respectively. For each DMU_j, the decision variables q_{ij} and p_{rj} denote the weights assigned to the intervals of input index i and output index r, respectively.

[P1] Imprecise-CCR (multiplier form)

$$\text{Maximize} \quad h_o = \sum_{r=1}^{s} [u_r y_{ro}^L + p_{ro}(y_{ro}^U - y_{ro}^L)] \tag{1}$$

$$\text{subject to} \quad \sum_{i=1}^{m} [v_i x_{io}^L + q_{io}(x_{io}^U - x_{io}^L)] = 1; \tag{2}$$

$$\sum_{r=1}^{s} [u_r y_{rj}^L + p_{rj}(y_{rj}^U - y_{rj}^L)] - \sum_{i=1}^{m} [v_i x_{ij}^L + q_{ij}(x_{ij}^U - x_{ij}^L)] \le 0, \; j = 1, 2, \ldots, n \tag{3}$$

$$p_{rj} - u_r \le 0, \; q_{ij} - v_i \le 0, \; i = 1, 2, \ldots, \; m \;, r = 1, 2, \ldots, s \;, \; j = 1, 2, \ldots, n; \tag{4}$$

$$u_r, v_i \ge \varepsilon, \forall r, i; \; p_{rj} \ge 0, \; q_{ij} \ge 0, \forall r, i, j. \tag{5}$$

3 Automated-Guided Vehicle Material Handling System with Deadlock-Free Control

Liu and Hung (2001) studied a deadlock-free control strategy for a single multi-load vehicle on a job shop with finite queuing size. This is depicted in the following Figure 1. Some assumptions are as follows: (1) the vehicle traverses bi-directionally on the single-loop guide-path; (2) I_i and O_i respectively represent the input queue and output queue at processing machine M_i. For the sake of simplicity, the entrance/exit centre is denoted as machine M_0, and (3) when one new task is assigned the vehicle takes the shortest path from its current location to its destination.

There are four factors considered in designing the material handling system: (1) P/T ratio, which is the ratio of the total scheduled processing time to the total essential minimum transferring time and is used as an index to represent the criticality of processing capacity, (2) mean job arrival rate γ units/per hour, (3) local input/output queue capacities C, and (4) vehicle control strategies S. The authors described a hypothetical example and designed a set of simulation experiments to investigate the effectiveness of the proposed vehicle control strategies. The paper does not provide a systematic approach to assess which is the best of the following three vehicle control strategies: (1) *the proposed strategy*, (2) *the strategy with limited shop WIP level*, and (3) *the strategy with a central buffer*. The proposed control strategy uses global shop real-time information to achieve the objectives as well as to avoid shop deadlocks caused by inappropriate job movement. Moreover, this satisfies the system transport requirement.

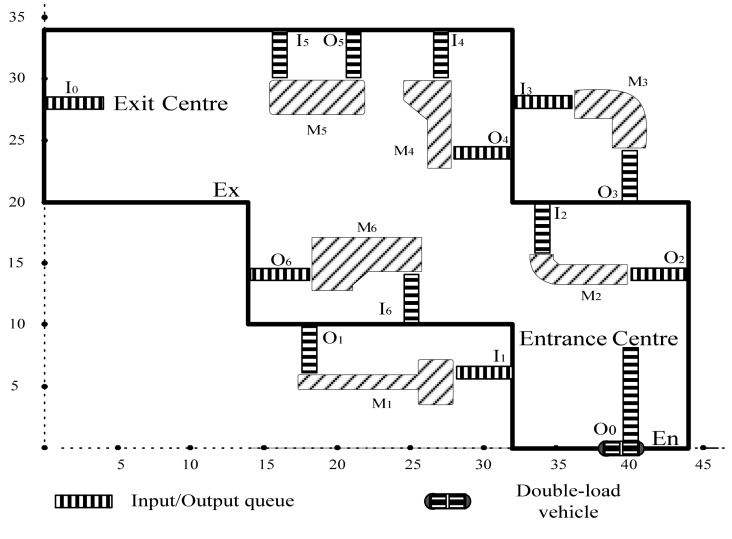


Fig. 1. The job shop layout with a multi-load automatic guided vehicle

There are 12 combinations of P/T ratio and arrival rate γ, namely, k= 1, 2, …, 12. Each of the combination has 15 combinations of queue capacity C and control strategies S. Therefore, there are 180 combinations (design alternatives) of the four design parameters. The simulation experiments for the material handling system are shown in Table 1.

Table 1. The factor levels of experimental design

Level	design factor			
	P/T ratio	arrival rate γ	queue capacity C	control strategies S
1	6	10	1	I
2	5	15	2	II
3	4	20	3	III
4		22	4	
5			5	

We developed a simulation model for each of the 180 design alternatives. Each simulation model was run for 24 hours. To eliminate the initial transit bias, the output during the first two hours was discarded in all of the runs. In each simulation run, the maximum and minimum values for the above five performance indices, PI_1, PI_2, PI_3, PI_4, and PI_5, were recorded.

The next table depicts the simulation results of the 15 design alternatives with P/T ratio=6 and arrival rate γ=10, k=1. The first row signifies the five performance indices, PI_1, PI_2, PI_3, PI_4, and PI_5, and the two design factors S and C. To convert the notations used for material handling system design problems to the notations used in the literature of DEA models, the seven notations were renamed as X_1, X_2, X_3, X_4, X_5, Y_1, and Y_2, respectively. The fourth row illustrates the notations of the upper bound and lower bound values of the seven DEA indices.

As indicated by Sarrico and Dyson (2004), the expressions DMU and decision maker could be misleading. The material handling system designer as the decision maker is, in fact, at a different level doing the evaluation. In this case, the DEA assessment becomes the designer as well as the higher level decision maker, imposing his/her preference structure to have the most favorable measurement for a design alternative, called the unit of assessment (UOA), against the other design alternatives.

The 15 simulation runs are recorded in the following rows of Table 2. Each row is a UOA in the DEA models that will be presented in the succeeding section of this paper. Indices X_4 and X_5 are non-discretionary indices (Banker and Morey, 1986) in our DEA model. These two indices are not included in our assessment model. Instead, we will examine the assessment results to see the effects of the control strategies S and buffer capacity C to the aggregate performance of the UOAs.

Generally, a good system design shall yield low riding time X_1, low machine blockage frequency X_2, low machine blockage duration X_3, high throughput Y_1, and high machine utilization Y_2. We use the following equation to calculate the aggregate performance score θ_j for each design alternative if the associated indices weights v_i and u_r are determined. x_{ij} and y_{rj} respectively denote the values of UOA_j falling

Table 2. The simulation output data of the 15 design alternatives with P/T ratio=6 and γ=10, k=1

	Material handling performance index						
	PI_1	PI_2	PI_3	S	C	PI_4	PI_5
UOA_j	X_1	X_2	X_3	X_4	X_5	Y_1	Y_2
$(UOA_{q,k})$	x_{1j}^U, x_{1j}^L	x_{2j}^U, x_{2j}^L	x_{3j}^U, x_{3j}^L	x_{4j}	x_{5j}	y_{1j}^U, y_{1j}^L	y_{2j}^U, y_{2j}^L
$UOA_{1,1}$	127.24, 64.82	97, 2	256.50, 37.63	I	1	279, 213	0.62, 0.45
$UOA_{2,1}$	116.46, 51.35	23, 0	316.48, 0.00	II	1	279, 214	0.62, 0.45
$UOA_{3,1}$	115.89, 49.57	5, 0	462.91, 0.00	III	1	279, 213	0.62, 0.45
$UOA_{4,1}$	116.65, 49.49	2, 0	99.77, 0.00	I	2	279, 213	0.62, 0.45
$UOA_{5,1}$	114.52, 49.15	4, 0	138.52, 0.00	II	2	279, 212	0.62, 0.45
$UOA_{6,1}$	61.93, 49.61	12, 1	322.57, 25.74	III	2	279, 182	0.48, 0.42
$UOA_{7,1}$	68.51, 52.14	13, 0	351.32, 0.00	I	3	279, 213	0.62, 0.45
$UOA_{8,1}$	75.49, 51.23	11, 0	422.32, 0.00	II	3	279, 213	0.62, 0.45
$UOA_{9,1}$	80.03, 51.15	16, 0	482.34, 0.00	III	3	279, 213	0.62, 0.45
$UOA_{10,1}$	78.23, 51.23	9, 0	498.79, 0.00	I	4	279, 213	0.62, 0.45
$UOA_{11,1}$	65.41, 53.94	209, 2	158.90, 38.15	II	4	279, 214	0.62, 0.45
$UOA_{12,1}$	68.75, 50.63	34, 0	343.99, 0.00	III	4	279, 213	0.62, 0.45
$UOA_{13,1}$	66.69, 49.95	5, 0	246.50, 0.00	I	5	279, 213	0.62, 0.45
$UOA_{141,}$	65.47, 49.85	4, 0	324.00, 0.00	II	5	279, 213	0.62, 0.45
$UOA_{15,1}$	65.43, 49.60	3, 0	188.55, 0.00	III	5	279, 213	0.62, 0.45

between its upper and lower bounds, (x_{ij}^U, x_{ij}^L) and (y_{rj}^U, y_{rj}^L). The counter j is formed by two tuples (q, k) where q and k range from 1 to 15 and 1 to 12, respectively. In total, 180 design alternatives were simulated.

$$\theta_j = (y_{1j}u_1+y_{2j}u_2) / (x_{1j}v_1+x_{2j}v_2+x_{3j}v_3+x_{4j}v_4+x_{5j}v_5) \tag{6}$$

The lower value in the denominator, x_{1j}, x_{2j}, and x_{3j}, and the higher value in the numerator, y_{1j} and y_{2j}, would cause a higher θ_j value. In DEA literature, the indices in the denominator and numerator are called *input* and *output* indices. A production system with a low input but generates a high output is desired. The problem at hand is to determine the weights of v_1, v_2, v_3, v_4, u_1, and u_2. We employ the concept of data envelopment analysis (DEA) to assess each UOA_j at its most favorable weights to obtain the maximum score. To deal with the interval data, we develop a new DEA model with slack-based measurement.

4 DEA-SBM Model with Imprecise Data

We transform [P1] to its dual form [P2]. The multiplier λ_j is the weight assigned to UOA_j in assessing the object UOA_o; j is ranged from 1 to n. For each UOA_j, α_{rj} and β_{ij}

are the dual prices of the difference p_{rj}-u_r and q_{ij}-v_i, respectively. s_i^- and s_r^+ respectively denote the slack variables of i^{th} input and r^{th} output. The radial efficiency score of UOA_o, θ_o, which could be determined as the second portion of the objective function, slack variables, is excluded. As the optimal value θ_o^* replaces the decision variable θ_o, the second phase is to calculate the maximum of the second portion and total slack values without introducing the positive non-Archimedean infinitesimal ε.

4.1 Imprecise-CCR (Dual Model)

[P2]

$$\text{Minimize} \quad \theta_o - (\varepsilon \sum_{r=1}^{s} s_r^+ + \varepsilon \sum_{i=1}^{m} s_i^-) \tag{7}$$

$$\text{subject to} \quad x_{io}^L \theta_o - s_i^- = \sum_{j=1}^{n} x_{ij}^L \lambda_j + \sum_{j=1}^{n} \beta_{ij}, i = 1, 2, \ldots, m; \tag{8}$$

$$\sum_{j=1}^{n} y_{rj}^L \lambda_j - \sum_{j=1}^{n} \alpha_{rj} - s_r^+ = y_{ro}^L, r = 1, 2, \ldots, s; \tag{9}$$

$$\sum_{j=1}^{n} (y_{rj}^U - y_{rj}^L)\lambda_j + \alpha_{rj} \geq (y_{ro}^U - y_{ro}^L), r = 1, 2, \ldots, s \ , \quad j = 1, 2, \ldots n; \tag{10}$$

$$-\sum_{j=1}^{n} (x_{ij}^U - x_{ij}^L)\lambda_j + (x_{io}^U - x_{io}^L)\theta_o + \beta_{ij} \geq 0, i = 1, 2, \ldots, m \ , \quad j = 1, 2, \ldots n; \tag{11}$$

$$\theta_o \quad \text{free in sign}; s_i^- \geq 0, \forall i; s_r^+ \geq 0, \forall r; \ \lambda_j \geq 0, \forall j; \tag{12}$$

$$\alpha_{rj} \geq 0, \forall r, \ j; \ \ \beta_{ij} \geq 0, \forall i, \ j. \tag{13}$$

4.2 Interval-SBM Model

The redial efficiency score θ_o^* obtained in [P2] excludes the portion of slack values. θ_o^* tends to overestimate the true efficiency score. The objective function of [P3] was substituted for the objective function of [P2] in order to compute the slack-based measurement (SBM) efficiency score. All the inefficiencies are expressed in a slack form. The score ρ_o^* obtained in [P3] would not be greater than θ_o^*. The theory of SBM is well introduced in Tone (2001).

[P3]

$$\text{Minimize} \quad \rho_o = [(1/m)\sum_{i=1}^{m} \frac{x_{io}^L - s_i^- - \sum_{j=1}^{n}\beta_{ij}}{x_{io}^L}][(1/s)\sum_{r=1}^{s} \frac{y_{ro}^L + s_r^+ + \sum_{j=1}^{n}\alpha_{rj}}{y_{ro}^L}]^{-1} \tag{14}$$

$$= [1 - \frac{1}{m}\sum_{i=1}^{m} \frac{s_i^- + \sum_{j=1}^{n}\beta_{ij}}{x_{io}^L}] / [1 + \frac{1}{s}\sum_{r=1}^{s} \frac{s_r^+ + \sum_{j=1}^{n}\alpha_{rj}}{y_{ro}^L}]$$

$$\text{subject to} \quad x_{io}^L - s_i^- = \sum_{j=1}^{n} x_{ij}^L \lambda_j + \sum_{j=1}^{n}\beta_{ij},\ i = 1, 2, \ldots, m \tag{15}$$

$$\sum_{j=1}^{n} y_{rj}^L \lambda_j - \sum_{j=1}^{n}\alpha_{rj} - s_r^+ = y_{ro}^L,\ r = 1, 2, \ldots, s; \tag{16}$$

$$\sum_{j=1}^{n} (y_{rj}^U - y_{rj}^L)\lambda_j + \alpha_{rj} \geq (y_{ro}^U - y_{ro}^L),\ r = 1, 2, \ldots, s,\ j = 1, 2, \ldots, n; \tag{17}$$

$$-\sum_{j=1}^{n} (x_{ij}^U - x_{ij}^L)\lambda_j + (x_{io}^U - x_{io}^L) + \beta_{ij} \geq 0,\ i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, n; \tag{18}$$

$$s_i^- \geq 0, \forall i;\ s_r^+ \geq 0, \forall r;\ \lambda_j \geq 0, \forall j;\ \alpha_{rj} \geq 0, \forall r, \forall j;\ \beta_{ij} \geq 0, \forall i, \forall j. \tag{19}$$

In the first term on the right-hand side, the ratio $x_{io}^L - s_i^- - \sum_{j=1}^{n}\beta_{ij} \Big/ x_{io}^L$ evaluates the relative reduction rate of input i and its interval range; therefore, the first term represents the mean reduction rate of inputs and their interval range or *input inefficiency*. Similarly, in the second term, the ratio $y_{ro}^L + s_r^+ + \sum_{j=1}^{n}\alpha_{ij} \Big/ y_{ro}^L$ evaluates the relative expansion rate of the r^{th} output and its interval range. Hence, $(1/s)\sum_{r=1}^{s}(y_{ro}^L + s_r^+ + \sum_{j=1}^{n}\alpha_{rj} \Big/ y_{ro}^L)$ is the mean expansion rate of outputs and its interval range or *output inefficiency*. Thus, SBM ρ_o can be interpreted as the product of input and output inefficiency. [P3] can be transformed into a linear program using the Charnes-Cooper transformation in a similar way as the CCR model (Charnes et al., 1978).

First, let us multiply a scalar variable t (>0) for both the denominator and the numerator in the objective function of [P3]. This causes no change in ρ_o. Second, we adjust t so that the denominator becomes 1. Then this term is moved to constraints. The objective is to minimize the numerator. Third, because the present model is a nonlinear programming problem (ts_r^+), we transform it into a linear program as follows. Let us define $S_i^- = ts_i^-$, $S_r^+ = ts_r^+$, $\Lambda_j = t\lambda_j$, $A_{rj} = t\alpha_{rj}$, $B_{ij} = t\beta_{ij}$, then [P4] becomes the following linear program in t, S_i^-, S_r^+, Λ_j, A_{rj}, and B_{ij}.

[P4]

$$\text{Minimize} \quad \tau = t - \frac{1}{m}\sum_{i=1}^{m}\frac{S_i^- + \sum_{j=1}^{n} B_{ij}}{x_{io}^L} \tag{20}$$

$$\text{subject to} \quad t + \frac{1}{s}\sum_{r=1}^{s}\frac{S_r^+ + \sum_{j=1}^{n} A_{rj}}{y_{ro}^L} = 1; \tag{21}$$

$$tx_{io}^L = \sum_{j=1}^{n} x_{ij}^L \Lambda_j + \sum_{j=1}^{n} B_{ij} + S_i^-, \; i = 1,2,\ldots,m; \tag{22}$$

$$-ty_{ro}^L = -\sum_{j=1}^{m} y_{rj}^L \Lambda_j + \sum_{j=1}^{n} A_{rj} + S_r^+, \; r = 1,2,\ldots,s; \tag{23}$$

$$\sum_{j=1}^{n}(y_{rj}^U - y_{rj}^L)\Lambda_j + A_{rj} \geq t(y_{ro}^U - y_{ro}^L), \; r = 1,2,\ldots,s; \; j = 1,2,\ldots,n; \tag{24}$$

$$-\sum_{j=1}^{n}(x_{ij}^U - x_{ij}^L)\Lambda_j + B_{ij} + t(x_{io}^U - x_{io}^L) \geq 0, \; i = 1,2,\ldots,m; \; j = 1,2,\ldots,n; \tag{25}$$

$$S_i^- \geq 0, \; \forall i; \; S_r^+ \geq 0, \forall r; \; \Lambda_j \geq 0, \forall j; \; A_{rj} \geq 0, \forall r, \forall j; \; B_{ij} \geq 0, \forall i, \forall j; t > 0. \tag{26}$$

4.3 Dual Model of Imprecise-SBM Model

The dual model of [P4] is derived as [P5], with the dual variables $\delta \in \Re$, $v \in \Re^m$, and $u \in \Re^s$.

[P5]

$$\text{Maximize} \quad \delta_o \tag{27}$$

$$\text{subject to} \quad \delta_o + \sum_{i=1}^{m} x_{io}^L v_i - \sum_{r=1}^{s} y_{ro}^L u_r - \sum_{r=1}^{s}(y_{ro}^U - y_{ro}^L)p_{rj} + \sum_{i=1}^{m}(x_{io}^U - x_{io}^L)q_{ij} = 1; \tag{28}$$

$$-\sum_{i=1}^{m} x_{ij}^L v_i + \sum_{r=1}^{s} y_{rj}^L u_r + \sum_{r=1}^{s}(y_{rj}^U - y_{rj}^L)p_{rj} - \sum_{i=1}^{m}(x_{ij}^U - x_{ij}^L)q_{ij} \leq 0, \; j = 1,2,\ldots,n; \tag{29}$$

$$v_i \geq \frac{1}{m}[1/x_{io}^L], \; i = 1,2,\ldots,m; \tag{30}$$

$$u_r \geq \frac{\delta_o}{s}[1/y_{ro}^L], \; r = 1,2,\ldots,s; \tag{31}$$

$$q_{ij} - v_i \le -\frac{1}{m}\frac{1}{x_{io}^L},\ i = 1,2,\ldots,m,\ j = 1,2,\ldots,n; \tag{32}$$

$$p_{rj} - u_r \le -\frac{1}{s}\frac{\delta_o}{y_{ro}^L},\ r = 1,2,\ldots,s,\ \ j = 1,2,\ldots,n; \tag{33}$$

$$u_r, v_i \ge \varepsilon, \forall r,i;\ \ p_{rj} \ge 0, q_{ij} \ge 0, \forall r,i,j. \tag{34}$$

Through the first constraint of [P5], we can eliminate δ_o. In addition, we integrate some constraints into one constraint. Then this problem is equivalent to [P6].

[P6]

$$\text{Maximize} \quad \sum_{r=1}^{s} y_{ro}^L u_r + \sum_{r=1}^{s}(y_{ro}^U - y_{ro}^L)p_{rj} - \sum_{i=1}^{m} x_{io}^L v_i - \sum_{i=1}^{m}(x_{io}^U - x_{io}^L)q_{ij} \tag{35}$$

$$\text{subject to} \quad -\sum_{i=1}^{m} x_{ij}^L v_i + \sum_{r=1}^{s} y_{rj}^L u_r + \sum_{r=1}^{s}(y_{rj}^U - y_{rj}^L)p_{rj} - \sum_{i=1}^{m}(x_{ij}^U - x_{ij}^L)q_{ij} \le 0, \quad j = 1,2,\ldots,n; \tag{36}$$

$$v_i - q_{ij} \ge \frac{1}{m}[1/x_{io}^L],\ i = 1,2,\ldots,m,\ j = 1,2,\ldots,n; \tag{37}$$

$$u_r - p_{rj} \ge \frac{1 - \sum_{i=1}^{m} x_{io}^U v_i + \sum_{r=1}^{s} y_{ro}^L u_r + \sum_{r=1}^{s}(y_{ro}^U - y_{ro}^L)p_{rj} - \sum_{i=1}^{m}(x_{io}^U - x_{io}^L)q_{ij}}{s}[1/y_{ro}^L], \tag{38}$$

$$r = 1,2,\ldots,s,\ \ j = 1,2,\ldots,n;$$

$$p_{rj} \ge 0, q_{ij} \ge 0\ \forall r,i,j. \tag{39}$$

The dual variables v_i and u_r can be interpreted as the virtual cost and price of the i^{th} input and the r^{th} output, respectively. For UOA_j, the dual variables q_{ij} and p_{rj} are interpreted as the virtual costs v_i multiplied by the weights of input interval and prices u_r, then multiplied by the weights of output interval, respectively. The goal of this dual model is to find the optimal virtual costs and prices for UOA_o so that the profit $\sum_{r=1}^{s} y_{rj}^L u_r + \sum_{r=1}^{s}(y_{rj}^U - y_{rj}^L)p_{rj} - \sum_{i=1}^{m} x_{ij}^L v_i - \sum_{i=1}^{m}(x_{ij}^U - x_{ij}^L)q_{ij}$ does not exceed zero for any UOA and maximizes the profit $\sum_{r=1}^{s} y_{ro}^L u_r + \sum_{r=1}^{s}(y_{ro}^U - y_{ro}^L)p_{rj} - \sum_{i=1}^{m} x_{io}^L v_i - \sum_{i=1}^{m}(x_{io}^U - x_{io}^L)q_{ij}$ for UOA_o. Obviously, the optimal profit is at best zero and hence δ_o =1 for the SBM efficient UOAs.

4.4 The Simulation Output Analysis

Using the 15 simulation outputs depicted in Table 2, UOA_j (or denoted as $UOA_{q,k}$, q=1, 2, …, 15 and k=1) and the other 165 simulation outputs ($UOA_{q,k}$, q=1, 2, …, 15, and k=2, 3,…,12) that are not shown here, we model [P4] to assess the aggregated performance of

UOA_o where o is ranged as 1, 2, …, 180. These data were divided by the optimal value t^* to return to the optimal solutions of [P3]. For the case in which x_{io}^L is zero in the above table, the term s_i^- / x_{io}^L is removed from the objective function of [P4]. That is, the slack s_i^- =0 is neglected and m is replaced by (m-1). Their optimum objective function values, ρ_o^*, are listed in the most right-hand column of Table 3. For instance, where k=1, $UOA_{4,1}$, $UOA_{5,1}$, and $UOA_{15,1}$ have ρ_o^* values equal to one and are non-dominant (efficient) units. The other UOAs are dominated or inefficient units.

Table 3. Objective value of [P3], ρ_o^* of the 180 simulation models

C	S	UOA_j	P/T = 6				P/T = 5				P/T = 4			
			γ=10	γ=15	γ=20	γ=22	γ=10	γ=15	γ=20	γ=22	γ=10	γ= 15	γ= 20	γ= 22
		$UOA_{q,k}$	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11	k=12
1	I	$UOA_{1,k}$	0.218	0.185	0.182	0.145	0.226	0.233	0.239	0.259	0.218	0.298	0.298	0.301
	II	$UOA_{2,k}$	0.860	0.285	0.263	0.327	0.756	0.271	1	0.270	0.35	0.291	0.265	0.317
	III	$UOA_{3,k}$	0.912	1	1	1	0.253	1	1	0.379	0.439	0.267	0.296	0.965
2	I	$UOA_{4,k}$	1	0.225	0.209	0.622	0.791	0.874	1	0.271	0.888	0.905	1	1
	II	$UOA_{5,k}$	1	1	1	1	0.794	0.959	0.997	1	1	0.939	0.980	1
	III	$UOA_{6,k}$	0.259	1	1	0.409	0.778	0.319	0.273	1	0.371	0.323	0.328	0.268
3	I	$UOA_{7,k}$	0.858	0.810	0.739	0.702	0.821	0.958	0.983	0.905	0.980	0.962	0.969	0.958
	II	$UOA_{8,k}$	0.865	0.915	0.902	0.859	0.805	1	1	1	0.913	1	1	1
	III	$UOA_{9,k}$	0.866	1	1	1	0.894	0.969	1	1	1	0.989	0.325	1
4	I	$UOA_{10,k}$	0.866	0.862	0.818	0.770	0.987	0.985	1	0.948	1	0.983	1	1
	II	$UOA_{11,k}$	0.559	1	0.945	1	0.845	0.979	0.284	1	0.496	0.977	1	0.290
	III	$UOA_{12,k}$	0.869	1	1	1	1	1	1	1	1	0.972	1	1
5	I	$UOA_{13,k}$	0.965	1	1	0.814	1	1	1	1	1	1	1	1
	II	$UOA_{14,k}$	0.985	1	1	0.912	1	1	1	1	0.991	0.966	0.999	1
	III	$UOA_{15,k}$	1	1	1	1	1	1	1	1	1	0.991	0.991	0.989

The 180 UOAs' performance measurement data depicted in Table 3 were analyzed by the three-ways factorial experiment (Montgomery, 1997) in terms of having the effect extent of the control strategy S, the buffer capacity C, and mean job arrival rate γ. The effective hypothesis decomposition is shown in Table 4. The null hypothesis of each item in the effective is not significant. As depicted in Table 4, only buffer capacity C rejects the hypothesis and is seen to have a highly significant impact on ρ_o^*. The individual control strategy S and mean job arrival rate γ are not significant at the 5 percent significance level. Among the four types of interaction, only the interaction of C*S has a significant effect.

Since C has a highly significant impact on ρ_o^*, we further perform a Fisher Least Significant Difference Test (Montgomery, 2001) for the five levels of C in Table 5.

The procedure uses the F statistic for testing each pair of levels, i and j, with the hypothesis that their means are equal, H_o: $\mu_i=\mu_j$. Clearly, the only pairs of means C =3, C =4, and C =5 do not differ significantly. C=1 and C=2 produces a significantly low ρ_o^* than the other levels of C.

Because C*S has a highly significant impact on ρ_o^*, we likewise further perform a Fisher Least Significant Difference Test (Montgomery, 2001) for the 15 combinations of C*S in Table 3. For the results depicted in Table 6, Combination No. 1 produces a significantly lower mean efficiency than the other combinations. The next group containing Combinations No. 2 and No. 6 has a higher mean efficiency than Combination No. 1, but they do not differ significantly from each another. The third group contains Combinations No. 3, No. 4, and No. 11, and the remaining combinations belong to the fourth group. The third group has a higher mean efficiency than the second group, and the forth group has a higher mean efficiency than the third group. However, it does not differ significantly within groups.

Table 4. Univariate result for each decision variable

General Effect	Sigma-restricted parameterization, effective hypothesis decomposition				
	Deg. of Freedom	Sum of Square	Mean of Square	F-test	p
Intercept	1	115.7197	115.7191	2722.328	0.000000
γ	3	0.0163	0.0054	0.128	0.943210
C	4	***6.4015***	***1.6004***	***37.650***	***0.000000***
S	2	0.2080	0.1040	2.446	0.090942
γ* C	12	0.1418	0.0118	0.278	0.991728
γ* S	6	0.1244	0.0207	0.488	0.816365
C *S	8	***2.6400***	***0.3300***	***7.763***	***0.000000***
γ* C * S	24	0.6035	0.0251	0.592	0.931469
Error	120	5.1009	0.0425		
Total	179	15.2364			

Table 5. Fisher Least Significant Difference Test for the cell of C

Level of C	C; Least Significant Means Current effect: F(4, 120)=37.650, p=0.00000 Effective hypothesis decomposition					
	C	Mean	Std. Err.	-95.00%	+95.00%	N
1	1	0.460086	0.034362	0.392052	0.528121	36
2	2	0.743886	0.034362	0.675851	0.811921	36
3	3	0.915155	0.034362	0.847121	0.983190	36
4	4	0.900935	0.034362	0.832900	0.968970	36
5	5	0.988939	0.034362	0.920904	1.056974	36

Table 6. Fisher Least Significant Difference Test for the cells of *C* *S*

Combination No.	*C* *S*; Least Significant Means Current effect: F(8, 120)=7.7632, p=0.00000 Effective hypothesis decomposition						
	C	*S*	Mean	Std. Err.	-95.00%	+95.00%	N
1	1	I	0.233555	0.059517	0.115716	0.351395	12
2	1	II	0.437607	0.059517	0.319768	0.555447	12
3	1	III	0.709096	0.059517	0.591257	0.826936	12
4	2	I	0.732004	0.059517	0.614164	0.849844	12
5	2	II	0.972270	0.059517	0.854431	1.090110	12
6	2	III	0.527384	0.059517	0.409545	0.645224	12
7	3	I	0.887049	0.059517	0.769210	1.004889	12
8	3	II	0.938262	0.059517	0.820423	1.056102	12
9	3	III	0.920154	0.059517	0.802315	1.037994	12
10	4	I	0.934888	0.059517	0.817048	1.052727	12
11	4	II	0.781208	0.059517	0.663369	0.899048	12
12	4	III	0.986709	0.059517	0.868869	1.104548	12
13	5	I	0.981560	0.059517	0.863721	1.099400	12
14	5	II	0.987660	0.059517	0.869820	1.105500	12
15	5	III	0.997597	0.059517	0.879758	1.115437	12

The simulation output analysis results show that setting the buffer capacity for the incoming and outgoing job at 5 outperforms the others. All in all, the higher the buffer capacity *C*, the higher the performance value. The control strategy *S* and mean job arrival rate γ are not the factors of change in the DEA performance measure ρ_o^*.

5 Conclusions

The proposed DEA model provides a systematic approach to deal with the simulation outputs with multiple indices that have interval data. The DEA model assigns weights for the output indices and discriminates the material handling alternatives. The approach leads to a new avenue for simulation output analysis under various circumstances such as bounded variables and virtual weight constraints on the indices. The aggregated scores of the simulation runs with multiple performance indices are analyzed by statistical test methods.

A number of future research issues remain to be addressed. We used six performance indices for different material handling systems. Furthermore, the number and context of indices were not limited. For a specific system and certain practical managerial considerations, we may add virtual assurance region restrictions to the DEA model. Other interesting indices could be identified and subjected to the characteristics of an individual material handling system.

Acknowledgement

This research was support by the National Science of Council of Taiwan under the project 952221-E-009-164.

References

Banker, R.D., Morey, R.C.: Efficiency analysis for exogenously fixed inputs and outputs. Operations Research 34(4), 513–521 (1986)

Charnes, A., Cooper, W.W., Rhodes, E.: Measuring efficiency of decision making units. European Journal of Operational Research 1, 429–444 (1978)

Cooper, W.W., Park, K.S., Yu, G.: IDEA and AR-IDEA: Models for dealing with imprecise data in DEA. Management Science 45, 597–607 (1999)

Halme, M., Joro, T., Korhonen, P., Salo, S., Wallenius, J.: A value efficiency approach to incorporating preference information in data envelopment analysis. Management Science 45, 103–115 (1999)

Jablonsky, J., Fiala, P., Smirlis, Y., Despotis, D.K.: DEA with interval data: An illustration using the evaluation of branches of a Czech bank. Central European Journal of Operational Research 12, 323–337 (2004)

Liu, F.H., Hung, P.C.: Real time deadlock-free control strategy for single multi-load automated guided vehicle on a job shop manufacturing system. International Journal of Production Research 39, 1323–1342 (2001)

Lin, J.T., Wang, F.K., Yang, C.J.: The performance of the number of vehicles in a dynamic connecting transport AMHS. International Journal of Production Research 43, 2263–2276 (2005)

Montgomery, D.C.: Design and Analysis of Experiments, 4th edn. John Wiley & Sons, New York (1997)

Sarrico, C.S., Dyson, R.G.: Restricting virtual weights in data envelopment analysis. European Journal of Operational Research 159, 17–34 (2004)

Tone, K.: A slacks-based measure of efficiency in data envelopment analysis. European Journal of Operational Research 130, 498–509 (2001)

Synchronization between Two Different Hyperchaotic Systems Containing Nonlinear Inputs

Her-Terng Yau[1], Cheng-Chi Wang[2], and Meei-Ling Hung[1]

[1] Department of Electrical Engineering, Far-East University, Tainan 744, Taiwan, R.O.C.
pan1012@ms52.hinet.net
[2] Department of Mechanical Engineering, Far-East University, Tainan 744, Taiwan, R.O.C.
wccpipn@yahoo.com.tw

Abstract. This paper introduces a variable structure technology for the synchronization of chaos between two different hyperchaotic systems with input nonlinearity. Based on Lyapunov stability theory, a sliding mode controller and some generic sufficient conditions for global asymptotic synchronization are designed such that the error dynamics of the hyperchaotic Rössler and hyperchaotic Chen systems satisfy stability in the Lyapunov sense in spit of the input nonlinearity. The Numerical simulation results demonstrate the validity and feasibility of the proposed controller.

1 Introduction

Since the ideal of synchronizing two identical chaotic systems from different initial conditions was first introduced by Carroll and Pecora, chaos synchronization has gained a lot of attention among scientists from variety of research fields over the last few years [1-3]. Chaos synchronization can be applied in the vast areas of physics and engineering science, especially in secure communication [4-5]. In order to achieve the synchronization, a nonlinear controller that obtains signals from the master and slave systems and manipulates the slave system should be designed. Recently, many control methods have been developed to achieve chaos synchronization between two identical chaotic systems with different initial conditions [6-11]. However, most of these methods are only applicable to the chaos synchronization of two systems that are identical in every respect and which contain only low dimensional attractors. This is in stark contrast to many real-word applications of the technology. In fact, in systems such as laser array, biological systems and cognitive processes, it is hardly the case that every component can be assumed to be identical. In the area of communications security for example, the adoption of higher dimensional chaotic systems as well as systems with more than one positive Lyapunov exponents has been proposed for use to generate more complex dynamics. Methods are therefore needed to synchronize chaotic systems that are both different and are of high dimensions. Moreover, when the controller is realized in practical physical systems, due to physical limitations of actuators, the nonlinearities in control input do exist. The presence of nonlinearities in control input may cause serious influence of system performance and decrease the system response. Besides, the nonlinearity in control input may cause the chaotic system

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 133–141, 2007.

perturbed to unpredictable results because the chaotic system is very sensitive to any system parameters. Therefore, its effect cannot be ignored in analysis of control design and realization for chaos synchronization. Thus the derivation of controller with input nonlinearity for chaos synchronization is an important problem.

In the literatures [12-15] used active control techniques to synchronize two different chaotic systems are either only concerns some low dimension chaotic systems or the input nonlinearity is not discussed. In this case of input nonlinearity, the applications of above method in Refs. [12-15] are hard to achieve

In this paper, the goal is to force the two different hyperchaotic Rössler system and hyperchaotic Chen system to be synchronized even if they are subjected to input nonlinearity. The method of active sliding mode control law is applied to control the chaos synchronization system. The technique requires two stages. The first stage is to select stable sliding surfaces for the desired dynamics, and the second stage is to design a switching control law to achieve the stable sliding surfaces. Finally, numerical simulation is carried to confirm the validity of the proposed theoretical approach.

2 System Description and Problem Formulation

In this paper, two different hyperchaotic systems included Rössler system and Chen system are described in the follows. In order to observe the synchronization behavior in these two systems, it is assumed that the hyperchaotic Rössler system drives the hyperchaotic chen system. Therefore, the master and slave systems are shown in the follows:

Master system:

$$\begin{cases} \dot{x}_1 = -x_2 - x_3 \\ \dot{x}_2 = x_1 + a_1 x_2 + x_4 \\ \dot{x}_3 = b_1 + x_1 x_3 \\ \dot{x}_4 = -c_1 x_3 + d_1 x_4 \end{cases}. \tag{1}$$

Slave system:

$$\begin{cases} \dot{y}_1 = a_2(y_2 - y_1) + y_4 + \phi_1(u_1(t)) \\ \dot{y}_2 = d_2 y_1 - y_1 y_3 + c_2 y_2 + \phi_2(u_2(t)) \\ \dot{y}_3 = y_1 y_2 - b_2 y_3 + \phi_3(u_3(t)) \\ \dot{y}_4 = y_2 y_3 + \kappa y_4 + \phi_4(u_4(t)) \end{cases}. \tag{2}$$

Where $\phi_1(u_1), \phi_2(u_2), \phi_3(u_3), \phi_4(u_4)$ are the nonlinear control inputs attached in the slave system. Let the synchronization error vector state is

$$e = [e_1 \quad e_2 \quad e_3 \quad e_4]^T = [x_1 - y_1 \quad x_2 - y_2 \quad x_3 - y_3 \quad x_4 - y_4]^T.$$

Substitution equations (1) and (2) into the error state, the error dynamic equations can be obtained as follows

$$\begin{cases} \dot{e}_1 = -x_2 - x_3 - a_2(y_2 - y_1) - y_4 - \phi_1(u_1) \\ \dot{e}_2 = x_1 + a_1 x_2 + x_4 - d_2 y_1 + y_1 y_3 - c_2 y_2 - \phi_2(u_2) \\ \dot{e}_3 = b_1 + x_1 x_3 - y_1 y_2 + b_2 y_3 - \phi_3(u_3) \\ \dot{e}_4 = -c_1 x_3 + d_1 x_4 - y_2 y_3 - \kappa y_4 - \phi_4(u_4) \end{cases}. \tag{3}$$

The $\phi_i(u_i(t)) \in C^1(R^n \to R)$ is a continues nonlinear function with $\phi_i(0) = 0$, and $u_i(t) \to \phi_i(u_i(t))$ is inside sector $[\varsigma_i, \rho_i]$ ($i = 1,2,3,4$), i.e.

$$\varsigma_i u_i^2 \le u_i \phi_i(u_i) \le \rho_i u_i^2, \tag{4}$$

where ς_i and ρ_i are nonzero positive constants. A nonlinear function $\phi_i(u_i(t))$ is shown in figure 1.

Now, the sliding surfaces suitable for the application can be defined as

$$S_i = e_i + \int_0^t \lambda_i e_i(\tau) d\tau,\ i = 1,2,3,4, \tag{5}$$

where $S_i(t) \in R$ and λ_i is the design parameters which can be determined later. For the existence of the sliding mode [16], it is necessary and sufficient that

$$S_i = e_i + \int_0^t \lambda_i e_i(\tau) d\tau = 0,\ i = 1,2,3,4, \tag{6}$$

and

$$\dot{S}_i = \dot{e}_i + \lambda_i e_i = 0,\ i = 1,2,3,4. \tag{7}$$

Therefore, the following sliding mode dynamics can be obtained as

$$\dot{e}_i = -\lambda_i e_i, i = 1,2,3,4. \tag{8}$$

Obviously, if the design parameters $\lambda_i > 0, i = 1,2,3,4$, the stability of (6) are surely guaranteed, that is $\lim_{t \to \infty} e_i(t) \to 0$. Thus, the slave system will be derived to master system by designing the appropriate signal control inputs $u_i(t)$, $i = 1,2,3,4$.

3 Sliding Mode Control Law with Input Nonlinearity

We choose a control law of the form

$$u_i = \gamma_i \eta_i sign(S_i), \gamma_i > \frac{1}{\varsigma_i}, i = 1,2,3,4, \tag{9}$$

where $\begin{cases} \eta_1 = \left| -x_2 - x_3 - a_2(y_2 - y_1) - y_4 + \lambda_1 e_1 \right| \\ \eta_2 = \left| x_1 + a_1 x_2 + x_4 - d_2 y_1 + y_1 y_3 - c_2 y_2 + \lambda_2 e_2 \right| \\ \eta_3 = \left| b_1 + x_1 x_3 - y_1 y_2 + b_2 y_3 + \lambda_3 e_3 \right| \\ \eta_4 = \left| -c_1 x_3 + d_1 x_4 - y_2 y_3 - \kappa y_4 + \lambda_4 e_4 \right| \end{cases}$.

Based on the control law (9), the reaching condition $s(t)\dot{s}(t) < 0$ is guaranteed in the following theorem, that is, the proposed scheme (9) will derive the system (3) with nonlinear inputs onto the sliding mode $s(t) = 0$.

Theorem 1. Consider the error dynamics system (3) with input nonlinearities. The hitting condition of the sliding mode is satisfied, if the control $u_i(t)$ is given by (9) for $i = 1,2,3,4$.

Proof:

Letting the Lyapunov function of the system be $V = \frac{1}{2}\left(S_1{}^2 + S_2{}^2 + S_3{}^2 + S_4{}^2\right)$, then its derivative with respect to time is

$$
\begin{aligned}
\dot{V} &= S_1\dot{S}_1 + S_2\dot{S}_2 + S_3\dot{S}_3 + S_4\dot{S}_4 \\
&= S_1(\dot{e}_1 + c_1e_1) + S_2(\dot{e}_2 + c_2e_2) + S_3(\dot{e}_3 + c_3e_3) + S_4(\dot{e}_4 + c_4e_4) \\
&= S_1\left(-x_2 - x_3 - a_2(y_2 - y_1)y_4 + \lambda_1e_1 - \phi_1(u_1)\right) \\
&+ S_2\left(x_1 + a_1x_2 + x_4 - d_2y_1 + y_1y_3 - c_2y_2 + \lambda_2e_2 - \phi_2(u_2)\right) \\
&+ S_3\left(b_1 + x_1x_3 - y_1y_2 + b_2y_3 + \lambda_3e_3 - \phi_3(u_3)\right) \\
&+ S_4\left(-c_1x_3 + d_1x_4 - y_2y_3 - \kappa y_4 + \lambda_4e_4 - \phi_4(u_4)\right) \\
&\le \left|S_1\right|\left|-x_2 - x_3 - a_2(y_2 - y_1)y_4 + \lambda_1e_1\right| - S_1\phi_1(u_1) \qquad (10) \\
&\quad + \left|S_2\right|\left|x_1 + a_1x_2 + x_4 - d_2y_1 + y_1y_3 - c_2y_2 + \lambda_2e_2\right| - S_2\phi_2(u_2) \\
&\quad + \left|S_3\right|\left|b_1 + x_1x_3 - y_1y_2 + b_2y_3 + \lambda_3e_3\right| - S_3\phi_3(u_3) \\
&\quad + \left|S_4\right|\left|-c_1x_3 + d_1x_4 - y_2y_3 - \kappa y_4 + \lambda_4e_4\right| - S_4\phi_4(u_4) \\
&\le \eta_1|S_1| - \varsigma_1\gamma_1\eta_1|S_1| + \eta_2|S_2| - \varsigma_2\gamma_2\eta_2|S_2|\eta_2|S_2| - \varsigma_3\gamma_3\eta_3|S_3| + \eta_4|S_4| - \varsigma_4\gamma_4\eta_4|S_4| \\
&\le (1 - \gamma_1\varsigma_1)\eta_1|S_1| + (1 - \gamma_2\varsigma_2)\eta_2|S_2| + (1 - \gamma_3\varsigma_3)\eta_3|S_3| + (1 - \gamma_4\varsigma_4)\eta_4|S_4|
\end{aligned}
$$

where

$$
\begin{aligned}
& u_i\phi_i(u_i) \ge \varsigma_i u_i^2 \\
& \Rightarrow \gamma\eta_i sigu(S_i)\phi(u_i) \ge \varsigma_i\gamma^2\eta_i^2 sign^2(S_i) \\
& \Rightarrow \gamma\eta_i|S_i|S_i\phi(u_i) \ge \varsigma_i\gamma\eta_i|S_i|\left|S_i\right| \qquad \text{, for i=1, 2, 3,4.} \\
& \Rightarrow S_i\phi(u_i) \ge \varsigma_i\gamma_i\eta_i|S_i| \\
& \Rightarrow -S_i\phi(u_i) \le -\varsigma_i\gamma_i\eta_i|S_i|
\end{aligned}
$$

Therefore, if

$$
\gamma_i > \frac{1}{\varsigma_i}\text{, for i=1, 2, 3, 4,} \qquad (11)
$$

then $\dot{V} < 0$, confirming the presence of reaching condition. Thus the proof is achieved completely.

4 Numerical Simulations

In this simulation, the 4th order Runge-Kutta algorithm was used to solve the sets of differential equations related to the master and slave systems with a time grid of 0.0001. We selected the parameters of the hyperchaotic Rössler system as $a_1 = 0.25$, $b_1 = 3$, $c_1 = 0.5$, $d_1 = 0.05$ and the parameters of the hyperchaotic Chen systems as $a_2 = 35$, $b_2 = 3$, $c_2 = 12$, $d_2 = 7$, $\kappa = 0.5$. The initial values of hyperchaotic Rössler and Chen systems are $x(0) = [x_1(0) \;\; x_2(0) \;\; x_3(0) \;\; x_4(0)] = [-15 \;\; -10 \;\; -20 \;\; 15]$, $y(0) = [y_1(0) \;\; y_2(0) \;\; y_3(0) \;\; y_4(0)] = [10 \;\; 15 \;\; 10 \;\; 5]$. In the synchronization example, we selected $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 2$ to result in stable sliding modes and the nonlinear inputs are defined as

$$\phi_i(u_i(t)) = [0.6 + 0.3 \cdot \sin(u_i(t))]u_i(t), i = 1,2,3,4 . \tag{12}$$

Furthermore, it is assumed that the slope of nonlinear sectors in these three synchronization examples are $\varsigma_1 = \varsigma_2 = \varsigma_3 = \varsigma_4 = 0.3$ and $\rho_1 = \rho_2 = \rho_3 = \rho_4 = 0.9$, and the parameters $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 5$ are selected to satisfy the condition (11). The time responses of the hyperchaotic Chen system controlled by the hyperchaotic Rössler system is shown in Fig. 2(a-d). It can be see that the slave system synchronizes with the master system in spite of input nonlinearity. Obviously, the synchronization errors converge asymptotically to zero after the control is active at time t=10 second in Fig. 3..The template sv-lncs.dot and its documentation can be downloaded from the LNCS Web page at http://www.springer.de/comp/lncs/authors.html.

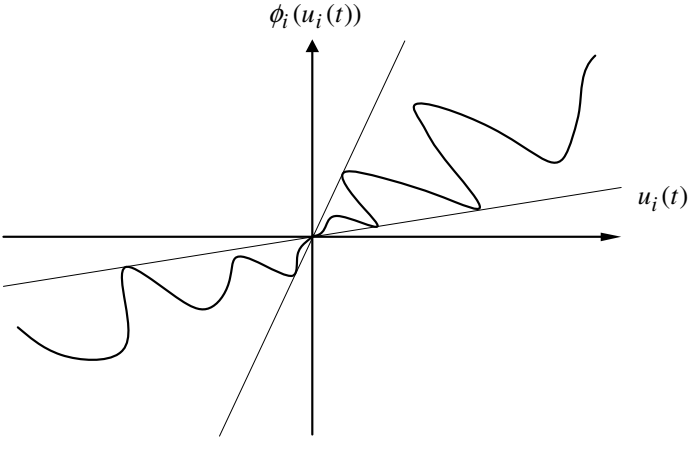
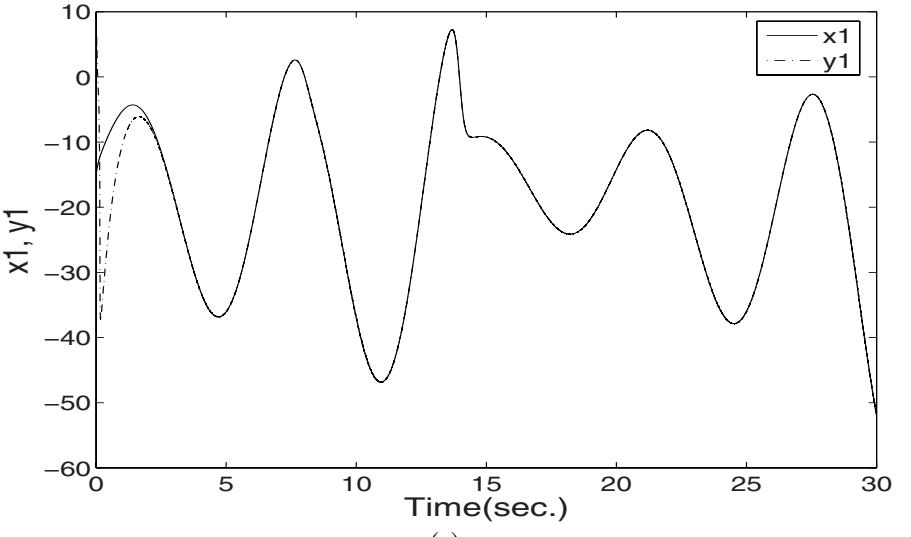
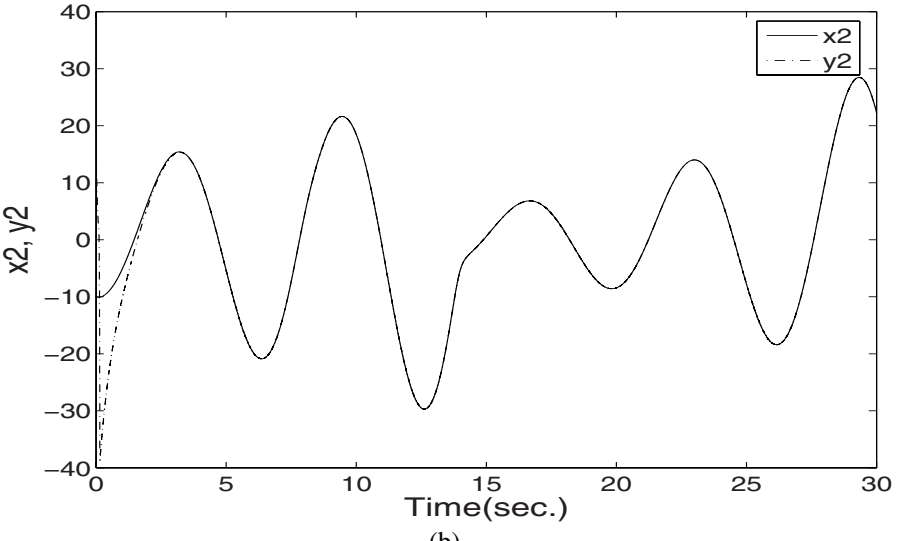


Fig. 1. A scalar nonlinear function $\phi_i(u_i(t))$ inside sector $[\varsigma_i, \rho_i]$

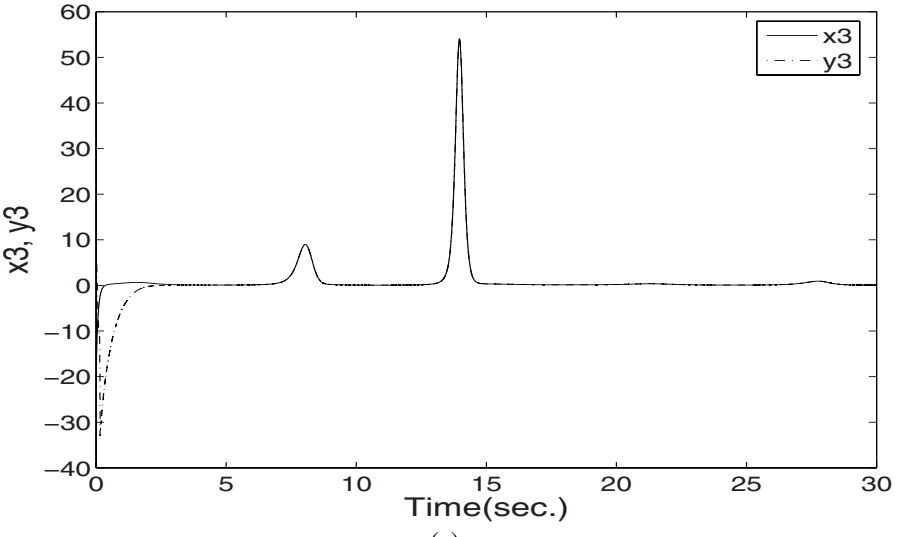


(a)

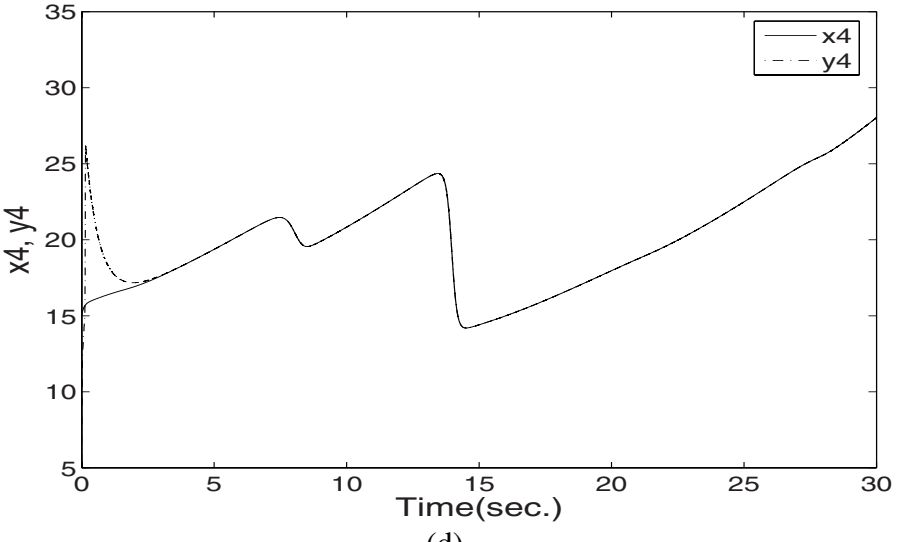


(b)

Fig. 2. The time history of controlled hyperchaotic Rössler (x_1, x_2, x_3, x_4) and Chen (y_1, y_2, y_3, y_4) chaotic systems: (a) x_1, y_1 versus time t; (b) x_2, y_2 versus time t; (c) x_3, y_3 versus time t; (d) x_4, y_4 versus time t



(c)



(d)

Fig. 2. (*Continued*)

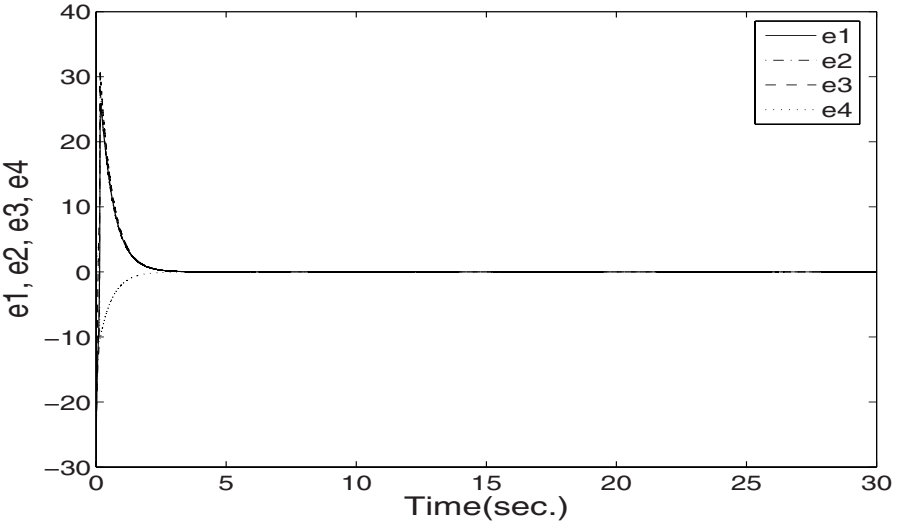


Fig. 3. The synchronization time response of error dynamics of controlled hyperchaotic Rössler and Chen systems

5 Conclusions

In this paper, we introduced a sliding mode control technique to synchronize the hyperchaotic Rössler system and hyperchaotic Chen system. Based on Lyapunov stability theorem, an effective control method for synchronizing different chaotic systems has been proposed using variable structure design. The proposed sliding mode control enables stabilization of synchronization error dynamics to zeros asymptotically in spit of input nonlinearity. Numerical simulation results are presented to verify the effectiveness of the proposed synchronization technique. The main feature of this approach is that it gives the flexibility to construct a control law so that the control strategy can be easily extended to any dimensional chaotic systems.

References

1. Carroll, T.L., Pecora, L.M.: Synchronizing chaotic circuits. IEEE Trans. Circ. Syst. 38, 453–456 (1991)
2. Carroll, T.L., Pecora, L.M.: Synchronization in chaotic systems. Phys. Rev. Lett. 64, 821–824 (1990)
3. Chen, G., Dong, X.: From chaos to order: methodologies, perspectives and applications. World Scientific, Singapore (1998)
4. Kocarev, L., Parlitz, U.: General approach for chaotic synchronization with application to communication. Phys. Rev. Lett. 74, 5028–5031 (1995)
5. Murali, K., Lakshmanan, M.: Secure communication using a compound signal from generalized synchronizable chaotic systems. Phys. Lett. A 241(6), 303–310 (1998)

6. Yassen, M.T.: Adaptive control and synchronization of a modified Chua's system. Applied Mathematic and Computation 135, 113–128 (2003)
7. Liao, T.L.: Adaptive synchronization of two Lorenz systems. Chaos Solitons & Fractals 9, 1555–1561 (1998)
8. Fang, J.Q., Hong, Y., Chen, G.: Physical review E 59, 2523–2526 (1999)
9. Yau, H.T., Lin, J.S., Yan, J.J.: Synchronization control for a class of chaotic systems with uncertainties. International Journal of Bifurcation and Chaos 15(7), 1–12 (2005)
10. Yau, H.T.: Design of adaptive sliding mode controller for chaos synchronization with uncertainties. Chaos, Solitons & Fractals 22, 341–347 (2004)
11. Yau, H.T., Kuo, C.L., Yan, J.J.: Fuzzy Sliding Mode Control for A Class of Chaos. International Journal of Nonlinear Sciences and Numerical simulation 17(3), 333–338 (2006)
12. Ho, M.C., Hung, Y.C.: Synchronization of two different systems by using generalized active control. Phys. Lett. A 301, 424–428 (2002)
13. Yassen, M.T.: Chaos synchronization between two different chaotic systems using active control. Solitons & Fractals 23, 131–140 (2005)
14. Zhang, H., Huang, W., Wang, Z., Chai, T.: Adaptive synchronization between two different chaotic systems with unknown parameters. Phys. Lett. A 350, 363–366 (2006)
15. Agiza, H.N., Yassen, M.-T.: Synchronization of Rössler and Chen chaotic dynamical systems using active control. Phys. Lett. A 278, 191–197 (2001)
16. Slotine, J.E., Li, W.: Applied Nonlinear Control. Prentice-Hall, Englewood Cliffs, New Jersey (1991)

Adaptive Chaos Synchronization of FitzHugh-Nagumo Neurons

Tsung-Wen Lai[1], Jui-Sheng Lin[2,3], Teh-Lu Liao[3], and Jun-Juh Yan[1]

[1] Department of Computer and Communication,
Shu-Te University, Kaohsiung 824, Taiwan, R.O.C.
jjyan@mail.stu.edu.tw
[2] Department of Electrical Engineering,
Far East University, Tainan 744, Taiwan, R.O.C.
[3] Department of Engineering Science,
National Cheng Kung University, Tainan 701, Taiwan, R.O.C.

Abstract. The remarkable system of FitzHugh-Nagumo (FHN) neurons in external electrical stimulation is studied from the view of chaos synchronization in this paper. An effective adaptive sliding mode controller is derived to achieve chaos synchronization even when the parameters of the drive and response FHN neurons are fully unknown. An illustrative example is presented for the purpose of verification and illustration.

1 Introduction

Chaos has already had an enduring effect on engineering science, and it has random behavior that extremely depends on initial conditions of nonlinear dynamical systems[1-4]. Recently, a few models of neuron have been developed to interpret the chaos of neurons in external stimulation [5,6]. The well-known FitzHung-Nagumo (FHN) model [7] is one of the simplified modifications of Hodgkin-Huxley model. Also the whole FHN has recently been almost carried out with an analog simulator [8] and the application of FHN will be more interesting in the future.

This paper addresses the development of an adaptive sliding mode control (ASMC) for synchronizing the state trajectories of FHN neurons with unknown parameters. A proportional-integral (PI) switching surface, which makes it easy to ensure the stability of the error dynamics in the sliding mode, is first proposed. And then based on this switching surface, an ASMC is derived to guarantee the occurrence of the sliding motion. In particular, the limitation of knowing system parameters in priori is also released due to this adaptive scheme.

This paper is presented as follows. Section 2 describes the dynamics of FHN neurons in external electrical stimulation and formulates the synchronization problem for two FHN neurons. In this section, a stable PI switching surface is first derived. Then an ASMC is designed to guarantee the occurrence of the sliding mode and ensure synchronization between master and slave FHN neurons. Numerical simulations that confirm the validity and feasibility of the proposed method are shown in Section 3. Finally, conclusions are presented in Section 4.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 142–150, 2007.

Throughout this paper, it is noted that, $|w|$ represents the absolute value of w and $\|w\|$ represents the Euclidean norm when w is a vector. $sign(\sigma)$ is the sign function of σ, if $\sigma > 0$, $sign(\sigma) = 1$; if $\sigma = 0$, $sign(\sigma) = 0$; if $\sigma < 0$, $sign(\sigma) = -1$.

2 Problem Formulation and ASMC Design

Now we consider a master-slave synchronization scheme for two identical FHN neurons but with fully unknown parameters. The master system is

$$\begin{aligned} \dot{x}_{m1} &= x_{m1}(x_{m1} - 1)(1 - rx_{m1}) - x_{m2} + I_0(t) \\ \dot{x}_{m2} &= bx_{m1} \end{aligned} \tag{1}$$

In (1), $I_0(t)$ represents the external electrical stimulation

$$I_0(t) = \frac{A}{\omega} \cos \omega t \tag{2}$$

where A and ω are the strength and frequency, respectively, of the applied field and $\omega = 2\pi f\,(rad/s)$, $f(Hz)$ is the stimulus frequency. We will show that the master FHN neuron (1) displays chaotic behavior in the later example.

The slave system is

$$\begin{aligned} \dot{x}_{s1} &= x_{s1}(x_{s1} - 1)(1 - rx_{s1}) - x_{s2} + I_0(t) + u \\ \dot{x}_{s2} &= bx_{s1} \end{aligned} \tag{3}$$

The added term u in the first equation of (3) is the control input. This input is to be determined for the purpose of synchronizing the two FHN neurons with the same but unknown parameters r, b, A, ω. Define the synchronization errors between the slave system (3) and the drive system (1) as follows:

$$e_1 = x_{s1} - x_{m1},\ e_2 = x_{s2} - x_{m2}, \tag{4}$$

we can obtain the dynamics of synchronization error between the master-slave FHN neurons as follows:

$$\begin{aligned} \dot{e}_1 &= x_{s1}(x_{s1} - 1)(1 - rx_{s1}) - x_{m1}(x_{m1} - 1)(1 - rx_{m1}) - e_2 + u \\ \dot{e}_2 &= \dot{x}_{s2} - \dot{x}_{m2} = be_1 \end{aligned} \tag{5}$$

It is clear that the synchronization problem is replaced by the equivalent of stabilizing the error dynamics (5) using a suitable choice of the control scheme u. Thus the considered goal of this paper is that for any given master-slave FHN neurons as (1) and (3), respectively, an ASMC is designed to ensure the asymptotical stability of the resulting error system (5) in the sense that $\|e(t)\| \to 0$ as $t \to \infty$, where $e(t) = [e_1, e_2]$. Generally speaking, using the sliding mode control technique to achieve the synchronization for a pair of chaotic systems involves two major steps. The first step is to select an appropriate switching surface which can guarantee the stability of the equivalent dynamics in the sliding mode such that the synchronization error dynamics (5) can converge to zero. The second step is to determine an ASMC to guarantee the

hitting of the switching surface. As mentioned above, we first need to design a proper switching surface to guarantee the stability of the system in the sliding mode. To achieve this aim, a proportional-integral (PI) switching surface is suggested as

$$\sigma = e_1 + \int_0^t (a_1 e_1(\tau) + a_2 e_2(\tau)) d\tau \tag{6}$$

where $\sigma(t) \in R$ and a_1 and a_2 are design parameters which can be easily determined later. As well known, when the system operates in the sliding mode, it satisfies the following equations [9,10]

$$\sigma(t) = \dot{\sigma}(t) = 0 \tag{7}$$

From (7), one has

$$\dot{e}_1 = -a_1 e_1 - a_2 e_2 \tag{8}$$

Therefore, the following equivalent sliding mode dynamics can be obtained as

$$\begin{bmatrix} \dot{e}_1 \\ \dot{e}_2 \end{bmatrix} = \begin{bmatrix} -a_1 & -a_2 \\ b & 0 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \tag{9}$$

Obviously, if the design parameters $a_1 > 0$ and $a_2 > 0$ are selected, the characteristic polynomial $p(s) = s^2 + a_1 s + a_2 b$ has all eigenvalues with negative real part and the stability of equivalent sliding mode dynamics (9) is surely guaranteed.

Having established the appropriate switching surface (6), as described above, the next step is to design an ASMC to guarantee the hitting of the switching surface $\sigma(t) = 0$. To guarantee the occurrence of the sliding mode, an ASMC scheme is proposed as

$$u(t) = -\eta(t) - \alpha(\theta |\psi(t)|) sign(\sigma) \tag{10}$$

Where $\alpha > 1$, $\eta(t) = e_1(x_{s1} + x_{m1} + a_1 - 1) - (1 - a_2)e_2$, $\psi(t) = x_{s1}^3 - x_{m1}^3 - e_1(x_{s1} + x_{m1})$

The adaptive law is

$$\dot{\theta} = |\psi||\sigma|; \theta(0) = \theta_0 \tag{11}$$

where θ_0 is the positive and bounded initial values of θ.

The proposed adaptive control scheme above will guarantee the occurrence of the sliding mode, and is proven in the following theorem.

Theorem 1: Consider the error dynamics (5), if this system is controlled by $u(t)$ in (10) with adaptation law (11). Then the system trajectory converges to the switching surface $\sigma(t) = 0$.

Before proving Theorem 1, the Barbalat's lemma is given below.

Lemma 1 (Barbalat's lemma) [11]**:** If $w: R \to R$ is a uniformly continuous function for $t \geq 0$ and if $\lim_{t \to \infty} \int_0^t |w(\lambda)| d\lambda$ exists and is finite, then $\lim_{t \to \infty} w(t) = 0$.

After introducing Lemma 1, we are ready to prove Theorem 1.

Proof of Theorem 1: Let

$$\hat{\theta} = \theta - |r| \tag{12}$$

where r is a unknown but bounded constant. Thus the following expression holds.

$$\dot{\hat{\theta}} = \dot{\theta} \tag{13}$$

Consider the following Lyapunov function candidate

$$V(t) = \frac{1}{2}(\sigma^2 + \hat{\theta}^2) \tag{14}$$

Evaluating the time derivative of $V(t)$ gives:

$$\dot{V}(t) = \sigma\dot{\sigma} + \hat{\theta}\dot{\hat{\theta}} \tag{15}$$

Introducing (5), (6), (10),(11) and (13) into (15) yields

$$\begin{aligned}
\dot{V}(t) &= \sigma(\dot{e}_1 + a_1 e_1 + a_2 e_2) + \hat{\theta}\dot{\hat{\theta}} \\
&= \sigma[e_1(x_{s1} + x_{m1} + a_1 - 1) - (1 - a_2)e_2 - r(x_{s1}^3 - x_{m1}^3 - e_1(x_{s1} + x_{m1})) + u] + \hat{\theta}\dot{\theta} \\
&\le |r| \cdot |\psi| \cdot |\sigma| - \alpha(\theta|\psi|)sign(\sigma)\sigma + \hat{\theta}\dot{\theta} \\
&\le (\overbrace{|r| - \theta}^{-\tilde{\theta}}) \cdot |\psi| \cdot |\sigma| - (\alpha - 1)(\theta|\psi|)|\sigma| + \hat{\theta}\dot{\theta} \\
&= -(\alpha - 1)(\theta|\psi|)|\sigma|
\end{aligned} \tag{16}$$

Since $\alpha > 1$ has been specified in (10), we obtain the following inequality

$$\dot{V}(t) \le -(\alpha - 1)(\theta|\psi|)|\sigma| \tag{17}$$

Now if we define $|w(t)| = (\alpha - 1)(\theta|\psi|)|\sigma| > 0$, and integrating the above equation from zero to t, it yields

$$V(t) \le V(0) - \int_0^t |w(\tau)| d\tau \Rightarrow \quad V(0) \ge V(t) + \int_0^t |w(\tau)| d\tau \ge \int_0^t |w(\tau)| d\tau \tag{18}$$

Since $V(t) \ge 0$ and $V(0)$ is bounded, taking the limit as $t \to \infty$ on both side of (18) yields

$$\infty > V(0) \ge \lim_{t \to \infty} \int_0^t |w(\tau)| d\tau \tag{19}$$

Thus according to Barbalat's lemma (see Lemma 1), we obtain

$$\lim_{t \to \infty} |w(t)| = \lim_{t \to \infty} (\alpha - 1)(\theta|\psi|)|\sigma| \to 0 \tag{20}$$

Since $(\alpha-1)>0$ and$(\theta\cdot|\psi|)>0$ for all $t>0$, (20) implies $\sigma(t)\to 0$ as $t\to\infty$. Hence the proof is achieved completely.

Remark: If the error system (5) is driven by the control input $u(t)$ given in (10) with adaptation law (11), according to Theorem 1, the trajectory converges to the sliding mode. When the error dynamics (5) enters the sliding mode, the equivalent error dynamics system in the sliding mode is obtained as shown in (9). As discussed previously, in (9), the design parameters $a_1>0$ and $a_2>0$ are selected to guarantee the asymptotical stability of the error dynamic system (9). Consequently, the asymptotical stability of the closed-loop error system is also ensured.

3 Numerical Example

In this section, simulation results are presented to demonstrate the effectiveness of the proposed adaptive synchronization algorithm. All the simulation procedures are coded and executed using the software of MATLAB. The system parameters are chosen as follows: $r=1, b=1, A=0.1$ and $f=129$. The initial states of the drive system (1) are $x_{m1}(0)=-0.7$, $x_{m2}(0)=1$ and initial states of the response system (3) are $x_{s1}(0)=0.3$, $x_{s2}(0)=0.5$. As mentioned in Section 2, the ASMC design procedure for synchronizing the drive and response FHN neurons can be obtained as follows. First, according to (9), we select $a_1=4>0$ and $a_2=4>0$ such that characteristic polynomial $p(s)=s^2+4s+4$ has eigenvalues $-2,-2$ and the stability of (9) is ensured. Consequently, the PI switching surface $\sigma(t)$ is constructed as:

$$\sigma(t)=e_1+\int_0^t(4e_1(\tau)+4e_2(\tau))d\tau \tag{21}$$

Then, from (10), (11) and (21), the ASMC is determined as

$$u(t)=-\eta(t)-1.1(\theta|\psi(t)|)sign(\sigma)\,;\dot{\theta}=|\psi||\sigma|;\theta(0)=\theta_0=0.4 \tag{22}$$

The simulation results are shown in Figs. 1-6 under the proposed ASMC (22) with the adaptation algorithm (23). The chaotic attractor of master FHN neurons is shown in Fig. 1. The response of the phase plane for the error dynamics is shown in Fig. 2. Fig. 3 shows the corresponding $\sigma(t)$ which converges to zero and the state responses for the controlled master-slave FHN neurons are shown in Fig. 4. The time responses of synchronization error and adaptation parameter are shown in Fig. 5 and Fig. 6, respectively. From the simulation results, it reveals that the trajectory of error dynamics converges to $\sigma=0$ and the synchronization error also converges to zero. Thus the proposed ASMC works well and master-slave FHN neurons from different initial values are indeed achieving chaos synchronization even when the system's parameters are fully unknown.

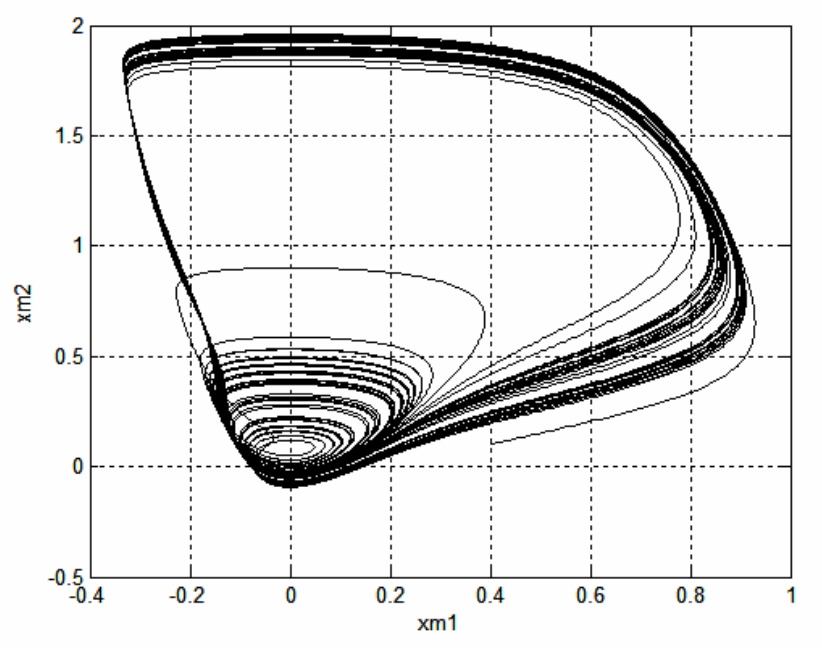


Fig. 1. The chaotic attractor

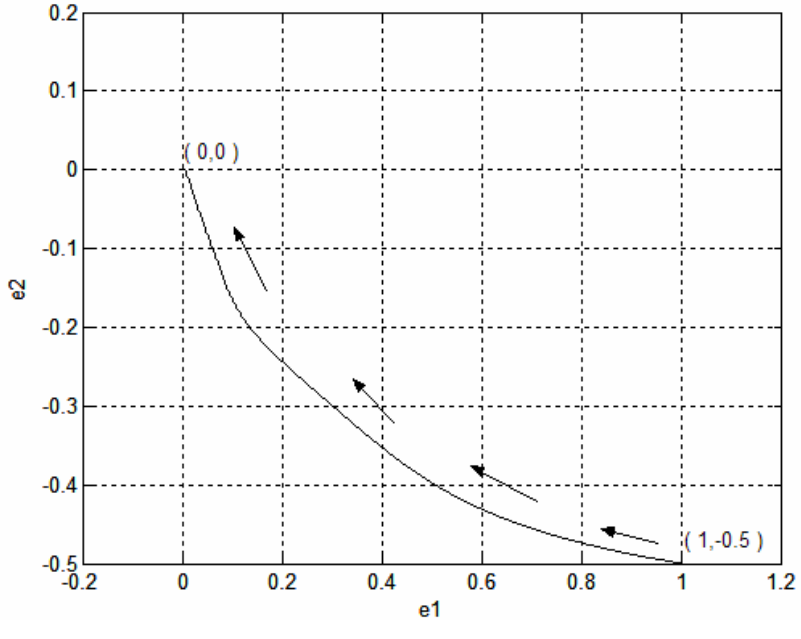


Fig. 2. Responses of the phase plane

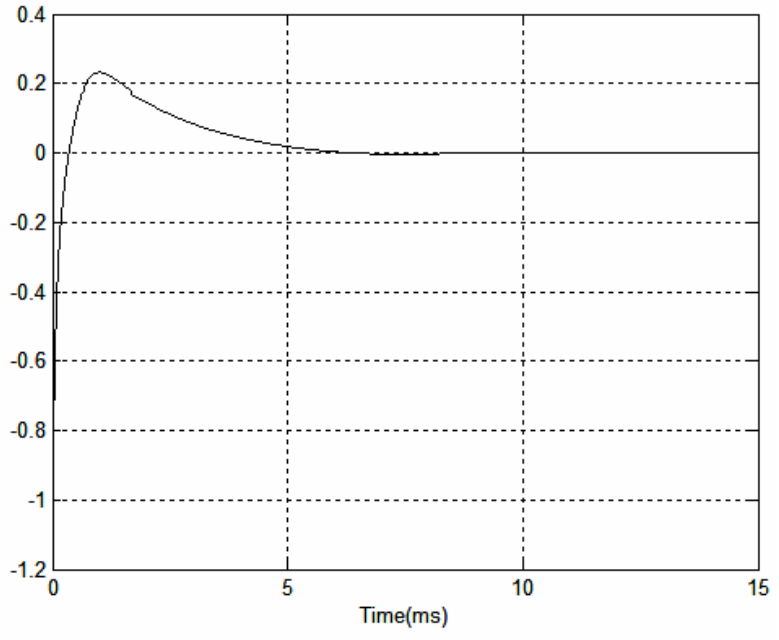


Fig. 3. Response of PI switching surface

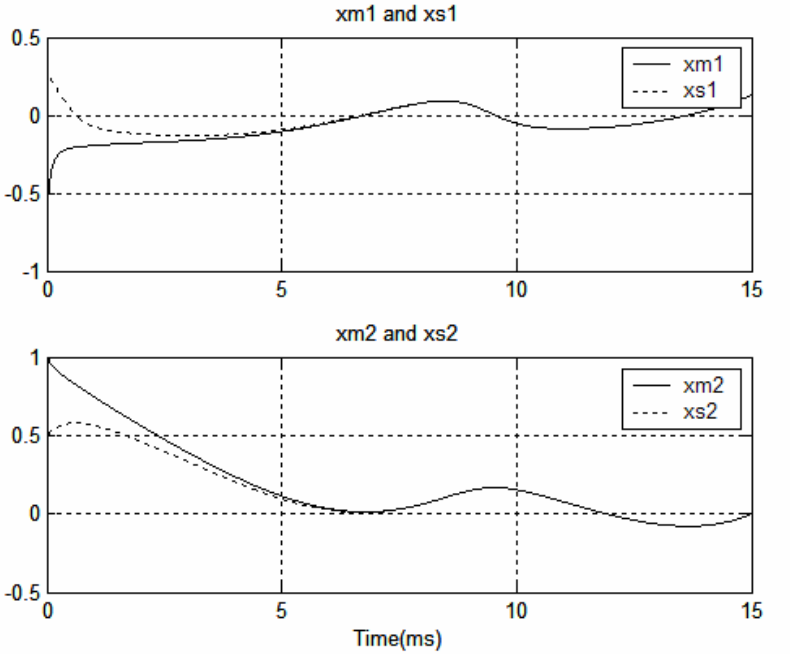


Fig. 4. The time response of FHN neurons

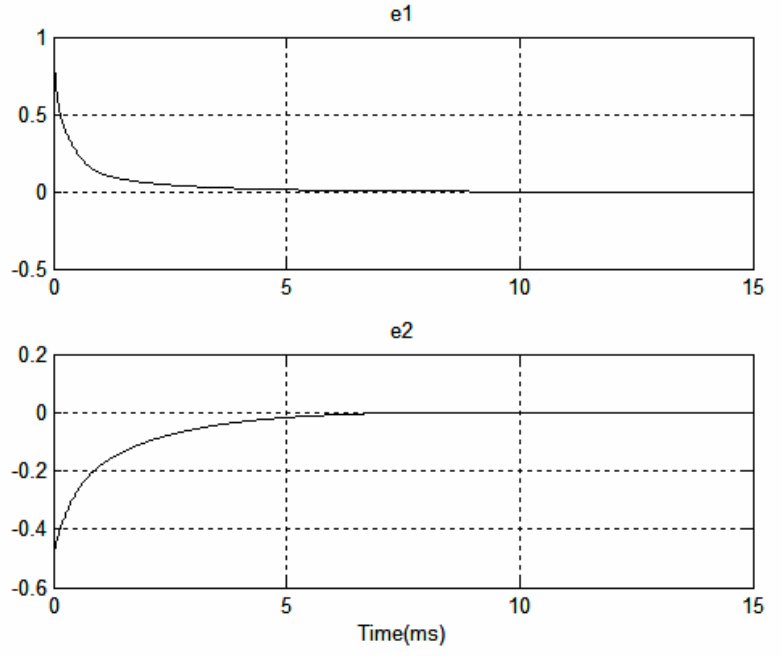


Fig. 5. Response of synchronization error state

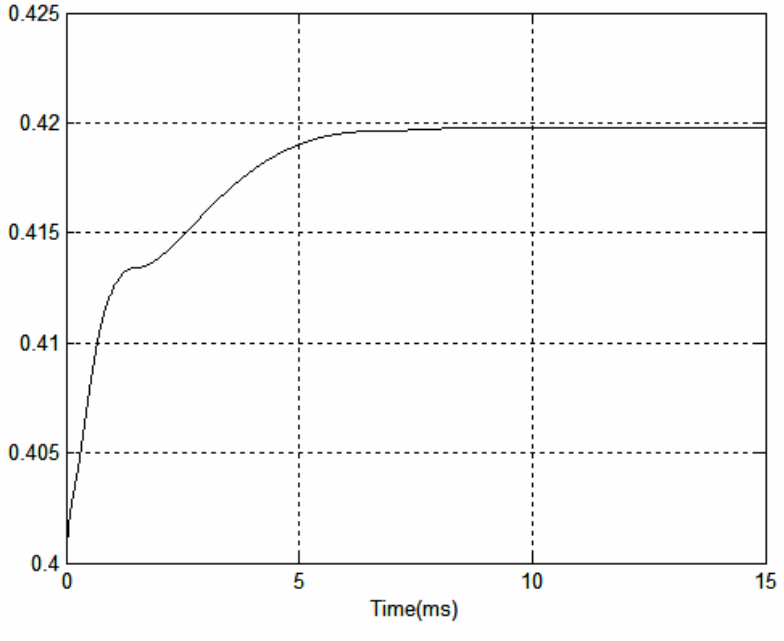


Fig. 6. Response of adaptation parameters

4 Conclusions

In this paper, an adaptive synchronization control for master-slave FitzHugh-Nagumo neurons in external electrical stimulation is demonstrated. Considering the fully unknown system parameters, a newly developed adaptive sliding mode controller has been proposed. Numerical simulations have verified the effectiveness of the proposed method.

References

1. Chua, L.O., Itah, M., Kosarev, L., Eckert, K.: Chaos synchronization in Chua's circuits. J. Circuit Syst. Comput. 3, 93–108 (1993)
2. Hegazi, A.S., Agiza, H.N., El Dessoky, M.M.: Synchronization and adaptive synchronization of nuclear spin generator system. Chaos, Solitons and Fractals 12, 1091–1099 (2001)
3. Han, X., Lu, J.-A., Wu, X.: Adaptive feedback synchronization of Lu system. Chaos, Solitons and Fractals 22, 221–227 (2004)
4. Park, J.H.: Adaptive synchronization of Rossler system with uncertain parameters. Chaos, Solitons and Fractals 25, 333–338 (2005)
5. Shoji, F.F., Lee, H.H.: On a response characteristics in the Hodgkin–Huxley model and muscle fiber to a periodic stimulation. IEEE 3, 2035–2041 (2000)
6. Thompson, C.J., Bardos, D.C., Yang, Y.S., Joyner, K.H.: Nonlinear cable models for cells exposed to electric fields. I. General theory and space-clamped solutions, Chaos, Solutions & Fractals 10, 1825–1842 (1999)
7. FitzhHugh, R.: Trashholds and plateaus in the Hodgkin–Huxley nerve equations. J. Gen. Physiol. 43, 867 (1960)
8. Shuai, J.W., Durand, D.M.: Phase synchronization in two coupled chaotic neurons. Phys. Lett. A 264, 289–297 (1999)
9. Itkis, U.: Control system of variable structure. Wiley, New York (1976)
10. Utkin, V.I.: Sliding mode and their applications in variable structure systems. Mir Editors, Moscow (1978)
11. Lin, J.S., Yan, J.J., Liao, T.L.: Chaotic synchronization via adaptive sliding mode observers subject to input nonlinearity. Chaos, Solitons and Fractals 24, 371–381 (2005)

Robust and Cost-Efficient Communication Based on SNMP in Mobile Networks

Sang-Hoon Ryu and Doo-Kwon Baik

Software System Lab., Dept. of Computer Science and Engineering, Korea University,
5 Anam-dong, Sungbuk-gu, Seoul, 136-701, Korea
{parake,baikdk}@korea.ac.kr

Abstract. A main challenge in the design of this mobile network is the development of dynamic routing protocols that can efficiently find routes between two communicating nodes. Multimedia streaming services are receiving considerable interest in the mobile network business. An entire mobile network may change its point of attachment to the Internet. The mobile network is operated by a basic specification to support network mobility called Network Mobility (NEMO) Basic Support. However, NEMO basic Support mechanism has some problem in continuous communication. In this paper, we propose robust and cost-efficient algorithm. And we simulate proposed method and conclude some remarks.

1 Introduction

A main challenge in the design of this mobile network is the development of dynamic routing protocols that can efficiently find routes between two communicating nodes. Multimedia streaming services are receiving considerable interest in the mobile network business. Supporting reliable real-time services is a decisive factor for the increasing migration toward packet-based mobile networks. The streaming clients become incorporated into advanced wireless communication devices.

An entire mobile network may change its point of attachment to the Internet. In this case, all communication between the nodes inside the mobile network and the global Internet must be maintained, although the mobile router (MR) connecting the mobile network to the global Internet must change their IPv6 address. The mobile network is operated by a basic specification to support network mobility called Network Mobility (NEMO) Basic Support [1].

In case of the NEMO, the more a mobile network node (MNN) moves to another area several times continually, the heavier the network overload, Round Trip Time (RTT) and disconnection are. These have an effect on real time multimedia communication between a correspondent node (CN) and a MNN. Therefore, this paper proposes keyword procedure for guaranteeing QoS and reducing disconnection which occurs in mobile network.

This paper is organized as follows. Section 2 represents some essential technology required at suggested the keyword connection mechanism. In section 3, we propose the SNMP information routing mechanism over wired network. In section 4, we

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 151–159, 2007.

describe some problems in NEMO and propose SNMP Information-based connection mechanism. In section 5, we evaluate the suggested new mechanism by calculating the effective cost in mathematical equations. In Section 6, we simulate the suggested mechanism and legacy mechanism and show result. Lastly in section 7 we conclude by placing our research in the context of further study.

2 Related Works

2.1 Mobile Network

The wireless environment, a mobile node would move in a random direction. Therefore reachability and connectivity for the mobile node become more important in the topology. When the mobile node moves to region of adjacent Access Router (AR), it seeks for new connection to Internet or CN so that both HA and CN are able to transmit all packets destined to the mobile node. Furthermore, all MRs that mobile node attaches have their own HAs. It causes pinball problem and handoff latency.

This arises when a router connecting a network to the Internet dynamically changes its point of attachment to the fixed infrastructure, thereby causing the reachability of the entire network to be changed in relation to the fixed Internet topology. Such a network is referred to as a mobile network. A mobile network acts as an access network and allows visiting nodes to attach to it [2].

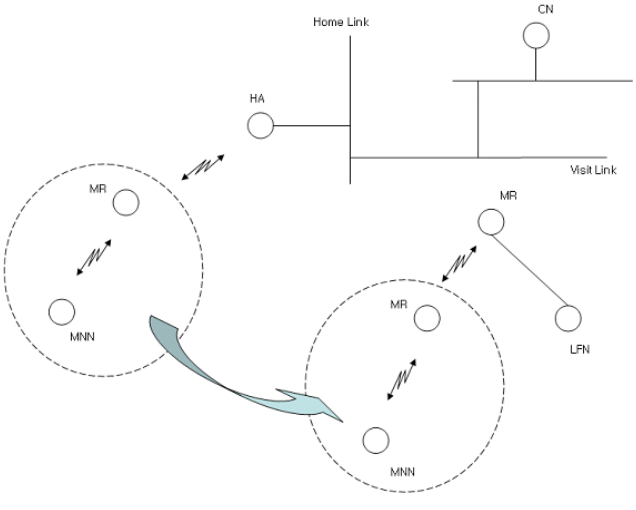


Fig. 1. A group of mobile nodes moves into new coverage under another AR in mobile networks

Figure 1 indicates a topology of Internet infrastructure and two Access Routers. The coverage of left AR contains a HA, a Local Fixed Node (LFN) and a sub-NEMO. And the sub-NEMO includes a MR, a LFN, a Local Mobile Node (LMN) and a MNN. And the coverage of right AR has a MR.

A mobile network is said to be multihomed when a mobile network (sub-NEMO) is attached to a larger mobile network (parent-NEMO). The aggregated hierarchy of mobile networks becomes a single multihomed mobile network. The root-NEMO is the mobile network at the top of the hierarchy connecting the aggregated multihomed mobile networks to the Internet. The parent-NEMO is the upstream mobile network providing Internet access to another mobile network further down the hierarchy. The sub-NEMO is the downstream mobile network attached to another mobile network up in the hierarchy [3]. It becomes subservient of the parent-NEMO. The sub-NEMO is getting Internet access through the parent-NEMO and does not provide Internet access to the parent-NEMO. The root-MR is the MR(s) of the root-NEMO used to connect the multihomed mobile network to the fixed Internet [4].

2.2 Hierarchical MIPv6

This hierarchical MIPv6 scheme introduces a new function, the Mobility Anchor Point (MAP), and minor extensions to the MH and the Home Agent operations. The CN operation will not be affected.

The introduction of the MAP concept minimizes the latency due to handoffs between access routers. Furthermore, the addition of bicasting to a MAP allows for Fast Handoffs, which will minimize the packet losses due to handoffs and consequently improve the throughput of best effort services and performance of real time data services over a radio interface. Just like MIPv6, this solution is independent of the underlying access technology, allowing Fast Handoffs within or between different types of access networks.

The introduction of the MAP concept will further diminish signaling generated by MIPv6 over a radio interface. This is due to the fact that a MN only needs to perform one local Binding Update (BU) to MAP when changing its layer 3 access points within the MAP domain.

The MAP will receive all packets on behalf of the MN it is serving and will encapsulate and forward them directly to the MN's current address. If the MN changes its current address within a local MAP domain, it only needs to register the new address with the MAP since the global CoA does not change. This makes the MN's mobility transparent to the CNs it is communicating with. The MAP can also be used to execute a Fast Handoff between ARs [5][6].

3 SNMP Information Routing Mechanism

It is very important to guarantee the cost-efficient transmission over wired network as well as wireless network. Because a MNN can receive continuous communication when the wired nodes have to process and support packet transmission. This section presents SNMP Information Based Hierarchical Routing Mechanism consisting of Keyup and Keyrout procedure over the wired network. Suggested mechanism abide by the fundamental assumption of beacon based routing. Namely, Beacon is a specific active node, and each active node should be connected to one beacon or more. Also in order to make a decision routing path, beacon broadcasts specific information to all of beacons in the network. The routing information of routing table existed at adjacent nodes can be transmitted by the Smart Packets [7].

3.1 Keyrout Procedure (Routing procedure by Keyword)

As shown in Figure 2, for example, by SNMP routing mechanism, service providers (Source in Figure 2) can distribute application program patches to their customers (Destination 1 and 2 in Figure 3) needing the patch fragments. For doing this, even though service providers do not know the exact IP addresses of their customer's nodes, as service providers just send any keyword (e.g. MS_Defense) to beacon (Beacon 1 in Figure 2), the application patch can be delivered to all of customer being in need of receiving the patch program.

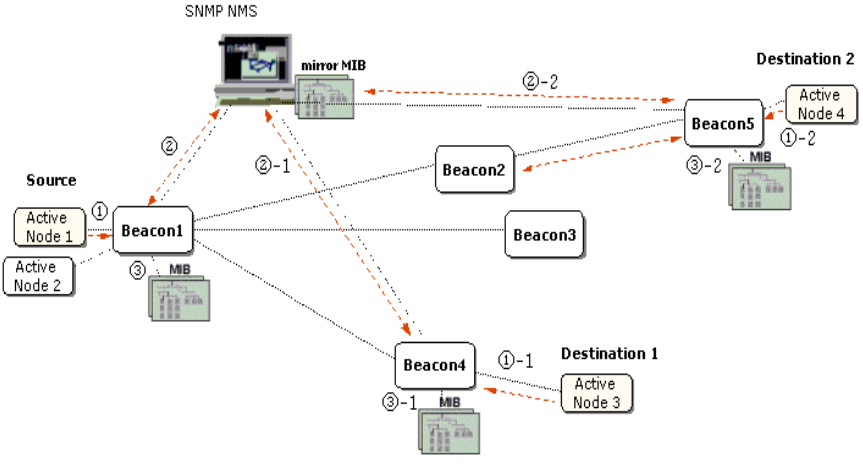


Fig. 2. Keyup Procedure

3.2 Keyrout Procedure (Routing procedure by Keyword)

Keyrout procedure in Figure 3 is routing path establishment procedure and in this procedure, the authenticated keyword derived from previous Keyup procedure is used. This procedure, for example, has next following steps.

Link setup for the path, Active Node 1 - Beacon 1 - Beacon 4 - Active Node 3 :

① Beacons(e.g. Beancon 1, 4 and 5) set up links to each active node, e.g. source node(Active Node 1) and destination node(Active Node 3 and 4), which already created keyword and made an request for authentication for the keyword.

② Beacon 1 which has been received keyword from source, Active Node 1, has got authentication over the keyword from the manager through the Keyup procedure. After doing that, Beacon 1 broadcasts the authenticated keyword to all of adjacent Beacons in the network.

③ Beacon 4 obtaining the keyword from Beacon 1 recognizes link information between Active Node 3 and itself. At the following step, link between Beacon 1 and Beacon 4 is established. Accordingly, full channel link is set up for the path, Active Node 1 - Beacon 1 - Beacon 4 - Active Node 3. Through this routing path, active packets are transmitted from Source to Destination 1.

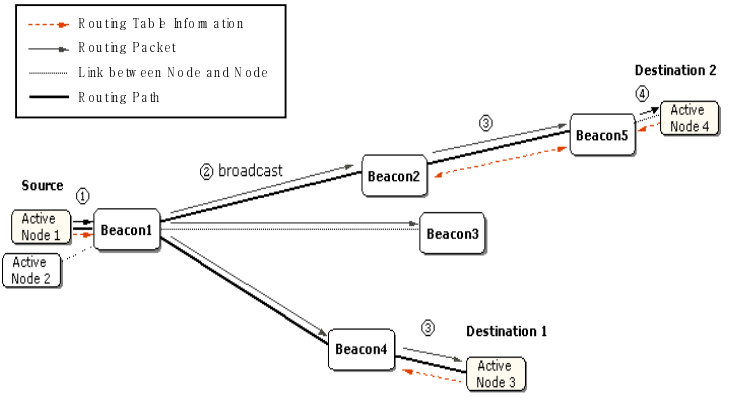


Fig. 3. Keyrout procedure

4 Communication Based on SNMP

4.1 Movement to Adjacent AR

When a MNN changes its points of attachment, it looks for new communication path using SNMP routing mechanism. Because MNN already has the same keyword as CN, it starts to distribute the authenticated keyword to adjacent MRs. If MRs do not have the same keyword in their own cache memory, they transmit the keyword in turn until the keyword message arrives at MR that have the matching one in cache memory. The new established connection avoids pinball problem and makes the most shortest path between HA and MNN. The HA also performs the same procedure to find new connection to MNN. HA distributes the authenticated keyword to adjacent MRs and MRs send the keyword to nearby nodes after checking if they have the keyword in cache memory. Hence, the time to find and set up new path is reduced by half as compared with general method. In addition, the CN processes the same procedure as well. As a result, two connection path for between MNN and HA, between MNN and CN are set up as quickly as possible. This means that CN and HA are able to send all packet destined to MNN which moves to new region of adjacent AR very rapidly.

Figure 4 represents that a MNN changes into other region for an adjacent AR. When the MNN which is located on the left HA migrates to the region of a right AR, the MNN should maintain current communication session between itself and CN. The HA1, HA2, HA3 and HA correspond to Home Agents of AR, MR2, MR3 and MNN respectively. The path that MNN sends BU message to HA using network mobility support is as follows: MNN → MR3 → MR2 → AR → HA1 → HA2 → HA3 → HA. However the SNMP information-based mechanism finds the effective connection. The path is as follows: MNN → MR3 → MR2 → AR → HA. Thus the packet from MNN to HA do not travel through home agents of MR3, MR2 and AR. The path that is set up by the proposed mechanism reduces handoff latency and end-to-end delay.

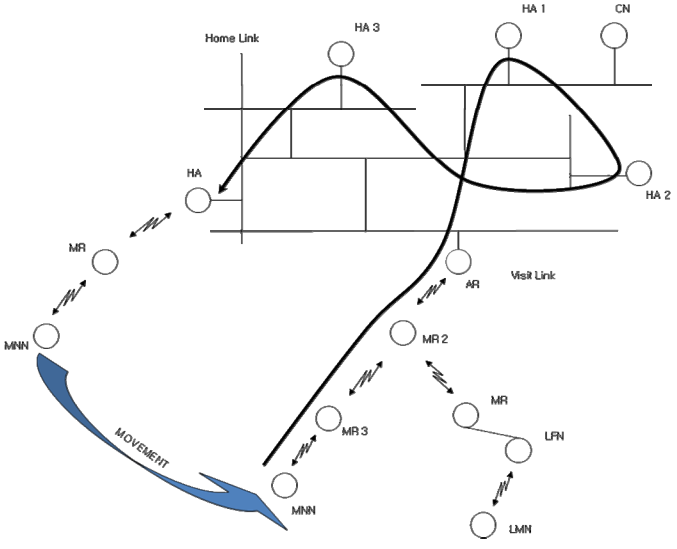


Fig. 4. A MNN attaches to sub-Network of neighbor AR in mobile networks

4.2 Movement to Adjacent MR within same AR

A MNN changes its point of attachment with the same region. The MNN attaches a different parent-MR, but it still belongs to a same grand parent-MR. In case of Network Mobility Support, the MNN should send BU message up to HA and CN. All of packets which head for MNN are forwarded to new connection path of MNN from CN. The delay takes place during connection configuration and causes connection disorder. Therefore, more efficient method is needed for streaming multimedia service.

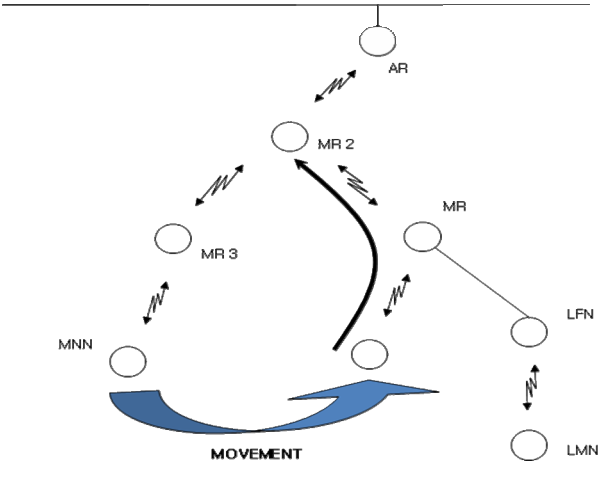


Fig. 5. A MNN moves to an adjacent MR within the same AR

Figure 5 shows the movement of a MNN forwards an adjacent MR within the same AR. After MNN changes its point of attachment, it distributes authenticated keyword to nearby MRs. The MR checks out if the same keyword is in cache memory. If MR does not have any information about received keyword, it sends the keyword to upper MR. The upper MR checks out the matching information after receiving keyword from lower MR. The upper MR, namely MR2, has the keyword in cache memory because it transmitted packets for MNN forwards MR3. New connection path for MNN is established and MR2 begins to send all packets destined to MNN through MR which is new attachment of MNN.

The proposed method only transmits BU message up to MR2 instead of HA. In cased of network mobility method, the BU message should go through All HAs of AR, MR2 and MR. It make big transmission cost and processing cost over Internet. Therefore, SNMP information-based mechanism is able to support packet transmission faster than legacy mechanism, network mobility support. And new method reduces the handoff latency, end-to-end delay and network overhead.

5 Evaluation and Analysis

5.1 Simulation Scenario

In order to investigate the impact of the proposed information-based connection mechanism and the traffic distribution mechanism on throughput and network traffic, the comparison targets are the throughput of TCP packet using internet protocol FTP and the exponential traffic in application.

In our simulation[8][9][10] by NS-2, a MNN attaches to a HA at first. The MNN starts to move and changes its point of attachment under the same AR after a moment. The MNN multicasts the RS message using the AgNemo module which we implement in NS-2. After the MNN receives the RA message with flag W, the MNN selects the most effective node of probable attachment nodes and connects to it. The MNN continues to move and at last the MNN migrates into the coverage of neighbor AR. The MNN should receive the ongoing communication packets continually. With this requirement supported, the MNN has to move in any direction freely.

5.2 Throughput

Each of 100 mobile nodes freely moves in any direction and keeps on changing its point of attachment while the MNNs communicate with the respective CNs. Therefore, the packet throughput depends on how effective method the MNNs alter its point of attachment by.

The graph of figure 6 shows the result of packet throughput between the mobile networks of the legacy mechanism and the proposed mechanism during simulation. When a MNN moves to the coverage of neighbor attachment node, the MNN sends or receives the packets to or from a CN fast if the MNN changes its point of attachment rapidly. Figure 6 shows the proposed mechanism is better than the legacy mechanism.

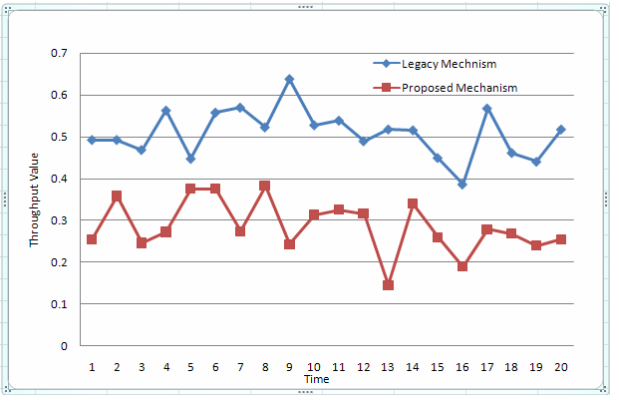


Fig. 6. Comparison result of throughput rate

In case of the legacy mechanism, the MNN chooses anyone of the attachment nodes. This legacy mechanism causes the heavy traffic and the unnecessary signals over the network. If the MNN changes its point of attachment and the upper node is full of many packets, the MNN dons not communicate with the CN well. Moreover it becomes more serious if the MNN wants to receive the multimedia streaming service in real time. Nobody wants to see the video which frequently stops for two or three seconds.

5.3 Traffic Analysis

Traffic occurs whenever intermediate nodes process the packets which become an exponential growth. Figure 7 shows that the traffic of the proposed mechanism is decreased less than the traffic of legacy mechanism. The reason is that network overhead is reduced by decreasing the messages both from MNNs to HA and from MNNs to CNs. Our proposed mechanism enables the attachment nodes to process the

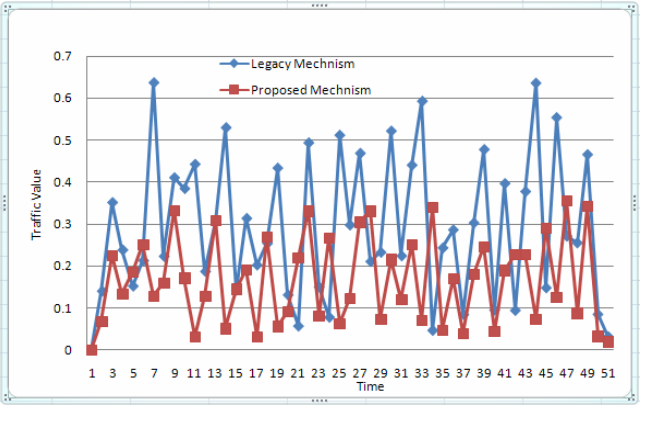


Fig. 7. Comparison Result of Traffic

packets faster. Therefore, the proposed mechanism optimizes the transmission path and reduces the delay time using the keyword connection algorithm after a MNN migrates to switching area and an attachment node can not process the packets in time.

Reduction of traffic overload over network is very important, because the more data packets are transmitted to a MNN, the more unnecessary signaling and traffic get larger. In Figure 7, an advantage of the proposed mechanism is that traffic of the proposed mechanism is lower than traffic of the legacy mechanism.

6 Conclusion

A MNN has to receive the multimedia streaming services and the real time application services in mobile networks while the MNN moves to any direction freely. The NEMO basic support algorithm only explains how a MNN changes its point of attachment. However in case that all nodes moves and the attachment node can not process the packet transmission for a new MNN, the legacy algorithm of the mobile networks is not enough.

To prove the QoS improvement in our approach, we model the proposed handoff mechanism and legacy NEMO handoff when setting up the routing convergence time, and then simulations were carried using NS-2.

The simulation results of new algorithm which is proposed in this paper show that it improves the packet throughput and the traffic rate. Therefore, the result of experiment validates that the proposed mechanism performs fast handoff and improves QoS by increasing the throughput rate and by decreasing the traffic rate.

Now we keep searching and simulating this research and we are going to apply the proposed mechanism to the heterogeneous systems.

References

1. Devarapali, V., Wakikawa, R., Petrescu, A., Thubert, P.: Network Mobility (NEMO) Basic Support Protocol. RFC3962 (2005)
2. Lamsel, P.: Network Mobility. Research Seminar on hot topics in Internet Protocols (2005)
3. Ernst, T., Lach, H.-Y.: Network Mobility Support Terminology. draft-ietf-nemo-terminology-03.txt (August 2005)
4. Ernst, T.: Network Mobility Support Goals and Requirements. draft-ietf-nemo-requirement-04 (August 2005)
5. Kupiainen, E.: Hierarhcical Mobile IPv6 Mobility Management (Feburary 2003)
6. Hagen, S.: IPv6 Essentials. O'reilly (July 2002)
7. Ryu, S.-H., Park, S.-H., Baik, D.-K.: SNMP Information Based Routing Mechanism for Fast Handoff in Mobile IP. Elsevier, Amsterdam (2005)
8. http://www.isi.edu/nsnam/ns/
9. The CMU Monarch Project's Wireless and Mobility Extension to ns, The CMU Monarch Project (August 1999)
10. Widmer, J.: Network simulations for A Mobile Network Architecture for vehicles. International Computer Science Institute Technology Report TR-00-009 (May 2000)

Stability of Backward Differential Formulae for Second Order Delay Differential Equations*

Y. Xu and J.J. Zhao

Department of Mathematics, Harbin Institute of Technology,
Harbin, 150001, China
yangx@hit.edu.cn,hit_zjj@hit.edu.cn

Abstract. This paper is concerned with the numerical stability of backward differential formulae for the second order differential equations with a lagging argument. A sufficient condition such that this kind of system is asymptotically stable is derived. Furthermore, it is proved that every A-stable BDF method preserves the stable properties of exact solutions for the second order delay differential equations. Finally, some numerical experiments are given to demonstrate the main conclusions.

Keywords: Delay Differential Equation, Numerical Method, Stability.

1 Introduction

Consider the delay integral differential equations

$$My'(t) = f(t, y(t), y'(t-\tau), G(t, y(t-\tau), y'(t-\tau), \int_{t-\tau}^{t} g(t, s, y(s))ds)), \quad (1)$$

where $f : R^+ \times R^d \times R^d \times R^d \to R^d$ is continuous, $M \in R^{d\times d}$ may be singular, and the constant τ is a positive lagging argument.

This kind of system can be found in a wide variety of scientific and engineering fields such as biology, physics, ecology and so on, which have a comprehensive list in [1]. Recently, there is a growing interest in developing numerical methods for solving the system (1).

As for the linear neutral Volterra delay integral differential equations

$$Au'(t) + Bu(t) + Cu'(t-\tau) + Du(t-\tau) + G\int_{t-\tau}^{t} u(x)dx = 0,$$

where $A, B, C, D, G \in R^{d\times d}$, and the matrix A may be singular. A sufficient condition such that the system is asymptotically stable was derived in [2]. It was

* This paper is supported by the Natural Science Foundation of Heilongjiang Province (A200602), the project (HITC2007) supported by Science Research Foundation and the project (HITQNJS.2006.053) supported by Development Program for Outstanding Young Teachers in Harbin Institute of Technology.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 160–165, 2006.

proved that every linear θ-method with $\theta \in (1/2, 1]$ and A-stable BDF method preserve the delay-independent stability of its exact solutions.

Reference [3] studied the nonlinear Volterra delay integral equations

$$y(t) = g(t) + \int_0^t f(\xi, y(\xi), y(\xi - \tau))d\xi,$$

the criteria on stability of BDF methods was given.

Up to now, few papers discuss the stability analysis of Volterra integral equations with a lagging argument (cf. [4,5]). However, this type of integral system arises in certain applications to impulse theory.

For the Volterra delay integral system of the second kind

$$y(t) = \varphi(0) + \int_0^t [ay(s) + b(t-s)y(s-\tau)]ds, \tag{2}$$

some suitable definitions for its stability of the numerical methods and plots of the stability regions for particular cases were included in [4].

Moreover, a sufficient condition was obtained for the linear θ-method to be TH-stable by differentiating the above Volterra equations with a lagging argument, i.e., the second order delay differential equations (cf. [5]).

In this paper, we focus on the numerical stability of backward differential formulae for the second order delay differential equations. A sufficient condition is given for the stability of this differential system. Furthermore, it is proved that the numerical methods preserve its stability if BDF method is A-stable. Finally, some numerical experiments are presented for illuminating the main conclusions.

2 Properties of Exact Solutions

Consider the second order differential equations with a lagging argument

$$\begin{aligned} y''(t) &= ay'(t) + by(t-\tau), \quad t \geq 0, \\ y(t) &= \varphi(t), \qquad\qquad -\tau \leq t \leq 0, \end{aligned} \tag{3}$$

where the parameters $a, b \in R$, the delay term $\tau > 0$ and the initial function $\varphi(t)$ is continuous. In fact, this second order delay differential system can be obtained by differentiating the Volterra integral equations (2) with a lagging argument.

Here, we assume that the conditions for the existence of a unique continuous solution are satisfied (cf. [6]).

Definition 1. *The second order differential system (3) with a lagging argument is said to be asymptotically stable if there exist a constant $r > 0$ such that, for any initial function value $\varphi(t)$ with $|\varphi(t)| < r$, its exact solutions $y(t)$ satisfy the condition*

$$\lim_{t \to \infty} y(t) = 0.$$

To study the stable properties of second order delay differential equations (3), we consider the corresponding characteristic polynomial

$$p(z) = z^2 - az - be^{-\tau z}. \tag{4}$$

A necessary and sufficient condition for its asymptotic stability of system (3) was given in [4].

Lemma 1. *Let λ be the root of the equation*

$$\cot \lambda = \lambda/(-a\tau),$$

in $(0, \pi)$, then all the zeros of characteristic polynomial (4) lie to the left of the imaginary axis if and only if

$$a < 0,\ b \geq 0,\ b\tau sin\lambda < -a\lambda. \tag{5}$$

Here, an important conclusion is quoted (see also [4]).

Lemma 2. *The second order delay di erential system (3) is asymptotically stable if and only if all the roots of its characteristic equation (4) have negative real parts.*

Furthermore, a sufficient condition can be easily obtained.

Theorem 1. *The second order delay di erential system (3) is asymptotically stable if it satisfies the condition*

$$b \geq 0,\ b\tau < -a. \tag{6}$$

3 Stability of BDF Methods

For the linear scalar autonomous test equation

$$y'(t) = \lambda y(t), \tag{7}$$

with $\Re\lambda < 0$, the BDF method (cf. Section 5.1 of [7]) gives out the recurrence relation

$$\sum_{i=0}^{s} \alpha_i y_{n-i} = h\lambda y_n, \tag{8}$$

where y_n can be interpreted as the approximation to $y(t_n)$.

Here we quote two important definitions (cf. Definition 8.2.2 in [8] or Definition 3.3 in [7]).

Definition 2. *The A-stability region of a numerical method for ordinary differential equation (7) is the set S_A of complex numbers $\alpha = h\lambda$ such that the numerical solution y_n obtained with the constant stepsize h satisfies*

$$\lim_{n\to\infty} y_n = 0.$$

Definition 3. *The numerical method for ordinary di erential equation (7) is A-stable if*

$$S_A \supseteq C^- = \{\alpha \in C | \Re(\alpha) < 0\}.$$

An important property of BDF methods is given (cf. Theorem 1.5 in [7]).

Lemma 3. *The BDF method satisfies the condition*

$$\Re(\sum_{i=0}^{s} \alpha_i z^{-i}) \geq 0,$$

for $\|z\| \geq 1$*, if it is A-stable.*

Applying the BDF methods (8) to solve second order delay differential equations (3), the recurrence relation could be written as

$$\begin{aligned} \sum_{i=0}^{s} \alpha_i y_{n-i} &= hY_n, \\ \sum_{i=0}^{s} \alpha_i Y_{n-i} &= haY_n + hby_{n-m}, \end{aligned} \tag{9}$$

where $h = \frac{\tau}{m}$ with $m \in Z^+$, y_n and Y_n can be interpreted as the approximations to $y(t)$ and $y'(t)$ at $t = t_n$, respectively.

Here, we formulate the definition of numerical stability.

Definition 4. *A numerical method applied to the second order delay di erential equations (3) is called stable, if for any initial function and fixed stepsize* $h > 0$*, the numerical solutions* y_n *satisfy* $\lim_{n\to\infty} y_n = 0$ *under condition (6).*

Through some simple computations, the difference equations (9) read

$$\sum_{i=0}^{s} \alpha_i \sum_{j=0}^{s} \alpha_j y_{n-j-i} = ha \sum_{i=0}^{s} \alpha_i y_{n-i} + h^2 b y_{n-m}. \tag{10}$$

To study the numerical stability of system (3), we consider its characteristic equation

$$(\rho(z))^2 - ha(\rho(z)) - h^2 b z^{-m} = 0, \tag{11}$$

where $\rho(z) = \sum_{i=0}^{s} \alpha_i z^{-i}$.

Theorem 2. *Every A-stable BDF method is stable for the second order delay di erential system (3) with condition (6) when fixed stepsize* $h < -\frac{a}{b\tau}$.

Proof. If the stepsize satisfies $h < -\frac{a}{b\tau}$, from condition (6), then all roots of equation

$$e^{\tau s}(s^2 - has) - h^2 b = 0,$$

have negative real parts. In other words, we have $e^{\tau s}(s^2 - has) - h^2 b \neq 0$ for $\Re s \geq 0$.

According to lemma 3, if the BDF method is A-stable, then $\Re(\rho(z)) \geq 0$ when $|z| \geq 1$. Thus, $z^m \cdot [(\rho(z))^2 - ha(\rho(z))] - h^2 b z^{-m} \neq 0$. which means that the modulus of every root is less than 1.

4 Numerical Experiments

Consider the second order delay differential equation

$$y''(t) = -200y'(t) + 500y(t-1), \quad t \geq 0, \tag{12}$$

with the initial value $y(t) = \cos t$ for $-1 \leq t \leq 0$. Here, the coefficients satisfy condition (6) in Theorem 1. Hence, this equation is asymptotically stable.

According to Theorem 2, the stepsize $h < 0.4$. Thus, we choose $h = 0.01$ in the following two examples.

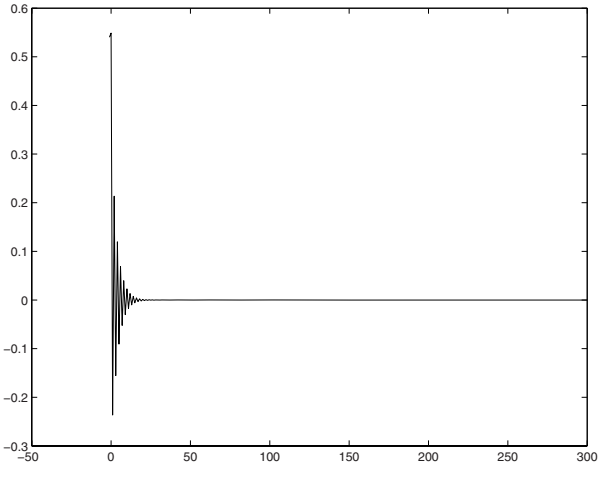


Fig. 1. BDF methods (13) with $h = 0.01$

Example 1. *Applying the BDF method (cf. [7])*

$$\frac{3}{2}y_n - 2y_{n-1} + \frac{1}{2}y_{n-2} = h\lambda y_n, \tag{13}$$

to equation (13) (see Fig. 1). Obviously, this method is A-stable, i.e., A(α)-stable with $\alpha = 90^0$.

Example 2. *Applying the BDF method (cf. [7])*

$$\frac{11}{6}y_n - 3y_{n-1} + \frac{3}{2}y_{n-2} - \frac{1}{3}y_{n-3} = h\lambda y_n, \tag{14}$$

to equation (13) (see Fig. 2). On the other hand, this method is A(α)-stable with $\alpha = 86^0 54'$, *but it is not A-stable.*

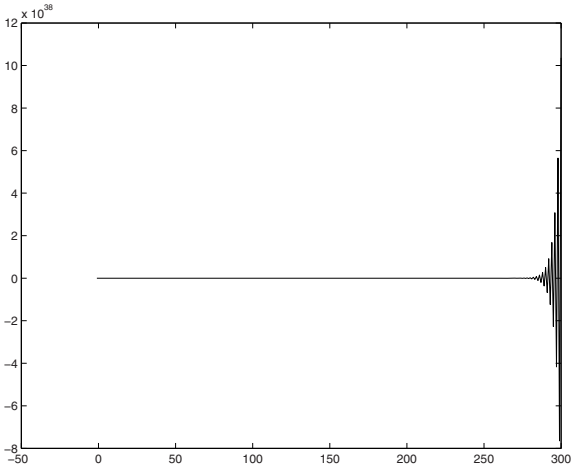


Fig. 2. BDF methods (14) with $h = 0.01$

References

1. Jerri, A.: Introduction to Integral Equations with Applications. Wiley, New York (1999)
2. Zhao, J.J., Xu, Y., Liu, M.Z.: Stability analysis of numerical methods for linear neutral Volterra delay integro-differential system. Appl. Math. Comput. 167, 1062–1079 (2005)
3. Zhang, C.J., Liao, X.X.: Stability of BDF methods for nonlinear Volterra integral equations with delay. Comput. Math. Appl. 43, 95–102 (2002)
4. Cahlon, B.: On the stability of Volterra integral equations with a lagging argument. BIT 35, 19–29 (1995)
5. Xu, Y., Zhao, J.J., Liu, M.Z.: TH-stability of θ-method for second order delay differential equation. Math. Numer. Sin. 26, 189–192 (2004)
6. Bownds, J.M., Cushing, J.M., Schutte, R.: Existence, uniqueness and extendibility of solutions of Volterra integral systems with multiple variable lags. Funkcial. Ekvac. 19, 101–111 (1976)
7. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations: Stiff and Differential-Algebraic Problems. Springer, Berlin (1991)
8. Bellen, A., Zennaro, M.: Numerical Methods for Delay Differential Equations. Oxford University Press, New York (2003)

Elastic Wave Propagation Simulation Using TLM Modeling

Lingyun Chai[1], Toshiyuki Meshii[1], and Yukio Kagawa[2]

[1] Graduate School of Engineering
Nuclear Power & Energy Safety Engineering Course, University of Fukui
3-931 Bunkyo, Fukui City, Fukui Prefecture, 910-8507 Japan
Chaily123@yahoo.co.jp,
meshii@mech.fukui-u.ac.jp
[2] Department of Electrical and Electronic Engineering,
College of Industrial Technology, Nihon University
Y2kagawa@cit.nihon-u.ac.jp

Abstract. A TLM nodal structure and its scatting matrix suitable for modeling the elastic wave propagation were developed by Langley et al [1]. In the present paper, extension is made to the inhomogeneous field in which an elastic wave propagation behavior between two different elastic media is considered for the normal and oblique incidences. The angles and amplitudes of the reflected and refracted waves are compared with those of the theoretical solution to confirm the validity of the present model.

1 Introduction

TLM stands for the transmission-line matrix modeling in analogy to the electrical transmission line which provides discretized model to the field. Unlike FEM and BFM, the field is modeled by a series of crossed transmission-lines. This results in the scattering matrix, which expressed the wave propagation in terms of the sequences of the impulses' scattering. The process is equivalent to what Huygens' principle states [2]. The modeling is suitable for the time-domain analysis. The TLM modeling was first developed by P.B.Johns for the electromagnetic wave propagation. The TLM nodal structure for the elastic wave propagation was developed by Langley et al [1] in two-dimensional elastic field. Its simplified version based on the displacement potential waves proposed by Kagawa et al[3]. With these modelings, two types of waves, the longitudinal wave and the shear wave, can be simulated in the homogeneous elastic medium. In this paper, with Langley's model is extended to the inhomogeneous field in which the behavior between two elastic media with different propagation velocity or different acoustic impedance is investigated. Both normal and oblique incidences are considered. The reflection and refraction on the plan interface are discussed for the angles and amplitudes.

2 Equations of Elastic Waves

The wave equations for the elastic waves are given for the in-plane deformation in $x-z$ plane as

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 166–174, 2007.

$$\frac{\rho}{\mu}\frac{\partial^2 u_x}{\partial t^2} = (2+\frac{\lambda}{\mu})\frac{\partial^2 u_x}{\partial x^2} + \frac{\partial^2 u_x}{\partial z^2} + (1+\frac{\lambda}{\mu})\frac{\partial^2 u_z}{\partial x \partial z}$$

$$\frac{\rho}{\mu}\frac{\partial^2 u_z}{\partial t^2} = \frac{\partial^2 u_z}{\partial x^2} + (2+\frac{\lambda}{\mu})\frac{\partial^2 u_z}{\partial z^2} + (1+\frac{\lambda}{\mu})\frac{\partial^2 u_x}{\partial x \partial z} \tag{1}$$

where u_x and u_z are the displacements in x and z direction. μ is the shear modulus, λ is Lame's constant in the plane stress situation and ρ is the material density. G and λ relate to the material properties according to

$$\mu = \frac{E}{2(1+\sigma)},\ \lambda = \frac{\sigma E}{(1+\sigma)(1-2\sigma)} \tag{2}$$

where E is Young's modulus and σ is Poisson's ratio.

The waves are propagated in two types of waves of the longitudinal wave and the shear wave with the propagation velocities of $c_L = \sqrt{(2\mu+\lambda)/\rho}$ and $c_S = \sqrt{\mu/\rho}$ respectively. The longitudinal wave travels faster than the shear wave as the velocity ratio is given by $\frac{c_L}{c_S} = \sqrt{\frac{2(1-\sigma)}{1-2\sigma}}$.

3 TLM Modeling

Langley et al built a pair of nodes stub-loaded with 8 branches as illustrated in Fig. 1, in which 9th branch is added for adjusting the wave propagation velocity. They are in pairs and coupled. The incident and reflected impulses are shown for $u_x(x,z)$ at (x,z) in Fig.2. The impedance of each branch is defined as fellows:

$$Z_{1x} = \frac{1}{K}Z_0\ ,\quad Z_{2x} = \frac{1}{K}Z_0\ ,\quad Z_{3x} = Z_0\ ,\quad Z_{4x} = Z_0\ ,\quad Z_{5x} = \frac{4}{H}Z_0\ ,\quad Z_{6x} = \frac{4}{H}Z_0\ ,$$

$$Z_{7x} = -\frac{4}{H}Z_0, Z_{8x} = -\frac{4}{H}Z_0,\ Z_{9x} = \frac{1}{\eta}Z_0. \tag{3}$$

where $K = 2+\frac{\lambda}{\mu}$, $H = 1+\frac{\lambda}{\mu}$ and the propagation velocity can be changed by the choice of parameter η in the 9th branch.

The incident pulses are ${}_kV^i_{1x}(x,z)$, ${}_kV^i_{2x}(x,z)$, ${}_kV^i_{3x}(x,z)$, ${}_kV^i_{4x}(x,z)$, ${}_kV^i_{5x}(x,z)$, ${}_kV^i_{6x}(x,z)$, ${}_kV^i_{7x}(x,z)$, ${}_kV^i_{8x}(x,z)$, ${}_kV^i_{9x}(x,z)$ at time step $t = k\Delta t$ and the response occurs on the same node at the next time step $t = (k+1)\Delta t$ and the reflected pulses are ${}_{k+1}V^r_{1x}(x,z)$, ${}_{k+1}V^r_{2x}(x,z)$, ${}_{k+1}V^r_{3x}(x,z)$, ${}_{k+1}V^r_{4x}(x,z)$, ${}_{k+1}V^r_{5z}(x,z)$, ${}_{k+1}V^r_{6z}(x,z)$, ${}_{k+1}V^r_{7z}(x,z)$, ${}_{k+1}V^r_{8z}(x,z)$, ${}_{k+1}V^r_{9x}(x,z)$. They are related by a scattering matrix in such a way as

$$
\begin{bmatrix} V_{0x} \\ V_{1x} \\ V_{2x} \\ V_{3x} \\ V_{4x} \\ V_{5x} \\ V_{6x} \\ V_{7x} \\ V_{8x} \end{bmatrix}^{k+1}_{(x,z)} = \frac{1}{2K+2+\eta} \begin{bmatrix} \eta-2K-2 & 2K & 2K & 2 & 2 & H/2 & H/2 & -H/2 & -H/2 \\ 2\eta & 2-\eta & 2K & 2 & 2 & H/2 & H/2 & -H/2 & -H/2 \\ 2\eta & 2K & 2-\eta & 2 & 2 & H/2 & H/2 & -H/2 & -H/2 \\ 2\eta & 2K & 2K & 2K-\eta & 2 & H/2 & H/2 & -H/2 & -H/2 \\ 2\eta & 2K & 2K & 2 & 2K-\eta & H/2 & H/2 & -H/2 & -H/2 \\ 2\eta & 2K & 2K & 2 & 2 & H/2-2K-2-\eta & H/2 & -H/2 & -H/2 \\ 2\eta & 2K & 2K & 2 & 2 & H/2 & H/2-2K-2-\eta & -H/2 & -H/2 \\ 2\eta & 2K & 2K & 2 & 2 & H/2 & H/2 & -H/2-2K-2-\eta & -H/2 \\ 2\eta & 2K & 2K & 2 & 2 & H/2 & H/2 & -H/2 & -H/2-2K-2-\eta \end{bmatrix}^{k}_{(x,z)} \begin{bmatrix} V_{0x} \\ V_{1x} \\ V_{2x} \\ V_{3x} \\ V_{4x} \\ V_{5x} \\ V_{6x} \\ V_{7x} \\ V_{8x} \end{bmatrix}^{k}_{(x,z)} \tag{4}
$$

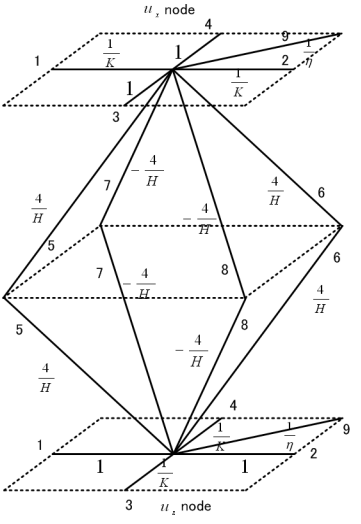


Fig. 1. A node structure for 2-dimensional field

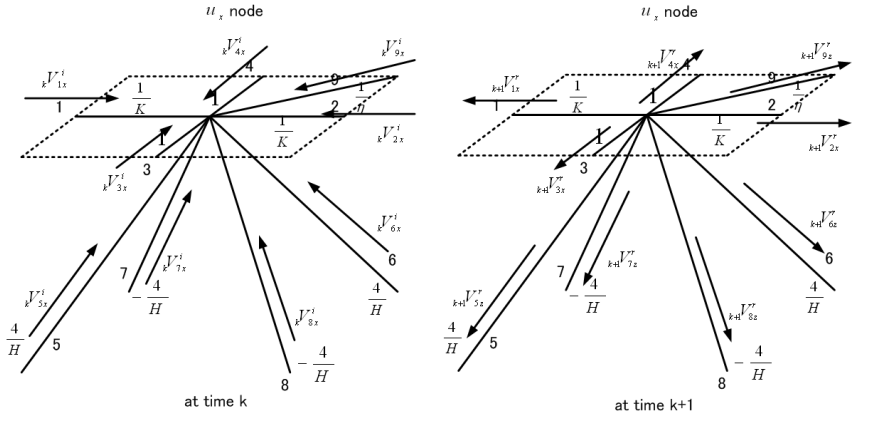


Fig. 2. Incident and reflected impulses at u_x node

$u_x(x,z)$ is evaluated by

$$
\begin{aligned}
{}_k u_x(x,z) &= \frac{1}{K+1+\eta/2}(\eta\, {}_k V_{9x}^i(x,z) + {}_k V_{1x}^i(x,z) + {}_k V_{2x}^i(x,z) + K\, {}_k V_{3x}^i(x,z) + K\, {}_k V_{4x}^i(x,z) \\
&+ \frac{H}{4}\, {}_k V_{5x}^i(x,z) + \frac{H}{4}\, {}_k V_{6x}^i(x,z) - \frac{H}{4}\, {}_k V_{7x}^i(x,z) - \frac{H}{4}\, {}_k V_{8x}^i(x,z))
\end{aligned} \tag{5}
$$

As the input pulses come from the reflected pulses of the adjacent elements, the compatibility conditions for connection are

$$
\begin{aligned}
&{}_{k+1}V_{1x}^i(x,z) = {}_{k+1}V_{2x}^r(x-1,z),\ {}_{k+1}V_{2x}^i(x,z) = {}_{k+1}V_{1x}^r(x+1,z),\ {}_{k+1}V_{3x}^i(x,z) = {}_{k+1}V_{4x}^r(x,z-1), \\
&{}_{k+1}V_{4x}^i(x,z) = {}_{k+1}V_{3x}^r(x,z+1),\ {}_{k+1}V_{5x}^i(x,z) = {}_{k+1}V_{6z}^r(x-1,z-1),\ {}_{k+1}V_{6x}^i(x,z) = {}_{k+1}V_{5z}^r(x+1,x+1), \\
&{}_{k+1}V_{7x}^i(x,z) = {}_{k+1}V_{8z}^r(x-1,z+1),\ {}_{k+1}V_{8x}^r(x,z) = {}_{k+1}V_{7}^r(x+1,z-1),\ {}_{k+1}V_{9x}^i(x,z) = {}_{k+1}V_{9x}^r(x,z),
\end{aligned} \tag{6}
$$

From Eqs. (2)–(5), one obtains the expression

$$
\begin{aligned}
&K({}_k u_x(x+1,z) - 2\, {}_k u_x(x,z) + {}_{k+1}u_x(x-1,z)) + ({}_k u_x(x,z+1) - 2\, {}_k u_x(x,z) + {}_{k+1}u_x(x,z-1) \\
&+ H/4({}_k u_z(x-1,z-1) - {}_k u_z(x-1,z+1) - {}_k u_z(x+1,z-1) + {}_k u_z(x+1,z+1)) \\
&= (K+1+\eta/2)({}_{k-1}u_x(x,z) - 2\, {}_k u_x(x,z) + {}_{k+1}u_x(x,z))
\end{aligned} \tag{7}
$$

This is a finite difference-time domain expression, which can be expanded in Taylor series about u_x to give the differential expression with higher order error terms removed

$$
\frac{\Delta t^2}{\Delta l^2}(K+1+\frac{\eta}{2})\frac{\partial^2 u_x(x,z)}{\partial t^2} = K\frac{\partial^2 u_x(x,z)}{\partial x^2} + \frac{\partial^2 u_x(x,z)}{\partial z^2} + H\frac{\partial^2 u_z(x,z)}{\partial x \partial z} \tag{8}
$$

Eq.(8) corresponds to the elastic wave equation (1) and the propagation velocities of the longitudinal wave and the shear wave can be given by $c_L = \sqrt{\dfrac{K}{K+1+\eta/2}}\,\dfrac{\Delta l}{\Delta t}$ and $c_S = \sqrt{\dfrac{1}{K+1+\eta/2}}\,\dfrac{\Delta l}{\Delta t}$ respectively.

With the similar procedure on the $u_z(x,z)$ node, the scattering matrix is given as

$$
\begin{bmatrix} V_{9z} \\ V_{1z} \\ V_{2z} \\ V_{3z} \\ V_{4z} \\ V_{5z} \\ V_{6z} \\ V_{7z} \\ V_{8z} \end{bmatrix}_{k+1}^{r} = \frac{1}{2K+2+\eta}
\begin{bmatrix}
\eta-2K-2 & 2 & 2 & 2K & 2K & H/2 & H/2 & -H/2 & -H/2 \\
2\eta & 2K-\eta & 2 & 2K & 2K & H/2 & H/2 & -H/2 & -H/2 \\
2\eta & 2 & 2K-\eta & 2K & 2K & H/2 & H/2 & -H/2 & -H/2 \\
2\eta & 2 & 2 & 2-\eta & 2K & H/2 & H/2 & -H/2 & -H/2 \\
2\eta & 2 & 2 & 2K & 2-\eta & H/2 & H/2 & -H/2 & -H/2 \\
2\eta & 2 & 2 & 2K & 2K & H/2-2K-2-\eta & H/2 & -H/2 & -H/2 \\
2\eta & 2 & 2 & 2K & 2K & H/2 & H/2-2K-2-\eta & -H/2 & -H/2 \\
2\eta & 2 & 2 & 2K & 2K & H/2 & H/2 & -H/2-2K-2-\eta & -H/2 \\
2\eta & 2 & 2 & 2K & 2K & H/2 & H/2 & -H/2 & -H/2-2K-2-\eta
\end{bmatrix}
\begin{bmatrix} V_{9z} \\ V_{1z} \\ V_{2z} \\ V_{3z} \\ V_{4z} \\ V_{5z} \\ V_{6z} \\ V_{7z} \\ V_{8z} \end{bmatrix}_{k}^{i}_{(x,z)} \tag{9}
$$

and the displacement $u_{z(x,z)}$ is evaluated by

$$
\begin{aligned}
{}_k u_z(x,z) &= \frac{1}{K+1+\eta/2}(\eta\, {}_k V_{9z}^i(x,z) + {}_k V_{1z}^i(x,z) + {}_k V_{2z}^i(x,z) + K\, {}_k V_{3z}^i(x,z) + K\, {}_k V_{4z}^i(x,z) \\
&+ \frac{H}{4}\, {}_k V_{5z}^i(x,z) + \frac{H}{4}\, {}_k V_{6z}^i(x,z) - \frac{H}{4}\, {}_k V_{7z}^i(x,z) - \frac{H}{4}\, {}_k V_{8z}^i(x,z))
\end{aligned} \tag{10}
$$

From the scattering matrices for two adjoining nodes, one again obtains the finite difference-time domain expression equivalent to eq.(2), that is

$$
\frac{\Delta t^2}{\Delta l^2}(K+1+\frac{\eta}{2})\frac{\partial^2 u_z(x,z)}{\partial t^2} = \frac{\partial^2 u_z(x,z)}{\partial x^2} + K\frac{\partial^2 u_z(x,z)}{\partial z^2} + H\frac{\partial^2 u_x(x,z)}{\partial x \partial z} \tag{11}
$$

4 Simulations

4.1 Normal Incidence

The simulation field for the normal incidence is shown in Fig.3, which consists of $600\Delta l \times 600\Delta l$ mesh and surrounded with non-reflective boundary termination. The Poisson's ratio of each medium is taken to be $\sigma = 0.3$ ($K = 3.5$). The whole field is divided into two half spaces with different propagation velocity. A line source made of the Gaussian waveform pulse is emitted at $x = 300\Delta l$. The displacement $u_a (= \sqrt{u_x^2 + u_z^2})$ are recorded alone the observation line $z = 300\Delta l$ at time step $t = 240\Delta t$, $280\Delta t$, $320\Delta t$, $360\Delta t$, as shown in Fig.4.

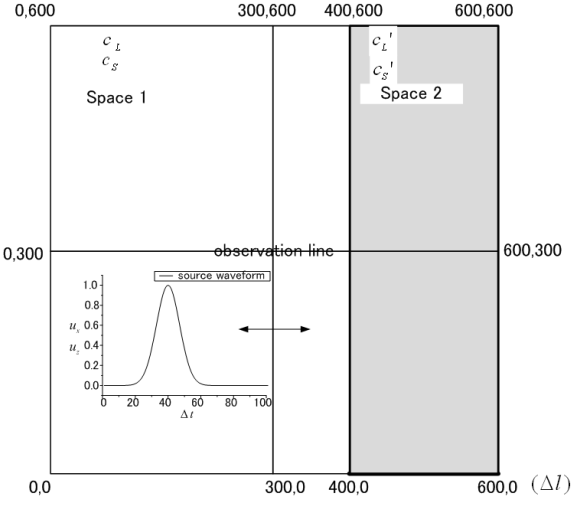


Fig. 3. Simulation field for normal incidence

4.1.1 Wave Propagation Velocities

We use parameter η to change the velocity, which is here chosen $\eta = 0$ in the space 1 and $\eta = 5$ in the space 2. The velocity ratio c_S / c_L must be 0.5345 for two kinds of the waves and velocity ratio between the two media $c_L' / c_L = 0.80178$ for the longitudinal waves. In the TLM modeling simulation, they can be evaluated from the propagation of the elastic waves along the observation line. During $40\Delta t$ time interval, the longitudinal waves are propagated by $35\Delta l$ in space 1 and $28\Delta l$ in space 2, while the shear waves are propagated by $19\Delta l$ in space 1, $15\Delta l$ in space 2. The velocity ratios are $c_S / c_L = 0.5428$, $c_L' / c_L = 0.8$, which are in good agreement with the values of the theoretical solution.

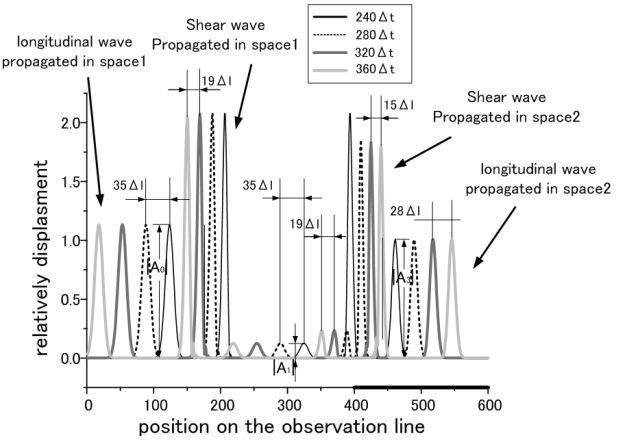


Fig. 4. Displacement $u_a (= \sqrt{u_x^2 + u_z^2})$ are recorded along the observation line $z = 300\Delta l$ at each time step

4.1.2 Amplitudes

As the reflection and refraction of the longitudinal waves are given along the observation line, the amplitude ratios of the reflected and refraction waves are evaluated to be

$$\left|A_1 / A_0\right| = 0.1109\,,\ \left|A_3 / A_0\right| = 0.88878 \tag{12}$$

which are again in good agreement with the theoretical solutions, as shown in the next section.

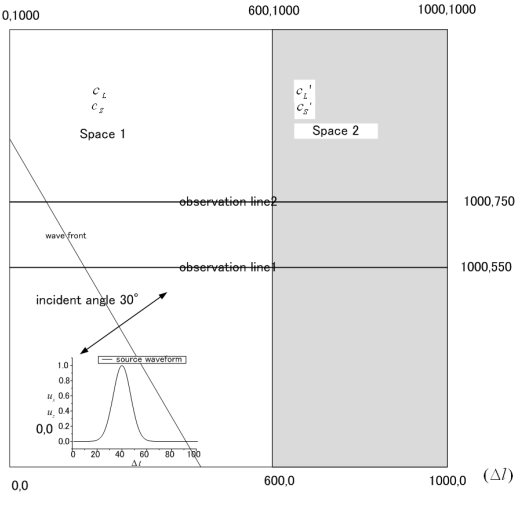


Fig. 5. Simulation field for oblique incidence

4.2 Oblique Incidence

The simulation field for the oblique incidence is shown in Fig.5, which consists of $1000\Delta l \times 1000\Delta l$ mesh and surrounded with non-reflective boundary. The Poisson's ratio of each medium is again taken to be $\sigma = 0.3$. The whole field is divided into two half spaces with different propagation velocity. A line source made of the Gaussian waveform pulse is emitted obliquely with angle $\theta_0 = 30^o$. The field of the wave are depicted in Fig.6 at time step $t = 600\Delta t$. The incidence, reflection and refraction waves are recorded along the observation line 1 and line 2 as shown in Fig.7. The angle and amplitude of each waves observed are given as below:

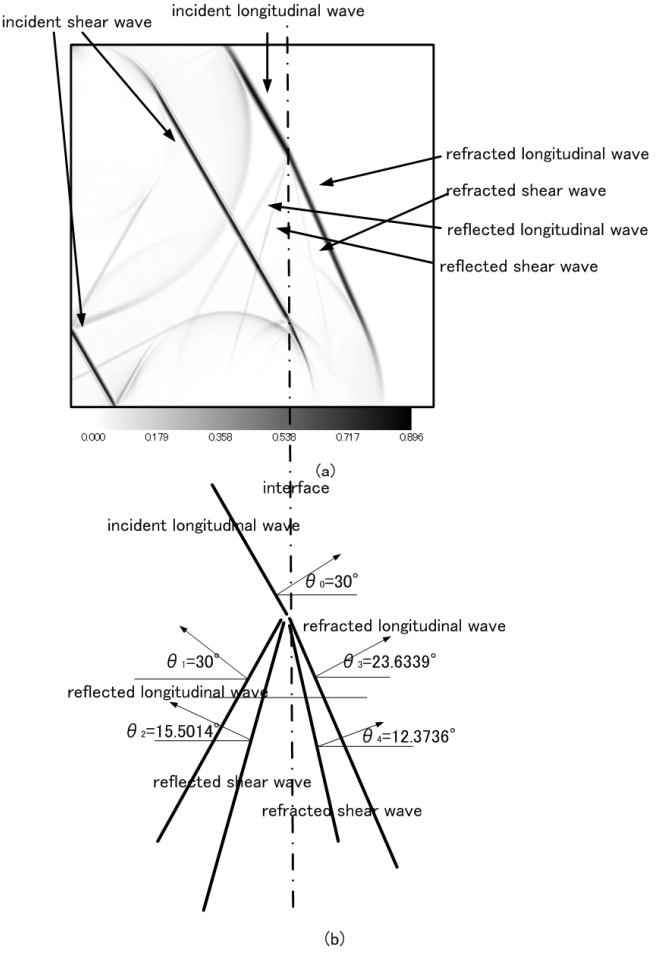


Fig. 6. Simulation for oblique incidence (a) Displacement $u_a (= \sqrt{u_x^2 + u_z^2})$ at time steps $t = 600\Delta t$ (b) Theoretical angle of each wave

The angle and amplitude of the incident longitudinal wave are $\theta_0 = 30^o$ and $|A_0| = 0.8493$.

For the reflected longitudinal wave, they are $\theta_1 = 30^o$ and $|A_1| = 0.064822$.

For the reflected shear wave, they are $\theta_2 = 15.2551^o$ and $|A_2| = 0.0.094677$.

For the refracted longitudinal wave, they are $\theta_3 = 23.6513^o$ and $|A_3| = 0.73048$.

For the refracted shear wave, they are $\theta_4 = 12.5288^o$ and $|A_4| = 0.074228$.

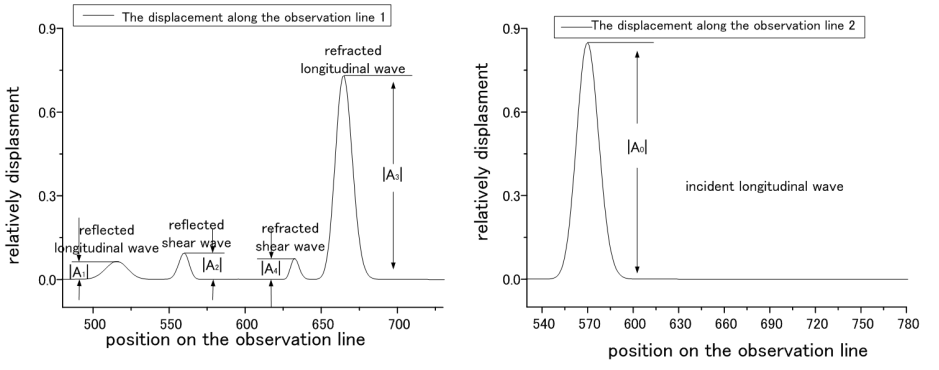


Fig. 7. Displacement $u_a = \sqrt{u_x^2 + u_z^2}$ along the observation lines 1 and 2

5 Theoretical Solution

We consider a plane interface separating the two elastic solid half spaces with different propagation velocity. The parameter $\eta = 5$ is taken for the second half space, which is the same as for TLM modeling simulation. When a longitudinal plane wave incident into the different velocity space with angle θ_0, the reflection and refraction of the longitudinal wave and the shear waves occur with the angle θ_1, θ_2, θ_3 and θ_4, respectively. In the plane $z = 0$ the displacements and the stresses must be compatible.

That is, the reflected and refracted waves must satisfy the continuity on the stresses and the displacements at the interface between the two spaces at $z = 0$.

From the Snell's law, one knows

$$\frac{\sin\theta_0}{c_L} = \frac{\sin\theta_1}{c_L} = \frac{\sin\theta_2}{c_S} = \frac{\sin\theta_3}{c_L'} = \frac{\sin\theta_4}{c_S'} \tag{13}$$

From the continuity condition, one obtains the following expression[4]

$$\begin{bmatrix} -\sin\theta_1 & -\cos\theta_2 & \sin\theta_3 & -\cos\theta_4 \\ \cos\theta_1 & -\sin\theta_2 & \cos\theta_3 & \sin\theta_4 \\ \sin 2\theta_1 & \frac{c_L}{c_S}\cos 2\theta_2 & \frac{c_L}{c_L'}\sin 2\theta_3 & -\frac{c_L}{c_S'}\cos 2\theta_4 \\ -\left(\frac{c_L}{c_S}\right)^2\cos 2\theta_4 & \frac{c_L}{c_S}\sin 2\theta_2 & \frac{c_L}{c_L'}\left(\frac{c_L}{c_S}\right)^2\cos 2\theta_4 & \frac{c_L}{c_S'}\sin 2\theta_4 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ A \end{bmatrix} = A_0 \begin{bmatrix} \sin\theta_0 \\ \cos\theta_0 \\ \sin 2\theta_0 \\ \left(\frac{c_L}{c_S}\right)^2\cos 2\theta_2 \end{bmatrix} \quad (14)$$

In a special case, under which the condition is $\theta_0 = 0$, $A_2 = A_4 = 0$, the amplitudes for the normal incidence of the longitudinal wave are

$$\frac{A_1}{A_0} = \frac{c_L' - c_L}{c_L' + c_L} = -0.111 \qquad \frac{A_3}{A_0} = \frac{2c_L}{c_L' + c_L} = 0.889 \quad (15)$$

For the oblique incidence with incident angle $\theta_0 = 30^o$ and incident amplitude $A_0 = 0.8493$, the reflected wave should be a longitudinal wave with angle $\theta_1 = 30^o$ and amplitude $A_1 = 0.0642035$ and a shear wave with angle $\theta_2 = 15.5014^o$ and amplitude $A_2 = -0.0933863$, while the refracted wave should be a longitudinal wave with angle $\theta_3 = 23.6339^o$, amplitude $A_3 = 0.732432$, a shear wave with angle $\theta_4 = 12.3736^o$ and amplitude $A_4 = -0.0748756$. They are well agreeing with those of the simulated values by TLM modeling as discussed in the section 4.1.2 and 4.2.

6 Conclusion

Langley's TLM node was applied to the case of inhomogeneous field. The elastic wave propagation behavior was examined between two media with different velocity. The reflected and refracted waves to longitudinal wave incidence were discussed for the angles and amplitudes. The simulated results are well agreed with the theoretical solution, which validates the modeling.

References

1. Langley, P., Pulko, S.H., Wilkinson, J.: A TLM Model of Transient, 2-Dimensional Stress Propagation. International Journal of Numerical Modelling: Electronic Networks, Devices and Fields 9, 429–443 (1996)
2. Christopoulos, C.: The Transmission-Line Modeling Method. IEEE Press, Los Alamitos (1995)
3. Kagawa, Y., Fujitani, T., Fujita, Y., Chai, L., Wakatsuki, N., Tsuchiya, T.: Discrete Huygens' Modelling Approach to Wave Propagations in Homogeneous Elastic Field. Journal of Sound and vibration 255(2), 323–335 (2002)
4. Achenbach, J.D.: Wave Propagation in Elastic Solids. North-Holland Publishing, Amsterdam (1973)

Research and Realization of Simulation Visualization Supporting Service

Ni Li, Xiaoyuan Peng, and Guanghong Gong

School of Automation Science and Electrical Engineering,
Beijing University of Aeronautics and Astronautics, Post Code 100083,
Beijing, P.R. China
lini@buaa.edu.cn

Abstract. X3D is currently the Web3D standard for modeling and rendering 3-Dimention visualization models on the Web. And Grid is the new-rising Internet computing technology to integrate and share different computing resources. In this paper, the simulation visualization environment architecture was built and the simulation visualization supporting service was realized by integrating X3D technology, grid technology and J2EE technology. It can support accessing, sharing and assembling different 3D models (such as airplanes, terrains, special effects, etc.) though Web browser. It also facilitates customizing 3D visualization simulation applications and the fast construction of simulation virtual environment. The simulation visualization environment architecture and the service realization method are introduced in detail.

Keywords: X3D; simulation visualization; grid technology; Web3D.

1 Introduction

The visualization technology is one of the important components of M&S supporting technology. 3-Dimension scene rending and driving is the common way in simulation visualization applications, which realize the 3D display in the runtime process of simulation. Simulation visualization supporting environment is a common framework to support the collaborative modeling and simulation of 3D visualization applications, depending on the different visualization needs in M&S applications.

Grid technology is a new-rising Internet computing technology to build the dynamic virtual organization based on distributed, heterogeneous and autonomous network resources for resource sharing and collaboration. Globus Project is one of the most representative grid projects, and at present it provides GT4 (Globus Toolkit 4.x) to support building grid applications [1].

X3D is the open standard for real-time 3D communication [2]. Based on VRML (Virtual Reality Modeling Language), X3D integrated XML (eXtensible Markup Language), Java and network streaming technology. So it can achieve more effective 3D computing ability, rendering quality and transfer velocity. It also has such benefits as platform-independent, component-based, more extensible and so on. All these facts of X3D technology provides an effective support for the research and realization of Web-based simulation visualization supporting service.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 175–183, 2007.

In the traditional simulation visualization application, the sharing and reusing of visualization models are not enough. There is no common supporting environment for developing simulation visualization applications. And for each application which has different application needs the development process always begins from scratch. So in this paper, the simulation visualization environment architecture and the visualization supporting service were researched and realized by integrating X3D technology, grid technology and J2EE technology. It can support users to access, share and assembly different 3D models (such as airplanes, terrains, special effects, etc.) though Web browser in Internet. It can also facilitate customizing 3D visualization simulation applications and building virtual simulation environment quickly. In the following sections, the simulation visualization environment architecture and the service realization will be introduced in detail.

2 Research on Web-Based Visualization Environment Architecture

The visualization environment architecture is shown in Fig.1. It consists of visualization resources layer, visualization service application logic layer and visualization portal layer.

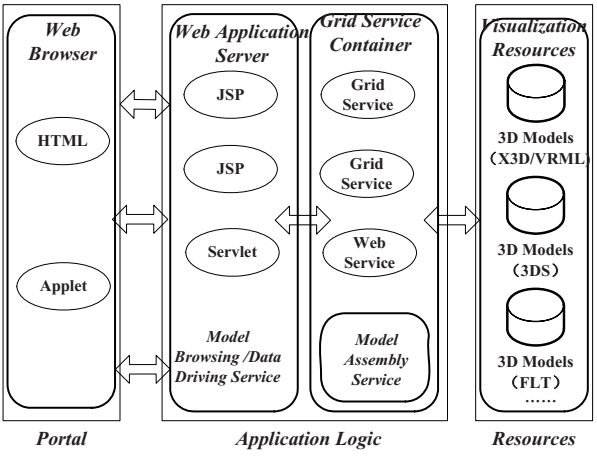


Fig. 1. Visualization supporting environment architecture which consists of visualization resources layer, visualization service application logic layer and visualization portal layer

Visualization resources layer. contains visualization model resources of different kinds (such as airplanes, terrains, special effects, etc.) and with different formats (such as 3D models in the format of X3D, VRML, 3DS, FLT, etc.). The 3D model resources which were developed with X3D-Edit, 3DMax, Maya, Multigen Creator and other 3D modeling software can be shared and reused.

Visualization service application logic layer. includes grid service container and Web application server. Grid service container connects with the visualization resources layer. Visualization model assembly function is provided as grid service/Web

service to realize the fast assembly and construction of large simulation visualization application scenes based on the bottom model resources. In the Web application server, components such as JSP and Servlet are used to process users' input and realize the browsing and data driving services of visualization models. In the phase of browsing models, the Web application server displays the 3D model through Web browser. In the phase of assembling models, it interacts with grid service container and revokes the visualization model assembly grid service to share and reuse the visualization models of different formats. In the phase of driving models, it communicates with Applet deployed in visualization portal layer and provides the simulation data to drive models.

Visualization portal layer. provides HTML pages and Applet, the two kinds of application interfaces to users in WAN to share and reuse visualization resources. Users can browse the visualization model resources and select the models meeting the application needs. The application portal provides the application custom page. It assists users to assembly and to generate the 3D virtual scene they need dynamically and quickly based on the bottom visualization model resources and X3D component specification. It also helps to realize the real time driving and simulation.

3 Assembling and Driving Visualization Models Based on X3D

3.1 Building and Browsing X3D-Format Visualization Models

In the visualization resource layer there are different formats of 3D model files. Since in the Web environment, X3D files has the advantages of better rendering quality, transfer velocity and extensibility. X3D is selected as the model file format to be transferred and shared in the visualization environment.

X3D/VRML browser software such as BS Contact [3] and XJ3D [4] can support rendering and displaying X3D/VRML-format models on Web pages. BS Contact is the commercial software which can be embedded in Web Browser to display X3D/VRML files. XJ3D is promoted by Web3D consortium. It is based on Java and it can work independently without Web Browser.

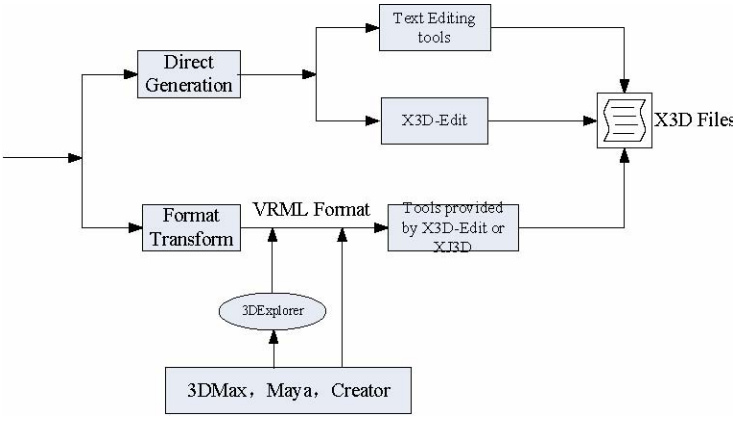


Fig. 2. Two ways to generate X3D-format visualization models

As shown in Fig.2, there are two ways to build X3D-format visualization models. One is to generate X3D files directly and the other is to transform models developed with other 3D modeling software to X3D models. For the first way, since X3D is a pure text format, developers who are familiar with X3D specifications can use general text edit software to develop X3D-format visualization models. X3D-Edit can also used to assist generating X3D files. It is platform-independent free software proposed by Web3D consortium. Through its human-machine interfaces the basic X3D components and nodes are provided to assist modeling basic geometry nodes and materials. However, for those 3D models which have complex geometry shapes, multi-textures and special effects (such as shadow, fog effects), it's better to take the second modeling way. Models can be developed with graphical modeling software such as 3DMax, Maya, Multigen Creator, etc. The models can be saved as VRML-format files and the texture pictures should also be saved together. Or the original format models (such as .3ds, .flt) can be transformed to VRML format first by using 3D software such as 3DExplorer. At last X3D-format model files can be generated by using VRML-X3D format transform tools provided by X3D-Edit or XJ3D.

Based on the X3D-format visualization models, BS Contact software is integrated here for displaying and rendering X3D/VRML-format visualization model files on Web pages.

3.2 Visualization Scene Assembly Based on X3D

The node of Inline is provided by X3D to support embedding other X3D models. And this function helps to realize the dynamic assembly of visualization models. The embedded X3D models can be operated as components. The component models' data such as position, orientation information can be set and revised. Since X3D file is in the form of XML, the X3D file can be dynamically generated and revised in JSP by adopting XML process tool. Different model files can be imported and embedded by using the node of Inline. The visualization scene can be dynamically assembled by indicating the initial position and orientation information of the embedded models. The main content of an X3D file is shown as follows, it implements the assembly of a terrain scene and an airplane model.

Example of a X3D file which implements the assembly of a terrain scene and an airplane model.

```
<Scene>
<Transform translation="0 0 0"><!--Setting the initial
postion-->
      <Inline url="beijing.x3d"/> <!--Embeding a terrain
scene file-->
    </Transform>
 <Transform translation="0 0 0"><!--Setting the initial
postion-->
      <Inline url="airplaneA.x3d"/><!--Embeding an
airplane model file-->
    </Transform>
    <Viewpoint position="0 0 50"/>
  </Scene>
```

X3D also provides Import and Export syntax function to operate the nodes inside the embedded models.

3.3 Driving Visualization Models Based on X3D SAI

For most simulation applications, in every simulation period the position and orientation data of simulation entities (such as airplanes, ships, and vehicles) should be updated and displayed. The event transfer function and SAI (Scene Access Interface) provided by X3D can be adopted to drive visualization models. SAI can be used to operate the objects in 3D scenes, such as loading scenes, revising fields' values, creating nodes, listening in the change of fields' values, adding event transfer function and walking around in virtual scenes. There are two ways provided by SAI to drive and update the display contents. The two driving ways are from inside X3D files and outside X3D files.

The driving way from inside X3D files

The node of Script is provided by X3D to support using SAI. Just as shown in Fig.3, the steps include editing the X3D file and editing the outside SAI-based programs.

The editing steps in X3D files are as follows:

– Firstly, defining the node of TimeSensor to set the data display and updating period.

Example of defining the node of TimeSensor: the field of cycleInterval determines the loop period and the field of loop determines whether running periodically or just once.

```
<TimeSensor DEF="TS" loop='true' cycleInterval='1.0' />
```

– Secondly, defining the node of Script to indicate the embedded program code file path, the interaction data and the data type.

Example of defining the node of Script: the field of accessType determines the data transfer direction. SFTime, SFVec3f, SFRotation are separately the field data type in X3D and they are used here to transfer the time, position and orientation data. The mark of <![CDATA[ ]]> is for indicating the path of executable file (.class file). JavaScript statements can also be embedded in the node of Script.

```
  <Script DEF="S" >
      <field     accessType='inputOnly'      type='SFTime'
name='pulse'/>
      <field    accessType='outputOnly'     type='SFVec3f'
name='location'/>
      <field   accessType='outputOnly'   type='SFRotation'
name='orientation'/>
<![CDATA[file:///C:/MyFile/X3D/Test3D/test/SAITest/MoveOb
jectScript.class]]>
   </Script>
```

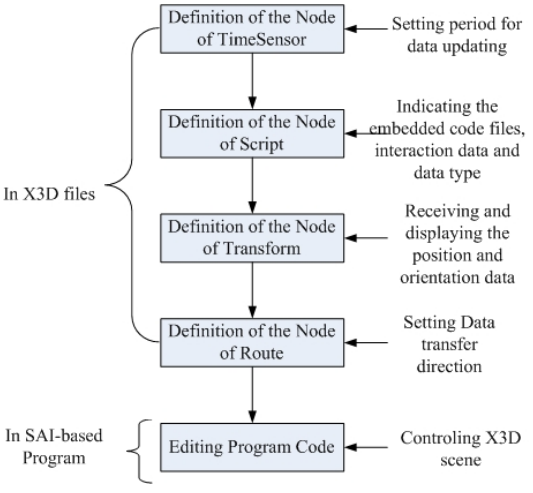


Fig. 3. The steps of driving way from inside X3D files based on SAI

– Thirdly, defining the node of Transform to receive the position and orientation data from executable file (.class file) or JavaScript codes for display.

Example of defining the node of Transform in X3D files.

```
<Transform DEF="TX" translation="0 0 0"/>
```

– Lastly, defining the event of Route to set the data transfer direction. A value is sent from one field of a node to one field of the other node so as to complete the event transfer process. And the dynamic behaviors will be caused by sending and receiving an event.

Example of defining the event of Route in X3D files.

```
<ROUTE fromNode="TS" fromField="cycleTime" toNode="S"
toField="pulse"/>
```

X3DScriptImplementaion interface should be implemented in the outside SAI-based program. The main steps include:

– Firstly implementing the operation of Initialization () to do some initialization operations such as connecting the driving data.
– Secondly implementing the operation of setFields () to add event listening functions for the input fields such as the field of pulse in the above example.
– Thirdly implementing the operation of shutdown () to release the resources such as destroying the event listening functions.
– Lastly the operation of readableFieldChanged () should be implemented to add the control function for the X3D scene such as updating the position and orientation data.

The driving way from outside X3D files

The driving way from outside X3D files is relatively more flexible and convenient. Firstly the nodes to be driven and updated should be identified with DEF field inside the X3D file. Then the node's data can be accessed and updated from outside program. An example is shown as follows:

Example of the segment in an X3D file.

```
<Transform DEF="TF" translation="0 0 0" rotation="0 0 1 0">
```

A timer can be set in the outside SAI-based program referring to the update period requirement. After identifying with DEF field, in each period of the timer, the handles of the Transform node and its fields such as translation and rotation can be attained by SAI mechanism. And the data value can be revised by the operation of setValue ().

4 Realization of Simulation Visualization Service Based on Web/Grid Technologies

4.1 Web-Based Model Browsing and Assembling

As illustrated in Fig.2, different kinds of X3D-format visualization models such as airplanes, ships, terrains and vehicles were generated. And a visualization modelbase was built to store and share these models (Shown in Fig.4). X3D-format model files in the modelbase can be browsed, uploaded and downloaded through Web Browser. And MIME (Multipurpose Internet Mail Extensions) should be configured properly when the application server starts, in order to support encoding X3D-format models.

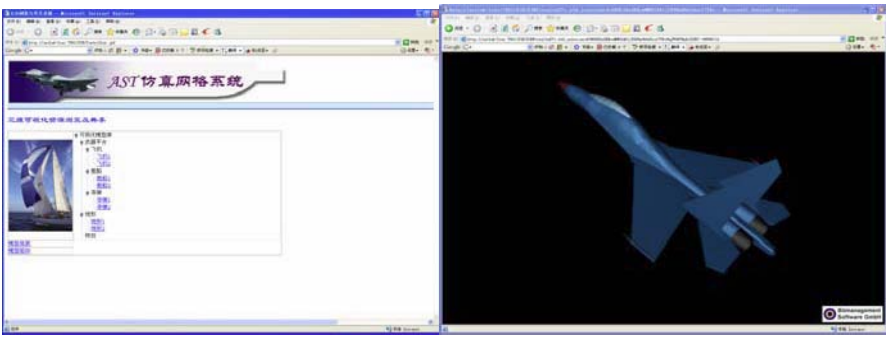


Fig. 4. Visualization Modelbase and Model Browsing Pages. The left figure shows the model tree structure listing all the models stored in visualization modelbase. And the right figure shows the X3D model of an airplane provided by modelbase.

Referring to the assembly method stated in section 3.2, JAXP (Java API for XML) provided by Sun was adopted here for construction and parsing of X3D files. The function of generating user-defined X3D model files was encapsulated as a common grid service, which was deployed in the GT4 grid service container. Such a common

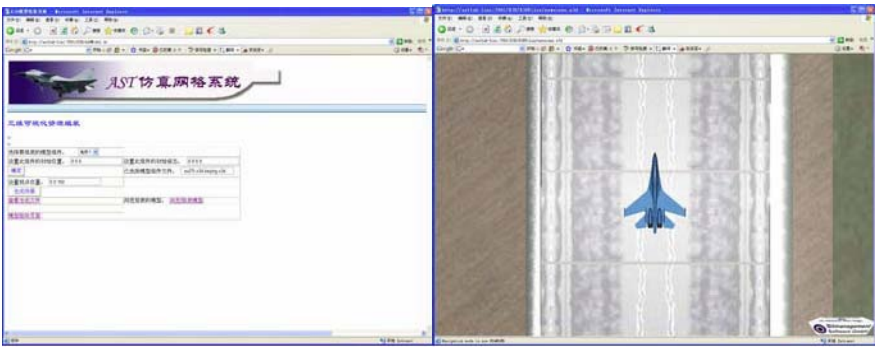


Fig. 5. Dynamic assembly of visualization models through web pages. The left figure is the configuration page for users to custom their simulation scene. After users select their needed models and indicate the initial position and orientation information of each model component, the new generated scene will be displayed on the Web page, just as the right figure shows.

grid service can also be accessed by other applications so that the model assembly function can be further shared and reused.

The visualization supporting service provides a portal for assembling X3D resources dynamically. Through the portal, users are able to custom their own visualization applications, by selecting different moving objects (such as airplanes, vehicles) or fixed models (such as a terrain), setting initial position and orientation data of these objects, and also doing other initializations such as setting illumination and cameras. The Web application server invokes the grid service which is working underneath and responsible for rebuilding X3D files implementing the user-defined application. And the rebuilt scene will be shown on the Web page immediately.

4.2 Web-Based Model Driving and Display

Since BS Contact mainly provides the function of browsing X3D/VRML files in Web Brower, it can not interact with external data to driving scenes and models. It is much

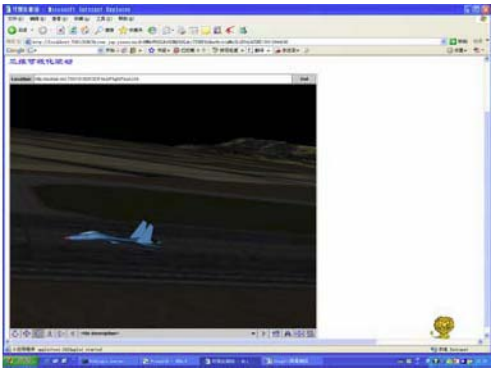


Fig. 6. Driving virtual scenes on Web pages. Applet communicates with Servlet to get the runtime data and drive the airplane on Web page based on SAI.

suitable for model browsing. While Xj3D provides the toolkit in java language for controlling and driving objects in the virtual scene through X3D SAI. So the simulation visualization supporting service adopts Xj3D to achieve interaction between models and external data. Xj3D APIs and SAI-based general drivers were packaged in Java Applet which was embedded in JSP so that user-specific visualization applications are able to run dynamically in web browser (Shown in Fig.6).

Driving objects in the virtual scenes with simulation runtime data was realized by the communication between the Servlet deployed in Web application server and the Applet embedded in Web page. Users upload the data files recording the movement of objects. Servlet is responsible for parsing these data files to get runtime data and deliver to Java Applet by HTTP protocol. The Applet receives these data, refresh and render the virtual scene periodically.

5 Conclusion

A Web-based visualization supporting architecture was researched and built in this paper and the service prototype was realized. By accessing the simulation visualization supporting service, users can share and reuse different kinds of model files though Web Browser without installing any 3D visualization modeling software. And a visualization simulation application can be quickly built on the base of visualization model resources by using the model assembly and data driving function. However, there are still a lot of research and development work to do to improve the service function and the visualization application display effects.

References

1. (2007), http://www.globus.org/toolkit/presentations/
2. (2007), http://www.web3d.org/x3d/
3. (2007), http://www.bitmanagement.de/products/bs_contact_vrml.en.html
4. (2007), http://www.web3d.org/x3d/xj3d/
5. Yin, G., Li, H., Fu, J.: Research on the Collaborative Virtual Products Development Based on Web and X3D. In: Proceedings of 16th International Conference on Artificial Reality and Telexistence-Workshops, pp. 141–144 (2006)
6. Li, B.H., Chai, X., Di, Y., et al.: Research on Service Oriented Simulation Grid. In: Proceedings of Autonomous Decentralized Systems, pp. 7–14 (2005)
7. Bouras, C., Panagopoulos, A., Tsiatsos, T.: Advances in X3D multi-user virtual environments. In: 7th IEEE International Symposium on Multimedia, p. 7 (2005)

Research on Simulation Application Mode and Run-Time Framework in Simulation Grid

Chang Feng Song[1], Xudong Chai[2], Bo Hu Li[1], and Zhen Tang[1]

[1] Beijing University of Aeronautics and Astronautics, Beijing 100083, China
[2] Beijing Simulation Center, Beijing 100854, China
song_chang_feng@yahoo.com.cn

Abstract. The combination of M&S with grid technology is an important developing direction for network-based modeling and simulation. Based on this combination, many relevant research projects have been launched all over the world, and Simulation Grid project we are developing is one of these projects. With the further research in Simulation Grid project, we believe simulation grid (SimGrid) is a new platform for modeling and simulation. In this paper, we introduce the simulation application mode in SimGrid environment. This mode puts emphasis on **simpleness** for user to implement simulation application, and requires SimGrid supplying ability on **self-organization** and **adaptiveness** in implement of simulation system. To support this simulation application mode in SimGrid, we propose a simulation run-time framework. Based on this simulation run-time framework, it is easy for user to implement simulation system, execute simulation, and achieve simulation results. In addition, the construction, execution and management of the simulation system is **perspective** to user.

Keywords: Simulation Grid (SimGrid); simulation application mode; simulation run-time framework.

1 Introduction

Networkitized modeling and simulation means a set of activities, including system modeling, simulation experiment, execution evaluation and so on, which are based modern network technologies[1,2]. After processing several typical stages such as DIS[3], ALSP[4] and HLA[5] in network-based modeling and simulation, with the new emerging grid technology[6] join, it promotes the rapidly developing in M&S.

Nowadays, many projects have been launched all over the world to research the combination of M&S with grid technology such as Cross Grid[7] and Alice Grid[8]. Among them Simulation Grid project[9] is one we are developing in China. With the further research and understanding on SimGrid, we believe it is not only a development in technology of M&S, but also a new platform for network-based M&S. In this paper, we analyze the simulation application mode and the process of simulation system implement in SimGrid environment. In order to support the creation and execution for this simulation application mode, we propose a simulation run-time framework.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 184–194, 2007.

The remainder of this paper is organized as follows: Section 2 contains the simple background of SimGrid. Section 3 introduces the simulation application mode in SimGrid. In the section 3, we first analyze the previous simulation application modes for network-based M&S, then we descript the simulation application mode in SimGrid and the details of implement of simulation system based on this mode, and finally we make a summary on the characters of this mode. In the section 4, we propose a simulation run-time framework for the simulation application mode mentioned in section 3. In this section we put research emphasis on the architecture, relationship of the components and on features of the framework. A concise conclusion is given in the last section.

2 Simulation Grid

Under the support of information domain of China national high performance 863 plan, high performance computer and core software projects, we process the project of the application research and development of Simulation Grid. This project is an important part of China national grid, and its background is to satisfy the demands of collaboration simulation based on complex product lifecycle.

SimGrid is a synthesized multidisciplinary technology and important tool, with the background of application field, synthetically applying complex system modeling technology, advanced distributed simulation technology, VR (virtual reality) technology, grid technology, management technology, system engineering technology and application field related specialized technology, to realize the dynamic share and reuse of various resource in grid/federation safely (including model/computing/ storage/network/data/ information/knowledge/software resource relative to unit involved in simulation application/project, and application related physical effect device and simulator, etc.), collaboratively cooperation, dynamic optimization of scheduling executions, etc., to support the various activities in full life cycle (from argumentation, research, design, development, test, execution, evaluation to maintenance and disposal) for existing or tentative complex system/project in engineering or non-engineering field" [9].

The architecture of SimGrid is illustrated as fig. 1.

We consider SimGrid as a new platform for modeling and simulation. As a new platform, an important point is that it provides a new mode for simulation application. This mode makes full use of the large mount of simulation resources and services provided by SimGrid environment, simulation system is implemented by SimGrid using these resources and services without participation of user, and SimGrid also manage the execution of the simulation application. All these work done by SimGrid decrease the difficulty of development of simulation application for user. Because this simulation mode is different from the past ones, the previous simulation run-time system can not support the new mode. We need a new run-time system to satisfy the needs required by the new mode.

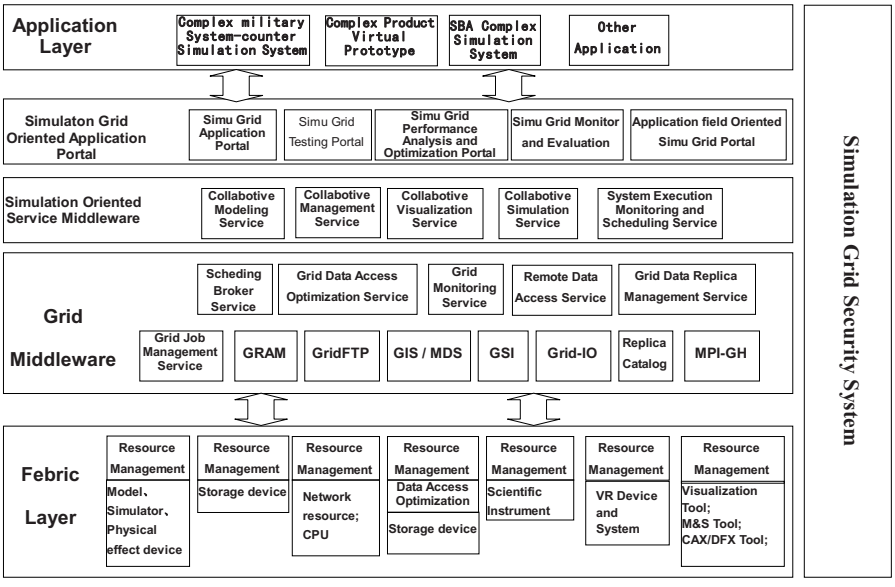


Fig. 1. Architecture of Simulation Grid

3 Simulation Application Mode in Simulation Grid

3.1 Traditional Simulation Application Mode

Traditionally a simulation application implement is as the following steps. After the analysis of the physical system for simulation, the physical system is decomposed into several sub-systems. Sub-system can be disassembled into a couple of simulation models which are abstracted as logical processes (LPs) according to the function. LPs are described by the computer programming language to implement the simulation function and to be compiled as executable codes, and then user assigns these codes on the distributed CPUs which are connected by the network. Based on the cooperative capacity of the LPs themselves or supported by the simulation run-time infrastructure such as RTI[10], a digital simulation system is constructed. Finally user can start up the simulation system which then advances according to causality to simulate the real system. From the steps we can see it requires user to design system, write programs, and deploy codes.

3.2 Simulation Application Mode in Simulation Grid

In SimGrid environment there exits lots of simulation resources including software resources and hardware resources, and these resources are shared as grid services[11]. These resources are the base components for construction of the simulation system

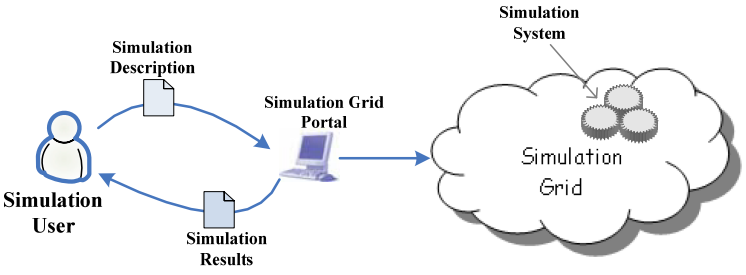


Fig. 2. Simulation Application Mode in Simulation Grid

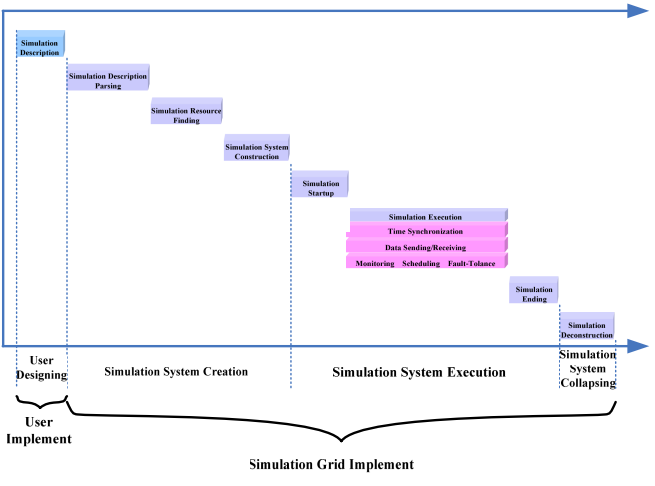


Fig. 3. Process of Simulation Application in Simulation Grid

which user expects. We expect the new simulation application mode in SimGrid makes easy to use, and this expectation is based on that SimGrid achieves implement and execution of simulation system by utilization of these resources/services without manual assistance. A simple illustration of the new mode is as fig. 2.

Under the mode, the process of simulation application is shown as the following fig. 3. Firstly, according to the analysis of the simulation application and decomposition of the simulation system, user abstracts the function components which are parts of the simulation system. User also needs to decide the relationship of communication between the components, the content and format of data for exchange, constrains for the components, and so on. After the end of design of the simulation system, user describes the system by using standard language which SimGrid can understand. Then user accesses SimGrid by desktop program or the portal of SimGrid, and submits the description script to SimGrid. At this time the work for user is finished and the following work is afforded by SimGrid. SimGrid parses the description submitted by user to understand the design ideas

of user and the understanding will go through all the following process. Then SimGrid searches the simulation model resources, the simulation tool resources, computing resources and other simulation services from registration center and models database for creation of the simulation system. When all needed resources or services are found, SimGrid organizes these resources and services to construct the simulation system as user expects and starts up the system. In the process of execution of the simulation system, SimGrid not only manages time synchronization and data receiving/sending, but also can provide the abilities to monitor the status and performance of each part component of the system. If exception, fault or decreasing of performance was detected, SimGrid would adjust the system according to the policies offered itself. These abilities decrease the possibility of the system collapse and increase the efficiency of the system. In addition, SimGrid will collect the middle state data in the execution phase as parts of the simulation results returned to user for further analysis after simulation. When achieving the condition defined by user to end simulation, SimGrid ends the execution of the simulation system, returns simulation results to user and finally deconstructs the system.

3.3 Characters of the New Simulation Application Mode

From the above introduction of the simulation application mode in SimGrid, we summarize the following points of the characters of the mode:

1. This mode decreases the requirement for user to implement a simulation application. User does not need have deep comprehension on the simulation function of components, write one word of codes, and originate and manage simulation system. All what user need to do is to design simulation application and describe simulation system in a manner by which SimGrid can understand, and all other works are done by SimGrid. This mode decreases the difficulty of development for user.
2. The construction and execution of the simulation system is perspective to user. User puts forward demands for simulation application, and SimGrid searches the required resources and services (including generation new service codes), arranges the sites for the resources and services, and maintain the resources and services through the process of the execution phase.
3. It promotes simulation reuse in the more range than the before ones. The computing resources belonging to different departments and the simulation models developed by different organizations can be safely and fully shared and reused. It decreases the costs of simulation development.

4 Simulation Grid Run-Time Framework

4.1 Components of SimGrid Run-Time Framework

Simulation Grid run-time framework (**SimGrid Run-Time Framework**) is comprised of **Simulation Description Specification** and **Simulation kernel Services** as shown in fig. 4. A simple introduce to the all components is as the following.

- **Simulation Description Specification**

This specification is the standard rules for the description of function, communication, information format, time management manner, system constrains, and so on. User utilizes this specification as the content requirement and syntax for the description script and **Simulation Kernel Services** uses this specification to understand the intent and demands of the simulation application which user wants. Obviously, **Simulation Description Specification** builds a bridge between user and SimGrid, and it supports decreasing the difficulty of the simulation system development and is the precondition to implement simulation application by SimGrid automatically.

- **Simulation Kernel Services**

This component includes a serial of services for construction, start-up, execution, maintenance, end and deconstruction of the simulation system.

- **Simulation Description Parsing Service.** This service parses the description script which user submitted. This service can be seen as the base service for the following other services, because this service goes through the whole process of the simulation.
- **Simulation Execution Controlling Service.** This service controls the creation, searching resources, set-up, start-up, execution, pause, end, and collapse of the simulation system.
- **Simulation Resources Searching Service.** This service searches simulation model resources required in the description script, and searches computing resources which fit the models to execute on.
- **Simulation Resource Semantic Service.** In order to describe simulation model precisely, semantic information is add into the description of the model. This service help understand the meaning of semantic information.
- **Simulation Resources Combination Service.** This service offers the ability of decomposing the function module of large grain into several modules of small grain.
- **Simulation Model Shell Service.** For every model resource instance which is part of the simulation system, there exits a shell service corresponding to the model instance. This service manages advance of the model, helps the model cooperate with other parts of the simulation system and records the performance of execution of the model.
- **Data Sending/Receiving Service.** All communication and exchange of the information, data and events is achieved by using this service. This service knows the relationship and constrains of publication and subscription among the components of the simulation system, and uses a safe and efficient manner such as GridFTP to transmit data.
- **Simulation Advance Service.** This service manages the request of time advance of simulation models.

- **Monitoring Service.** This service monitors the computing resources for fault-tolerance. The range of monitoring includes exceptions and loads.
- **Simulation Models Scheduling Service.** Scheduling service includes static scheduling and dynamic scheduling. Static scheduling exits in the phase of construction of simulation system, when it assigned computing resources for simulation model instances. Dynamic scheduling exits in the phase of execution of simulation system, when it adjusts the relationship of assignment between computing resources and models instances in order to increase the efficiency of the simulation application.
- **Simulation Execution Fault-Tolerance Service.** When there exits broken in network or faults in computing resources, this service requests to frozen parts of or all simulation execution, and gives supports to create new simulation model instance and assigns new computing resource for the model instance, and then recovers the states of the simulation system to the just before fault occurred, and finally resumes the simulation execution.
- **Simulation Model Migration Service.** In the process of the simulation execution, the dynamic re-assign between simulation model and computing resource is achieved by this service.

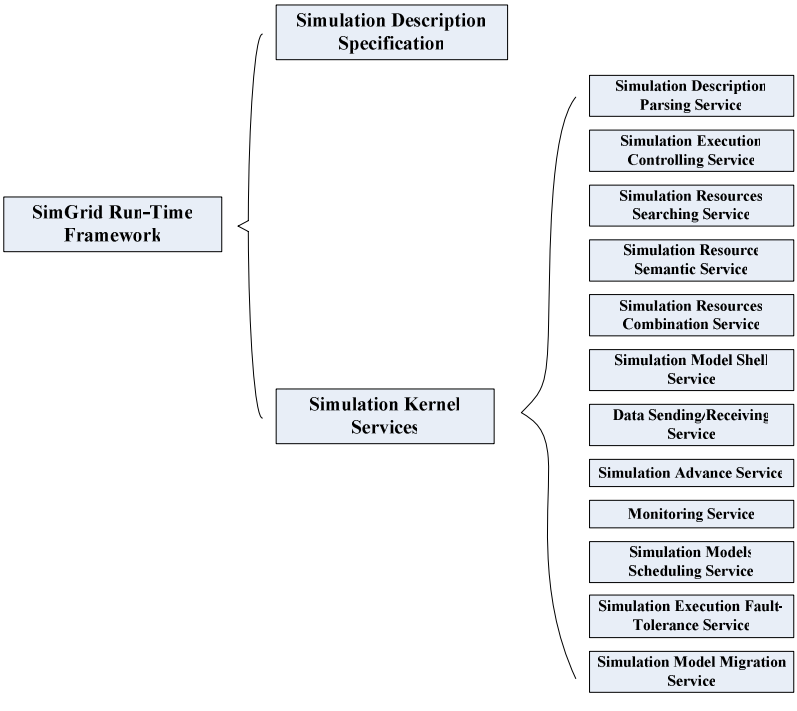


Fig. 4. Components of SimGrid Run-Time Framework

4.2 Relationship among Components of SimGrid Run-Time Framework

The relationship among components of **SimGrid Run-Time Framework** is illustrated as fig. 5.

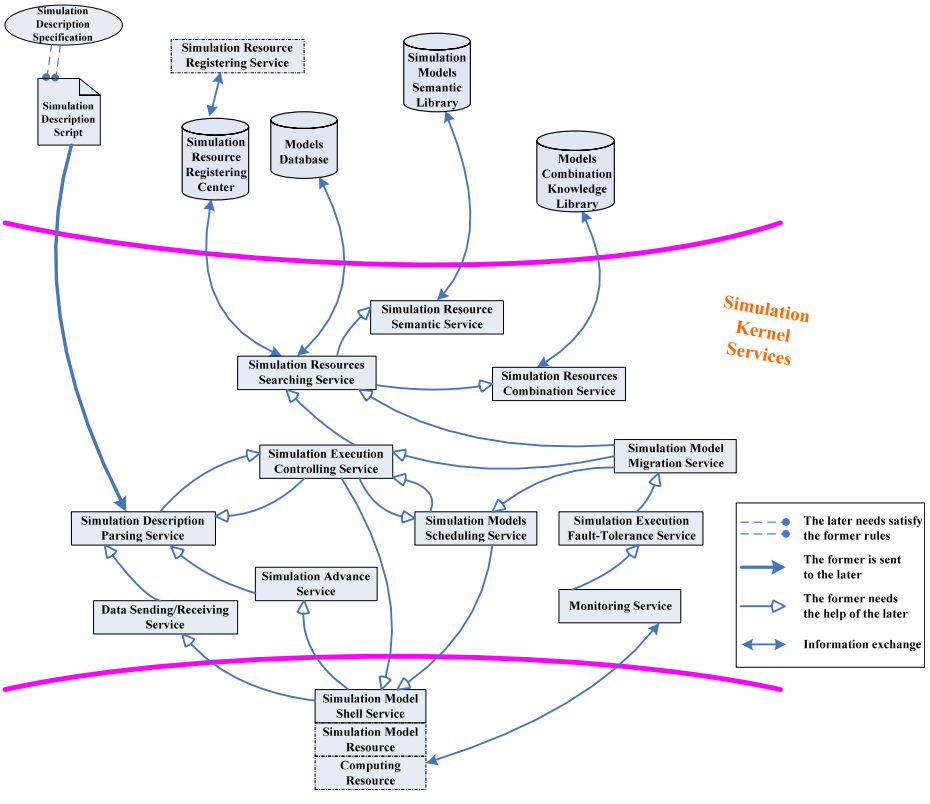


Fig. 5. Relationship among Components of SimGrid Run-Time Framework

Simulation description script is created by user according to **Simulation Description Specification**, and the script is submitted to **Simulation Description Parsing Service**.

Simulation Description Parsing Service parses the script to understand the function, architecture and other constrains of the simulation application user defined and then actives **Simulation Execution Controlling Service** to create simulation system.

Simulation Execution Controlling Service refers to **Simulation Description Parsing Service** to have the function profile and uses **Simulation Resources Searching Service** to find simulation models which satisfied with the function required by the simulation application.

Simulation Resources Searching Service submits the searching requests to the simulation resources registering center and simulation models database, and before

sending requests **Simulation Resources Searching Service** may need the help of **Simulation Resource Semantic Service** to understand the semantic information which supplied by user. Sometime **Searching Service** can not find the model resource for some function, and that may be because of the function is too complicated. So with the help of **Simulation Resources Combination Service**, a complex function can be disassembled several simple and single functions, and **Simulation Resources Searching Service** can search several models corresponding to the small functions to substitute searching one model satisfied with the original request. Function disassembly is with assist of models combination knowledge library.

After acquiring all needed simulation model resources, **Simulation Execution Controlling Service** refers again to **Simulation Resources Searching Service** to search computing resources which is free, and refers to **Simulation Models Scheduling Service** to create instances for simulation models on fitted computing resources. After the end of assignment, **Simulation Execution Controlling Service** creates **Simulation Model Shell Service** for every model instance. Then **Simulation Execution Controlling Service** gets information from **Simulation Description Parsing Service** to initialize every model instance. The information includes time management of model, the events for model to receive and send the format and fidelity of data and so on. An important thing is that for every model instance there is a **Simulation Model Shell Service.**

Every simulation model presents part function of the simulation application, and **Simulation Model Shell Service** represents the logical behavior on simulation model. **Simulation Model Shell Service** builds connection between simulation models and **Simulation Kernel Services**, and builds a bridge between simulation models and other parts of the simulation system. **Simulation Model Shell Service** is under the control of **Simulation Execution Controlling Service** to decide whether to start up, initialize, execute, pause or finish model which corresponding to the **Shell Service**. **Simulation Model Shell Service** refers to **Data Sending/Receiving Service** to exchange the information and events. **Simulation Model Shell Service** refers to **Simulation Advance Service** to achieve time advance. In addition, **Simulation Model Shell Service** is beard recording performance data of simulation models and those data will be sent to **Simulation Models Scheduling Service**.

In the execution phase of simulation system, **Simulation Models Scheduling Service** refers to **Simulation Model Shell Service** to get performance data in the past of all simulation models. With the analysis of the performance data, **Simulation Models Scheduling Service** may re-assign the relationships of models and computing resources to promote increasing on efficiency of simulation system.

The task of **Monitoring Service** is to monitor the computing resources in the simulation system. When fault is detected and validated in computing resource, **Monitoring Service** notifies **Simulation Execution Fault-Tolerance Service**. **Simulation Execution Fault-Tolerance Service** refers to **Simulation Model Migration Service** which then refers to **Simulation Resources Searching Service** to search new computing resource, and then requests **Simulation Execution Controlling Service** to pause simulation execution fully or partly. After migrating the model to the new computing resource, **Simulation Execution Fault-Tolerance Service** refers to

Simulation Execution Controlling Service to recover the states of all models to a synchronized point just before fault occurred and finally to resume simulation execution.

4.3 Features of SimGrid Run-Time Framework

From the analysis of **SimGrid Run-Time Framework**, we draw conclusion that there are the following features of the framework:

- It provides a standard manner to describe simulation application. The framework can understand the description. The understanding to the description runs across the whole lifetime of simulation application.
- It supports to search simulation resources for the needs of simulation application in SimGrid. The searching includes two aspects of which one aspect is focused on the syntax description such as interface description and another aspect puts emphasis on the semantic description such as function description and field description.
- It delegate user to construct, execute and maintain simulation system. All these works are perspective to user, namely user does not know what and where services are offered.
- It can detect disconnection on network and fault in computing resources, and carry fault-tolerance and migration without re-start simulation system. This guarantees the robustness of the simulation.
- It monitors the performance of the simulation system, and supplies dynamically scheduling algorithm according to the performance data. This increases the efficiency of the simulation.
- It provides security for the services, data and communication.

5 Conclusion

Under the mode of simulation application introduced in this paper, user can put his or her energy on the design of the simulation application and on description of the simulation system. After the description as the form of script was submitted to the SimGrid, all works is done for user and the remainder is to wait simulation results from SimGrid returning. So this simulation application mode is easier for user than the past modes in simulation application. In the whole process, user does not need to write one word of code, which lows the threshold of requirement for user. According to the description of the simulation system, SimGrid can search simulation resources required for the simulation system, organize these resources to construct simulation system, then execute the simulation system, return results after the execution is over, and finally deconstruct simulation system and release simulation resources.

In this paper we propose a framework, namely **SimGrid Run-Time Framework**, to the new mode of simulation application in SimGrid. **Simulation Description Specification** in the framework provides the base for SimGrid to understand the design intent of user, and it builds a bridge between user and SimGrid. **Simulation Kernel**

Services in the framework gathers a serial of services to support building a simulation application automatically in SimGrid. In addition, **Simulation Kernel Services** maintains the simulation system aiming at promoting the stability and robustness, decreasing time of simulation execution and waste usage of the computing resources, and adding security in simulation execution.

References

1. Li, B.H., Chai, X., et al.: The Development of Modern Modeling & Simulation. In: Front of China Science (Chinese Academy of Engineering), vol. 6, Higher Education Press, Beijing (2003)
2. Li, B.H., Chai, X., et al.: Some Focusing Points in Development of Modern Modeling and Simulation Technology. In: Asia Simulation Conference 2004, Korea (2004)
3. Hofer, R.C, Loper, M.L.: DIS today [distributed interactive simulation]. Proceedings of the IEEE 83(8), 1124–1137 (1995)
4. Weatherly, R.M., Wilson, A.L., Griffin, S.P.: Alsp – Theory, Experience, and Future Directions. In: 1993 Winter Simulation Conference Proceedings, pp. 1068–1072 (December 1993)
5. Kuhl, F., Dahmann, J., Weatherly, R.: Creating Computer Simulation Systems: An Introduction to the High Level Architecture, PTR (September 1999)
6. Foster, I.: The Grid: A New Infrastructure for 21st Century Science. Physics Today, 42–47 (2002)
7. CrossGrid – Development of Grid Environment for Interactive Application, EU Project, IST-2001-32243, http://www.eu-crossgrid.org
8. Alice Grid System. http://www.comp.nus.edu.sg/ teoym/alice.htm
9. Li, B.H., Chai, X., et al.: Research and Application on CoSim(Collaborative Simulation) Grid. In: 2006 Summer Simulation Multiconference. M&S-MTSA, Calgary, Alberta, Canada (2006)
10. HLA-RTI, Defense Modeling and Simulation Office, http://www.dmso.mil/public/transition/hla/
11. Foster, I., Kesselman, C., Nick, J., et al.: Grid Services for Distributed System Integration. Computer (2002)

Web-HLA and Service-Enabled RTI in the Simulation Grid

Jijie Huang[1], Bo Hu Li[1], Xudong Chai[2], and Lin Zhang[1]

[1] Beijing University of Aeronautics and Astronautics, Beijing 100083, China
chxhjjco-p@163.com
[2] Beijing Simulation Center, Beijing 100854, China
xdchai@263.net

Abstract. HLA-based simulations in a grid environment have now become a main research hotspot in the M&S community, but there are many shortcomings of the current HLA running in a grid environment. This paper analyzes the analogies between HLA and OGSA from the software architecture point of view, and points out the service-oriented method should be introduced into the three components of HLA to overcome its shortcomings. This paper proposes an expanded running architecture that can integrate the HLA with OGSA and realizes a service-enabled RTI (SE-RTI). In addition, in order to handle the bottleneck problem that is how to efficiently realize the HLA time management mechanism, this paper proposes a centralized way by which the CRC of the SE-RTI takes charge of the time management and the dispatching of TSO events of each federate. Benchmark experiments indicate that the running velocity of simulations in Internet or WAN is properly improved.

Keywords: Grid, HLA/RTI, OMT, local RTI component (LRC), core RTI component (CRC).

1 Introduction

The reusability and interoperability of simulation resources are the goals of the High Level Architecture (HLA). Currently, resources in simulation grid are encapsulated as grid services using Service-Oriented (SO) methods, but HLA is designed by adopting the object-oriented (OO) method. So, HLA-based simulations accessing grid services have become a main studying hotspot in current M&S community.

When a HLA-based simulation runs in grid environment, HLA can neither support it to share the simulation resource service, especially to share the model service, nor provide a fault-tolerance mechanism. In order to overcome these shortcomings, many organizations introduce web and grid technology into distributed simulation, such as the XMSF [1][2], the CrossGrid [3] and the DS-Grid [4][5] etc. The Defense Information Systems Agency (DISA), which is the organization responsible for building the Global Information Grid (GIG), has applied for founding a distributed simulation architecture based on the HLA [6]. It plans

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 195–204, 2006.

to modify the HLA from three aspects: Service-Oriented Architecture (SOA), joint command and control and information assurance.

This paper will propose a service-oriented software architecture for HLA-based simulations that expands the three components of HLA and rebuilds the Run-Time Infrastructure (RTI) as a service-enabled RTI using the SO method.

2 Web-Enabled HLA Using the SO Method

HLA is a layered software architecture that is divided into the RTI layer and the federate layer so that federates can not be affected when the RTI is modified to adapt to a new net environment. HLA is also data-abstract software architecture that encapsulates data and functions of object instances in a federate using the OO method. Grid adopts the Open Grid Services Architecture (OGSA) that is divided into four layers: fabric layer, Grid service layer, application toolkit layer and grid application layer. A service in OGSA is the abstract of a sharing resource. It provides some interfaces to allow clients to access it.

The analogies between HLA and the Grid make it possible to conduct a HLA-based simulation in a grid environment. One common way is that the federate layer corresponds to the grid application layer and the RTI of HLA corresponds to the other three layers of Grid [7][8][9]. Thereby, a federate of this simulation can not only change data with other federates by using the RTI layer, but also access a grid resource by using the grid infrastructure. This conducting mechanism can enable a federation to use the gird resources, but it can not integrate the HLA with the Grid from the software architecture point of view. In order to found an expanded architecture to support the development and running of HLA-based simulations in a grid environment, the HLA must introduce grid core technologies into its components, namely the SO technologies.

2.1 Introducing the SO Method into Rules of HLA

The rules of HLA are the foundation of the High Level Architecture, they describe the responsibility of simulations and federates to ensure the correct interaction in a federation execution. One of the rules limits a federate to interact with other federates only by RTI, but does not specify how to realize the RTI. So a federate can run in grid environment by a RTI that has realized all the server interfaces of HLA using Web technologies.

A RTI consists of the local RTI component (LRC) and the central RTI component (CRC). The LRC is embedded into every federate, it must be realized using Web technologies. However, the CRC can be handled in two ways:

1. It remains running in a local net. An agent, which is realized using Web technologies, is added to it and takes the responsibility for communication with the LRC and the CRC. The Web-Enabled RTI of XMSF and the Grid-Enabled RTI of the DS-Grid are realized in this way.

2. It is rebuilt as a grid service and is shared by any grid user. It communicates directly with the LRC using Web technologies. This Service-Enabled RTI can be integrated into grid environment better than the above two kinds of RTI.

The last rule demands every federate to manage its local time in order to cooperate the data interaction among federates. HLA provides a perfect time management mechanism which demands the LRC and the CRC to communicate frequently. Because the transmission latency varies from several hundred milliseconds to several seconds and the multicast technology is not mature in gridhow to efficiently realize the HLA time management mechanism is a bottleneck problem when a HLA-based simulation is planted from a local net to the grid environment. To solve this problem, the number of communication times between the LRC and the CRC must be reduced by all possible means.

2.2 Introducing the SO Method into the OMT

HLA defines the rules and specifications to support reusability and interoperability amongst the simulation federates. The Object Model Template (OMT) is an important part of HLA specifications. It defines the way in which federations and federates have to be documented. The HLA object models are the formal definition of the data that is transferred between federates and thus are one of the main vehicles for interoperability in HLA. But the OMT does not specify how to create an object model and what contents should be recoded. Thus, some new contents can be added to OMT according the demand of simulation activities in grid environment.

The simulation in grid environment is an activity based on model services. Registration, searching, matching and assembling of model services should be added to the Federation Development and Execute Process Model (FEDEP). Object instances in a federation may update their attributes by accessing model services and be permanently attached with them during the federation execution. If a model service fails to provide a object instances with service for the sake of the net or the grid server on which the service is deployed, the simulation will terminate because the middle stat information of the object instance, which is needed by the object instance to calculate its attributes in next time, remain in the model service. So HLA should provide several model service URLs for an object class and a dynamic access mechanism to let an object instances select one from these URLs to access the corresponding service. In addition, the model service should have the ability to migrate the middle stat information of the object instances to other model service.

From the view of object instances in a simulation running in grid environment, the URL of a model service should be an attribute of its object class. An object class demands simulation developers to find several model services for it and adds their URLs to OMT. HLA can demand its RTI layer to parse these URLs for an object class according the fed file and inform each of these peer model services with the others URLs. So, when one of these model services is found by a grid user, it can inform the others URLs to the user.

According to above analysis, the model service URL is an important attribute of an object class and it should be added to the OMT. The Quality of Service (QoS) of a used model service should be evaluated when simulation developers begin to analyze data and evaluate the results. If developers do not satisfy with the results, they will repeat the simulation. At the same time, the RTI recodes the QoS value of used model services and informs their peer model services so that a model service can arrange its peer model services in order of QoS values.

Apparently, some interfaces should be added to the services of HLA in order to fulfill above purpose.

3 The SO Architecture of HLA-Based Simulations

According to the definition given by Shaw and Garlan[10], a software architecture comprises of component, connector and Task-flow. The component of HLA is the elements defined in the rules and interface specifications, namely federates, RTI and OMT. The connector of HLA is the data interaction between a federate and RTI, which is defined in the rules and interface specifications. This kind of interaction is based on a FOM or SOM file, which is produced according to the OMT. The Task-flow of HLA is the federation development process that is constrained by the rules of HLA.

When a HLA-based simulation in the grid environment runs in a service-oriented way, its federates access the Service-Enabled RTI to exchange data and its object instances access model services to calculate their attributes. Model services are the foundation of simulation activities in grid environment and should be added to the component of HLA. The Service-Enabled RTI and model services need to exchange information by accessing interfaces with each other, so these interfaces should be added to the connector of HLA too. Because the Grid is a more complicated environment than the local net, a fault-tolerance mechanism should be introduced into the Task-flow of HLA [11].

In addition, federates of this federation have become dispatchers of grid services and the main calculating work are done by grid services. So the possibility of occurring an error in a federate decreases, but the possibility of failing to access a grid service increases. Services deployed in the third generation grid are demanded to be of the adaptive ability [12]. They can apperceive the environment changes and adopt corresponding tactics to ensure it can still provide service correctly. In this paper, the self-adaptive model service obtains the apperception ability by communicating with a supervisor agent, which directly accesses the hardware resources. The self-adaptive ability of model services can be changed into the fault tolerant ability of the simulation, so it will become the fault-tolerance foundation of HLA-based simulations in the grid environment.

According to above analysis, the service-oriented architecture of HLA-based simulations is shown in Figure 1:

The next part of this paper is focused on the discussion of the SE-RTI. The fault-tolerance of HLA-based simulation in grid environment and the expanded FOM will be discussed in another paper.

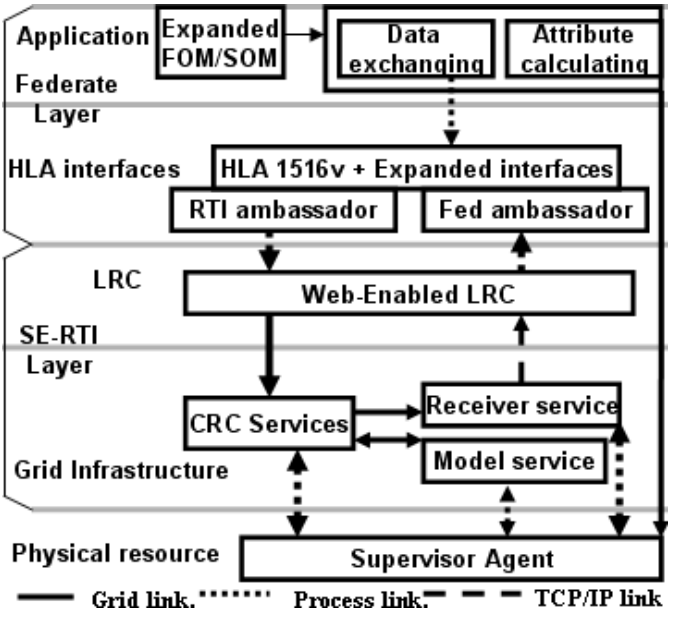


Fig. 1. The SO architecture of HLA-based simulations

4 The Realization of the SE-RTI

The realization of the SE-RTI can be based on the local RTI server that receives net messages from federates and selects corresponding entry functions with respect to the contents of the messages. As a result, the CRC only needs to encapsulate entry functions to grid service interfaces and package the implementation codes of HLA services to a C dynamic library. The CRC calls the C dynamic library by JNI to respond LRC's requests, which is illustrated in Figure 2:

During the process of the realization, the SE-RTI must solve two key problems: data dispatching and time advancing mechanism.

4.1 Data Dispatching of the SE-RTI

The HLA-based simulation in the grid environment requires peer to peer communications between the LRC and the CRC, but http is uni-directional transmission protocol that makes it unsuitable for supporting simulation communication patterns. XMSF realized the federate callback communications between the LRC and the CRC over the Web by using the SOAP and the BEEP protocols, and DS-Grid adopted GT3.0 notification technologies [10]. Since they do not use the public socket ports such as 80 or 8080, their federate callback communications can not pass through firewalls over the web.

In order to pass through firewalls over the web, we use the SOAP and the Http protocols to realize the federate callback communications. As a result, a data

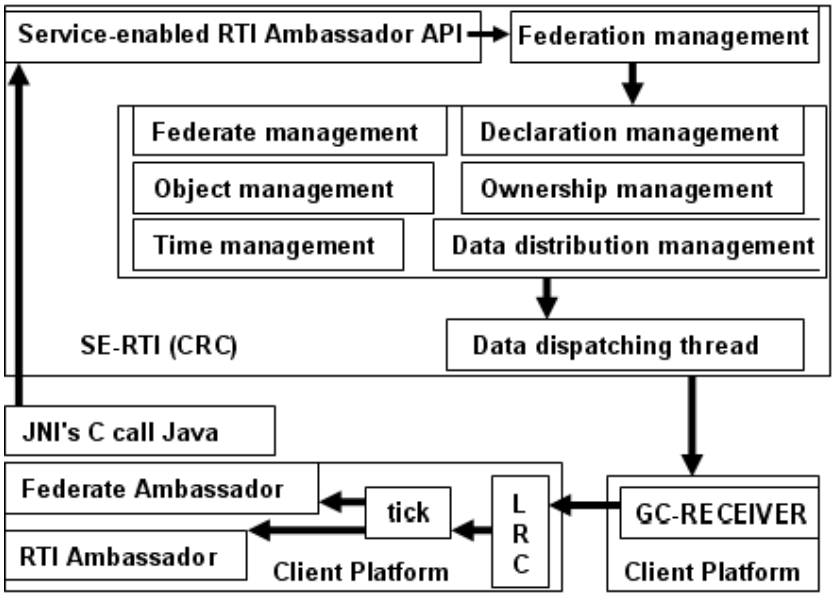


Fig. 2. The SE-RTI based on two http links; 1: RTI ambassador link; 2: Fed ambassador link

dispatching thread is added to the CRC and a receiver grid service is deployed to the LRC local net. The two added components can communicate by using the Http protocol according to the following way:

Firstly, the CRC main thread executes the RTI ambassador request and sends generated federate callback information to a shared list. Secondly, the data dispatching thread circularly reads and analyses the federate callback information from the shared list, and sends it to the receiver-grid service of which the URL is given by the callback information. Lastly, the receiver grid service gets the federate callback information from its interface and sends it to the LRC using TCP/IP protocols.

The federate callback information is generated by the CRC main thread and is asynchronously sent to the LRC by the CRC data dispatching thread. The transmitting order of the federate callback information is guaranteed by both the FIFO shared list and the SE-RTI time management.

4.2 Time Management of the SE-RTI

The SE-RTI software uses the LBTS algorithm to control the advancement of federates and ensure the correct consequence of received messages. In the local network, the LBTS algorithm is realized in the libRTI library which is linked to each federate. The libRTI exchanges a large amount of time-updating messages with a RTI server to compute the LBTS of the federation; meanwhile it fetches

the federate callback message from a FIFO list and a TSO list according to the LBTS value.

As the multicast technology in the grid environment hasn't matured, the communication between the LRC and the CRC can only be realized with the Http protocol, which can not meet the needs that support a large amount of time-updating message exchanges. Therefore, the SE-RTI must use a centralized method to manage time-updating messages, which moves the LBTS calculation and the TSO list to the CRC component.

The labels of 1,2,3 in Figure 3 indicate the process that the CRC handles Receive Order (RO) events coming from a federate. Firstly, the CRC receives and handles request messages from the LRC. Secondly, the CRC sends the generated federate callback message to the shared FIFO list. Lastly, the CRC dispatching thread constantly reads a callback message from the shared list and sends it to the LRC. The tick function in the main thread of a federate fetches callback messages and calls the corresponding handling function until a granting time advancement message is received.

If the CRC receives Timestamp Order (TSO) events coming from a federate, it sends the generated federate callback TSO message to TSO lists of federates which need this TSO message and informs the LRC to end this access.

The labels of 4,5,6 in Figure 3 indicate the process that the CRC handles the RTI ambassador time-advance-requesting interface. When the CRC receives the message sent by this interface from the federate 1, it firstly calculates the federation LBTS that includes some federates (labeled from F1 to Fn) and updates the corresponding time data of federate F1. Then the CRC scans each

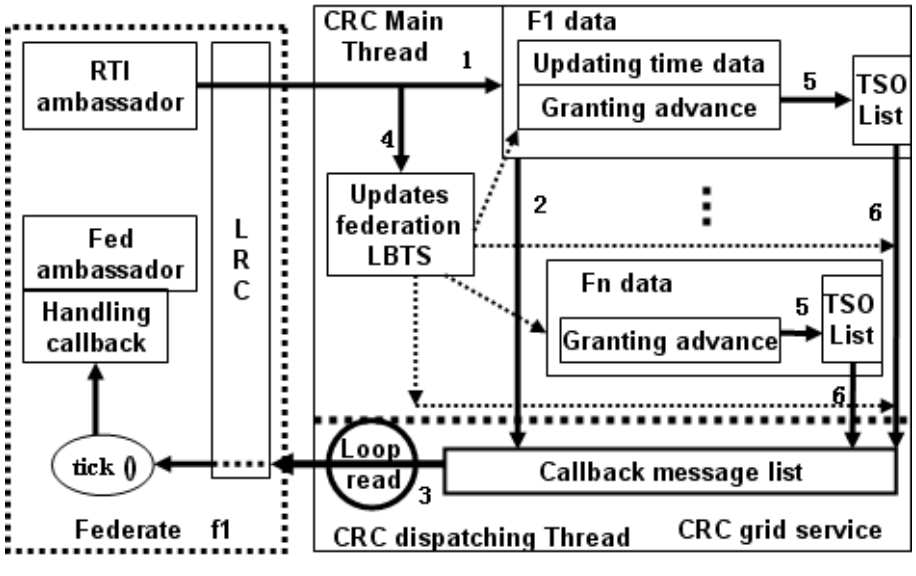


Fig. 3. Handling RO and TSO messages in SE-RTI

federate to check whether it satisfies the advancement according to the federate time-managing type. For a regulation federate, a Time Advance Granted (TAG) message is immediately generated and sent to the TSO list by the CRC, but for a constraint federate, the CRC must compare its local time and the LBTS value to decide if it can be satisfied or not. At last, the CRC checks the TSO list of these federates and only sends callback messages whose number of times is less than the LBTS value.

As the CRC uses thecentralized and active way to handle the federate callback message, the number of communications between the LRC and the CRC will be reduced, and the CPU load will be increased at the same time. With development of computers and the complexity of the grid environment, the CRC should take over more tasks from the LRC.

5 The Performance Analysis of the SE-RTI

Experiments have been carried out to test the performance of the SE-RTI in Internet environment. The CRC service is deployed in a cluster computer which is located in Beijing University of Aeronautics and Astronautics (BUAA). Federates in experiments have run in two computers, one is located in BUAA, and the other is located in Tsinghua University (TsingHua).The federation in these experiments is derived from a DMSO example, namely the foodfight federation. Each federate in this federation runs in both constrained and regulating time management way.

Experiments have been carried out to compare the performance in several kinds of net environment in terms of the speed of the simulation time advancement. Table 1 presents the experimental results reported as the TAG rate.

The TAG rate of the federation running in BUAA is faster than the rate of the TsingHua's federation, because transmission latency between a BUAA federate and the CRC grid server is 0.1626 second, but the latency between a TsingHua federate and the CRC is 0.2677 second. Apparently, the transmission speed in Grid is more slower than the speed in local net, and the transmission latency

Table 1. The SE-RTI performance comparison

number of federate pairs	*Federation Time advance Granted velocity (TAG/sec)*		
	SE-RTI in LAN	*SE-RTI in Internet Cluster grid server (at BUAA)*	
	SMP grid server	BUAA accesses SE-RTI	TsingHua accesses SE-RTI
2	2.17	1.721	0.665
3	1.47	1.636	0.521
4	1.06	1.504	0.381
5	0.83	1.264	0.259
6	0.81	1.203	0.216
7	0.568	1.002	0.183
8	0.49	0.881	0.169

in Grid increases remarkably while the transmission distance is increasing. This results in that how to efficiently realize the HLA time management mechanism is the bottleneck problem when HLA-based simulations run in grid environment. This also results in that the WE-RTI and Grid-RTI could not give experiment data when federates is more then two.

The rate decreases rapidly when the number of federate pairs is increased, because object instances and interaction instances in this federation both publish and subscribe their attributes or parameters. In addition, when the federate number n equals 8, the federation spends 418.3 seconds to finish the front 40 times advancement. After this, all object instances are destroyed and the communication times between the LRC and the CRC decrease, so the federation just spends 173.0 seconds to finish the remainder 60 times advancement. Apparently, decreasing the number of data exchange times in one simulation loop can remarkably increase the TAG rate of a HLA-based simulation in grid environment.

By using the central time management way, the SE-RTI has decreased data exchange times in one simulation loop. For example, if the footfight federation has n federates, it can decrease (n^2 - n) times at best. So, the running velocity of simulations in Internet or WAN is properly improved. The SE-RTI can support eight federates to run in Internet. Thereby, it can efficiently realize the HLA time management mechanism comparing with above two kinds of RTI. But the SE-RTI can still improve its performance from the following aspects:

1. When the CRC service selects callback messages from a TSO list of a federate according to the LBTS value, it should packet these messages into a buffer and send this buffer to the federate as a package.
2. The CRC should create a dispatching thread for every federate which accesses it, so that each dispatching thread takes charge of sending callback message to a corresponding federate.

6 Conclusions

By introducing the service-oriented idea into the HLA, the OMT needs to add the model service URL to object class attributes and the CRC of RTI needs to be realized as a grid service that can communicate with a model service by accessing its interfaces. In this way, the sharing mechanism of HLA can be expanded to support the sharing of the grid resources, and a fault-tolerance mechanism based on model services can be added to HLA-based simulations in grid environment.

The Service-enabled RTI should adopt a centralized management method in CRC to reduce the number of communication times among federates in a federation so that the federation can efficiently run in an Internet environment whose transmission latency varies from several hundred milliseconds to several seconds. While the service-enabled RTI should use the parallel arithmetic to dispatch information and packet several callback messages of the same federate into a package.

References

1. Morse, K.L., Brunton, R., Drake, D.: Developing and Integrating Web-Enabled RTIs - an XMSF Update[C]. In: Proceedings of the 2003 Fall Simulation Interoperability Workshop, Orlando, FL (September 2003)
2. Morse, K., Drake, D., Brunton, R.: Web Enabling HLA Compliant Simulations to support Network Centric Applications[R]. In: Proceedings of The 2004 Symposium on Command and Control Research and Technology, San Diego, CA (2004)
3. Zajac, K., Tirado-Ramos, A., Zhao, Z., et al.: Grid Service for HLA-based Distributed Simulation Frameworks[A]. In: Proceedings of First European Across Grids Conference[C], Santiago de Compostela, Spain, pp. 147–154 (2003)
4. Xie, Y., Teo, Y.M., Cai, W.T., et al.: A Distributed Simulation Backbone for Executing HLA-based Simulation Over the Internet[C]. Workshop on Grid Computing & Applications. In: International Conference on Scientific and Engineering Computation(IC-SEC 2004) (June 2004)
5. Xie, Y., Teo, Y.M., Cai, W., Turner, S.J.: Service. Service Provisioning for HLA-based Distributed Simulation on the Grid[C]. Provisioning for HLA-based Distributed Simulation on the Grid. In: Proceedings of the 19th ACM/IEEE/SCS Workshop on Principles of Advanced and Distributed Simulation, pp. 282–291. IEEE Computer Society Press, Monterey, USA (2005)
6. In, J.-U., Avery, P., Cavanaugh, R., et al.: Policy based scheduling for simple quality of service in grid computing[C]. In: Proceedings of the 18th International Parallel and Distributed Processing Symposium, Santa Fe, USA, pp. 315–324. IEEE Computer Society Press, Los Alamitos (2004)
7. Rycerz, K., Bubak, M., Malawski, M., Sloot, P.: A Framework for HLABased Interactive Simulations on the Grid. Simulation 81(1), 67–76 (2005)
8. Pullen, J.M., Brunton, R., Brutzman, D., Drake, D., Hieb, M., Morse, K.L., Tolk, A.: Using Web Services to Integrate Heterogeneous Simulations in a Grid Environment[R]. In: Proceedings of the International Conference on Computational Science 2004, Krakow, Poland (to appear)
9. Chen, D., Turner, S.J., Cai, W.: A Framework for Robust HLA-Based Distributed Simulation. In: 20th ACM/IEEE/SCS Workshop on Principles of Advanced and Distributed Simulation (PADS 2006) (May 2006)
10. Shaw, M., Softwar, G.D.: Architecture: Perspectives on an emerging discipline[M], pp. 1–168. Prentice Hall, Englewood Cliffs (1996)
11. Möller, B., Löfstrand, B., Karlsson, M.: Developing Fault Tolerant Federations Using HLA Evolved[C]. In: Proceeding of 2005 Spring Simulation Interoperability Workshop, paper no. 05S-SIW-048, San Diego, California, USA (April 2005)
12. Pratap, P., Ekanadham, K., Jann, J.: Autonomic Computing and Grid[M]. Grid Computing. Wiley, Chichester (2003)

Encryption in TECB Mode: Modeling, Simulation and Synthesis

M.B.I. Reaz[1], M.I. Ibrahimy[1], F. Mohd-Yasin[2], C.S. Wei[2], and M. Kamada[3]

[1] Department of Electrical and Computer Engineering, International Islamic University Malaysia, 53100 Kuala Lumpur, Malaysia
[2] Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia
[3] Department of Computer and Information Sciences, Ibaraki University, Hitachi, Ibaraki 316-8511, Japan

Abstract. The growth of the Internet as a vehicle for secure communication has resulted in Data Encryption Standard (DES) no longer capable of providing high-level security for data protection. Triple Data Encryption Standard (3DES) is a symmetric block cipher with 192 bits key proposed to further enhance DES. Many applications crave for the speed of a hardware encryption implementation while trying to preserve the flexibility and low cost of a software implementation. This project used single core module to implement encryption in Triple DES Electronic Code Book (TECB) mode, which was modeled using hardware description language VHDL. The architecture was mapped in Altera EPF10K100EFC484-1 and EP20K200EFC672-1X for performance investigations and resulted in achieving encryption rate of 102.56 Mbps, area utilization of 2111 logic cells (25%) and a higher maximum operating frequency of 78.59 MHz by implementing on the larger FPGA device EP20K200EFC672-1X. It also suggested that 3DES hardware was 2.4 times faster than its software counterpart.

1 Introduction

In the wake of advancement in computer technology and increasingly volatile information flow, we are faced with challenges of safeguarding information that is not meant for public knowledge [1]. It is common to see all sorts of electronic inventions such as the cellular phone, various devices in the military system and smart cards today [2].

The growth of the Internet has contributed to the increase in the amount of data transferred daily across regions. These data transmissions may contain funds amounting to millions of dollars or government records. However, these applications require high data security. The ease in obtaining and duplicating these data through resourceful parties [e.g. hackers] has resulted in a decline in confidence amongst Internet users towards online transaction. As such, it is essential to ensure the privacy and authenticity of these data. One of the existing methods that can be used to guard the security and authenticity of data through the Internet is through cryptography.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 205–215, 2007.

Data Encryption Standard, DES has been the world wide standard for more than 20 years [3]. DES is used in IPSec protocols, secure socket layer (SSL) protocol and ATM cell encryption. During those years, bundles of software and hardware had been developed to implement this algorithm. However due to the need of higher security, 3DES had been chosen based on its close relationship to DES [4]. Triple DES is an improved version of DES and provides better security compared to DES. This is due to its longer key length and more rounds of DES encryptions. DES only has an effective key length of 56 bits, which is insufficient to resist any brute force attack today [3]. Research has shown that a key-breaking machine that costs less than $1 million can find a key in an average of 3.5 hours and the cost is estimated to drop by a factor of 5 every 10 years [3]. Even though 3DES is three times slower than DES, if used properly, it can be as strong as the 2304-bit public key algorithm because it has longer key length. With an increase in its security standards and compatibility to the DES software and hardware, 3DES is clearly a better choice compared to other algorithms such as RSA and ECC [5].

Due to its symmetric nature, 3DES is a better choice in encrypting bulk data and is therefore less expensive [1, 6]. 3DES uses only 128 or 196 bits symmetric keys and has simpler algorithm. It is less complicated, less computationally intensive and does not introduce much overhead. Thus, it requires relatively inexpensive hardware [1, 3]. 3DES is faster than RSA. Due to its much longer key length, RSA causes high-level resource utilization and is not suitable to be used in mobile or wireless devices as these devices have underpowered processors [6].

In comparison to AES, 3DES is faster [7]. The limitation of AES exists because the cipher and its inverse use different codes and/or tables. As such, it does not have bidirectional architecture for encryption and decryption as that of the 3DES. The inverse cipher can only partially reuse the circuitry that implements the cipher, resulting in a larger hardware presumably.

Hardware realization of cryptography has the advantages of being more secured and faster in speed [8]. It gives a higher performance as desired [9]. Even though software implementation of cryptography uses general-purpose processors that offer enough power to satisfy the needs of individuals, hardware realization is the only way to achieve speeds that is more significant than the general-purpose microprocessor [10]. This feature is important for commercial and communication purposes as this is shown in [9] that security related processing can consume up to 95% of a server's processing capacity. By using a dedicated hardware to run encryption application, more computing can be done within a stipulated period due to parallel processing. Field Programmable Gate Array (FPGA) offers a potential alternative to speed up the hardware realization. From the perspective of computer-aided design, FPGA comes with the merits of lower cost, higher density, and shorter design cycle. The programmability and simplicity of FPGA made it favorable for prototyping digital system.

In this paper, the framework of FPGA-based hardware realization of cryptography using 3DES is proposed. With this approach, both the speed and performance are preserved without the need to trade-off between these two important criteria in encryption and decryption. In this method, VHDL (Very High Speed Integrated Circuit Hardware Description Language) is selected as the hardware description language to realize the system.

2 Triple DES Algorithm

Triple DES encrypts a block of 64-bit data using two or three unrelated 64 bits keys [5]. The internal operation done on these data is similar to that of DES where the only difference is that DES consists of 16 iterations whereas 3DES consists of 48 iterations. In other word, 3DES contains three successive DES operations. Out of the 64-bit key used in DES, the effective key size is only 56 bits. The eighth bit in each byte is used for odd parity checking and is thus ignored. As such the total effective key size for 3DES is 168 bits.

A DES encryption operation is divided into two stages involving the key and the data. In the first stage, 16 subkeys are created from the key whereas the encryption of data message is occurred in the second stage.

During the first stage, permutation is initially done on the 64-bit key and resulting in a 56-bit permutated key. After key permutation, this 56-bit permutated key is divided into left and right halves C_0 and D_0, where each half has 28 bits. With C_0 and D_0 defined, sixteen blocks C_n and D_n, where $n = 1, 2, 3,\ldots, 16$ are formed by left shifting C_{n-1} and D_{n-1} (once or twice). C_n and D_n are then concatenated to form a 56-bit data, C_nD_n. This 56-bit data is then permutated and resulted in 48-bit subkeys formed. After 16 iterations, 16 sets of subkeys are created. These subkeys are used for data encryption during the second stage.

During the second stage, permutation is done on the 64-bit message data block, M. As this is the first permutation process being done on the data, it is called Initial Permutation. The permutated data are then divided into left half L_0 and right half R_0, each having 32 bits. It is followed by 16 iterations of operations, using function f, which operates on two blocks: data block of 32 bits and subkey block of 48 bits to produce an output block of 32 bits.

$$L_n = R_{n-1} \tag{1}$$

$$R_n = L_{n-1} + f(R_{n-1}, K_n) \tag{2}$$

where $n = 1, 2, 3,\ldots, 16$

As shown in (1) and (2), during each of the 16 iterations, the right 32 bits of the previous iteration, R_{n-1} is used as the left 32 bits of the current iteration, L_n. The right 32 bits in the current iteration, R_n is obtained by implementing XOR to the left 32 bits of the previous step with f function.

To calculate f function, each block R_{n-1} is expanded from 32 bits to 48 bits and the expanded R_{n-1}, $E(R_{n-1})$ is then XORed with the block of subkey K_n, i.e.,

$$K_n + E\,(R_{n-1}) = B_1B_2B_3B_4B_5B_6B_7B_8 \tag{3}$$

where $n = 1, 2, 3,\ldots, 16$ and B_i is a group of 6 bits. This results in a 48-bit block, which is then divided into $B_1B_2B_3B_4B_5B_6B_7B_8$. Each B_i gives an address in a different S box, S_i. The 4-bit blocks for the entire eight S boxes are combined to form a 32 -bit block.

Function f is obtained by implementing permutation on the group output such as,

$$f = P(S_1(B_1)S_2(B_2)\ldots S_8(B_8)) \tag{4}$$

At the end of the sixteenth iteration, the order of the two blocks $L_{16}R_{16}$ is reversed to $R_{16}L_{16}$ before applying the permutation on the reversed block. This is the last permutation to be done on the data, thus being called the Final Permutation.

Decryption in DES uses the same process as the encryption operation. The only difference lies in the order in which the subkeys are used. In the decryption process, the subkeys are used in reverse order, meaning that K_{16} is applied first with K_1 being applied last.

Triple DES shows a high level of similarity in operation to that of DES. Encryption and decryption in 3DES are done by compounding the operation of DES encryption $E_k(I)$ and decryption $D_k(I)$ operations. Encryption operation in 3DES is defined by,

$$Encryption = E_{K3}(D_{K2}(E_{K1}(I))) \tag{5}$$

where as the decryption operation is defined by,

$$Decryption = D_{K1}(E_{K2}(D_{K3}(I))) \tag{6}$$

From equation (5), it shows that the plaintext is first encrypted by $K1$ using DES. The encrypted data is then decrypted by $K2$ before being encrypted by $K3$. In contrast to that, equation (6) indicates that the 3DES cipher text is initially decrypted by $K3$ using DES, whereby the result is then being encrypted by $K2$. The plaintext is recovered by decrypting the output from second DES operation by $K1$.

Final permutation is actually the inverse operation of initial permutation. As such in a 3DES operation, the initial permutation of the second DES round cancels the final permutation of the first DES round. This is the same in the third DES round where its initial permutation cancels off the final permutation of the second DES round, leaving only an initial permutation and a final permutation during the whole 3DES operation.

3 Design Flow of 3DES Single Core Module

The specification of the 3DES core is set prior to the start of the design process. Different 3DES operation mode could result in different design complexity and different level of security. As such, a trade off between these two conditions must be taken into consideration during the design stage. As to avoid complicated design, 3DES Electronic Code Book (TECB) had been chosen as the mode of operation in this project. This resulted in reduced area utilization and compromised security level in the core design.

Due to varying number of bits being shifted during the different iteration rounds, normal shift register could not be used. A counter had to be added in the design so as to determine the current iteration round. The input signals to the shifting module were shifted appropriately depending on the output of the counter. The output of counter must be passed correctly to the shifting module. Error in connections such as MSB of the counter output being connected to the LSB of the shifting module input could result in error in bit shifting.

During DES encryption operation, the subkeys were transmitted in the sequence of 1 to 16 whereas during DES decryption operation, the subkeys were transmitted in the sequence of 16 to 1. The initial design in mind was to have a multiplexer and a

demultiplexer. The 16 subkeys were multiplexed. This was then followed by demultiplexing these multiplexed subkeys either in the sequence of 1 to 16 or from 16 to 1. The sequence in which the subkeys were sent out was determined by the select signal of the demultiplexer. However, this design was difficult to implement as complicated control signals were needed to obtain the 16 subkeys in the correct sequence. These subkeys had to be stored in registers before being demultiplexed.

Based on this, the design of the full implementation of 3DES encryption engine was produced as shown in Figure 1. The multiplexers and demultiplexers in Figure 1 played the role of realising the 16 iterations in a DES operation and the 48 iterations in the 3DES operation.

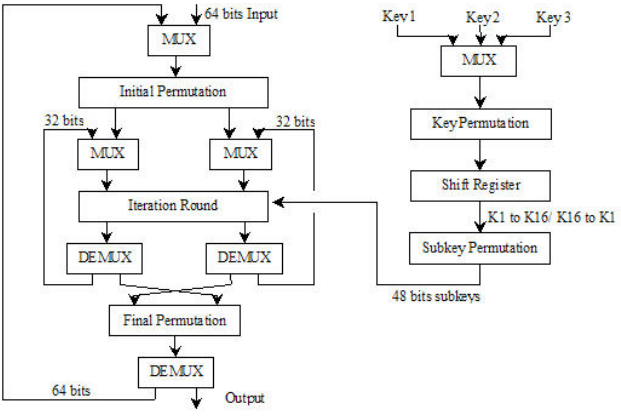


Fig. 1. Full implementation of 3DES operation

4 Simulations and Synthesis

4.1 Functional Simulation

Functional simulation was done to verify that the design behaved as expected on the VHDL coding using MAX PLUS II software. Since the delay of the combinational logic and wires were not known, the signal suffered only a constant signal change delay of 0.1ns. This delay must be taken into consideration. As the design operated on positive clock edge, this delay could cause the response of the circuit to be delayed by 1 clock cycle. As such, the processes of the other modules were also delayed 1 clock cycle so as to synchronize the operation of the whole design.

The validation of the 3DES operation was done by referring to [11], where the data message was encrypted based on DES. Validation of the above design after 16 iterations showed the same result as that of DES in [11]. To further verify the design, a simulation done for the whole 3DES operation showed that the encrypted input data could be decrypted to recover the original data message. The two aforementioned verification methods indicated that this design implemented 3DES correctly. The timing diagrams obtained from the simulation that was done during the verification process are shown in Figure 2 and Figure 3 for encryption and decryption respectively.

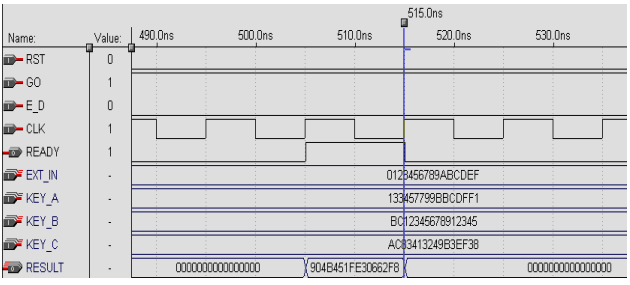


Fig. 2. Functional simulation of encryption operation

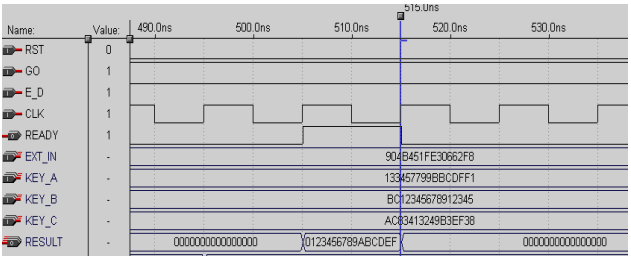


Fig. 3. Function simulation of decryption operation

From Figure 2, the keys used in the encryption of data message 0123456789ABCDEF were 133457799BBCDFF1, BC12345678912345 and AC83413249B3EF38. The cipher resulted text was 904B451FE30662FB. During the decryption operation, the cipher text was decrypted to obtain the original data message of 0123456789ABCDEF as shown in Figure 3.

4.2 Synthesis and Optimization

With the functional simulation showing the correct behavioural result, synthesis was done using Altera Quartus II 4.0 software on the core design implemented into FPGA. Device family that could fit the design into it was chosen and the timing requirements were set.

Different FPGA could result in different maximum frequency obtained. A larger FPGA device family such as APEX20KE gave a higher maximum clock frequency than FLEX10KE. This was an important criterion to be considered while deciding on the FPGA to be used even though the design could be fitted into both. The smaller device family had a higher resource utilization percentage. By deciding to use a larger device family, speed optimization had been given priority in view of excessive amount of resource in the FPGA selected.

There was a trade off between area and speed. A higher number of logic elements used that resulted in higher maximum operating frequency. Initial synthesis of the design on APEX20KE family gave a maximum frequency of 72.77 MHz. However, after switching off the ‘Remove Duplicate Registers’ and ‘Remove Duplicate Logic’ setting, the maximum operating frequency achieved approximately 77MHz. The logic

cells used that summed up to be 25%, which was 2% more than the previous setting. By setting the maximum frequency requirement to 80MHz, a higher value of 78.59 MHz was achieved. This was the highest maximum clock frequency value with APEX20KE family that could be obtained from the optimization process.

4.3 Timing Simulation

Timing simulation was performed to verify that the module functioned correctly and there were no timing violations in the implemented design. The functional simulation was done using MAX PLUS II software but the timing simulation was done using Quartus II 4.0 software.

During timing simulation, the total delay of the wires and combinational logic was taken into account. Initial testing using the clock signal having frequency that was higher than that of maximum operating frequency resulted in erroneous output. The result obtained was not the encrypted data message. This was because the encrypted data cannot be decrypted to recover the initial data message. The total delay had exceeded one clock cycle period.

The clock signal period was then set to 13ns. This clocking period was larger than the total wire and combinational logic delay. Different sets of keys and input data blocks were used during the simulation. It was found that the encrypted data could be decrypted to recover the original data. Besides that, the reset pin had also been tested. Reset signal was set to 'high' to reset the design.

4.4 Synthesis Results

Table 1 shows the synthesis results of the 3DES encryption engine. The FPGA family that had been selected for the realization of 3DES encryption engine was APEX20KE (more precisely, EP20K200EFC672-1X).

Table 1. Font sizes of headings. Table captions should always be positioned *above* the tables. The final sentence of a table caption should end without a period.

Family	APEX20KE
Device	EP20K200EFC672-1X
Name	Core
Total logic elements	2111 / 8320 (25%)
Total I/O pins	325 / 376 (86%)
Total memory bits	2048 / 106496 (1%)
Total PLLs	0 / 2 (0%)
Total combinational functions	2110
Total registers	408
Performance, f_{max}	78.59 MHz
Clock period	12.724 ns

Out of the 8320 logic elements contained in the device, a total of 2111 logic cells were used. A total of 325 I/O pins were utilized, which is equivalent to 86 percent of the total pins in the device. Out of these 325 pins, 65 pins were output pins while the remaining

260 pins were input pins. Out of a total of 106496 memory bits in the device, 2048 of them were utilized. This is equivalent to 1 percent of the total memory bits resource. Besides that, the total number of registers used in the EP20K200EFC672-1X device summed up to be 408. A maximum clock frequency of 78.59 MHz was obtained. The clock signal that was used in the device must have a period of at least 12.724 ns. Any period below this value gave a faulty result.

4.5 Timing and Area Analysis

The results for timing and area analysis of the main modules are presented in terms of maximum operating frequency and logic cell (LC). The analysis was done using Quartus II software. The devices chosen for the implementation were EP20K200EFC672-1X of APEX20KE family and EPF10K100EFC484-1 of FLEX10KE family. Comparison was done between the two devices.

Table 2. Effect of registers and logic cells duplication in EP20K200EFC672-1X

'Remove Duplicate Registers' and 'Remove Duplicate Logic'	Area (LC)	Clock Period (ns)	Maximum Operating Frequency (MHz)
On	1984/8320 (23%)	13.742	72.77
Off	2111/8320 (25%)	12.724	78.59

Table 3. Effect of registers and logic cells duplication in EPF10K100EFC484-1

'Remove Duplicate Registers' and 'Remove Duplicate Logic'	Area (LC)	Clock Period (ns)	Maximum Operating Frequency (MHz)
On	1924/4992 (38%)	17.5	57.14
Off	2080/4992 (42%)	17.6	56.82

Tables 2 and 3 show the effect of registers and logic cells duplication in EP20K200EFC672-1X and EPF10K100EFC484-1 respectively when the full 3DES architecture was mapped into them. To implement these features, the 'Remove Duplicate Registers' and 'Remove Duplicate Logic' settings were selected or deselected.

When the 'Remove Duplicate Registers' and 'Remove Duplicate Logic' settings were selected during the hardware implementation of the encryption module in EP20K200EFC672-1X, this resulted in lower area utilization of 1984 logic cells and lower maximum operating frequency of 72.77 MHz. When these settings were deselected, higher area utilization of 2111 logic cells and higher maximum operating frequency of 78.59 MHz was obtained.

However, that is not the case when EPF10K100EFC484-1 was used. Selecting the 'Remove Duplicate Registers' and 'Remove Duplicate Logic' setting resulted in lower area utilization but higher maximum operating frequency.

Table 4. Synthesis results

Device	Area (LC)	Clock Period (ns)	Maximum Operating Frequency (MHz)
EP20K200EFC672-1X	2111/8320 (25%)	12.724	78.59
EPF10K100EFC484-1	1924/4992 (38%)	17.5	57.14

Table 4 shows the synthesis results for the final design of the project. Two devices were used, namely EP20K200EFC672-1X and EPF10K100EFC484-1. EP20K200EFC672-1X is a larger device compared to EPF10K100EFC484-1.

When the larger device was used, it was found that the final design had a higher maximum operating frequency of 78.59 MHz. It utilized more logic cells. However, when the smaller device from FLEX10KE family was used, it only had a maximum operating frequency of 57.14 MHz. Besides that, the design used only 1924 logic cells of the resource, which was lesser than the 2111 logic cells used in EP20K200EFC672-1X.

With this, it can be concluded that the mapping of the design architecture on different devices can result in different maximum operating frequency and area utilization. A larger device results in higher maximum operating frequency and larger area utilization. As such, considerable decision must be taken on whether a faster operation is needed or a smaller device is required.

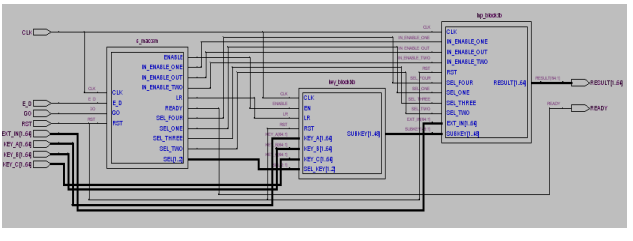


Fig. 4. RTL view of core entity

Figure 4 demonstrates the RTL view of the core entity. It is shown that core was formed by three smaller entities, namely s_mac, key_block and inp_block. Each of these entities had its own unique function. S_mac controlled and synchronized the operations of the other two entities while key_block processed the three keys, producing the subkeys needed before sending them to inp_block. Inp_block was the entity where the actual encryption and decryption of the plaintext occurred.

Table 5. Comparisons between hardware and software implementation

	3DES (FPGA)	3DES (software)
Key size (bits)	192	192
Data rate (Mbps)	102.56	42.9

Table 5 shows the comparisons done on the performances of the hardware and software implementation of 3DES. Triple DES was implemented into FPGA and as well as MATLAB using an Intel Pentium III 866 MHz machine. It shows that 3DES hardware was significantly (2.4 times) faster than its software counterpart. The 3DES software could only manage a data rate of 42.9 Mbps compared to 102.56 Mbps of 3DES hardware.

5 Conclusions

The hardware implementation of 3DES encryption engine on FPGA chip was realized. The chip selected was EP20K200EFC672-1X of APEX20KE family. It could encrypt data at a rate of 102.56 Mbps, with a maximum operating frequency of 78.59 MHz and area utilization of 2111 logic cells.

The throughput of 102.56 Mbps in the current full implementation of 3DES core can be considered as low by industry standard. As such, to improve the throughput of the design, pipelining of the iterations process can be implemented. Registers can be added to store data during the pipelining process. This will invariably reduce the maximum clock frequency; however the number of clock cycles being used for one complete 3DES operation can be greatly reduced, thus reducing the latency.

To allow more secured encryption process, additional 3DES operation modes can be added to the core module. Currently, the encryption hardware only operates under TECB mode. By including more modes of operation, users can choose to operate under certain mode, depending on their preference.

Acknowledgement. The authors would like to express sincere gratitude to the Research Centre, International Islamic University Malaysia for providing fund for the research under IIUM Long-Term Research Grant (IIUM/504/RES/G/14/3/01/LT38).

References

1. Aladdin Knowledge System, The enduring Value of Symmetric Encryption, White Paper, pp. 5–8 (August 2000)
2. Harper, S., Athanas, P.: A Security Policy Based Upon Hardware Encryption. In: System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference, Virginia, January 5-8, 2004, pp. 190–197 (2004)
3. Runje, D., Kovac, M.: Universal Strong Encryption FPGA Core Implementation. In: Design, Automation and Test in Europe, 1998, Proc. France, February 23-26, 1998, pp. 923–924 (1998)
4. Cheung, O.Y.H., Leong, P.H.W.: Implementation of an FPGA Based Accelerator for Virtual Private Networks. In: Proceedings of IEEE International Conference on Field-Programmable Technology (ICFPT), Hong Kong, pp. 34–41 (2002)
5. Data Encryption Standard, Federal Information Processing Standards (FIPS) Publication 46-7, National Institute of Standards and Technology (NIST), USA (1999)
6. Kim, Y.S., Kang, W.S., Choi, J.R.: Implementation of 1024-bit Modular Processor for RSA Cryptosystem. In: AP-ASIC 2000, Proceedings of the Second IEEE Asia Pacific Conference, Korea, August 28-30, 2000, pp. 187–190.

7. Sanchez-Avilla, C., Sanchez-Reillo, R.: The Rijndael Block Cipher (AES Proposal): A Comparison with DES. In: Security Technology, 2001 IEEE 35th International Carnahan Conference, London, October 16-19, 2001, pp. 229–234 (2001)
8. Raghuram, S.S., Chakrabarti, C.: A Programmable Processor for Cryptography. In: Proceedings. The 2000 IEEE International Symposium on Circuits and Systems, Geneva Switzerland, May 28-31, 2000, vol. 5, pp. 685–688 (2000)
9. Wu, L., Weaver, C., Austin, T.: Crypto Maniac: A Fast Flexible Architecture for Secure Communication. In: Computer Architecture, 2001. Proceeding. 28th Annual International Symposium, Goteborg, Sweden, 30 June-4 July 2001, pp. 110–119 (2001)
10. Chodowiec, P.R.: Comparison of the Hardware Performance of the AES Candidates Using Reconfigurable Hardware. Master Thesis, 150 pages, George Mason University (2002)
11. Orlin Grabbe, J.: (January 5, 2004), www.aci.net/kalliste/des.htm

Finite Element Simulation of a Biomimetic Olfactory Microsystem for Spatio-temporal Signal Generation

S.L. Tan[1,2], J.A. Covington[2], J.W. Gardner[2], and T.C. Pearce[3]

[1] School of Computer Engineering, Nanyang Technological University, Singapore (current)
[2] Sensors Research Laboratory, University of Warwick, Coventry, CV4 7AL, UK
[3] Centre for Bioengineering (NeuroLab), University of Leicester, Leicester, LE1 7RH, UK

Abstract. The sense of smell is a powerful biological tool although it is the least understood. Attempts to mimic this feature over the last two decades have resulted in the creation of the electronic nose. In comparison to the biological system, its ability to distinguish complex odours is poor. This has mainly been attributed to the lack of sensors and their diversity compared to the human in the order of 10^5 and 10^2 respectively. In our efforts to improve the performance of the electronic nose, here we have used a different approach using a unique feature of the biological olfactory system. This technique is analogous to a multi-dimensional gas chromatography (MD-GC) technique that is capable in generating spatial and temporal signals to aid odour discrimination. As the physical realisation requires expensive and time consuming micro- nano fabrication processes, finite element method simulations have been used to validate the proposed design and aid optimisation. This paper describes the finite element modelling process and compares these simulation results to that of the well-established analytical model. Preliminary results of the optimised system are also presented; these results are in good agreement to the simulated outputs.

Keywords: Spatio-temporal; Multi-dimensional gas chromatography (MD-GC).

1 Introduction

Gas chromatography (GC) is a widely used chemical analysis technique. Though successful it suffers from several disadvantages such as long analysis time, large instrumentation size and high cost. To this end there have been efforts to circumvent these drawbacks as seen by the introduction of the micro GC systems by Terry et al. [1], Kolesar et al. [2], Sacks et al. [3] and Hudson et al. [4]. To predict and optimise the performance of these systems prior to fabrication, simulation is usually performed using the well-established plate theory model developed by Golay [5]. Numerous commercial GC simulation packages such as GC-SOS [6] and ChromWIN [7] are readily available to allow system optimisation. These packages employ either a theoretical plate model or a stochastic model based on the simulation of Gaussian peaks or a combination of both [1-9]. This has served the community well given that the calculations have a low simulation overhead and the requirement is to predict the response only at the outlet of the column.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 216–226, 2007.

Nasal chromatography is commonly used to describe the process that occurs within the biological olfactory system to generate spatio-temporal signals. This process is considered akin to MD-GC. The spatio-temporal signals can be visualised as a multi-GC columns system each coated with a different stationary phase material, and each column having a number of sensors placed along the column. To investigate and replicate this effect, we have created a design incorporating a microstructure geometry consisting of 5 GC columns and a common injection and extraction system. The spatio-temporal measurement is a technique to extract the signals differences in space and time. The spatial (space-difference) signal can be obtained by comparing signals from sensors among different columns. Signals from sensors along each column are compared to produce the temporal (time-difference) signal.

2 Spatio-temporal Geometry

To investigate these spatio-temporal signals our initial design consisted of multiple GC columns (spatial) with each column containing an array of sensors (temporal). With this design it would also be desirable to have a common inlet and outlet to reduce the number of pumps and valves required. This common injection and flow system must ensure equal concentration distribution across all the columns at the same time. Several geometries were simulated and they are shown in Figure 2.1. It was found that the "WHEEL" architectural geometry provides the most satisfactory design in terms of concentration distribution across all columns simultaneously, and had the least band broadening. Band broadening is the effect of widening pulse as the test analyte transverse from the inlet to the outlet. This is an important parameter dictating the performance of the system.

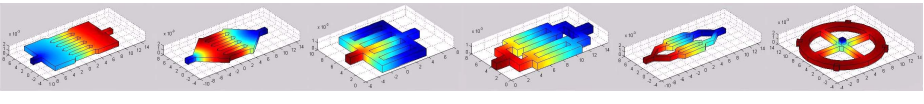


Fig. 2.1. Different column geometries for spatio-temporal signal generation

As the design evolved, an optimized geometry was achieved. So far, we concentrated on the geometry designs to derive a common inlet/outlet system. Next, the geometry for the GC separation process will be described. All the columns are the same though each column is coated with a different stationary phase. To reduce the simulation overhead it is possible to simulate just one-eighth of the columnby exploiting the symmetry of its cross-section. It is feasible to simulate the transport flow as a 2-D parabolic profile and combine it with the convection and diffusion along the column. The 2-D parabolic flow profile reduces the computation complexity without any compromise in resolution because a uniform cross-section column has the same flow profile. The one-eighth column symmetry increases the accuracy of the model as more nodes (computation points) are used to compute the geometry rather than representing redundant nodes, as they are symmetrical.

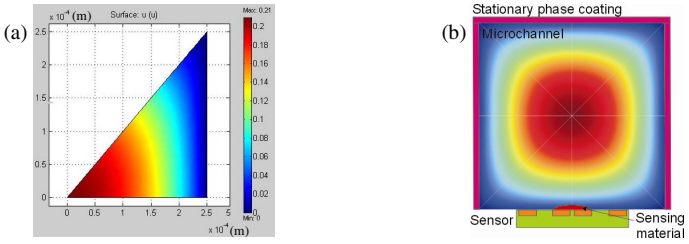


Fig. 2.2. 2-D parabolic velocity profile of micro-channel. (a) 1/8 cross-sectional velocity profile. (b) Complete micro-channel velocity profile with sensor placement.

Figure 2.2(a) shows the 2-D velocity flow profile. The dimension matches the cross-section of the 3-D column shown in Figure 2.3. Figure 2.2(b) shows the symmetrical property of the column where one-eight of the geometry is simulated. The 2-D geometry corresponds to the cross-section of the column inlet, and the solution is exported to the inlet in Figure 2.3.

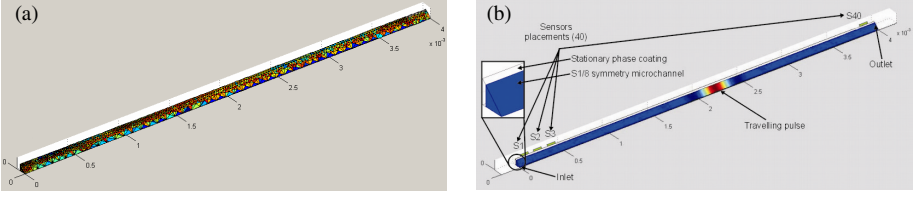


Fig. 2.3. Geometrical model of a stationary phase coated micro-channel. (a) Micro-channel mesh (computational nodes). (b) Simulated surface profile of a travelling ethanol vapour pulse in air.

Figure 2.3 shows the isometric view of the 3-D 1 m long GC column scaled by 250 times coated with a 5 µm stationary phase. The cross-section of the column is 100 µm by 100 µm. The geometry shown in Figure 2.3(a) creates the computational nodes or the finite points that the simulation will be based upon. Figure 2.3(b) illustrates an ethanol vapour pulse travelling down the coated micro-channel at time $t = 7$ sec.

3 Mathematical Model and Properties

Numerous researchers have identified the physical phenomenon dominating the separation process of GC as the laminar flow with convection and diffusion occurring within the column, and only diffusion occurs within the stationary phase coating [15-16]. A 2-D *Navier-Stokes* (NS) is used to simulate the laminar transport. The result of this parabolic velocity profile is then exported to the inlet of the column. Within the column, a multi-physics mode with *Convection and Diffusion* (CD) in mobile phase and *Diffusion* (DI) in stationary phase are used. The NS simulates the laminar transport of carrier gas and test analyte along the column, while CD accounts for the flux transport and dispersion of the injected pulse. A separate DI mode is required to

simulate diffusion within the stationary phase. The retentive characteristic of an analyte with the stationary phase material is simulated by modifying boundary conditions as described later. The flow is laminar because the Reynolds number at optimum velocity is only 82, which is far less than the transition region from laminar to turbulent flow of 2000. In this model, the diffusion in the stationary phase is incorporated to improve model accuracy. This parameter is not usually included in most GC simulations [17], instead instantaneous diffusion is assumed. A small convective term also occurs within the column, but is neglected by [1-5], but described by Spangler [18] is also incorporated. The *Partial Differential Equations* (PDEs) used for the simulation are described as follows:

Navier-Stoke (NS) equation

$$\mu\nabla^2 V + \rho(V.\nabla)V + \nabla P = 0 \tag{3-1}$$

where μ is the viscosity, ρ is the density, P is the pressure and V is the velocity.

Convection and Diffusion (CD) equation

$$\frac{\partial C}{\partial t} + \nabla.(-D\nabla C + CV) = 0 \tag{3-2}$$

where D is the diffusion coefficient, C is the concentration.

Diffusion (DI) equation

$$\frac{\partial C}{\partial t} + \nabla.(-D\nabla C) = 0 \tag{3-3}$$

Boundary conditions for concentration gradient in stationary phase
At the inlet, we define the concentration conditions according to:

$$\mathrm{C} = C_i \times (t < t_{PW}) \text{ at } \partial\ \Omega_{C,in} \tag{3-4}$$

where C_i is the injected concentration of the analyte for a duration of t_{PW}.

At the outlet, we assume the convective contribution to the mass transport is much larger than the diffusive contribution.

$$(-D\nabla C_M + C_M v).n = (C_M v).n \text{ at } \partial\ \Omega_{C,out} \tag{3-5}$$

where n is the normal vector to the respective boundary and C_M is the concentration in mobile phase.

Furthermore, we assume there is no transport over the symmetry boundaries.

$$(-D\nabla C_M + C_M v).n = 0 \text{ at } \partial\ \Omega_{C,sym} \tag{3-6}$$

Figure 3.1 shows the symmetrical property of the column and the boundary conditions used in the simulation. To account for the partition coefficient c, which measures the equilibrium distribution of analyte molecules between the mobile (gas) phase and the stationary phase, the boundaries conditions are modified as in (3-7) and (3-8).

$$\left[\frac{\partial C_M}{\partial t} + \nabla.(-D\nabla C_M + C_M V)\right].n = M.(C_S - cC_M) \text{ at } \partial\,\Omega_{C,SP} \tag{3-7}$$

$$\left[\frac{\partial C_S}{\partial t} + \nabla.(-D\nabla C_S + C_S V)\right].n = M.(cC_M - C_S) \text{ at } \partial\,\Omega_{SP,C} \tag{3-8}$$

where C_S is the concentration in stationary phase.

As there will be discontinuities in the concentration profile at the boundaries between mobile and stationary phase, two separate variables, C_M and C_S are used to describe the concentration in these respective phases. In order to obtain continuous flux over the phase boundaries, a special type of boundary condition, using the stiff-spring method is employed. Instead of defining the Dirichlet concentration condition according to the partition coefficient c, which would destroy the continuity of the flux, a continuous flux condition is defined. At the same time, the concentration is forced to the desired values. M is a large number to let the concentration differences in the brackets (Equation (3-7) and (3-8)) on the right to approach zero. Note that these boundary conditions also give us continuity in flux, provided M is sufficiently large (≈ 10000) [19].

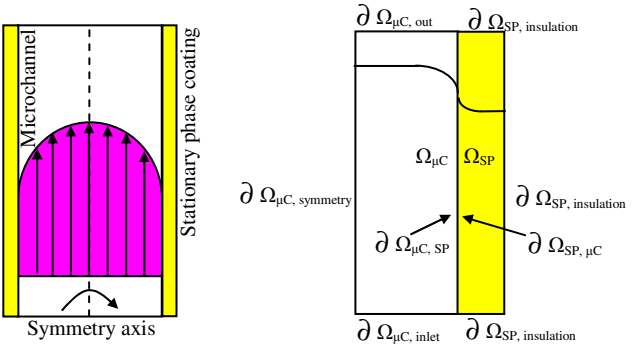


Fig. 3.1. Micro-channel GC boundary conditions and symmetry axis

c is the partition coefficient defined as

$$c = \frac{C_S}{C_M} \tag{3-9}$$

When using FEMLAB (COMSOL, Sweden), the modelling starts with the selection of approximate PDEs (if available) to represent the physical phenomena (otherwise the user has to formulate their own PDEs). The geometry with exact dimension is then created using a graphical user interface (GUI) with a library of basic geometries. Complex geometry can be created using basic set theory operations such as union, subtract and intersect, or imported from third-party software. Once the boundary conditions and domain parameters are defined, the structure has to be mesh, to define elements or specific nodes within the geometry where the mathematical computation will be performed. There is a trade-off between accuracy and computational time depending on the numbers of nodes and the PDE modes used.

Table 3.1. Mesh properties of the 2-D NS and 3-D column

Mesh property	2-D NS	2-D column
Number of nodes	221	5009
Number of boundary elements	54	9611
Number of elements	386	18667

Table 3.1 list the mesh properties of the two geometries used in the simulation. The 3-D column contributes the bulk of the simulation time as it employs a multi-physics mode. It takes about 7 hours on an Intel Pentium 4 2 GHz PC with 512 MB RDRAM to perform 100 time-step calculations. The solutions are readily available in graphical format using various plot modes within FEMLAB or can be easily exported to, for example, MATLAB for post-processing.

4 Simulation Parameters and Results

The simulations were conducted with 5 stationary phase materials and 2 test analytes, as listed in Table 4.1. Essential parameter such as the partition coefficient is calculated using the linear solvation energy relationship (LSER) equations [20-21]. To verify the accuracy and correctness of the simulated results, Golay [5] and Spangler [22] GC models were used for comparison, concentrating on two main parameters. The first parameter is the retention time, which specifies the time an injected pulse is being delayed by the column. This is affected by the partition coefficient between the analyte and the stationary phase material, the carrier gas velocity, the thickness of the stationary phase coating and the diffusion coefficient of the analyte in each phase. The second parameter is the separation factor that characterises the performance of the GC system.

Table 4.1. Simulated stationary phases and test analytes with 5 stationary phase materials

Stationary phase material	Test analyte
1. Polydimethylsiloxane (SE-30)	1. Ethanol
2. Polydimethylmethylphenylsiloxane 35 % mol phenyl groups (OV-11)	2. Toluene
3. Polyethylene glycol adipate (EGAD)	
4. Bis-3-allyl-4-hydroxyphenyl sulfone (H10)	
5. Polytrifluoropropylmethylsiloxane (QF1)	

The average velocity (Vel_{ave}) can be calculated analytically as follows:

$$Vel_{ave} = \frac{1}{1+k} \times v_c \tag{4-1}$$

where v_c is the velocity of the carrier gas and k is the partition ratio.

Therefore, the retention time is:

$$t_r = \frac{L}{Vel_{ave}} \tag{4-2}$$

where L is the length of the column.

Equation (4-1) and (4-2) can be used to determine the retention time analytically [5]. Table 4.2 shows that the retention time of the two models is a maximum of 3 % apart, with the numeric solution from the finite element model giving a higher value.

Table 4.2. Retention time for ethanol and toluene vapours

Ethanol vapour				
	c	Analytical	Numeric	Error
SE-30	4.76	18.9 s	19.0 s	0.53 %
OV-11	6.03	21.4 s	21.6 s	0.93 %
EGAD	14.97	38.7 s	39.5 s	1.96 %
H10	58.03	122.3 s	125.0 s	2.22 %
QF1	6.28	21.9 s	22.0 s	0.55 %

Toluene vapour				
	c	Analytical	Numeric	Error
SE-30	37.59	82.6 s	84.5 s	2.28 %
OV-11	56.74	119.8 s	123.0 s	2.69 %
EGAD	170.7	340.9 s	349.0 s	2.39 %
H10	48.21	103.2 s	106.0 s	2.68 %
QF1	28.57	65.1 s	66.5 s	2.10 %

$$HETP_{Golay} = 2\frac{D_m}{v_c} + \frac{4(1+9k+51k^2/2)}{105(1+k)^2}\frac{v_c z_0^2}{D_m} + \frac{8k}{3(1+k)^2}\frac{v_c w^2}{D_s} \tag{4-3}$$

$$HETP_{Spangler} = 2\frac{D_m}{v_c} + \frac{4(0.9+2k+35k^2)}{96(1+k)^2}\frac{v_c z_0^2}{D_m} + \frac{2k}{3(1+k)^2}\frac{v_c w^2}{D_s} \tag{4-4}$$

$$SF = \frac{L}{HETP}\left(\frac{k}{1+k}\right)^2 \tag{4-5}$$

The separation factor dictates the performance of a GC system and there are various analytically solutions available, evolving from the original Golays' equations [5]. Here the *height equivalent to theoretical plate* (HETP) and separation factors are used for comparison and are shown in Equation (4-3) to (4-5).

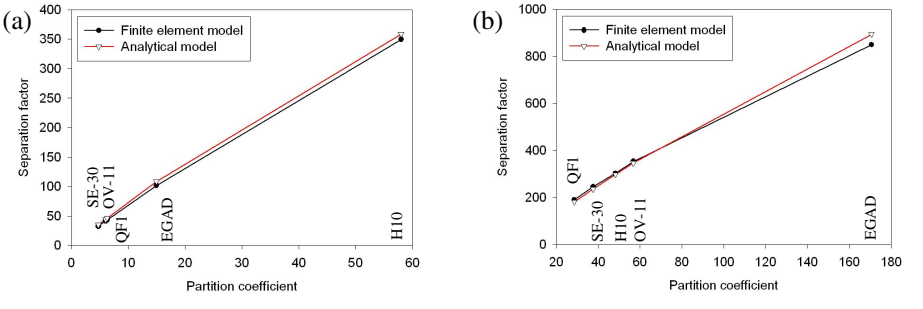


Fig. 4.1. Comparison of separation factor between analytical and finite element model. (a) Separation factor for ethanol vapour. (b) Separation factor for toluene vapour.

Comparing the numeric separation factor to Golay's model shows that it is about 3 times larger. When it is compared to the Spangler model, it is about 9 % lower. The significant difference in separation factor could be contributed by the input pulse width used. The results shown in Figure 4.2(a) and (b) have an input pulse width of 0.2 sec, which is significantly larger than the normal GC pulse width in the milli-second range. This parameter is not considered in both Golay and Spangler HETP derivation. Nevertheless, the finite element model provides a separation factor within the range of the various analytical solutions.

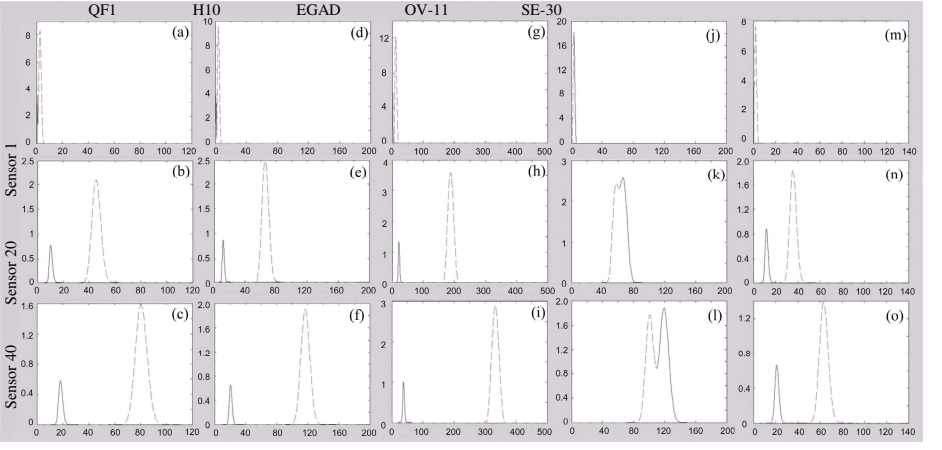


Fig. 4.2. Spatio-temporal signals, (a)-(c) Sensor responses to ethanol (solid) and toluene vapour (dotted) pulses in air at three locations (near inlet (S1), middle (S20) and near outlet (S40)) along a SE-30 coated column, (d)-(f) Sensor responses under the same condition with a column coated with OV-11, (g)-(i) Sensor responses under the same condition with a column coated with EGAD, (j)-(l) Sensor responses under the same condition with a column coated with H10, and (m)-(o) Sensor responses under the same condition with a column coated with QF1. (Plotted in percentage change in resistance versus time (sec)).

Besides predicting the retention time and separation factor of the GC system, the main aim of performing finite element simulation is to extract spatio-temporal signal at strategic locations along the column. Given the same analytes, it will propagate along different column at different average velocity due to the different affinity of the analytes with the stationary phases. Within a column, by comparing the signals at various locations along the column, we can obtain a slowly separated signal. This corresponds to the temporal information. Comparing across different columns, the signals will be different as each column has a different stationary phase coating. These responses produce the spatio-temporal signals as shown in Figure 4.2.

The results obtained above clearly indicate that the simulated geometry is capable of generating both spatial and temporal signals. Using a micro-channel coated with different materials will allow different test analytes to provide very contrasting responses. The next natural progression is to validate the simulated models with physical system realisation. Details pertaining to these realizations are published elsewhere [23-25]. Figure 4.3 shows a fabricated microsystem by assembly a 2.4 m long micro-channel coated with Parylene C assembled with an 80 element microsensor array.

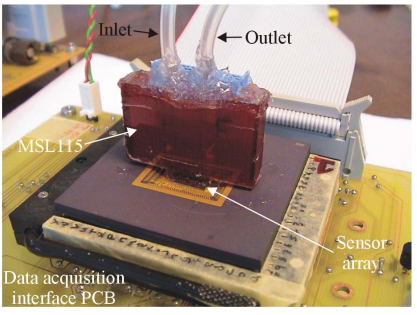


Fig. 4.3. Test setup of a MSL115 micro-channel package fabricated using stereolithography assembled with a silicon-based sensor array

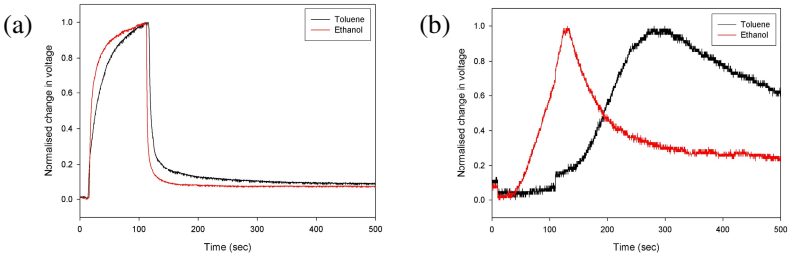


Fig. 4.4. Varying temporal signals for simple analytes on micro-nose system when assembled with MSL115 micro-channel package. (a) Sensor (30 mm from inlet) response. (b) Sensor (2280 mm from inlet) response.

Figure 4.4 shows some of the preliminary results obtained when the micro-nose system was tested with ethanol and toluene vapour in air. Figure 4.4(a) shows the sensor responses of the ethanol and toluene vapour pulses near the inlet where both arrived almost simultaneously while (b) shows the responses towards the outlet with the toluene vapour pulse showing a significant time delay over the ethanol vapour pulse. The other important component not discussed in this paper is the response model of the sensor. Here a polymer-composite sensor is used. Hence, the responses obtained in Figure 4.4(b) are different from the simulated results shown in Figure 4.2(a).

5 Conclusion

The rational for performing finite element simulation as opposed to analytical simulation is that this technique allows almost any geometry to be evaluated. For analytical analysis, only circular and rectangular columns models have been developed. Besides, to achieve multiple long columns within a miniature microstructure will require spiral columns fabricated adjacent to each other in which case circular and rectangular spiral networks are usually used. Furthermore, finite element simulation allows a non-perfect column to be analysed, this can take into account uneven stationary phase coating, dead volume of the system and irregular column shape as a result of manufacturing tolerance. In addition, there is the need to

predict the response at various points along the column to obtain the temporal signal. Such a requirement is not possible with analytical analysis as only the response at the outlet is available. Finite element modelling of the GC system provides an alternative simulation technique and has the added advantages necessary to obtain a spatio-temporal response that is not possible using theoretical or stochastic models. This model provides highly accurate results resembling those generated by the well-established plate theory model. Once the model is created, simulation with different test analytes and stationary phases can be performed with ease simply by modifying the variables.

The spatio-temporal signal measurement of a MD-GC provides a unique and novel approach towards addressing the key challenges posed by the human olfactory system towards artificial odour discrimination. Such an approach not only generates a new dimension towards solving this problem, it also provides a platform for improving traditional GC systems in terms of analysis time and capability. This creates a whole new arena of signal processing challenges to look into different ways of analysing this multi-dimensional problem. To bring this a step closer to reality, the research efforts at University of Warwick are currently focussed towards developing a prototype of this system using commercially available aVLSI technology. At the same time, other signal processing techniques, such as wavelet analysis is being used to efficiently process the vast amounts of information.

Acknowledgement

The authors would like to thank the technical support team of COMSOL for their numerous stimulating discussions and assistance during the development of the finite element model. We would also like to thank the EPSRC in the UK for the financial support of this project.

References

[1] Terry, S.C., Jerman, J.H., Angeli, J.B.: A gas chromatographic air analyzer fabricated on a silicon wafer. IEEE Transaction on Electron Devices, 1880–1886 (1979)

[2] Kolesar, E.S., Reston, R.R.: Silicon-micromachined gas chromatography system used to separate and detect ammonia and nitrogen dioxide. Part I & II, Journal of the Microelectromechanical Systems 3, 134–154 (1994)

[3] Whiting, J., Sacks, R.: Selectivity enhancement for high-speed GC analysis of volatile organic compounds with portable instruments designed for vacuum-outlet and atmospheric-pressure inlet operation using air as the carrier gas. Anal. Chem. 74, 246 (2002)

[4] Hudson, M.L., Kottenstette, R., Matzke, C.M., Frye-Mason, G.C., Shollenberger, K A., Adkins, D.R., Wong, C.C.: Design, testing and simulation of microscale gas chromatography columns. MEMS, ASME Int. Mechanical Engineering Congress and Exposition DSC66, 207–214 (1998)

[5] Golay, M.J.E.: Gas Chromatography: Theory of chromatography in open and coated tubular columns with round and rectangular cross-sections. In: Desty, D.H. (ed.) 1958 Amsterdam Symposium, Butterworths, London, pp. 36–55 (1958)

[6] GC-SOS Gas Chromatography Simulation and Optimization Software, Viewed on December 20, 2002 http://www.chemsw.com/

[7] Trapp, O., Schurig, V.: ChromWIN - A computer program for the determination of enantiomerization barriers in dynamic chromatography. Computer and Chemistry 25, 187–195 (2001)
[8] Snijders, H., Janssen, H.-G., Cramers, C.: Optimization of temperature-programmed gas chromatographic separation, I. Prediction of retention times and peak widths from retention indices. Journal of chromatography A 718, 339–355 (1995)
[9] Wu, Z.Y.: Retention simulation in gas chromatography. Journal of chromatography A 840, 137–143 (1999)
[10] Mozell, M.M.: Evidence for a Chromatographic Model of Olfaction. Journal of General Physiology 56, 46–63 (1970)
[11] Nachbar, R.B., Morton, T.H.: A Gas Chromatographic (GLPC) Model for the Sense of Smell. Variation of Olfactory Sensitivity with Conditions of Stimulation. Journal of Theoretical Biology 89, 387–407 (1981)
[12] DeAngelis, G.C., Ohzawa, I., Freeman, R.D.: Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex, I & II. J. Neurophysiology 69 (1993)
[13] Briglin, S.M., Freund, M.S., Tokumaru, P., Lewis, N.S.: Exploitation of spatiotemporal information and geometric optimization of signal/noise performance using arrays of carbon black-polymer composite vapour detectors. Sensors and Actuators B82, 54–74 (2002)
[14] Mozell, M.M., Jagodowicz, M.: Mechanisms underlying the analysis of odorant quality at the level of the olfactory mucosa, I: Spatiotemporal sorption patterns. Annual New York Academy of Sciences 237, 76–90 (1974)
[15] Keyhani, K., Scherer, P.W., Mozell, M.M.: A numerical model of nasal odorant transport for the analysis of human olfaction. Journal of Theoretical Biology, 279–301 (1997)
[16] Hahn, I., Scherer, P.W., Mozell, M.M.: A Mass Transport Model of Olfaction. Journal of Theoretical Biology 167, 115–128 (1994)
[17] Wong, C.C., Adkins, D.R., Frye-Mason, G.C., Hudson, M.L., Kottenstette, R., Matzke, C.M., Shudid, J.N., Salinger, A.G.: Modelling transport in gas chromatography columns for the micro-ChemLAB. In: Proc. of SPIE- Microfluidic devices and System II, Santa Clara, California, pp. 120–129 (September 1999)
[18] Spangler, G.E.: Height equivalent to a theoretical plate theory for rectangular GC columns. J. of Analytical Chemistry 70, 4805–4816 (1998)
[19] Fontes, E., Byrne, P., Sundqvist, J., Bosander, P., Marklund, M.: FEMLAB: Chemical Engineering version 2.1 user manual, Comsol, Inc. (March 2001)
[20] Gardner, J.W., Bartlett, P.N.: Electronic nose: principles and applications. Oxford University Press, Oxford (1999)
[21] Abraham, M.H.: Scales of solute hydrogen-bonding: Their construction and application to physicochemical and biochemical processes. Chemical Society Reviews, 73–83 (1993)
[22] Spangler, G.E.: Relationships for modelling the performance of rectangular gas chromatographic columns. Journal of Microcolumn Separations 13(13), 285–292 (2001)
[23] Tan, S.L., Covington, J.A., Gardner, J.W.: Velocity-optimised diffusion for ultra-fast polymer-based resistive gas sensors. IEE- Science, Measurement and Technology 153(3), 94–100 (2006)
[24] Gardner, J.W., Tan, S.L., Covington, J.A., Pearce, T.C.: Enhanced discrimination of complex odours based upon spatio-temporal signals from a micro-mucosa. In: The 14 International conference on Solid-State Sensors, Actuators and Microsystems, Lyon, France, June 10-14, 2007 (2007)
[25] Tan, S.L., Covington, J.A., Gardner, J.W., Hesketh, P.J.: Ultra-fast/low volume odour delivery package for chemical microsystems. In: IEEE Sensor Conference, Vienna (October 2004)

Simulation Study of Port Container Terminal Quay Side Traffic

Y.K.H. Lau and M.Y.N. Lee

Dept. of Industrial and Manufacturing Systems Engineering,
The University of Hong Kong, Pokfulam Road, Hong Kong, PRC
hyklau@hkucc.hku.hk, h0171306@hkusua.hku.hk

Abstract. This paper studies the berth operation, in particular, the quay side traffic condition in one of the busiest container port in Hong Kong. The developed simulation model effectively incorporates both the traffic model and service model, built using the AutoMod simulation tool. Sensitivity analysis was performed on different scenarios and configurations at the quay side. The results reveal that an increase in the number of internal tractors increases the productivity of the quay cranes, and the deployment of a traffic controller can improve the traffic congestions but may not help to improve the utilization of the quay cranes. Apart from the flow control of the tractors, the closure of tractor lanes under a quay crane also reduces the performance of the quay crane. The simulation studies performed provides a definitive basis for evaluating the performance of the berth and for determining the potential for the operation improvements of a port container terminal in general.

1 Introduction

Resource planning and optimization is curial in port container terminal development. A large number of study was carried out on berth operation in the past. These studies have focused on berth planning of visiting vessels [5], ships loading and unloading operation [2] and the performance evaluation on resources planning in multiple berths [6], yard operation [8] and the traffic of straddle carrier [1]. Our study focuses on the traffic flow of internal tractors at the quay side of a berth. In the simulation case study, we have selected a typical berth with a quay length that is long enough to accommodate a mega vessel.

A number of traffic studies had been found in the aerospace system [3]. Foregoing simulation analysis on container terminal operations has been studied by using object-oriented software [2] [9] However, traffic model is not being considered in these studies. In particular, the traffic flow condition and animation of the traffic requirement was not explicitly modeled. To address the operation at the quay side with focus on the traffic flow of the internal tractors, it is therefore necessary to construct an integrated simulation model that combines the traffic model and service model. This integrated model enables the effective and efficient representation of the quay side traffic flow in a port container terminal, and hence provides a more accurate means to the analysis of the actual terminal operation.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 227–236, 2007.

The proposed integrated simulation model is developed using AutoMod[1] version 11.2 which provides built in mechanisms for modeling vehicle routing [4][7]. The movement of the internal tractors and the operation of the quay cranes are modeled and their operations are clearly revealed by the outputs of the simulation runs. Observations, potential operational problems, and in particular, the cause of tractor traffic problems such as congestions and deadlocks are investigated and resolved with the assistance of the simulation system implemented.

1.1 Objectives

To maximize the utilization of the resources such as cranes, tractors and other terminal equipment are crucial in modern port container terminals. Internal tractors and quay cranes are the major resources that are deployed in berth. In general, internal tractors are deployed in a container terminal solely for the delivery of containers between the yard and the berth whereas quay cranes are designed specifically to handle containers to and from a vessel that is berthed. They are by far the most efficient equipment for transferring containers between a vessel and the quay. The objective of our study is to investigate the traffic flow of internal tractors with corresponding interactions with the quay cranes at a typical berth. The developed model is used to analyze the effect of the quay crane rate and other key performance indicators with respect to the changes in the following parameters, namely,

- Number of internal tractors deployed for a vessel
- Number of traffic lanes under the quay cranes
- Number of internal tractors traveling to the quay side (the function of the proposed IT Controller)

2 Terminal Operation

Loading and discharging of containers are the main operation at the quay side of a typical port container terminal. These two operations occur simultaneously and interactively with the major resources including the quay cranes (QC) and internal tractors (IT). Loading operation (PS) is the movement of an outbound container from the quay side to a vessel. A container is picked up form the yard and is loaded on a dedicated IT, which then travels to and unloads the container at the quay side. Discharge operation (PL) is the movement of an inbound container from a vessel to the quay side. An empty internal tractor travels to the quay side and stopped at the dedicated location, call a service point (SP), then a container is loaded onto the tractor. The operation is said to be completed when an inbound container is unloaded at the yard or an outbound container is loaded onto the tractor at the quay side. Once the tractors has completed the quay side operation, they either travel to the service point of the next operation or exit the quay depending on the movetype of the current job (Ct) and the next job (Nt) as described in Table 1. In this study, the activities that take place in the yard is not explicitly modeled in the simulation model, they are modeled by a black box process with a processing time that relates to the corresponding yard operation.

[1] http://www.brookssoftware.com/pages/245_automod_overview.cfm

Table 1. A summary of the container movements at quay side and yard side

Ct	Nt	Quay activities (of Ct)	Yard activities (of Nt)
PS	PS	Unload an outbound container(s)	Claim new outbound container(s)
PS	PL	Unload an outbound container(s)	No activity involved (IT directly travels to the service point of Nt)
PL	PL	Claim inbound container(s)	Unload the inbound container(s)
PL	PS	Claim an inbound container	Unload the inbound container and claim a new outbound container for Nt

3 Simulation Model

The proposed traffic model has the capability to adapt to different configurations in the berth. A text file (configuration file) is introduced to determine the actual configuration of the quay side through the setting of different parameters. The file is directly imported to the model before the simulation run. The configuration file allows the user to change the configuration of the quay side and other parameters easily in the simulation study.

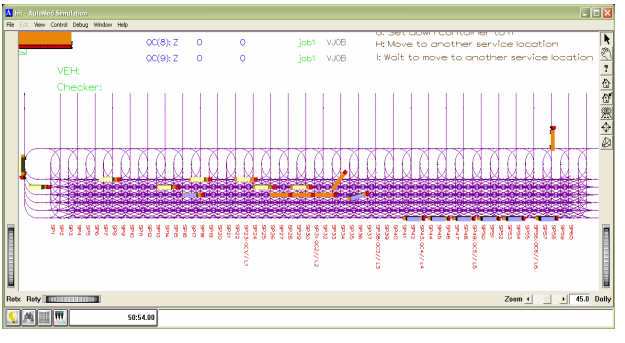


Fig. 1. Architecture of the traffic network implemented in AutoMod

3.1 Data Collection

Historical data of the selected berth is provided by the terminal operator. Data includes the cycle time of the quay crane of both land side and vessel side regarding to the quay operation. The estimated operation times at yard is identified by the appropriate statistical distributions.

3.2 Physical Layout

Path mover system in AutoMod is used to create a guide path, in which internal tractors move along these guide paths and carry a load between specific locations.

In the case study, the layout of the model (see Fig. 1.) is taken from a berth that represents one of the busiest terminals in Hong Kong. According to the actual configuration, six service lanes are built at the land side of the berth, which can be activated and deactivated by the users. Along each lane, there are number of service points which are matched with the bay number of the vessel. The layout design also provides the flexibility by mapping with different sizes of vessel (and with different configurations of the bays).

3.3 Control Logics

The proposed simulation model has twofold functionality – it controls the movement of internal tractors as well as the quay cranes as shown in Fig. 2.

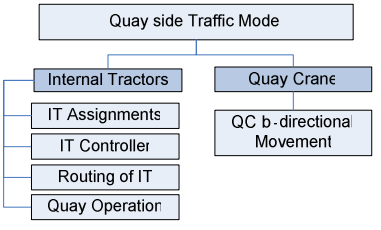


Fig. 2. An overview of all control logics applied in the Quay Side Traffic Model

Quay Side Traffic Control. Routing of the internal tractors (IT) is governed by the movetype of the current job (Ct) and the movetype of the next job (Nt). They are the key attributes of the internal tractors.

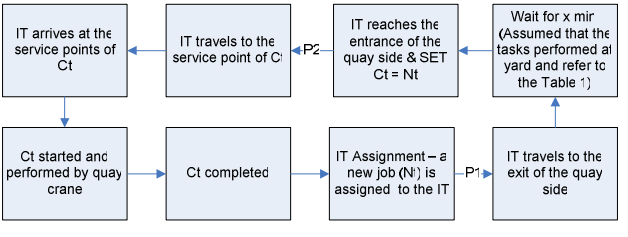


Fig. 3. Logistics flow of an internal tractors

Each tractor has two segments of path in order to complete a job. They are denoted as P1 and P2 as shown in Fig. 3. P1 refers to the tractors traveling from SP at quay to the exit point of the quay side, in which, the nearest exit point is selected. Once the tractor leaves the exit point, it is assumed that tractor performing the tasks includes (1) traveling from quay to yard, (2) performing yard operation and (3) traveling from yard to quay. Those activities will not be presented in the simulation model and stay for a short time in a queue known as "Yard". Time consumed at "Yard" is dependent on the activities defined in Table 1. The tractor then appears at the entrance of the

quay. P2 is the route from the entrance to the service point of Ct. In general, tractors are directly dispatched to the dedicated lane of the quay crane of Ct. The dispatching mechanism is defined based on the observation made at the actual quay side.

Quay Crane Bi-directional Movement. Movement is needed when the bay location is shifted. In other words, when the bay location of the current job is different from the bay location of the next job, a quay crane has to move to the appropriate location for completing the jobs. Quay cranes move bi-directionally in reality and this is designed in the traffic model. Two major features are controlled by the Quay Crane Bi-directional Movement Algorithm: (1) Movement of QC and (2) Collision detection. Collision detection refers to the conflicts between QC and the simulation is no longer being continued as illustrated in Fig. 4.

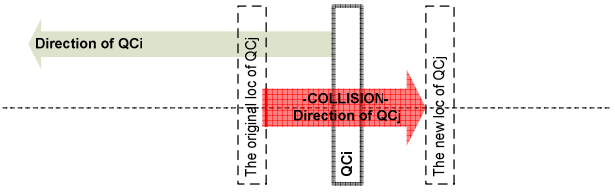


Fig. 4. QC conflict (Collision between QC)

The movement of quay crane is controlled by a series of What-If cases and is triggered by QCi which gets ready for taking the movement and triggers the algorithm. QCj is a quay crane which is being checked by QCi before such a movement.

The status of a quay crane is used to differentiate its behavior. Five statuses are introduced: (1) Reserved, (2) InService, (3) Pending, (4) Completed and (5) Moving. While the QC is in either of the status (1) – (4), the location of QCj is the main consideration before QCi moves as illustrated in Fig. 5. When QCi is in the Moving status, there are eight possible scenarios as summarized in Table 2. These eight scenarios are classified by the following attributes,

Di: Direction of the movement of QCi
Dj: Direction of the movement of QCj
Lj: Location of both Oj (Original Location of QCj) and Nj (Destination of the movement of QCj)

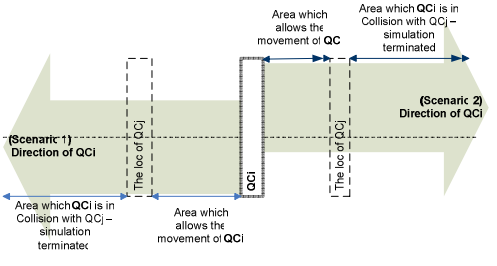


Fig. 5. Schematic diagram showing the influence of the QC_i when QCi is InService

Table 2. A summary of all possible scenarios when QCi is in the Moving status

Case	Di	Dj	Lj
M1	Left	Right	Right
M2	Left	Right	Left
M3	Left	Left	Right
M4	Left	Left	Left
M5	Right	Right	Right
M6	Right	Right	Left
M7	Right	Left	Right
M8	Right	Left	Left

Fig. 6 illustrates the decision logic of a quay crane, QCi when it meets a moving quay crane, QCj. In general, QC movement will proceed if either of the cases occurred: (1) The target location is not being blocked or (2) When the current location and the destination of QCj do not affect the movement of QCi. If there is any blockage of QCi path, internal tractors have to wait until the path is cleared and trigger the algorithm again until QCi is moved to the new location to continue its operation.

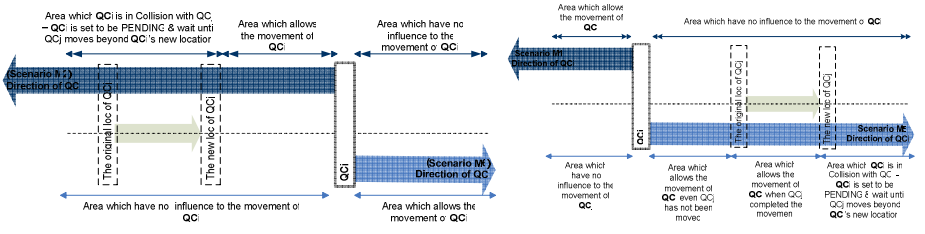


Fig. 6. A schematic diagram of scenarios M2, M6 (Left) and scenarios M1, M5 (Right)

When QCj is blocked by another QC, the path of QCj will further be checked to avoid any obstruction that blocks the movement of QCj and QCi.

3.4 Model Verification and Validation

The model is validated by comparing the actual QC Rate that is determined from the historical data and the rate obtained from the simulations. Other parameters are also investigated such as the dwell time of internal tractors. In addition to the verification of key performance indicator mentioned, the overall operation of the simulation model is validated with the terminal operators through detail review of the outputs, including the data and animation produced by AutoMod obtained from the simulation experiments.

4 Experiments Setup

Three experiments are undertaken to investigate the average QC Rate with respect to the changes of (1) Total number of internal tractors, (2) Traffic lanes deployed and (3) the introduction of an internal tractors controller. The configurations of the quay are listed in Table 3.

Table 3. Configuration of the quay (Experiment 1) – QCit = 7

Scenario	1	2	3
No. of IT deployed (VIT)	40	48	54

Table 4. Configuration of the quay (Experiment 2) – QCit = 12

Scenario	1	2	3	4	5
No. of IT deployed (VIT)	48	54	60	66	72

An actual Containers Work Program (CWP), which defines the sequence of container moves to be executed by a quay crane for a vessel provided by the port container terminal, was used in the simulation study. Six quay cranes are deployed and the simulation results were used for model validation and for the evaluation of the actual cases.

It is assumed that the number of IT assigned to each QC (QCit) is fixed and cannot exceed seven tractors in the simulation to prevent traffic congestion that is caused by the overcrowding of the tractors on particular lanes and avoid uneven allocation of tractors to each QC. Thus VIT is limited by QCit. Under this constraint, in scenario 3, forty-two tractors that is maximum number of IT deployed, are assigned for the vessel. In order to increase VIT, QCit have to increase accordingly as shown in Table 4.

For the study of traffic lanes deployment, the scenarios listed in Table 3 will be adopted and the number of traffic lanes closure (Lc) is varied from zero to three. If Lc is set to one, the lane near the sea side will be selected first and so QCs are then assigned to the corresponding lanes sequentially, i.e., the first QC is assigned to the first lane and the second to the next available lane until all quay cranes are being assigned. As a result, more than one quay crane will be assigned to a lane. There are six replications for each configuration.

5 Results

There are number of performance indicators often adopted to quantify the efficiency and productivity of a berth in a port container terminal. Amongst these indicators, the major performance indicator adopted in our study is the Quay Crane Rate (QC Rate). Other indicators that are used include the waiting time and queue length of QC as well as the waiting time and transit time of IT.

The study is mainly focused on the effect of the Average QC Rate under different configurations of the berth. This is defined as the number of boxes handled per hour by each QC. Average QC Rate is defined by:

$$\sum_{i=1}^{n}(B_{QCi})/\sum_{i=1}^{n}(T_{QCi}) \tag{1}$$

5.1 Total Number of Internal Tractors

From the simulation study (Experiment 1), it is found that the total number of internal tractors deployed for a vessel has a great impact to the QC Rate as shown in Fig. 7.

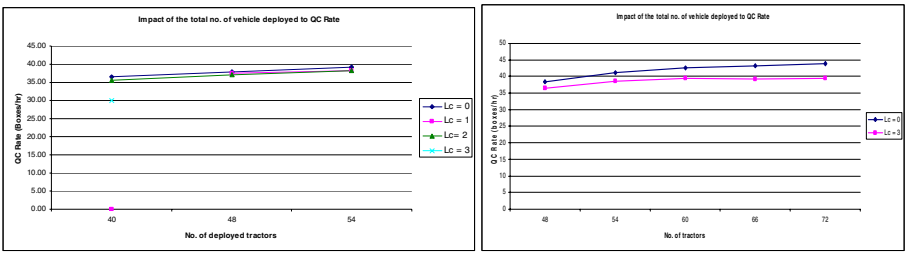


Fig. 7. Impact of the total number of internal tractors and the number of lanes closure to the QC Rate Experiment 1(Left) and Experiment 2 (Right)

The QC Rate is increased when the number of tractors serving the quay crane is increased. A big improvement is seen when there are eight additional tractors serving the vessel (Scenario 2). When six additional tractors are added (Scenario 3), QC Rate is approaching saturation. The analysis revealed that about two to three additional tractors deployed contributes to one more container handled hourly. As more tractors are deployed, the optimal QC Rate is observed (Fig. 7). QC Rate is becoming saturated as its upper limit of the cycle time of quay crane is reached. It shows that

- QC Rate will not be unlimitedly increased.
- An optimal configuration of berth is obtained when QCit = 12 and VIT = 72, the achieved QC rate is approximately 44 boxes/hr.
- The QC Rate is expected to growth by 2-3% when more than sixty tractors deployed for the vessel, which is comparatively lower than about 12% gained in QC Rate when less tractor deployed, say fifty (Fig. 7. (right)).

5.2 Traffic Lanes Deployment

IT Lanes closure causes a reduction in the QC Rate as shown in Fig. 7, about 2 per cent less (<= 1 box) in the QC Rate when a lane is closed. When three or more lanes are closed with more than one QC has to share a service lane, the simulation run is halted due to congestions at the quay side. Internal tractors accumulate in the service lanes due to inter-blockage and conflicts between QC are found. Here, QC conflicts refer to the scenario that QCi and QCj want to move in opposite direction as shown in Fig. 4. In addition, studies on controlling the number of tractors that travels to the berth are performed that aims at minimizing the conflicts between tractors. The results are presented in the following section.

5.3 Internal Tractor Controller Deployment

An IT Controller is designed to manage the flow of tractors near the entrance of the quay and to limit the number of tractors traveling to the berth (CTt). This is a concept

adapted from real operation where a physical traffic controller is deployed to direct the tractors to the assigned lanes. From the simulation study, it is found that by limiting the number of tractors traveling to the berth does not improve the QC Rate (See Fig. 8). Specifically, the following points can be concluded:

- No congestion happened in the previous cases and limiting the number of tractors is not necessary.
- Reducing CTt does not improve the traffic congestion in some cases due to some tractors blocking the paths of others while waiting for loading/discharging of containers.
- Congestion is induced by other traffic problems such as conflicts between QCs.

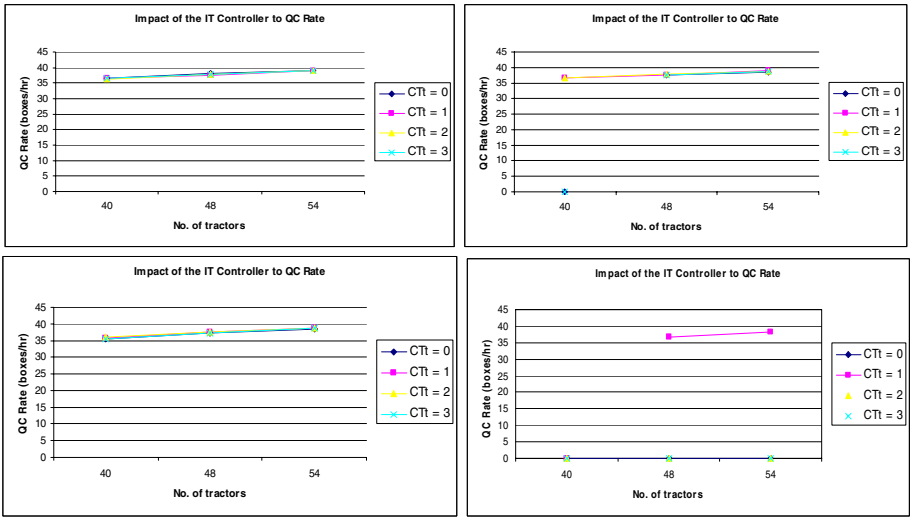


Fig. 8. Impact to the QC Rate after deployed the IT Controller when Lc = 0(top left), Lc= 1(top right), Lc = 2 (bottom left) and Lc = 3 (bottom right) (Experiment 1)

In this experiment, although QC Rate does not have a substantial improvement, congestion has been resolved when an IT Controller is deployed. Table 5 summarizes the occurrence of congestion when Lc is varied. It is seen that congestions are eliminated by deploying an IT Controller.

Table 5. Statistical findings on congestion problem in Experiment 1

Lc	No. of congestion	Congestion resolved
0	0	0
1	1	1
2	0	0
3	3	2

It is found that 75% of congestion problem has been resolved by controlling the internal tractors travels to the berth at the entrance.

6 Conclusions

To conclude our funding from the undertaking an integrated modeling approach and through extensive simulation experiments, the total number of internal tractors deployed for a vessel has an important impact to the QC Rate. The QC performance improves as the number of tractors increases (about 2-3 boxes to be handled hourly if about 8 more tractors serving the vessel) and gradually reaches a saturation point. On the other hand, a reduction of approximately 2% (<= 1 box) of the QC Rate when a lane is closed during quay side operation is observed. On the other hand, with the introduction of an IT Controller, no significant improvement to the QC Rate is observed, though the problem of IT congestion can been eliminated in some of the cases.

An optimal configuration of berth is picked from the integrated study of different configurations. The optimal QC Rate of 44 boxes/hr is achieved. This configuration also leads to a minimum berthing time of the vessel.

With the incorporation of that the traffic model into a traditional service model, other features of the traffic model can be further enhanced. For instance, the mechanism of overtaking which is performed quite often in a real container terminal can be added to the existing model. Given the capability of the model is being proven with detailed traffic requirements defined, our quay side traffic simulation model can become a generic traffic model that can be used to predict the traffic conditions in other part of the terminal including yard and gate.

References

1. Ballis, A., Abacoumkin: A container terminal simulation model with animation capabilities. Journal of Advanced Transportation 30(1), 37–57 (1996)
2. Bielli, M., Boulmakoul, A., Rida, M.: Object oriented model for container terminal distributed simulation. European Journal of Operational Research 175, 1731–1751 (2006)
3. Bulter, R., Maddalon, J., Geser, A., Munoz, C.: Formal analysis of air traffic management systems: the case of conflict resolution and recovery. In: Proceeding of Winter Simulation Conference, pp. 906–914 (2003)
4. Hugan Joseph, C.: Using simulation to evaluate site traffic at an automobile truck plant. In: Proceedings of Winter Simulation Conference, pp. 1013–1016 (2001)
5. Legato, P., Mazza, R.M.: Berth planning and resources optimization at a container terminal via discrete event simulation. European Journal of Operational Research 133, 537–547 (2001)
6. Nam, K.C., Kwak, K.S., Yu, M.S.: Simulation study of container terminal performance. Journal of Waterway, Port, coastal and Ocean Engineering, 126–132 (2002)
7. Michael, N.: A recursive method for traffic management through a complex path network. In: Proceeding of Winter Simulation Conference, pp. 537–541 (2002)
8. Sgourids, S.P., Makris, D., Angelides, D.C.: Simulation analysis for midterm yard planning in container terminal. Journal of Waterway, Coastal and Ocean Engineering, 178–187 (2003)
9. Yun, W.Y., Choi, Y.S.: A simulation model for container-terminal operation analysis using an object-oriented approach. Int. J. of Production Economics 59, 221–230 (1999)

A Framework for Verification and Validation of Simulation Models and Applications

Zhongshi Wang[1] and Axel Lehmann[2]

[1] Institut für Technik Intelligenter Systeme (ITIS e.V.)
[2] Institut für Technische Informatik
Fakultät für Informatik
Universität der Bundeswehr München
D-85577 Neubiberg, Germany
{zhongshi.wang,axel.lehmann}@unibw.de

Abstract. The permanently increasing performance of information and communication technologies enables the development and effective application of increasingly complex simulation models and their applications (M&S). As this trend will be an ongoing one, the demand for controlling and demonstrating the quality of a model and its applications by introducing adequate quality measures, techniques, and tools is obvious. Consequently, the introduction of accepted or even standardized verification and validation processes is an urgent requirement in this context. This paper will present an Extended M&S Verification and Validation (EV&V) framework based on multi-phase M&S-design, -development and –application processes. The basics of EV&V, related processes, and products will be discussed. In addition, we will show how this EV&V framework can be implemented and tailored according to predefined credibility or risk criteria, and under specified cost, time and application constraints.

Keywords: Modelling and Simulation (M&S), Verification and Validation (V&V), V&V Framework, Extended V&V Triangle (EV&V).

1 Introduction

Enabled by rapid advances of computer and network technologies, the development and application of increasingly complex simulation models (M&S) in various domains is inevitable. But with increasingly powerful visualization and simulation infrastructures and tools, quality assurance – especially correctness and validity of models and of modelling results - becomes a major concern. As a result of some well-known safety disasters or financial losses caused by modelling faults or failures, efforts towards model verification and validation (V&V) are getting more and more attention. Therefore, a consequent and accepted methodology, or framework, for M&S-verification and -validation (V&V) are still missing.

A proof of correctness and validity of a model and its experimental applications will be – in general – not feasible. Even mastering the quality of complex monolithic models (with respect to the number of input, state, and output parameters) is almost not feasible and would require tremendous efforts and costs for “testing”. Taking also

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 237–246, 2007.

into account requirements regarding time-constraints, availability of data and knowledge, current trends in M&S development include:

- **Component-based model design:** The goal of this approach is focusing on the reusability of model components (or submodels accessible through (application domain specific) model libraries. The reuse or refinement of well documented, verified and validated components decreases significantly efforts and costs for model development and quality assurance of composed models.
- **Cooperative modelling:** Model design and development is - in general - a team effort requiring time- and location-independent cooperation among team members having different competences, expertise and responsibilities (roles).
- **Parallel and distributed model execution:** Regarding real-time constraints or effective execution of federations of models, the distribution of model federates over different computing nodes can offer efficient solutions. Also with respect to reliability of modelling experiments, distributed or parallel execution of redundant models can be an important issue.
- **Visualization of simulation experiments:** Enhanced graphics, virtual and augmented reality development tools offer various powerful capabilities for animation and new application scenarios of simulation.

Considering rigorous analysis of divers existing concepts, techniques and tools for verification and validation (V&V) of simulation models and their experimental applications, we have summarized major requirements for V&V, and defined a structured V&V approach [5] also referred to as "V&V Triangle" [4], or as "Extended V&V Triangle"(EV&V) [17]. Based on some lessons learned from software engineering [7, 16], such as the German IT standard development process V-Modell XT [15], we propose a EV&V framework in this paper to facilitate the credibility assessment of M&S applications. The Extended V&V Triangle can be then directly integrated into an operational framework used in practice.

The remainder of this paper is organized as follows: Section 2 summarizes the essential requirements for successful model V&V. In Section 3, the major elements of the V&V framework are introduced. Further in Section 4, the Extended V&V Triangle is outlined. In addition, an advanced tailoring concept applied in the framework is presented. Finally, Section 5 concludes the contributions of this work to the field of M&S verification and validation.

2 Requirements for Successful Model Verification and Validation

Conducting M&S verification and validation is a complex process which must be integrated into the entire life cycle of a simulation project. Figure 1 shows a structured process used to create work products (intermediate products) during the M&S development life cycle. It indicates which progress stages of model building ought to be reached and what results will be yielded at the end of each process phase.

To achieve the defined intermediate products, an enormous amount of information must be gathered, analyzed, and modeled necessary for model development and its experiment application in terms of qualitative and quantitative data. This task is

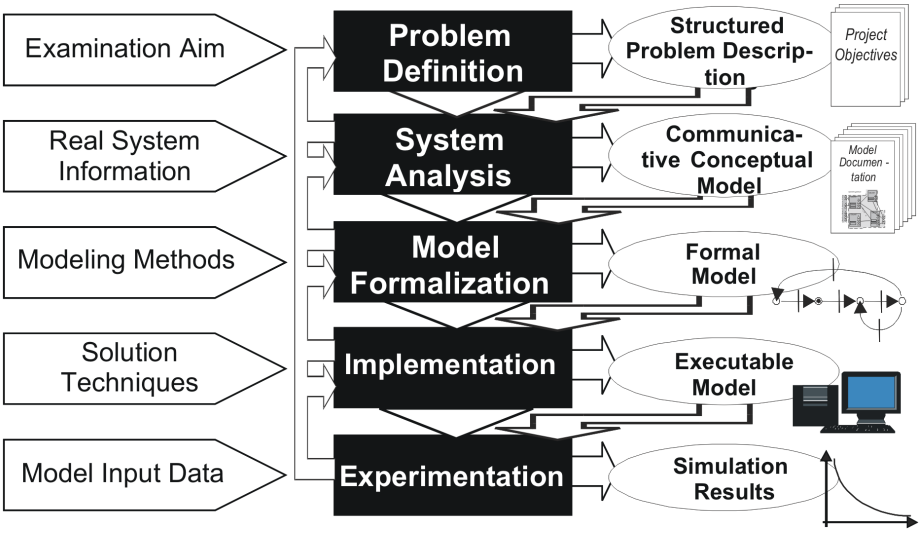


Fig. 1. A model development process [9]

referred to as data modelling. Data modelling is of prime importance for conducting a simulation study. Up to one third of the total model development time can be spent on it [14].

Since the quality of input data is also a crucial factor for the credibility assessment of M&S applications, correctness and accuracy of data acquisition, data analysis, data transformation, and data use must be estimated in accordance with model V&V for each model development phase. More detailed discussions about topics of data modelling can be found in the literature [3, 6, 8, 13, 17].

The design and development process of M&S directly affects the quality of the work (intermediate) products being created. Applying a well structured strategy based on the actual technologies discussed above certainly increases the process quality, which in turn increases the product quality, and consequently, the V&V efficiency as well. As depicted in Figure 2, we summarize the following essential requirements to be implemented for conducting successful verification and validation of M&S applications:

- Modelling is a well-defined, product-oriented M&S-process;
- An M&S-process is performed by personnel from the project sponsor and from the M&S-developer having different, but well-defined responsibilities (roles);
- Each process phase, including its required input data, is summarized by one or several (intermediate) products which are subject of V&V;
- V&V activities and results have to be documented, thus forming also V&V-products;
- Quality assurance builds a chain of evidence based on V&V products.

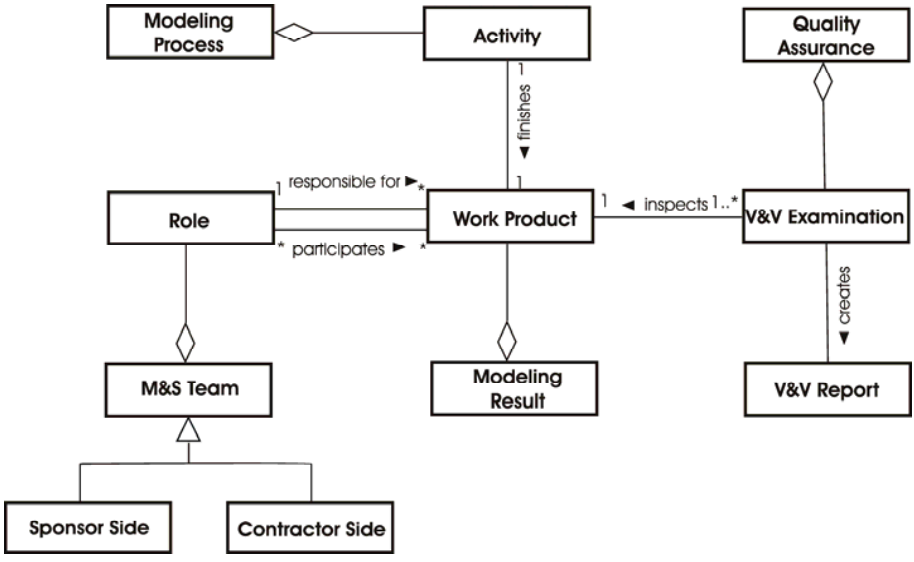


Fig. 2. Baseline for successful model V&V

3 Proposed V&V Framework

Figure 3 gives an overview of the proposed V&V framework intended to facilitate the successful credibility assessment of M&S applications in practice. Three major components for conducting a simulation study are a process and product concept, an associated role concept, and a quality assurance concept, are coupled together in a cooperative form.

The process and product concept adopted here is target- and result-oriented, that means the (intermediate) products are the main project results. To create products, the proper activities have to be well defined and structured. In addition, this concept extends the traditional approach by adding an extra viewpoint of a client (sponsor or user of a simulation application) with regard to model development, so that the particular activities to be performed and the intermediate products to be created can be clearly specified for a client and a M&S developer.

The introduction of a role concept is intended to define the responsibility of product creation. Within the scope of a M&S application a person or organizational unit is assigned to a role, according to the accordant competency. Although many participating roles can support a product's creation concurrently, only one role accepts being responsible for the product.

Concerning the aspect of quality assurance, the model development process is marked by several so-called decision points, which indicate that different progress stages of model initialized, or model designed, or model realized, or model applied have been achieved during a simulation study. Each decision point defines a set of V&V activities, which can be applied to evaluate the intermediate product(s) created during the current model progress stage.

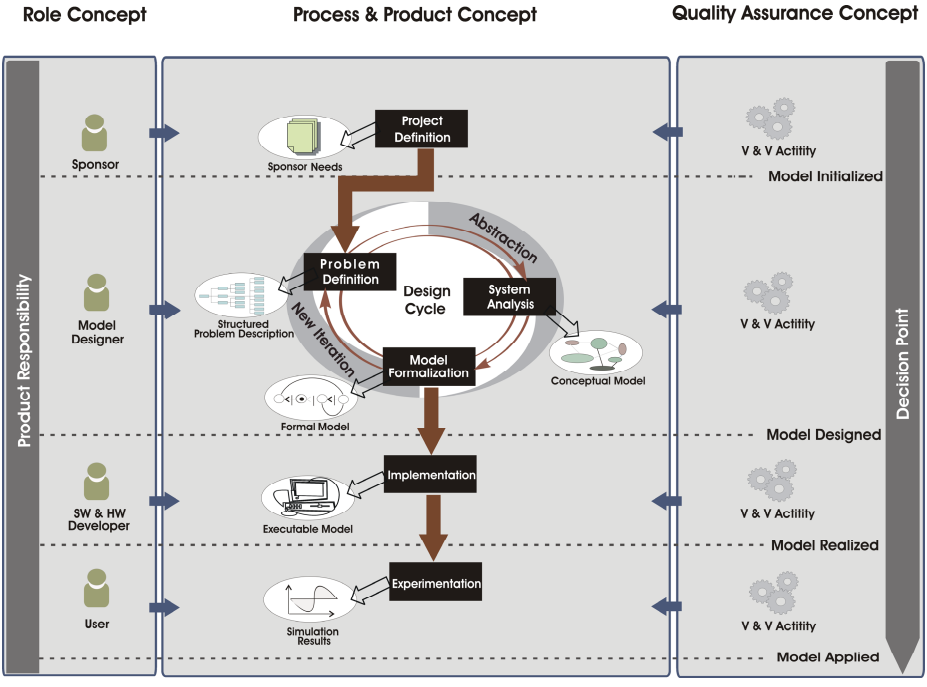


Fig. 3. Structure of the V&V framework

Unless the V&V process is completed for assessing the quality of each intermediate product, the further progress stage or new development iteration will not be continued. In this manner, all intermediate products of model development from the description of sponsor needs to the interpretation of simulation results can be evaluated respectively, and thus, assessing the product quality of an overall completed M&S application can be achieved consequently. This V&V concept is a typical product-centered assessment approach [2].

In order to facilitate the application of the V&V framework in practice, detailed documentation requirements of model development and V&V activities need to be defined, so that the created intermediate products, the responsibilities of different project roles, the V&V plan, and the V&V reports can be well-structured documented.

4 A Quality Assurance Process – The Extended V&V Triangle

In this section, we propose a V&V process which implements the quality assurance concept described by the V&V framework above. This approach extends the "classic" V&V process, the so-called V&V Triangle [4], by appending an extra consideration of data modelling [1, 11, 12], and therefore, is also referred to as "Extended V&V Triangle" presented in [17].

4.1 Overview of The Extended V&V Triangle

As Figure 4 illustrates, the Extended V&V Triangle (EV&V) consists of two closely associated parts of V&V activities: model V&V and data V&V.

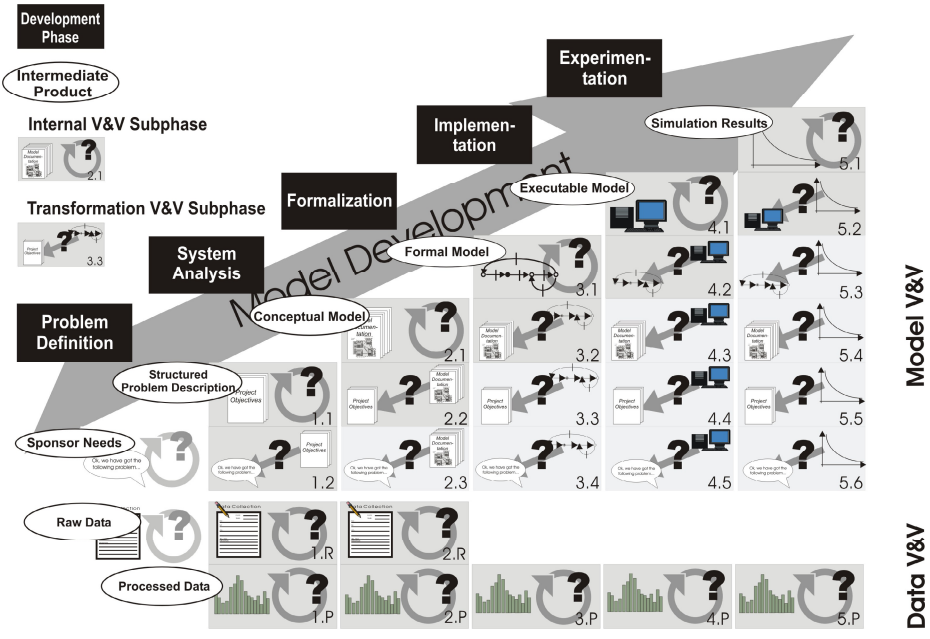


Fig. 4. The Extended V&V Triangle (EV&V)

In model V&V, the quality of the product sponsor needs obtained from the project initialization phase must be first evaluated. Thereby, the examination of correctness in form and internal consistency for this product is conducted. Any other intermediate product created during model development from "structured problem description" to "model results" is input to a V&V phase, numbered 1 through 5. Each V&V phase is again split into sub-phases, each with a defined sub-aim to detect the internal errors or transformation errors. In each sub-phase numbered as x.1, the absence of internal errors in each particular intermediate product should be demonstrated. For example, in sub-phase 1.1 it should be ensured that the problem description is free of misunderstandings and inconsistencies, and in sub-phase 3.1, a syntax check can be applied to the formal model for comparison of the chosen formalism. In any other sub-phase, the pairwise comparison between the current intermediate product and each previous intermediate product can be performed to confirm the absence of transformation errors. Such as in sub-phases 3.2, 3.3, and 3.4, the formal model could be compared with the conceptual model, the structured problem description, and the sponsor needs, one by one. By repeating the comparisons between intermediate products in this manner, an increasing intensity of V&V activities is achieved, and consequently, the credibility established so far can also be increased accordingly.

With respect to data V&V, each V&V phase is extended by additional V&V activities of data derived from the corresponding data modelling process. Two types of data in the data flow throughout model development should be distinguished: raw data and processed data. Raw data are obtained directly from different sources, which are – in general - unstructured and unformed data. Although historical (or empirical) data inherited from previous simulation applications and usually available at the beginning of a simulation study, could be available in a well-structured input format, but have to be considered in this regard also as raw data. It is by any means necessary to ensure prior to the use of such data whether these data are actually reasonable in the current context. Processed data are, however, created by editing, transforming, or adapting raw data during the modelling process. Thus, data V&V involves credibility assessment of raw data and of processed data used for creating each intermediate product. It should be noted that raw data are usually only relevant for specifying structured problem description and a conceptual model. However, it is not be directly applicable for creating formal models and other succeeding intermediate products. During V&V of raw data concerning each intermediate product, the following issues must be ensured:

- data sources have to be reasonable for the intended purpose of the model. It is especially critical if data are derived from a different system;
- quantitative as well as qualitative data have been measured precisely enough;
- the amount of data gathered is adequate for the further investigation;
- the context in which raw data have to match their application objectives.

V&V of processed data focus on ensuring that all data used for each intermediate product are correctly and accurately adapted or transformed, which requires:

- to evaluate the assumptions of independence and homogeneity made on the gathered data sets;
- to ensure that the probability distribution used and the associated parameters are reasonable for the data collection, for example, by using a goodness-of-fit test;
- to determine that data have been transformed in required form and structure;
- to ensure that enough independent simulation runs have been executed for a stochastic model

For documentation purposes, the results of data V&V of all phases are added to the V&V plan and V&V report, according to the well-defined document structures.

4.2 Tailoring Concept

As mentioned above, executing V&V activities in V&V sub-phases redundantly, such as comparing the formal model individually with the conceptual model, the structured problem description and the sponsor needs, takes several viewpoints of model assessment into consideration, so that the concluded V&V results should increase credibility. However, for many simulation applications in practice it is impossible to fully conduct all V&V activities recommended in the V&V Triangle due to time and budge constraints. In such cases, a slim set of V&V activities, which conducts the credibility assessment still at a certain acceptable confidence level in spite of the limitations, should be applied by tailoring.

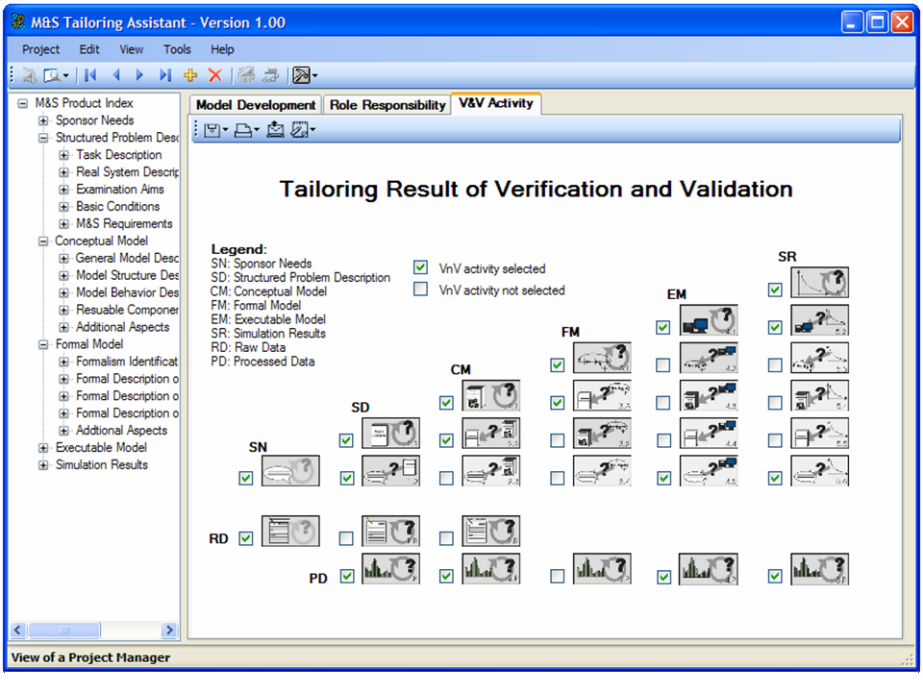


Fig. 5. M&S Tailoring Assistant

Based on the detailed documentation requirements of M&S and VV&A [10], in which project constraints, intermediate products and their dependencies, VV&A acceptance criteria, V&V activities, and project roles are well defined and specified, a multistage tailoring process is proposed not only for model development, but also for conducting VV&A, including the following stages:

1. Tailoring at the process level,
2. Tailoring at the product level, and
3. Tailoring at the role level.

At the beginning of a M&S project, tailoring of the model development process applied in the simulation study is of prime importance for preparing a M&S and a VV&A plan. According to project constraints, essential project-specific products (intermediate products or parts of them) and the associated V&V activities can be identified, and less important products are therefore ignored. For example, if determined that the formal model is not essential for the current project, the related products of the formal model and its defined V&V activities remain out of consideration. Considering specified product dependencies, project adaptation at the product level can be conducted. This is the reason that during the model development further products to be developed may be selected, whereas the products existing in the M&S plan may be removed because the obligations between the products are identified. Moreover, in a third step, tailoring can be conducted at the role level. It means each project role has only access according to its authority issued for this M&S project.

Based on this tailoring concept, a supporting tool, M&S Tailoring Assistant, is prototypically implemented. Figure 5 shows an example of tailoring V&V activities for an M&S application from a project manager's point of view.

5 Conclusion

Considering current technologies applicable for model development, this paper introduces a V&V framework to facilitate the quality assurance efforts for M&S design, development and applications. By incorporating the fundamental concepts of "process", "product", "role", and "quality assessment", this approach offers a platform for each project participant to create products and conduct model V&V in a cooperative manner. Any well-structured and product-oriented model development process can be directly applied in this framework. In addition, we present a quality assurance process - the Extended V&V Triangle (EV&V), which specifies detailed V&V activities for quality assessment of M&S applications, including model V&V and data V&V. Within the scope of project planning, a multistage tailoring concept for the purpose of project-specific adaptation is introduced. By arranging the tailoring efforts respectively at the levels of process, product and role, this generic tailoring concept offers a high degree of flexibility and feasibility for conducting M&S and VV&A under different constraints.

Further Research will concentrate on (1) applying the proposed framework and V&V approach in a real M&S project to make experiences and learn lessons from practical use; (2) extending the tool support for planning and conducting V&V activities. First, M&S Tailoring Assistant will be extended. In addition, a collaboration tool will be developed to support cooperative task execution assigned to each project member for model design, development and for quality assurance.

References

1. Balci, O.: Verification, Validation and Testing. In: Banks, J. (ed.) The Handbook of Simulation, ch. 10, John Wiley & Sons, Chichester (1998)
2. Balci, O.: Quality Assessment, Verification, and Validation of Modeling and Simulation Applications. In: Proceedings of the 2004 Winter Simulation Conference (2004)
3. Banks, J., Carson II, J., Nelson, B., Nicol, D.: Discrete-Event System Simulation, 4th edn. Pearson Education International (2005)
4. Brade, D.: Enhancing Modeling and Simulation Accreditation by Structuring Verification and Validation Results. In: Proceeding of the 2000 Winter Simulation Conference (2000)
5. Brade, D., Lehmann, A.: Model Validation and Verification. In: Modeling and Simulation Environment for Satellite and Terrestrial Communication Networks. In: Proceedings of the European COST Telecommunication Symposium, Kluwer Academic Publishers, Bosten, USA (2002)
6. Hoover, S., Perry, R.: Simulation: a problem-solving approach. Addison-Wesley Publishing Company, Reading (1989)
7. Kuhrmann, M., Niebuhr, D., Rausch, A.: Application of the V-Modell XT – Report from a Pilot Project. In: Li, M., Boehm, B., Osterweil, L. (eds.) Unifying the Software Process Spectrum, International Software Process Workshop, pp. 463–473. Springer, Heidelberg (2005)

8. Law, A., Kelton, W.: Simulation Modelling and Analysis, 2nd edn. McGraw-Hill, New York (1991)
9. Lehmann, A., Lüthi, J., Berchtold, C., Barde, D., Köster, A.: Zukunftsfelder der Modellbildung und Simulation. Abschlussbericht (in German). ITIS e.V. (2000)
10. Lehmann, A., Saad, S., Best, M., Köster, A., Pohl, S., Qian, J., Walder, C., Wang, Z., Xu, Z.: Leitfaden für Modelldokumentation. Abschlussbericht (in German). ITIS e.V. (2005)
11. Rabe, M., Spieckermann, S., Wenzel, S. (Hrsg.): Verifikation und Validierung für die Simulation in Produktion und Logistik (in German). Springer, Berlin (2007)
12. Spieckermann, A., Lehmann, A., Rabe, M.: Verifikation und Validierung: Überlegungen zu einer integrierten Vorgehensweise. In: Mertins, K., Rabe, M.(Hrsg.) Experiences from the Future (in German), Fraunhofer IRB, Stuttgart, pp. 263–274 (2004)
13. Shannon, R.: Systems Simulation: the Art and Science. Prentice-Hall, Englewood Cliffs (1975)
14. Shannon, R.: Introduction to the Art and Science of Simulation. In: Proceedings of the 1998 Winter Simulation Conference (1998)
15. The V-Modell XT Portal http://www.v-modell-xt.de
16. Wang, Z.: Eine Ergänzung des V-Modell XT zum Einsatz in Projekten der Modellbildung und Simulation. In: Beiträge der 35. Jahrestagung der Gesellschaft für Informatik (in German), GI-Edition. Lecture Notes in Informatics, Köllen Verlag, Bonn (2005)
17. Wang, Z., Lehmann, A.: Verification and Validation of Simulation Models and Application: A Methodological Approach. In: Ince, N., Bragg, A. (eds.) Recent Advances in Modeling and Simulation Tools for Communication Networks and Services, Springer, New York (2007)

A Decision Processing Algorithm for CDC Location Under Minimum Cost SCM Network*

N.K. Park, J.Y. Kim, W.Y. Choi, Z.M. Tian, and D.J. Kim

Distribution Management Department
Tongmyung University,
535, Yongdang-dong, nam-gu, Busan, Korea
nkpark@tu.ac.kr

Abstract. Location of CDC in the matter of network on Supply Chain is becoming on the high concern these days. Present status of methods on CDC has been mainly based on the calculation manually by the spread sheet to achieve the goal of minimum logistics cost. This study is focused on the development of new processing algorithm to overcome the limit of present methods, and examination of the propriety of this algorithm by case study. The algorithm suggested by this study is based on the principle of optimization on the directive GRAPH of SCM model and suggest the algorithm utilizing the traditionally introduced MST, shortest paths finding methods, etc. By the aftermath of this study, it helps to assess suitability of the present on-going SCM network and could be the criterion on the decision-making process for the optimal SCM network building-up for the demand prospect in the future.

1 Introduction

Accelerating headway of multi-national corporations powered by the market globalization and economic growth in Northeast Asia, it is no wonder anymore that the efficient logistics network building-up is on the core demand together with the securing of the international logistic center.

To decide where to locate their logistic center, it is preliminarily required to calculate for the costs concerned case by case. Here we suggest network cost modeling system which can deduce a method to calculate in the view of total logistic cost and referred to the data for the international trading collected from the "L" company to see the propriety of this model.

It consists of basic study, problem configuration, model building-up, solution providing, model examination and case study.2 Basic Study

1.1 Literature Review

There have been various kinds of approaches to logistic network solution. Takakuwa(2000) used a simulation method for the role of logistic center located in

* This research was supported by MIC(Ministry of Information and Communication) Korea, under the ITRC(Information Technology Research Center) Support program supervised by the IITA(Institute of Information Technology Advancement)" (IITA-2006-C1090-0602-0013).

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 247–258, 2007.

the distribution area and the factory. But this simulation method was found not appropriate for this kind of huge scaled matter due to the complexity while still available to define the matter. On the study of supply chain management and the simulation by Prof. S.J.Seo & K.S.Kim(2000), the level of logistic network management of a corporation could be classified as operating level, tactic level and strategic level. Furthermore in the point of complexity, importance, period of the matter, this simulation method would be the most appropriate for strategic intermediate and long-range subject. Jansen(2001) and others studied on the change of costs in accordance with the scenario after classifying all the products such as revaluation index, financial index, logistic index under the special conditions of catering supply chain. On the study of Bandinelli(2006), in the planning of SCM, simulation modeling & analysis is very efficient tools for the supply chain performance analysis. On Jahan Heinrich von thunen(1996), when considering facility location, transportation cost should be on priority. Under the same condition of the commodity price and unit price, the place of candidate logistic center should be considered in respect of minimum cost. On S.H.Kim, Y.H.Lee (2000), which studied conclusion model, examination model, statistics model and simulation model, found simulation model to be the most efficient method to solve the problem. Greenhut(1956) put the focus on the profit instead of transportation cost on the decision of network matters. The optimistic logistic place is located to produce maximum profit. In this regard, the minimum-cost location might not be the most suitable place because the demand and the price could be different according to the place. On the study of Fraser(2003), it showed that 500 corporations made profits through the planning of the supply chain. BAH consulting company(1987) simplified the logistic decision-making process by suggesting PILOT-SCOPE model. By this model, 4 stages of facility requirement recognition, strategic alternatives development and the assessment, decision on location and scale are followed. It could be simplified by the facility scheme model and operating scheme model. Fisher(1982), decision-making process could be simplified by classifying logistic matter into strategic matter and tactic matter in the case of Du Pont. Strategic matter is the decision-making process on the local logistic center moving, enlargement and establishment. Tactic matter is the efficient adjustment of distribution route by the inducement of new customers. Fisher utilized ROVER(Real-Time Optimizer for Vehicle Routing), developed by Pennsylvania Univ. to solve Du Pont's logistic problem. Kekre(1990) solved the real problem by building up the transportation model for Heinz corporation. Decision-making supporting model jointly developed with Carnegie Mellon Univ. was utilized for this transportation matter. This program supports various transportation alternatives and helps to develop optimistic operating plan on each logistic foothold including calculation of constant cost and variable cost.3 Problem Configuration

1.2 Objective Function

It includes CDC of export port of loading and import port. Logistics cost consists of the costs at nodes and transportation cost and the core purpose of this study is to find out the CDC and logistics network suitable for the minimum total cost. Optimal CDC is the port of call in the import area on the point of the logistic network of the minimum cost, and the formula concerned to compute out minimum cost of logistics cost is shown as follows.

$$\min(l\sum_{i=1}^{m} NCi + LCos + k\sum_{i,j=1}^{m} LCij) \tag{1}$$

NCi : cost of node i
(inventory cost, administration cost, depreciation cost, discharging cost)
LCij : transportation cost including backward node's demand from node i to j
LCos : transportation cost from supplier loading port to CDC candidate
Subject to :

$$l, k = 0 \, or \, 1$$

$$\sum_{i=2}^{n} rdcDi \leq cdcDi$$

$$NCi \geq 0$$

$$LCi \geq 0$$

1.3 Model Building-Up

To find out minimum cost CDC and network, here below shows the simple graph model. Network consists of 1 supply area, 4 CDC candidate areas and 2 RDC and served by the sea transportation from the supply area to CDC and road haulage from CDC to RDC. <figure 1>.

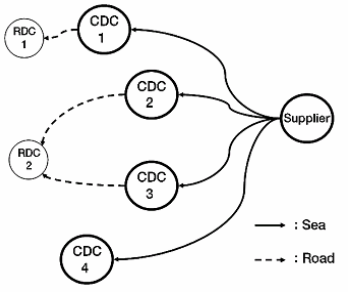


Fig. 1. Graph modeling as SCM network

2 Solution Providing

2.1 Traditional Methods Investigation

MST(Minimum cost Spanning Tree) method
It is the solution for the matter of calling all nodes in the network at the minimum cost, and MST is well known method. Although there are interesting algorithms of Kruskal and Prim, these 2 methods are applicable in the case of non-direction graph which isn't suitable for our study. Directive link should be considered to support our study, it is on account of cost difference according to the direction.

Floyd Algorithm
This method refers to the solution to find minimum spanning on the network. Dijkstra and Floyd methods are known well. Dijkstra method applies to the minimum distance solution from a node to a designated node and it skips some nodes which isn't applicable for this study. Floyd algorithm applies to the solution to find the minimum spanning for all the nodes except a node itself on the aggravated graph. It could confined only to show the aggravated route from a designated node to a node but that is not enough to compare all the costs on the network.

2.2 CDC Searching Algorithm

To overcome the limit of Floyd algorithm, this study suggests as follows as seen on below diagram.

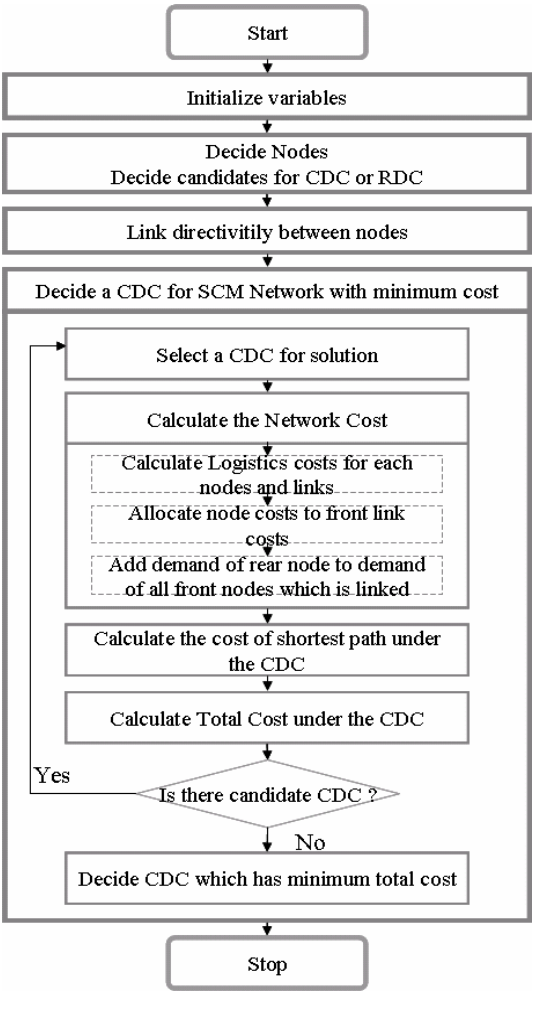


Fig. 2. Flow Chart to process solution

Variables Initializing
Build up basic data such as demands on each node, transportation rate, rental cost.

Network Build-up
Setting up the directive links and CDC candidate locations after allocating the nodes.

CDC Selection and Network Cost Calculation
Select one among CDC candidates and calculate logistics network cost based on the selected CDC. Logistic costs are divided into constant cost, stock cost & transportation cost and then constant cost and stock cost are added to the transportation cost to transfer into the link cost.(refer to phrase 4.3). Calculate the minimum network cost serving all the nodes on the network. After calculating total costs for all the candidate CDCs, conclude CDC Network which shows the minimum total logistic cost by the logistic network composed of minimum cost CDC.<figure 2>

2.3 Definition of Node, Link Cost and Transferring Method

To transfer node cost to link cost, it is necessary to define composition of node cost first. Node cost consists of constant cost, overheads, storage and stevedorage. (refer to figure 3). It is important to insert demand of rear node into front node, for example, when calculating ocean freight T1, T1 should be calculated by summing up the demand figure of node1, node2, node3 respectively. In case of node positioned in rear, T2 and T3 are applied in the same way.

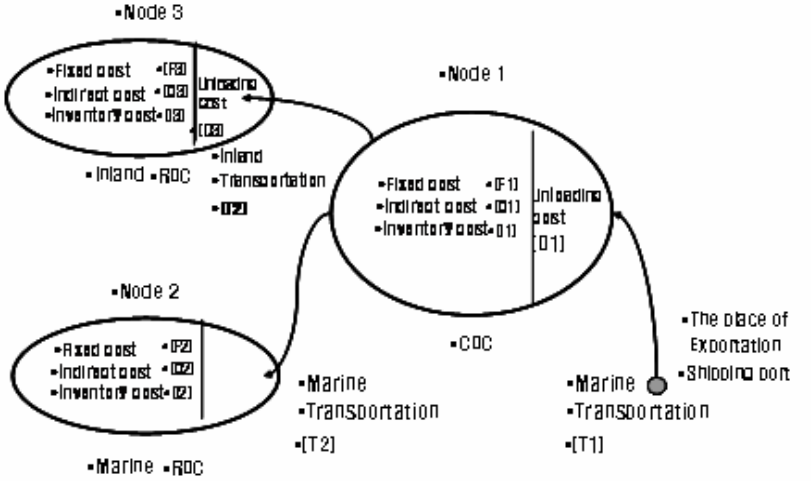


Fig. 3. Cost structure of nodes & links

It would be helpful to understand by introducing example of cost calculation. For example, if we put Cbs as a total cost from a loading port to CDC, Cbs=T1 +D1+ (F1+ O1+I1). If routed from the loading port to node2 via node1, total cost to node2, Cbh=Cbs+(T2+D2)+(F2+O2+I2).

3 Model Verification

3.1 Definition of Input Data for Verifying the MODEL

Demand quantity is assumed for verification and node and link costs are derived by demand quantity as <table 1>. In order to verify the algorithm we would like to transform the model 3.2 into <figure 4> which include the node and link costs on the diagram.

Table 1. Input data of each nodes

	Demand	Node Cost	Link Cost
C1	150	45	75
C2	170	51	85
C3	270	81	135
C4	210	63	105
R1	110	33	55
R2	90	27	45
TTL	1,000	300	500

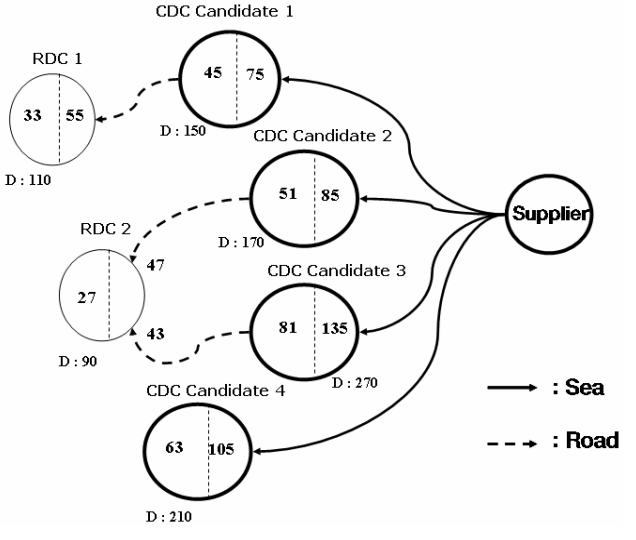


Fig. 4. Network modeling including node and link costs

In order to apply CDC searching algorithm, we have to reallocate some of the node costs into link costs as <figure 5>.

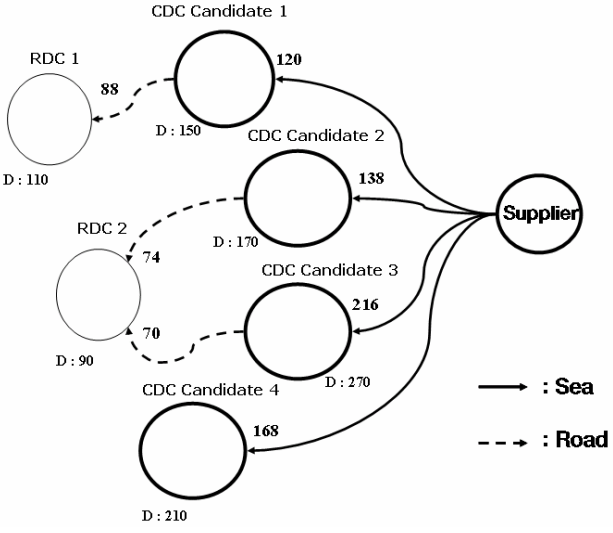


Fig. 5. Reallocation of the costs

Assuming that CDC candidate 1 is CDC, all of cargo are transported to CDC candidate 1 and then distributed to each RDC as figure 6.

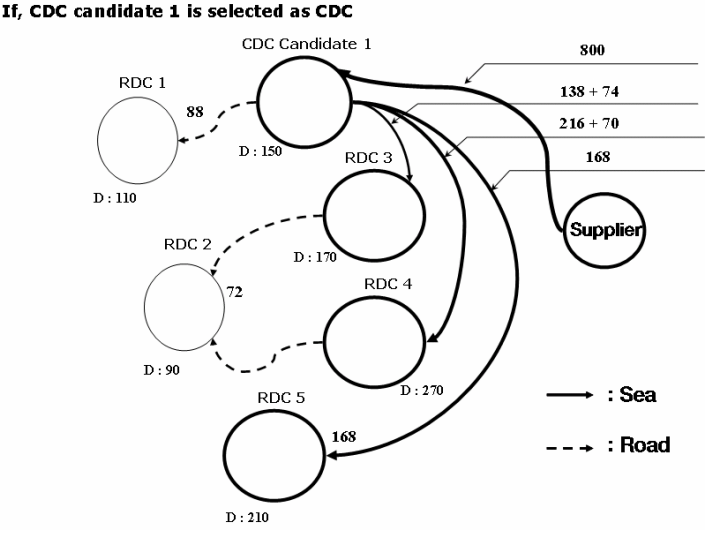


Fig. 6. Assuming model of CDC candidate 1

Table 2. Result of algorithm test

	C1	C2	C3	C4	R1	R2	Total Cost
C1	0	212	286	168	88	286	1,040
C2	208	0	144	168	296	74	890
C3	208	138	0	168	296	70	880
C4	208	212	286	0	296	286	1,288

3.2 Result of Algorithm Test

The result of applying algorithm which is described as CDC searching algorithm of 4.2 in this paper is shown as <table 2>. As the result of test, we can find candidate 3 as the best solution in terms of the minimum cost.

4 Case Study

4.1 Scenario Building-Up

It is an imaginary form of distribution system between a foothold of the main logistic center and consumers using the real data of "L" company which produces high quality TV panels. It is assumed in this Chinese market area which is the core of Northeast logistic market. "L" company is in the field of producing panels for PDP & LCD TV equipped with the facility which can produce 1.2 million panels per year and occupys 70% of market share in Chinese PDP panel market.

Concerned with logistic center locating matter, cost, distance and customer service level should be handled with priority to build up the optimistic logistic network on various scale demand locations.

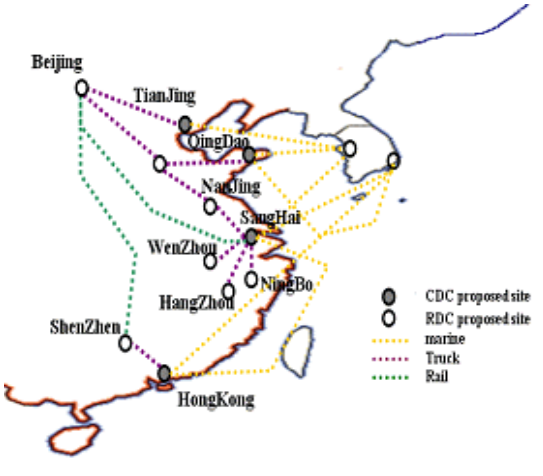


Fig. 7. Network model of the scenario

For the purpose of case analysis, it is assumed that production is made in Chang Won, Korea and destination is one of Tianjin, Qingdao, Shanghai or Hongkong in China via Pusan or Inchon, Korea.

The model of transportation from the port of discharging to inland distribution area or other port was used for the study analysis.

4.2 Scenario Input Data

Local Data Analysis

Table 3. The sales of HDTV in china 04-05

	2004	specific gravity	2005	specific gravity
LCD	22.9	13.1%	135.3	55.2%
PDP	16.4	9.4%	53.1	21.7%
R P	135.8	77.5%	56.5	23.1%
TOTAL	175.1	100%	244.9	100%

Logistic Cost Calculation

It needs to calculate total logistic cost to solve the matter of optimistic logistic network and it can be modeled through this formula.

Therefore systematical approach is necessary for the logistic cost calculation. Though there are distribution cost of finished products and raw material cost in the logistic cost according to R.Ballou(2005), only the distribution cost is mentioned in this study. Logistic cost is classified into constant cost and variable cost, and constant cost includes general management cost and facility and equipment depreciation cost, and variable cost consists of transportation cost, storage, quotation and order cost. Among these costs, transportation cost includes transportation cost between the locations, stevedorage. Storage consists of storage during transportation and in the warehouse. Quotation and order cost consists of information processing cost.

Demand Data

Total consumption figure was calculated based on 2006 China PDP TV market research report (CCID.2006) of total Chinese demand data. China demand and supply data were prospected and calculated referring to chart 2 by putting aggravation for each city consumption figure at the relative ratio for Shanghai GDP & disposition income.

Transportation Cost

Ocean freight is based on HANJIN shipping rate, and logistic cost in China is calculated by the (t/km) with the data of Chinese Logistic Center. Other costs of storage, packing cost, stevedorage, logistic information cost, general management cost were calculated following national logistic cost table and prospect.

Table 4. Logistic cost appraisal process

Transportation cost	Respective distribution transportation cost	Transportation rate by carriage patterns	
	Stevedorage	THC(Terminal Handling Charge) + (Port Charge)	
stock cost	(Stock cost during transportation)	Stock cost during transplantation=ICDT/365 I= Yearly stock utilization ratio) (20%) C= production price (USD) D=yearly demand (Unit) T=Transit Time (day)	
	warehouse storage	$AIL = \frac{dT}{2} + zS'd$ $Sd' = Sd\sqrt{T + LT}$ $T = \frac{EOQ}{D}$ d=Unit)(daily average demand) T=(stock checking period) z=(75%)(stock availability ratio during stock checking period)	
order cost	order-EDI cost	customer order handling cost)=time order handling cost) x yearly order f	
general management cost and indirect cost	selling management cost	salary	DC people × average salary
		others	Salary × 1.1 salary
	own warehouse depreciation cost	initial investment cost – remaining ost/economic durability initial investment cost = construction cost(㎡) x required area economical durability = 25 years	
	stevedore equipment depreciation cost	initial purchasing price – remaining value initial purchasing price = unit price x required machine unit economical durability = 8 years	

According to the scenario developed in the case study, it needs connectable route among nodes to select the place of central warehouse and the optimistic node for the minimum cost. If consider constant cost including transportation cost the distribution system shows as Figure 8.

If input simulation data considering transportation cost as shown Table 5, it shows as Figure 8. Respective data in the cell marks transportation cost from node to node and ∞ means that no-link exists between nodes.

Table 6 shows the minimum cost warehouse location deduced by the simulation with the concerned data input, and red sector shows yearly total cost for each node.

At the result of minimum cost calculation by the simulation, Shanghai is the optimistic place for CDC logistic network center in the aspect of distribution and cost saving and followed by Qingdao, Hongkong, Tianjin.

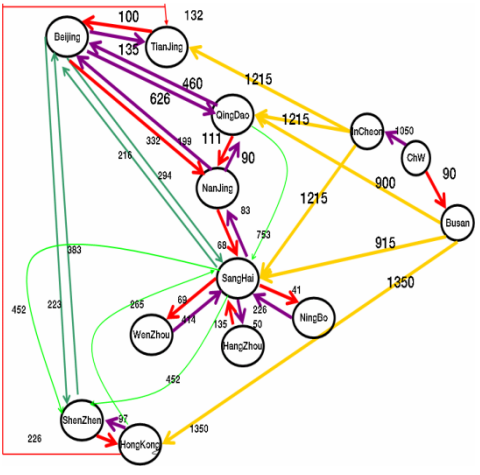


Fig. 8. Graph of the scenario

Table 5. Input data of the scenario

	Tian Jing	Bei Jing	Qing Dao	Nan Jing	Shang Hai	Ning B o	Hang Zhou	Wen Zhou	Hong Kong	Shen Zhen
TianJing	0	100	∞	∞	∞	∞	∞	∞	∞	∞
BeiJing	135	0	626	332	294	∞	∞	∞	∞	223
QingDao	∞	460	0		111	753	∞	∞	∞	∞
NanJing	∞	199	90	0	68	∞	∞	∞	∞	∞
ShanHai	∞	216	∞	83	0	41	50	69	∞	452
NingBo	∞	∞	∞	∞	226	0	∞	∞	∞	∞
HangZhou	∞	∞	∞	∞	135	∞	0	∞	∞	∞
WenZhou	∞	∞	∞	∞	414	∞	∞	0	∞	∞
HongKong	226	∞	∞	∞	265	∞	∞	∞	0	97
ShenZhen	∞	383	∞	∞	∞	∞	∞	∞	97	0

5 Conclusion

By the aftermath of this study, we developed the algorithm which could determine a single optimal CDC on SCM Network. In order to determine the minimum cost network under the various network conditions of lots of cases, this Decision Processing Algorithm refers to traditional methods of MST and shortest span algorithm.

Most of SCM network processing is concentrated on NP-Hard only, but this algorithm which makes building up the minimum cost network easily helps companies to decide where to locate their logistic centers.

Table 6. Out data of the scenario

	Tian Jing	Bei Jing	Qing Dao	Nan Jing	Shang Hai	Ning B o	Hang Zhou	Wen Zhou	Hong Kong	Shen Zhen	Total Cost
TianJing	0	100	522	432	394	435	444	463	420	323	3533
BeiJing	135	0	422	332	294	335	344	363	320	223	2768
QingDao	445	310	0	111	179	220	229	248	630	533	2905
NanJing	334	199	90	0	68	109	118	137	519	422	1996
ShangHai	351	216	173	83	0	41	50	69	536	439	1958
NingB o	577	442	399	309	226	0	276	295	761	665	3951
HangZhou	486	351	308	218	135	176	0	204	671	574	3123
WenZhou	765	630	587	497	414	455	464	0	950	853	5615
HongKong	226	326	438	348	265	306	315	334	0	97	2655
ShenZhen	323	383	535	445	362	403	412	431	97	0	3391

In the matter of locating of logistic center, on account that this study suggests a single CDC selecting solution without considering social and cultural factors which are of importance in some respects, it has limited considering all the complex real constrains.

This paper has limitation to decide single CDC in network decision problem . When we solve the multiple CDCs in network, more elaborated research is required in future.

References

1. Wisner, J.D., et al.: Principles of supply chain management A balanced approach (2005)
2. Shapiro, J.F.: Modeling the Supply Chain (2001)
3. Ballou, R.H.: Business Logistics Manament, 3rd edn. (1992)
4. Warnenburg tran, C.M., Hall, P. (eds.): von thunen,s Isolated state (1966)
5. Greenhut, M.L.: Plant Location in theory of the Location of Industries(1956)
6. Kekre, S., Nandakumar, P., Powell, T.A., Rajagopalan, S.: A Logistics Analysis at Heinz. Interfaces 20 (1990)
7. Jang, B.M.: A study on O.R model application for supply chain management. Logistics Research 9(2)
8. Jung, S.J., et al.: Simulation analysis for location and scale of building up logistics network Korea Simulation Institute thesis 14(3)
9. Park, J.H., et al.: Appraisal and tendency of Logistic cost in Seoul. Seoul Urban Planning Study 5(3)
10. Kim, S.H., et al.: Supply on time can capacity Production-distribution planning simulation and optimal model application on chain management. Korea Simulation Institute, A-Ju Univ. (2000)
11. Su., S.J., et al.: Industry Technology 13(3) (2000)
12. CCID Consulting, China PDP TV market research report, 2005-2006
13. Kim, T.H., et al.: Strategic logistics management, Byum Han (2006)

Simulation of Intersection Rivet at Non-signalized Intersection in Housing Scheme

Mohd Nazmi[1], Sadao Takaba[2], Sumio Ohno[2],
and Mohd Nazaruddin Yusoff[3]

[1] Graduate School of System Electronic, Tokyo University of Technology,
1404-1 Katakura, Hachioji, 192-0982 Tokyo, JAPAN
mnazmi@iium.edu.my, nazmi@so.cs.teu.ac.jp
http://staff.iiu.edu.my/mnazmi/
[2] School of Computer Science, Tokyo University of Technology,
1404-1 Katakura, Hachioji, 192-0982 Tokyo, JAPAN
{takaba,ohno}@cc.teu.ac.jp
[3] School of Environmental Engineering, University Malaysia Perlis,
Kubang Gajah Campus, 02600 Arau, Perlis, MALAYSIA
nazaruddin@uum.edu.my

Abstract. Accident in the residential area are become serious case in Malaysia. Most of the incidents occur among pedestrians, bicycles, motorcycles and vehicles. Our research purpose is to avoid collision at the non-signalized intersection in the housing scheme. We committed to reduce injuries and increase pedestrians' safety. Our research provides important information that can help driver predict common problems and take steps to prevent collisions. Intersection rivet is proposed for this matter. This type of signal system can prevent any accident in a dangerous non-signalized intersection. Simulation tools and systems are developed to find and solve the problem in order to decrease any fatal incident. Investigation data were used to simulate the situation more precisely. The result will be effective as reference to set the parameter of control system of the intersection rivet.

1 Introduction

Many incidents at the non-signalized intersection in Malaysia occur at night. The main problems are (1) Drivers intend to increase the speed of their vehicle even when they enter non signalized intersection. (2) Drivers do not detect the presence of the pedestrian or bicyclist in time to avoid collision. (3) Darkness and obstacle sight-lines will give effect to the driver when they enter the intersection, etc. In order to solve those problems, the implementation of the intersection rivet system[1] is proposed in this research that is capable of reacting as traffic signal at the non-signalized intersection. Intersection rivet is a device which is located in the middle of the intersection to notify the drivers the existence of intersection. It is useful when the vision of the driver is bad especially at night. Our system detects the existence of pedestrians and gives intensive warning to the driver approaching the intersection to decrease of their speed. It will reduce collision at the non-signalized intersection.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 259–268, 2007.

2 Research Objective

Traffic safety at the non-signalized intersection is now very important to decrease accident at the urban intersection. Many researches and projects to prevent the serious mistake occurred at the intersection have been considered. Some progresses have been proposed such using the mirrors and sensors[2]. We take a first step to investigate the accident at the residential area, where no significant improvements are for the past ten years. Design and architecture of the intersection rivet as well as layout of pedestrian sensors is shown. Three types of signal models to set the signal priority are considered. Simulation tool and system are developed for this purpose. Investigation results are used in the simulation and the results are compared. The effectiveness of the system is shown by the simulation.

3 Investigations

Investigations have been done to find out the research problem at non-signalized intersection. It took places at the area of housing scheme in the State of Kedah, Malaysia. Five places have been selected as all these intersections recorded as the highest collision among vehicle and pedestrian particularly at night. We used the camera to detect the movement at the problem intersection.

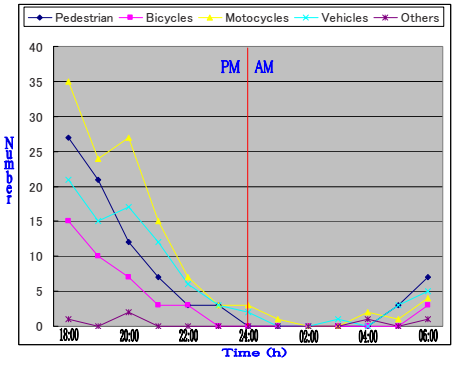


Fig. 1. Graph for Data Collection of users at the intersection

Table 1. Data of users

TIME	Pedestrians	Bicycles	Motorcycles	Vehicles	Others
18	27	15	35	21	1
19	21	10	24	15	0
20	12	7	27	17	2
21	7	3	15	12	0
22	3	3	7	6	0
23	3	0	3	3	0
24	0	0	3	2	0
01	0	0	1	0	0
02	0	0	0	0	0
03	0	0	0	1	0
04	0	0	2	0	1
05	3	0	1	3	0
06	7	3	4	5	1

The data of vehicles, motorcycles, bicycles and pedestrians are taken and the peak time for pedestrians who passed the intersections are recorded. We calculated the speed of vehicles when approaching and entering the intersection. The numbers of vehicles which are increasing and decreasing their speeds are recorded[2]. The investigations have been done for a month everyday from 18:00 (evening) until 06:00 (morning). One of the investigation results are shown on figure 1 and table 1. The data and the results are used as the input data for the simulation.

4 Outline Systems

Intersection Rivet has been designed and has 5 layers where:

(a) First layer - Contain the solar circuit to gain solar power on the day light
(b) Second layer - Contain the LED where Yellow and Red are installed to be signal.
(c) Third layer - Contain the system circuit
(d) Forth layer - Contain the communication and sensor layer.
(e) Fifth layer - Contain the battery to keep the energy for night used.

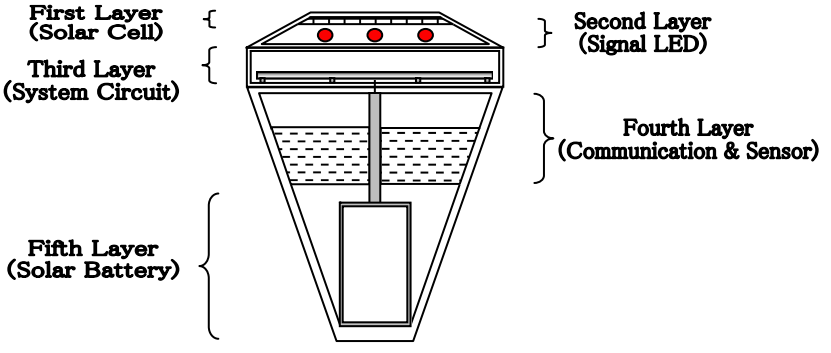


Fig. 2. Intersection Rivet Systems

Rivet is installed at the center of the non-signalized intersection[1]. It mainly works to inform the present of intersection for the users in the area. In some condition, it is also functioned as a signal for the intersection but it has only the red and yellow signal. Four sensors are installed 25 meters before the stop line at every direction. Rivet always blinks in a yellow as a default signal. When any approaching movement toward intersection is detected by the sensor, the yellow signal will change to red signal. This situation depends to the approaching vehicles or pedestrians at the intersection. The signal systems are capable with a decision to give a priority for the vehicle to stop or move into the intersection. The four-sides of rivet signal can change their signals from yellow to red or red to yellow depending to the side that has the priority.

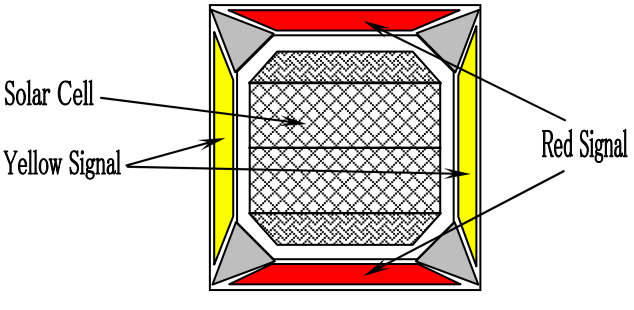


Fig. 3. First Layer and Second Layer of Rivet Image

The system can avoid any collision inside the intersection and close to the intersection with giving the information by blinking the rivet signal to inform the approaching vehicles or pedestrians. When it intersects with the arterial road, the priority will be given to the main road depending to the traffic condition. The system is designed to choose the priority. Figure 4 shows the position of rivet and all four sensors.

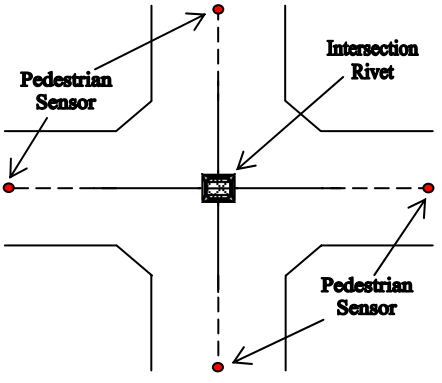


Fig. 4. Position of Intersection Rivet and Pedestrian Sensor

Four sensors are set 25 meters from the entrance (stop line) at the intersection. The sensor will detect the pedestrian and send the data to the sensor layer in the rivet. The rivet will change the LED signal to red with reference to the approach way of pedestrian. The system algorithm is shown as figure 5. It contains the rivet (whole) circuit, solar circuit, communication and sensor circuit, and the LED (signal light).

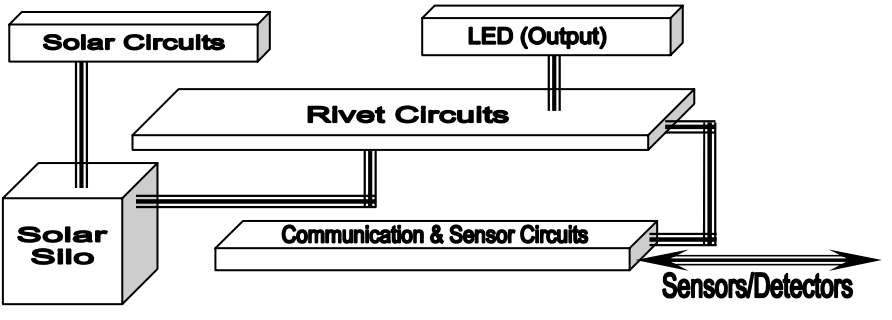


Fig. 5. System Architecture of Intersection Rivet

5 Signal Model

The concept of detection system and signal system is shown below. Analysis of the signal system has been precise to three main difficulties. The rivet signal parameter

change from the default condition when the sensors perceive any pedestrian approach to the intersection as shown below. Model A, B and C describe each signal parameter.

(1) Model A

In this model, there is no priority for the road. The rivet signal is blink with yellow signal as the default. When the sensor detected any pedestrian or vehicles approaching the intersection, it will turn the signal from yellow to red. For example, figure 6 show that the sensor detected a pedestrian from the north area approaching the intersection. The system will change the signal for east and west area to be blink in red.

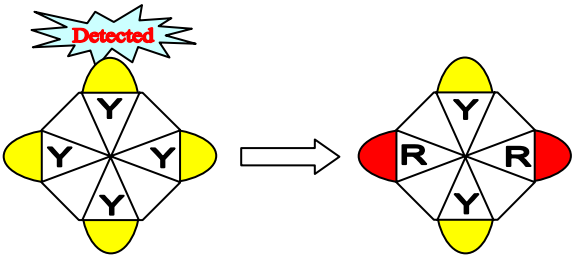


Fig. 6. Rivet System for Model A

(2) Model B

Model B shows how rivet system change the signal when the sensor detected movements from different directions. For example in figure 7, the sensor detects approaching pedestrian from both north and west area. Basically when there are multi-detection, the rivet signal will change to all red to warn the approaching vehicle.

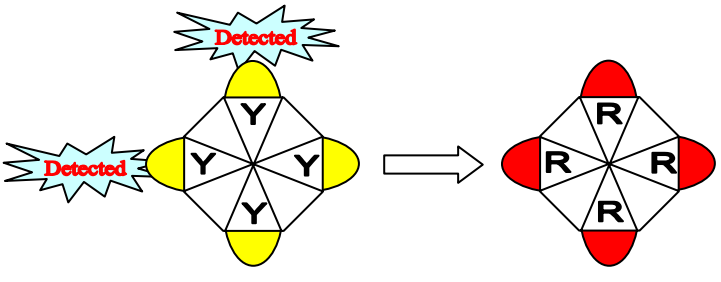


Fig. 7. Rivet System for Model B

Both model A and model B are assumed to be used at the housing scheme where both side intersected with minor road. Therefore, no priority signals are concerned for the road.

(3) Model C

In Model C, East-West bound are the main road, therefore priority signal are set to the main road where yellow signal for the main road and red signal for the minor road. When the sensor detects the pedestrian at the minor road, the signal will change from red to yellow to give a priority for the users until they cross the intersection as shown in figure 8.

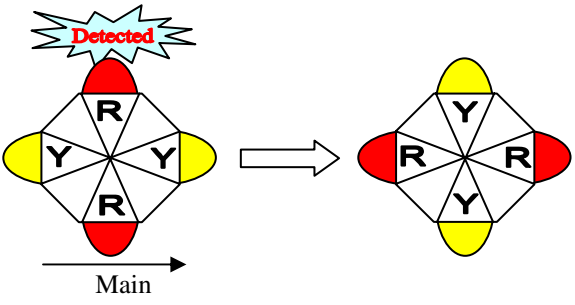


Fig. 8. Rivet System for Model C

If the sensor detect pedestrian from other direction, the system will change the signal by referring to the direction of the pedestrian.

6 Simulation

Simulation tools and its systems are developed to be more practical in order to find any problem at the non-signalized intersection. We developed the simulation tools by using the C++ language and Arena Simulation 11 (Rockwell). Five different situation models have been created for this research purpose. The models by using the intersection rivet and without using the rivet are compared. In the intersection rivet, we put the priority selection in the algorithm, where the rivet can make decision to give a signal priority for vehicle or to the pedestrian.

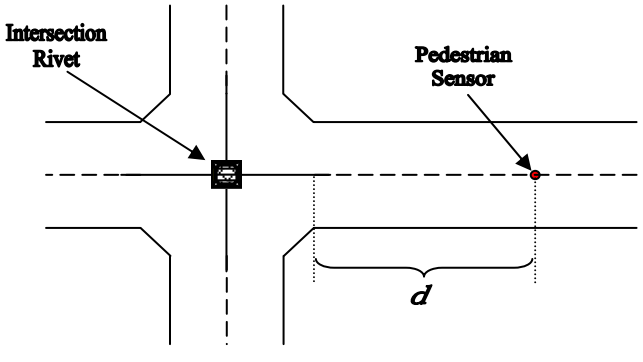


Fig. 9. Calculated Distance *d*

Simulations are done randomly by using the investigation data. Four scanning system located at the outside the intersection works as detector functions to sensor a vehicle or pedestrian approaching the intersection. All the scanning system connected to the intersection rivet and will transfer the data to the rivet to inform any movement in the detection range. We set the different speed for vehicle and pedestrian, so that the detector can recognize different between vehicles and pedestrians. Pedestrian walking

time is calculated by using the distance *d*, where *d* is the distance between sensor and intersection stop line.

6.1 Simulation Model

Pedestrian and vehicle are been modeled in this research. Pedestrians in the simulation are modeled as autonomous objects to be seen from above. Investigation data for pedestrian has been used to run the simulation. Each pedestrian has been set its own initial location, initial time, and initial velocity. The target location is specified randomly. The model uses a simple equation to calculate the pedestrian walking time from sensor until its reach the entrance of intersection [2]. When pedestrian reach the intersection, rivet will delay the signal for 3 seconds and the signal will be changed to the default one.

The time *t* represents the simulation clock. The current location (sensor) *s(t)*, velocity *v(t)*, and acceleration *a(t)*are denoted by vector respectively. The basic dynamical model is given by

$$s(t+1) = s(t) + v(t) \tag{1}$$

$$a(t) = v(t+1) - v(t) \tag{2}$$

An intended velocity is equivalent to a vector that will direct the pedestrian from the sensor position into the entrance position. Since acceleration can be seen as the difference of velocities, it is equal to the difference between intended velocity and the current velocity[3]. Estimated time for pedestrian from sensor to entrance is determined from the distance *d*. Figure 9 shows the distance between sensor and entrance. We set the model to scan, and then activate the signal when found the pedestrian. The logical model for sensor to detect and activate the rivet signal is shown in figure 10.

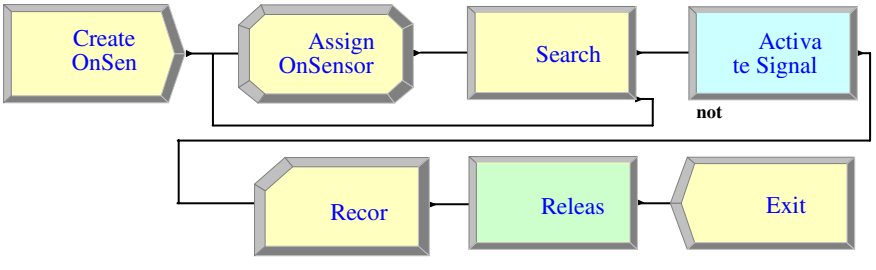


Fig. 10. Scanning and Signal Activation Model

6.2 Simulation Flow Algorithms

We create the flow chart of the simulation model and simulation code. Figure 11 shows the simulation model's flow chart where,

- There is a sensor to detect the approaching pedestrian.
- Calculation and time prediction
- Activate the rivet signal from default to red (notice to vehicle to stop)

- Extend the signal until pedestrian cross the intersection
- Back to he default signal when no detection.

Figure 12 shows the code that we use in "Arena Simulation" code to run the model.

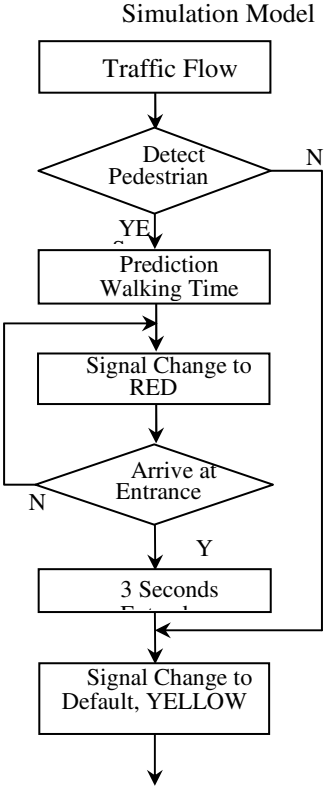


Fig. 11. Simulation Model Chart

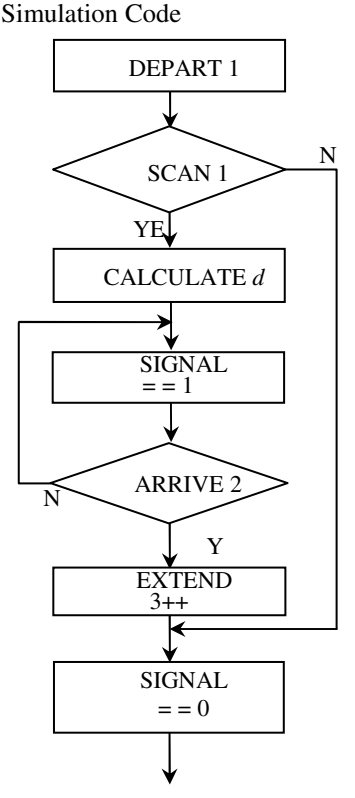


Fig. 12. Simulation Code Chart

6.3 Simulation Image

Simulations are done by using the investigation data randomly. The comparison results of stop vehicle temporarily at the stop line by using the rivet are recorded. Figure 13 shows the simulation model with logical part. The arrow in the picture shows the pedestrian sensor which has been set outside intersection. Yellow signal are set as a default signal which is located in the middle of the intersection. Figure 14 shows the image of one pedestrian (blue circle) entering the intersection from one direction. At this condition, the intersection rivet is in red signal.

Figure 15 show the pedestrian from two different directions. Signal is change when the first pedestrian and will turn back to default after the second pedestrian reach the intersection entrance. At this situation, the rivet signal is change to all red. Figure 16 shows the condition of one pedestrian and one vehicle. At this situation, priority is given to the pedestrian, where vehicle has to stop until pedestrian across the intersection.

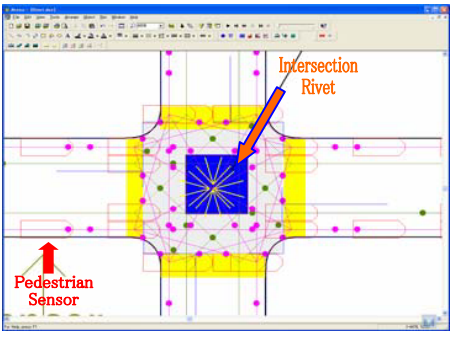


Fig. 13. Simulation Model (Image)

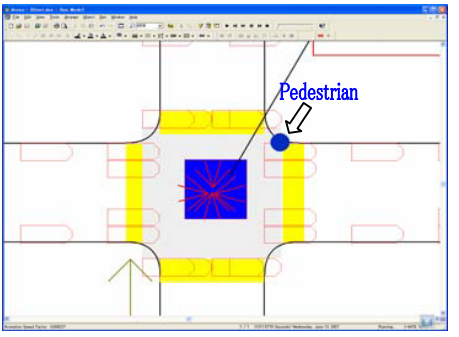


Fig. 14. A pedestrian from one direction

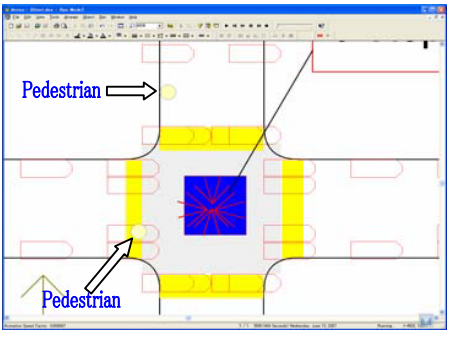


Fig. 15. two pedestrian from different direction

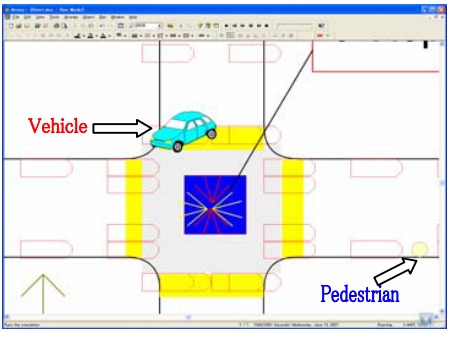


Fig. 16. Pedestrian and vehicle

6.4 Simulation Result

The simulation result shows that by using the intersection rivet in housing scheme, the drivers will reflect with the rivet signal if the rivet sign turn from yellow to red. The numbers of vehicle which alert of rivet will stop at the stop line temporary are increased

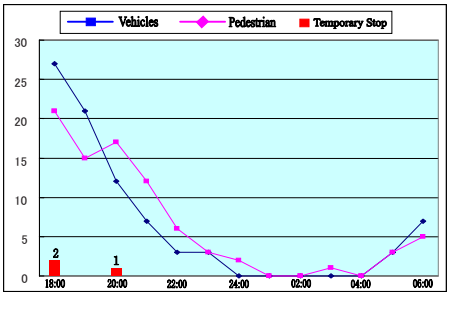


Fig. 17. Simulation without Rivet

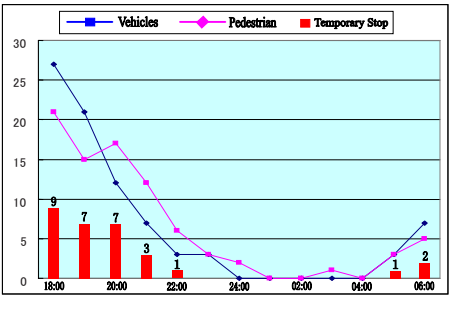


Fig. 18. Simulation with Rivet

(figure 18). The results show that by using the intersection rivet, the drivers will more watchfully when the signals begin to change to red. Rivet alerts the vehicle where they are approaching the non-signalized intersection and have to be more careful when entering the intersection. The result shows that stopping vehicle could avoid fatal incident at the problem intersection.

7 Conclusions

Intersection rivet is very useful to avoid any collision at the non-signalized intersection[1]. It is very helpful where the driver can notice presence of other vehicle or pedestrian at the intersection and allow the driver to decrease the speed of vehicle before entering the intersection. This system is considered to be introduced to Malaysian government, to be used in the resident and urban area. Rivet setup cost is less than the traffic system installation; it is practically effective to be used in the under-populated area since rivet also possibly can be functioned as traffic signal light. Simulation shows the effectiveness of intersection rivet at the non-signalized intersection in Malaysia for the traffic safety.

Drivers' behaviors are important to avoid the unnecessary incident at the intersection. The problem of pedestrian safety is being concerned worldwide in the activities of Intelligent Transport Systems. The design of roads and intersections which help pedestrians, and also facilities and equipments to activate drivers awareness are widely concerned. Effective integration of tools and action plans are recommended[2][3].

References

1. Kawatsu, H., Takaba, S., Souma, K.: Proposal and Evaluation of a System Which Aims Reduction in Unexpected Collision at Non-signal Intersection. In: Baik, D.-K. (ed.) AsiaSim 2004. LNCS (LNAI), vol. 3398, pp. 403–409. Springer, Heidelberg (2005)
2. Fitzpatrick, K., Turner, S., Brewer, M., Carlson, P., Ullman, B., Trout, N., Park, E.S., Whitacre, J.: Improving Pedestrian Safety at Unsignalized Crossings. TCRP Report 112/NCHRP Report 562, 100p, Transportation Research Board (TRB), 2006. Improving Pedestrian Safety at Unsignalized Crossing. Transport Research Board (NCHRP) (2006)
3. Redmon, T. (ed.): How to Develop a Pedestrian Safety Action Plan. Pedestrian and Bicycle Information Center Highway Safety Research Center (2006)

Adaptive DNA Computing Algorithm by Using PCR and Restriction Enzyme

Yuji Kon, Kaoru Yabe, Nordiana Rajaee, and Osamu Ono

Department of Electronics and Bioinformatics, Meiji University
Higashi-mita, Tama-ku, Kawasaki-shi, Kanagawa-ken, 214-8571 Japan
{ce77037,nrajaee,ono}@isc.meiji.ac.jp

Abstract. In this paper, we introduce an adaptive DNA computing algorithm by using polymerase chain reaction (PCR) and restriction enzyme. The adaptive algorithm is designed based on Adleman-Lipton paradigm[3] of DNA computing. In this work, however, unlike the Adleman- Lipton architecture a cutting operation has been introduced to the algorithm and the mechanism in which the molecules used by computation were feedback to the next cycle devised. Moreover, the amplification by PCR is performed in the molecule used by feedback and the difference concentration arisen in the base sequence can be used again. By this operation the molecules which serve as a solution candidate can be reduced down and the optimal solution is carried out in the shortest path problem. The validity of the proposed adaptive algorithm is considered with the logical simulation and finally we go on to propose applying adaptive algorithm to the chemical experiment which used the actual DNA molecules for solving an optimal network problem.

1 Introduction

In 1994, Leonard Adleman[1] successfully solved the Hamiltonian path problem of seven nodes using the actual DNA molecules. Following his succeed, a new research field called DNA computing has been established. However, in DNA computing, there is a problem is that the quantity of the molecules used for computation will increase exponentially with the scale of the problem. In this paper, we introduce an adaptive DNA-based computing algorithm by using polymerase chain reaction which is a new method aiming to improve the explosion problem of the DNA molecules.

The adaptive DNA computing algorithm which will be proposed here is based on the Adleman-Lipton paradigm tends to discover the solution of a problem by a series of biochemical operations. Based on this paradigm, for solving Hamiltonian Path Problem (HPP), the DNA molecules encoding all the possible routes are poured into a test tube. After the hybridization and ligation, based on the Watson-Crick complementary, it is expected that there will exist a combination among various kind of combinations that is represents the solution to the problem. Thus, before the computation begins, it is necessary to prepare beforehand sufficient DNA molecules in order to generate the wanted combinations. The main disadvantage of this procedure is that the quantity of

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 269–275, 2007.

required DNA molecules increases as the problem size increases. In order to solve this imitation, we propose an adaptive algorithm which consists of two concepts: study and strengthening. Study is memorizing the knowledge acquired from experience and building new knowledge structure. Acquisition of the knowledge by study enables the suitable response to the given input. The study function in DNA computing introduces the feedback structure of the detection result of a solution. By doing this, the detected molecules can be memorized as knowledge for the next computation.

The strengthening function in DNA computing is realized by concentration control of DNA molecules. That is to say, by amplifying a molecule using PCR, the concentration of the molecule is changed dynamically. Consequently, the concentration of the molecules suitable for detection of a solution can be made. By combining a feedback mechanism and dynamic concentration control, adaptive algorithm adjusts the concentration value of the molecules by 1 cycle computation, and reuses the results of concentration adjustment to the next cycles. Therefore, this algorithm does not need to search a solution in a series of operations. Furthermore, it is able to cut down the quantity of DNA molecules by learning in order to increase the accuracy of the computation. The proposed adaptive DNA computing algorithm based on PCR is applied to the shortest path problem, and the correctness is verified by experiment.

2 Adaptive DNA Computing Algorithm

The adaptive DNA computing algorithm consists of six steps as shown in Fig. 1.

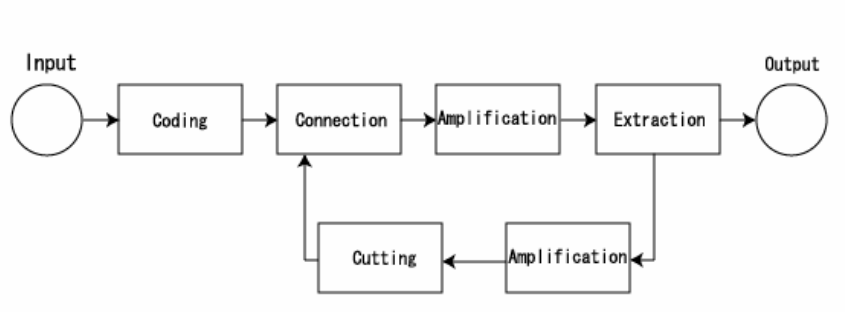


Fig. 1. The flow of adaptive DNA computing algorithm

- Step 1 Coding : The given problem is encoded by DNA molecules.
- Step 2 Connection , Amplification : The coded DNA molecules are made to connect and amplified by using PCR.
- Step 3 Extraction : The combination which is a solution candidate is extracted among the generated combinations.
- Step 4 Amplification : The extracted combination is amplified by using PCR.
- Step 5 Cutting : The amplified combination is cut by restriction enzyme.
- Step 6 Repetition : Step 2 to Step5 are repeated.

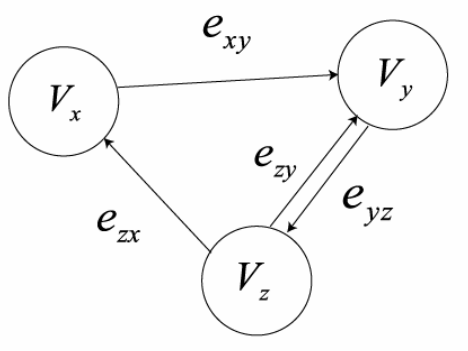


Fig. 2. A graph of 3 vertices and 4 edges

As such, a graph with three vertices and four edges as shown in the Fig. 2 is considered. This graph gives as an example, and the procedure of coding is explained in detail. For that reason, especially, this graph is not set up as a problem.

Firstly, oligonucleotides, or oligos for short, are assigned to the vertex v_x, v_y, and $v_□$, and edge e_{xy}, e_{zx}, e_{yz}, and e_{zy}. The oligos are made of 20 nucleotides in this graph. In order to simplify the notation of base sequence, the sign described below expresses the nucleotides assigned to each vertex.

$$\begin{array}{llll} v_x : \text{GCATTCGGAT} & \text{CTAGCATCGG} & v_x : v_x(p) & v_x(n) \\ v_y : \text{TACGGTTCCA} & \text{GCATTGGGTC} & v_y : v_y(p) & v_y(n) \\ v_z : \text{AGGGCATTCC} & \text{CATTCCGAGT} & v_z : v_z(p) & v_z(n) \end{array}$$

Each vertex $v_i(p)$ (i=x, y, z) expresses 10 nucleotides in the first half of the oligos which encodes the vertex v_i, and each vertex $v_i(n)$ (i=x, y, z) expresses 10 nucleotides in the second half of the oligos which encodes the vertex v_i. In the adaptive DNA computing algorithm, α and β segments are used as recognition part of a restriction enzyme between $v_i(p)$ and $v_i(n)$ as described below. In here, α and β segments are corresponded to a restriction enzyme EcoRV. That is to say, α is assigned GAT, β is assigned ATC. Consequently, if α and β segments connect with complementary α and β segment, it will be designed so that it may become the recognition part of a restriction enzyme EcoRV.

$$\begin{array}{lll} v_x : v_x(p) & \alpha\beta & v_x(n) \\ v_y : v_y(p) & \alpha\beta & v_y(n) \\ v_z : v_z(p) & \alpha\beta & v_z(n) \end{array}$$

Next, each edge consists of the complementary sequence in the first half of each vertex and in the second half of each vertex. Furthermore, each edge is encoded by adding a restriction enzyme recognition part.

$$\begin{array}{llll} e_{xy} : \sim\beta & \sim v_x(p) & \sim v_y(n) & \sim\alpha \\ e_{yz} : \sim\beta & \sim v_y(p) & \sim v_z(n) & \sim\alpha \\ e_{zy} : \sim\beta & \sim v_z(p) & \sim v_y(n) & \sim\alpha \\ e_{zx} : \sim\beta & \sim v_z(p) & \sim v_x(n) & \sim\alpha \end{array}$$

The tilde means complementary relation. Each encoded vertices and edges alternately connected by Watson-Crick complementary. The procedure of coding described above can apply to coding of the shortest path problem.

3 The Shortest Path Problem

We applied the proposed adaptive algorithm to the shortest path problem. The search space is identified by denoting the lower left of the space as search start point and the upper right of the space as a goal in a 4*4 square search space as described in Fig. 3.

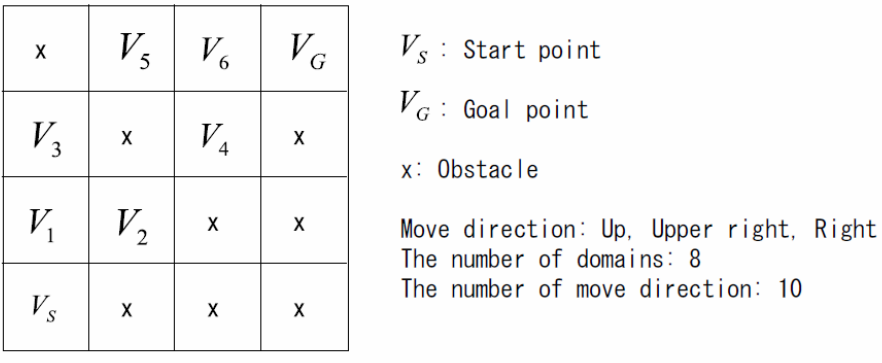


Fig. 3. 4*4 square search space

The move directions of the object are limited only in three directions, right, upper right, and up. Since the start point is lower left and the goal point is upper right, if the object moves to other directions, it may be a detour. Moreover, DNA molecules to prepare increase as the move directions increase. We limit the move directions to hold down cost when we experiment.

3.1 Coding and Connection

Firstly, the sequence for domain and the sequence for movement as described in Fig. 4 are prepared. The sequence for domain is encoded for each domain in the square search space. The sequence for movement is encoded for a path moving from a domain to other domain. $V_n(P)$ expresses nucleotides in the first half of the oligos and $V_n(N)$ expresses nucleotides in the second half of the oligos. α and β express sequence used as recognition part of a restriction enzyme. A bar above alphabet expresses complementary relation. Parallel Overlap Assembly (POA) is performed using all the designed sequence for domain and sequence for movement. Consequently, various combinations are generated. At this time, in the case of the Adleman technique, all combinations are needed to obtain an optimal solution. However, in the case of the proposed algorithm, all combinations are not needed. It is because, even if the optimal solution does not exist in generated combinations, by using feedback structure, the solution candidate is getting close to the optimal solution.

Fig. 4. Coding base sequence

3.2 Amplification and Extraction

The base sequence which serves as a solution candidate among all the combinations generated by connection operation is amplified by using primers shown in Fig. 7. There are two conditions for becoming a solution in the shortest path problem. Firstly, it is necessary to have the feature that the combination begins from the domain of a start point, and finishes with the domain of a goal point. Secondly, when the length of the combinations is shorter, it is clear that a possibility which the combinations serve as the optimal solution is high. Therefore, the combination which fulfills these two feature points is amplified, and the combination is considered as superior solution candidate sequence. After the solution candidates amplified, gel electrophoresis is performed. Extraction operation is performed by making gel electrophoresis, and by doing this, solution candidates are divided by their length. The shortest combination should be an optimal solution.

3.3 Amplification and Cutting

After extracted, the shortest combination as the shortest path is amplified by using primers shown in Fig. 7. The extracted combinations are detached to single-stranded DNA molecules by denaturaiton. By supplying primers described in Fig. 7 in the detached sequence, the primers cling to each sequence which has complementary relation, respectively. And only the coding sequence which constituted the combinations extracted as a solution candidate can be made to amplify by expanding the primers. Then, amplified sequences are cut by the restriction enzyme EcoRV as described in Fig. 5. It is important that one chain of detached the single-stranded DNA molecules serves as the same form as a sequence for movement as described in Fig. 6.

3.4 Repeat Calculation

Each sequence for movement which arise the concentration difference is returned to the process of connection again, and repeated calculation is performed. Under repeated calculation, since a concentration difference arises in each sequence for movement, the high sequence for movement of concentration generates many combinations. Conversely, the low sequence for movement of concentration does not generate many combinations. That is to say, under repeated calculation, generation probability of combinations serves as a solution candidate sequence becomes high compared with before calculation. Thus, by repeating the cycle of connection, extraction, cutting, and amplification, generation probability of combinations serves as a solution candidate sequence is raised, and finally the adaptive DNA computing algorithm detect an optimal solution.

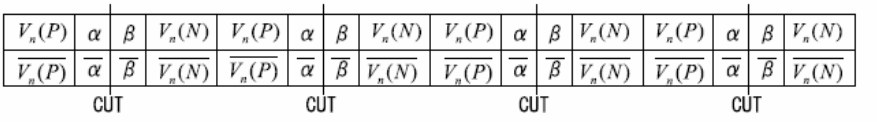


Fig. 5. Cutting of the base arrangement

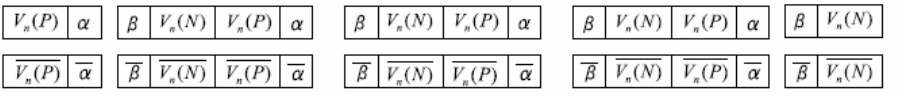


Fig. 6. Detachment of single-stranded molecule

Fig. 7. Primers

3.5 Expected Experiment Result

The experiment result is observed by performing gel electrophoresis. DNA molecules are tinged with the negative electric charge, that is to say, in an electric field, DNA molecules move to positive side.

$$(\text{negative electric charge of DNA molecule}) \propto (\text{length of DNA molecules}) \quad (1)$$

$$(\text{force to move DNA molecules}) \propto (\text{length of DNA molecules}) \quad (2)$$

These two offset, in an ideal gel, DNA molecules move at the same speed. Then, using gel, DNA molecules move at a different speed depending on their difference of length.

The solution candidates are

- V_S-V_2-V_4-V_G : length of 160 base pairs
- V_S-V_1-V_2-V_4-V_G : length of 200 base pairs
- V_S-V_2-V_4-V_6 -V_G : length of 200 base pairs
- V_S-V_1 -V_3 -V_5-V_6-V_G : length of 240 base pairs
- V_S-V_1 –V_2-V_4-V_6-V_G : length of 240 base pairs

In the viewpoint of length, the result will be observed as three highlighted bands: length of 160, 200, and 240. In the case of this experiment, it is obvious that the shortest path is V_S-V_2-V_4-V_G, and we design the length of each sequence for domain as length of 40 bases. So, the length of the shortest path is 160. An expected experimental result is shown in Fig. 8. If the experiment is successful, the concentration of the band of length 160 increases as the number of cycles increases.

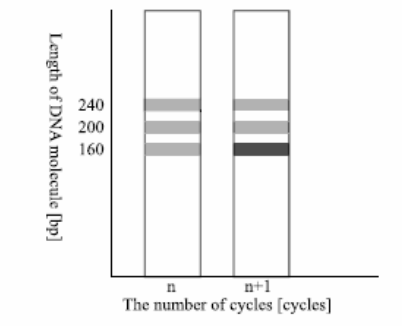


Fig. 8. Expected Experiment Result

4 Conclusion

It has been shown in literature that DNA computing provides a lot of potentials. However, there are many critical problems in this research area which must be overcome. One of the problems is the explosion possibility of DNA molecules with respect to the increasing of the size of the problem. In this paper, in order to solve this problem, an algorithm which is based on adaptive DNA molecules by using PCR was introduced. The validity and the correctness of the proposed adaptive DNA computing algorithm have been proved from our past simulation results. Moreover, this adaptive DNA computing algorithm is able to decrease the generation of useless paths which is caused by the connection of unexpected molecules. It is hope that the proposed adaptive DNA computing algorithm will provides great enhancements from the molecules explosion point of view for a more applicable DNA-based computing approach.

References

1. Adleman, L.: Molecular Computation of Solutions to Combinatorial Problems. Science 266, 1021–1024 (1994)
2. Watson, J.D., Crick, F.H.C.: Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature (London) 171, 737–738 (1953)
3. Lipton, R.J.: DNA Solution of Hard computational Problems. Science 268, 542–545 (1995)

Elementary Studies on Pervasive Simulation Grid

Tang Zhen[1], Li Bo-hu[1], Chai Xu-dong[2], and Song Chang-feng[1]

[1] School of Automation & Electric Eng., Beihang Univ., Beijing 100083, China
[2] Beijing Simulation Center, Beijing 100854, China

Abstract. Simulation grid has well solved the problem of sharing, reusing, cooperating and dynamic scheduling, optimizing and executing of simulation model and resources in the Internet. But it still has some disadvantages, such as the capabilities of moving, independence, adaptability and context-aware. If the pervasive computing technology could be introduced into the simulation grid, it can meet our requirement of gaining simulation services anywhere anytime in pervasive computing mode in the future. The requirement for pervasive computing technology in simulation grid was firstly discussed. Then the connotation and architecture of the pervasive simulation grid were raised, and some approaches for realizing pervasive simulation grid were presented. Finally, the conclusion and some further work were also provided.

Keywords: pervasive simulation grid, pervasive computing, simulation grid, context-aware computing, intelligent agent, service mobile.

1 Introduction

The networkitized M&S technology [1] is one of the studying hotspots existing in current M&S community. The development process of networkitized M&S includes several typical development phases [2], such as DIS, ALSP and HLA [3]. Outlining the development process of networkitized M&S, We can draw the conclusion that the technology connotation and application mode for networkitized M&S are always continuously extended and enriched with the development of network technology.

With the appearance and rapid development of grid computing technology, the combination of grid and simulation becomes a new trend of the development of networkitized M&S. Simulation grid is a new distributed modeling and simulation system with the background of application field. It synthetically applies complex system modeling technology, advanced distributed simulation technology, VR (virtual reality) technology, grid technology, management technology, system engineering technology and application field related specialized technology, to realize the dynamic share, reuse and cooperation of various resources in grid/federation safely (including model, computing, storage, network, data, information, knowledge, software resources relative to unit involved in simulation application/project, and application related physical effect devices and simulators, etc.), collaborative scheduling executions and optimization, etc., to support the various activities in full life cycle (from argumentation, research, design, development, test, execution, evaluation to maintenance and disposal) for existing or tentative complex system/project in engineering or non-engineering field [4],[5].

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 276–286, 2007.

At present, the rapid development of network technology and computing technology lead us to enter a new era of pervasive computing through wide distributed global calculating devices. Pervasive (or ubiquitous) computing was firstly proposed by Mark Weiser in 1991, which was a conception about future computing pattern [6]. Pervasive computing technology [7] is a combination of computing technology, communication technology and multimedia technology. It provides a brand-new computing pattern. The goal of pervasive computing is coming into smart space, which is combined with the day-life physical space and information space. The people can transparently obtain the computation and the information service in this smart space any where any time. The appearance and the arrival of pervasive computation will bring huge innovation for the traditional networkitized M&S technology. The networkitized M&S technology will develop towards pervasivization direction.

2 Introduce Pervasive Computation into Simulation Grid

The "pervasive simulation technology" [1], which is combined with pervasive computing technology, will realize one kind of new simulation pattern, and impel the research, development and application of modern M&S to enter a brand-new era. Face the future pervasive computing environment, which is complex, heterogeneous and dynamic, the pervasive simulation system will be supposed to have the basic characteristics as follows:

1. Pervasive: The simulation resources are omnipresent;
2. Any where any time: The people can obtain the simulation services in work or life scene, but do not have to sit in front of a special computer;
3. Auto-adaptive: The simulation information space is able to provide the coherent simulation services with suiting the user's way and adapting the changes of computation environment;
4. Transparent: The user does not need to pay more attentions when obtains the simulation services, and the visit way of simulation services is extremely natural even the user himself pays no attention, namely the implicit interaction so-called.

At present, the simulation grid has solved the problems of sharing, reusing, cooperation, dynamic optimization and implementation of the simulation model resources in Internet (including LAN and WAN) very well. Because of the characteristics and the superiorities, the simulation grid is a good method and technical way for realizing the pervasive simulation. First, the simulation grid with the aid of the grid technology, uses any kinds of simulation resources (various encapsulated model resources, simulation tools, computing resources, storage resources, data resources, information resources and knowledge resources, etc.) in the form of services, shields the complex and heterogeneous pervasive computing environment for users, and causes the simulation resources to be omnipresent. It has solved the problem of 'pervasive' very well. Next, the grid technology causes the terminal of simulation application to extend to each corner of the network. The users are allowed to visit the simulation resources and services through any networking (wired or wireless, mobile or fixed) equipments, and thoroughly get rid of the fetter of time and the space. It has satisfied the requirement of 'any time any where'.

But in view of the pervasive simulation demand, the simulation grid also has its limitations. First, considering these common problems of pervasive computation environment, such as the mobility of computation resources and the user terminal, the instability and the band width limitation of the network, the simulation grid does not have appropriate solution strategies. Complex and dynamic environment of pervasive computing sets a higher requirement for auto-adaptability of the simulation grid. Next, the simulation grid has not still gotten rid of 'the tabletop computation pattern', namely the simulation information space and the physical space are isolated.

Analyzing from the aspect of technology, the simulation grid lacks some essential characteristics for pervasive computing environment:

1. Context-aware: Lacks the ability of gaining the user and the simulation services' context information;
2. Seamless migration: Has not realized the seamless migration of simulation service and task in the mobile computing environment;
3. Spontaneous interaction: Has not realized spontaneous interactive among the simulation services, and between the simulation services and the user.

These problems will be solved well, if introducing pervasive computing technology into simulation grid. At present, the combination of grid and pervasive computing in many application scenes becomes a kind of tendency. Nigel Davies etc. have proposed an idea that realizing the pervasive computation by combining grid computing technology as early as in 2003 [8]. Manish Parashar and Jean-Marc Pierson proposed a vision of Pervasive Grid [9] together. The Electronics and Computer Science Institute of University of Southampton has found a Grid and Pervasive Computing Group. Their research topics include: Service-based Grid and Pervasive computing and so on.

The pervasive simulation grid is the specifically embodiment of the combination of pervasive computation and grid computation in the distribution interactive simulation domain. The research on pervasive simulation grid is needs to analyze the characteristics of the grid technology, the pervasive computation technology and the distribution simulation technology as well as their uniformity in idea. The research emphasis is how to use the technology idea of the grid and the pervasive computation for reference to construct 'the harmonious man-machine relationship' in the M&S domain, and further to seamlessly embed the M&S technology in people's daily life. At present, the project named "the application research and development of simulation grid" which the author participates in, has acquired phased research achievements, including completed simulation grid prototype named CoSim-Grid 1.0v [10], parts of solved key technologies and some typical application demonstration systems of CoSim-Grid. The research and the applications of pervasive computation have also had the certain technology foundation. These all have built the solid foundation and provided the advantageous condition for the research on pervasive simulation grid.

3 The Pervasive Simulation Grid

3.1 The Technical Connotation of Pervasive Simulation Grid

The pervasive simulation grid is a new distributed M&S system, which is complex, heterogeneous and dynamic. The comprehensive, independent simulation resources, such as computer software and hardware, communication software and hardware,

sensors, equipments, simulators and so on, are integrated together by the grid technology seamlessly. The core technologies of pervasive computing, which is including the sensor network, mobile computing, context-aware computing, wearable computing, intelligent agent, semantics and so on, are applied synthetically to establish the "man-machine" simulation environment to let the simulation information space and physical space for people lives amalgamate harmoniously and provide effective, the implicit simulation services for the user.

3.2 Architecture of Pervasive Simulation Grid

The architecture of pervasive simulation grid are consist of simulation resources, grid core service level, pervasive simulation service middleware level, simulation application level and pervasive simulation grid portal level.

1. Simulation resources level: Provides each kind of resources in the form of services, which is including model\service resources, tools\software, network resources, computing resources, memory resources, data resources, information resource, knowledge resources, simulation application related physical effective equipments and simulators and so on, in order to serve the pervasive simulation grid application system dynamically.

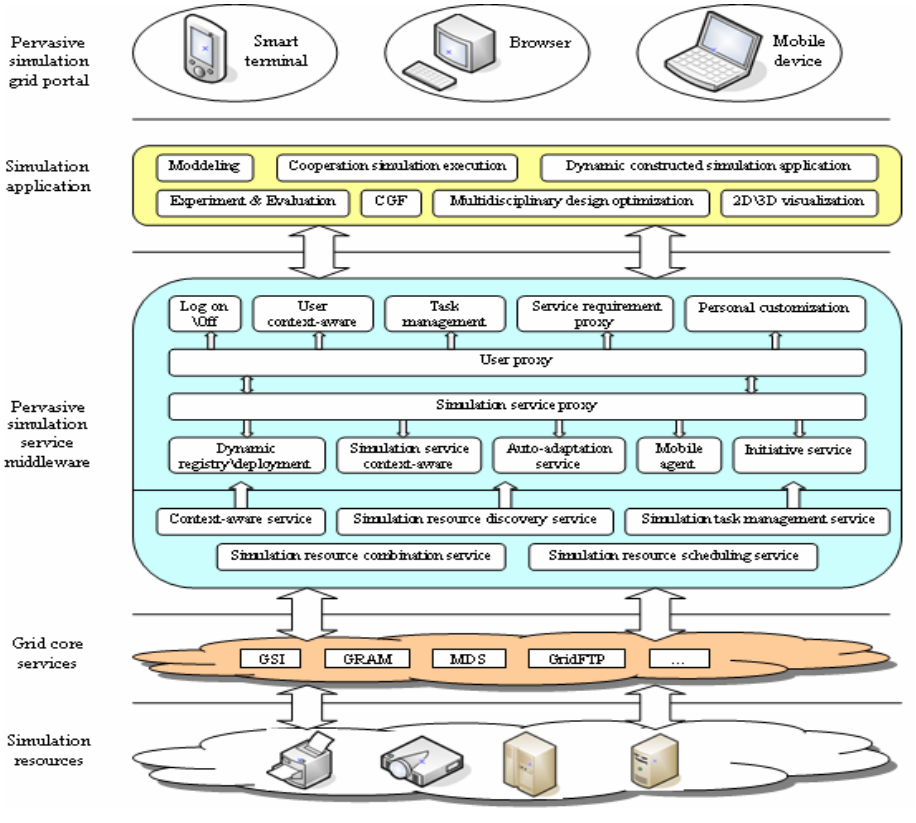


Fig. 1. The architecture of pervasive simulation grid

2. Grid core service level: Provides the grid core services based on the open grid service environment, which was provided by the grid middleware. The grid core services are consist of resources deployment and registration services, resources scheduling service, resources monitoring and catalog services, and the core management services focused on the job and the resources data.
3. Pervasive simulation service middleware level: The pervasive simulation service middleware is the core service level for construction of pervasive simulation application. It downwards invocate the core service interface which the grid middleware provides and shields the complex and heterogeneous resources, computing environment, network environment to realize the virtualization, service encapsulation, transparentness of the simulation resources. Upwards it provided one kind pervasive and transparent management mechanism for the simulation service and task. The management mechanism is including the simulation service dynamic discovery, combination, scheduling service, the simulation service run-time environment and the user context-aware service, the simulation service auto-adapted service, the simulation service proxy, as well as provides the services of personal customizing simulation service and simulation task management service for the user.
4. Simulation application level: All kinds of simulation applications, which are complying with the grid service standard, orienting the pervasive computing environment, developed and deployed based on the pervasive simulation service middleware, share all kinds of simulation resources from the simulation grid, and provide dynamically constructed, highly effective, transparent, collaborative and distributed simulation run-time environment.
5. Pervasive simulation grid portal level: Provides not only the traditional accessing tool\ portal based on the browser and the desktop to application personnel, but also an unified mechanism, through which people can access to each kind of simulation resources from pervasive simulation grid using each kind of terminals (including mobile equipment, intelligent terminal and so on). It is the key component to realize the new pervasive simulation application pattern, which is human-centered, autonomous interaction and transparent invocation. The pervasive simulation grid portal not only provides the visiting function, but also may aware the context information, understand the user's natural language and intention, describe and definite the simulation demand automatically, and generate intelligent proxy for the user, help the user to complete the simulation service request, invocation and the execution of simulation task.

4 Technical Challenges and Approach

In the project named "the application research and development of simulation grid", we have acquired important research achievements, including completed simulation grid prototype named CoSim-Grid 1.0v, and parts of solved key technologies [10] as follows:

1. The technology of Problem-Solving Environments (PSE)

The Simulation Grid PSE provides simulation services for user as follow: to establish the high level model of simulation systems, which include the static structure and dynamic behavior; to carry out the multidisciplinary optimization design; to develop

and deploy the simulation model services; and to submit the simulation jobs and control its execution.

2. The technology of Web enabled HLA/RTI
COSIM-RTI is developed based on the technology of Web enabled HLA/RTI. It acts as the run time infrastructure for the simulation systems, which are deployed in the grid/WAN environment. COSIM-RTI realizes the simulation time management, data communication and cooperation among federates,

3. The technology of resource service middleware GOS
Simulation grid uses GOS as its resource service middleware. It hides the heterogeneity and dynamic features of legacy resources (computing resources, storage resources, software resources, data resources, and etc.), provides virtualization of grid resources, and application interface and user application/development environment, which are low level technology independent and hardware independent.

4. The technology of service encapsulation for simulation resources
The software tools, simulation model resources, as well as computation resources are shared through the service encapsulation technology.

5. The technology of simulation models discovery based on semantic
The ontology technology is introduced and the semantic-based, universal and standard description ontology is defined to implement the semantic description for simulation models. The semantic matching algorithm based on the ontology inference is presented. And the simulation model discovery service based on semantic is realized.

6. The technology of dynamic simulation models composition
In simulation grid, the simulation tasks are dynamic organized from distributed simulation models around. In CoSim-Grid 1.0v, the simulation tasks are described using High Level Modeling Language provided by the problem-solving environments, and executed by the simulation engine service, which is the controller to implement organization and scheduling of the simulation systems.

7. The technology of security mechanism and user management for simulation grid
Security mechanism of Cosim-Grid 1.0v is based on GSI, which is accomplished by CA system (Gridshield) of GOS including providing certificate, user management compatible to Globus format, accessing control of resources and etc.

At present the Cosim-Grid 1.0v has had some pervasive characteristic, but it is insufficient to help us to construct a vigorous and healthy pervasive simulation application system. Above this foundation we need to introduce some advanced pervasive computation technologies for strengthening the mobility, independency, intelligence and fault-tolerant of the simulation grid, and providing the implicit interaction pattern for the user. The following article will discuss some technical challenges and the approaches which are involved in the process of the PSG realization.

4.1 Context-aware Technology

The earliest definition of context-aware was proposed by Schilit and Theimer [11]: The context-aware system is that can sense changes of user environment and make the corresponding adjustment. The researchers call this kind of technology which can

sense the user's context information then regulate system's behavior as context-aware technology.

In pervasive simulation grid, the system not only needs to sense the context of the user, but also the context of the run-time environment of the simulation resources or services. User context includes user's own information (such as body temperature, language, posture, mood and so on) and user's environment information (such as temperature, humidity, latitude and longitude, light, noise and so on). Simulation resources and services context include the computation ability of controllable computation resources and equipment, network connection situation, network load situation and so on.

The applications of context-aware technologies in pervasive simulation grid are presented as follows:

1. Context inquiry: The user can inquire the related context information by tracking the changes of simulation resources context;
2. Service trigger: Triggers the execution of specific service when the context satisfies some kinds of situation;
3. Service selection: Uses context as the condition for selecting the simulation resources and services;
4. Initiative service: Provides the simulation services and information for the user initiatively based on the context;
5. Implicative input: Uses the context information automatically as service input without the user's participation.

At present, the tendency of context-aware technology is developing towards middleware. There are some representative context-aware middleware, such as: CoBrA [12] based on intelligent agent infrastructure, service-oriented SOCAM [13], object-oriented ContextTookit [14] and so on.

In pervasive simulation grid, we use the mature context-aware middleware for gathering the context information of the user and the simulation resources, and providing context information as a service for the pervasive simulation service middleware and application system. Research contents of context-aware technology in pervasive simulation grid include:

1. Context modeling based on ontology;
2. Context-aware and reasoning service.

4.2 Simulation Service Migration Technology

Simulation service migration technology can not only be used in simulation service dynamical fault tolerance, but also can reduce the network band width, overcome the network time delay, support the off-line operation, and provide local service for the user.

1. Fault tolerance migration: By sensing the changes of simulation services run-time environment context, and observing or predicting the problems that will appear, migrates the simulation service and its run-time state parameters to another service container immediately and initiatively, then registers and deploys the simulation service again, restores the original run-time state parameters, and continues the movement of the entire simulation application system.

2. Local service: When the simulation invocation causes the mass network data transmission, and engenders a serious dependence on network bandwidth, the simulation service can be dynamically migrated to the terminal of service consumer and executed. This makes the simulation service invocation to avoid network transmitting, and enhances the efficiency and the reliability of the simulation application system.
3. Off-line service: To some simulation tasks, which only care the final simulation results, but the on-line man-machine interaction, the simulation application execution can get rid of the dependence on the network connection using the simulation service migration technology. The custom-made simulation services can be migrated passively or initiatively from the connection-unstable user ends to the connection-stable network environment, interacting with other simulation resources/services. It will withdraw to the user ends and return the simulation results when the user is on-line.

Internationally, the researchers launch the various researches focused on the service migration problem in the grid computing and pervasive computing. The mobile agent technology is adopted mostly. Mobile agent has not only the generic attribute of the agent, such as autonomy, initiative and intelligence, but also the mobility, auto-adaptation. Mobile agent can be migrated to the resources centralized network nodes or mobiles among those network nodes. For example, the NOMADS [15]designed by Niranjan Suri etc., provides mobile agent services for the CoABSGrid, supports strong and forced mobility, also supports interoperability of mobile agents through GMAS (Grid Mobile Agent System). ZHANG De-gan etc. have presented and designed a seamless transfer strategy [16] based on mobile agent, which solves the problems of seamless transferring method, transferring delay, transferring failure.

Introducing mobile agent technology into the pervasive simulation grid can solve the simulation service migration between each node in grid environment very well. The pervasive simulation service middleware creates a mobile agent for each simulation service, which is deployed in grid. The simulation resource scheduling service makes a decision and informs the mobile proxy based on information about the response time of simulation service, the status of simulation task, and the simulation service run-time context (such as the CPU load, memory load). The mobile agent forbids this simulation service to be used, and saves the simulation service's run-time state information and the 'gar' package. After finding the nearby useable grid container by router service, it migrates to the destination with carrying the run-time state information and the 'gar' package. Then it re-deploys and starts the simulation service, and return the run-time state information and the new URL to the simulation resource scheduling service. The simulation resource scheduling service resumes the run-time information of the simulation service, and continues the execution of the whole simulation system.

The simulation service migration also involves some key issues, for example:

1. The opportunity and the condition of the original node hang up simulation service;
2. The mechanism and method for target node choice;
3. The dynamic deployment of grid service;
4. The simulation time synchronization matter caused by the simulation service migration;

5. The obtainment and preservation of simulation service run-time status;
6. The break point recovery mechanism of the simulation service.

4.3 Autonomous Interaction Technology

The most important characteristic of pervasive simulation grid is to obtain simulation services "transparently". "Transparent" means the simulation services could be obtained breezily and easily. The visit pattern for the simulation services could be provided by system initiatively, even the user may not be conscious. It requires the system has the ability of searching and filtering for appropriate simulation services initiatively to meet the users' demand, or the simulation service can sense the existing demand, and provides the service initiatively. To realize the implicit man-machine interaction pattern, we must study the autonomous interaction technology.

Intelligent Agent [17] is a software entity which has the characteristics of autonomy, social ability, reactivity and pro-activities, etc.

Bring some Agent concepts to pervasive simulation grid will change the service pattern from the traditional human-driving to intelligent man-machine autonomous interaction. The pervasive simulation service middleware will help the user to construct an intelligent multi-agent pervasive simulation environment, which is shown in chart 2.

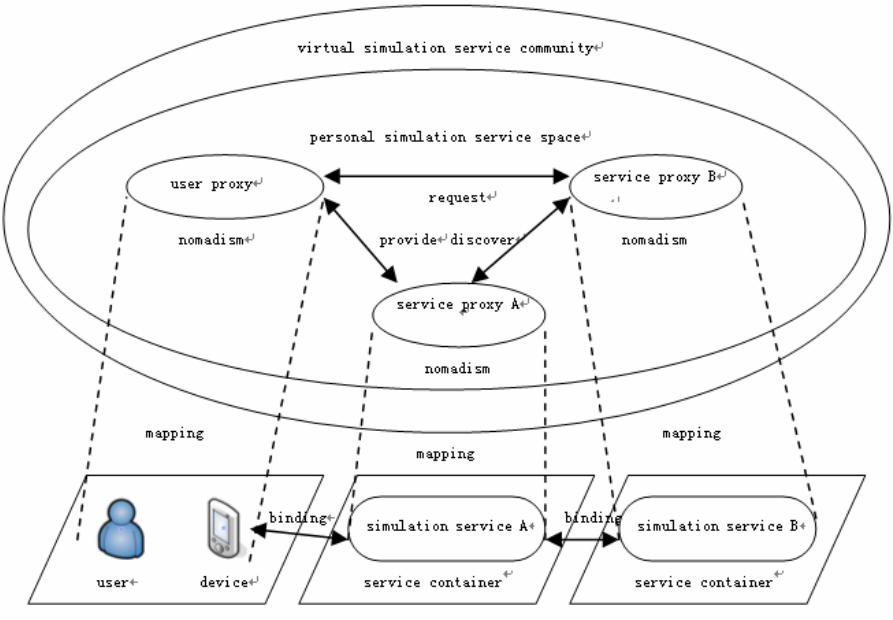


Fig. 2. The multi-agent pervasive simulation grid environment

In this chart, all kinds of simulation related physical resources are encapsulated for grid services. Grid services are signed with semantics information in the registration process. It causes the grid service to be understood possibly. The pervasive simulation

service middleware automatically creates a service proxy for every simulation service. The service proxy is an intelligent software entity, which maps the physical service resource dynamically, can sense the context information of the physical service resource, and has the autonomous ability, the auto-adaptability, can exchange information with other service proxies.

After the user or mobile terminal devices registering in the pervasive simulation grid application portal, the pervasive simulation service middleware automatically creates a user proxy for the user or termination. The user proxy is a smart software entity, which can sense the user's context and automatically makes the response, can find and call the simulation service for the user, can create simulation task for the user, and can exchange information with other service proxies.

The virtual simulation service community is a virtual container of simulation services. It conducts the main functions, such as registration, discovery, and semantic management of service proxies and so on. The user proxy selects corresponding services from a set of services in the simulation service community according to the user's preference, context information, process and the inherent relativity between services. This collection of simulation services is called the current user's individual simulation service space.

In the individual simulation service space, the user proxy carries the user's simulation requirement semantics information, and the simulation service proxy carries the service semantics information. The user proxy corresponds and interacts with the simulation service proxy based on the news transmission mechanism. Based on the semantic information, the service demander can search and discover the service provider, and the service provider can also sense the existence of the service demander, and provides service initiatively. After services matching based on semantic, the simulation service proxy provides the physical address to the user proxy. The user proxy informs the client programs or services to bind the simulation service through this address, and then realizes the invocation for this simulation service.

5 Conclusions and Further Works

Pervasive computing oriented simulation grid technology is a new subject. The discussion and research on this subject is at the beginning stage in domestic and overseas. The pervasive simulation grid also has some problems waiting for further studies. For example, the security mechanism in the pervasive simulation grid, seamless integration with the sensor network, the simulation task dynamic management, simulation service dynamic discovery in pervasive computing environment, simulation resource dynamic deployment for supporting the auto-adaptation, simulation service dynamic combination and scheduling, the simulation grid terminal oriented pervasive computing, and the application pattern of the pervasive simulation grid and so on.

Along with the development of the pervasive computing technology and the extension of the scale and scope of the simulation grid application, the technical connotation and the extension, specially the correlation key technologies and the architecture of the pervasive simulation grid should be continuously extended and enriched.

This article has launched the preliminary studies on the architecture and partial essential technologies of pervasive simulation grid, and has obtained some study fruits.

In the following work, we will realize a pervasive simulation grid application demonstration based on a pervasive simulation scenario, validate these key technologies mentioned in this paper, and launch comprehensive research on the technical issues involved in the pervasive simulation grid.

References

1. Bo-hu, L., et al.: The Development of Modern Modeling & Simulation, Front of China Science (Chinese Academy of Engineering), vol. 6. Higher Education Press, Beijing (2003) (in Chinese)
2. Bo-hu, L., Xing-ren, W., Ke-di, H., Xing-ping, X., Du-zheng, Q., Xu-dong, C.: Further Research on Synthetic Simulation Environment. In: Proceedings of SCSC 2001, San Diego (2001)
3. Kilby, M.: Overview of the DOD High Level Architecture (2000), http://www.dmso.mil
4. Bo-hu, L., Xu-dong, C., Yan-qiang, D., et al.: Research on Service Oriented Simulation Grid. In: Proceedings of ISADS, China, pp. 7–14 (2005)
5. Bo-hu, L., Xu-dong, C., Bao-cun, H., et al.: A New infrastructure of networkitized M&S. M. L. Dertouzos. The future of computing Scientific. American 281(2), 52–55 (1999)
6. Weiser, M.: The Computer for the 21st Century. Scientific American, 265(3), 94–104 (1991)
7. Guang-You, X., Yuan-Chun, S., Wei-Kai, X.: Pervasive/ Ubiquitous Computing. Chinese Journal of Computers 26(9) (2003)
8. Storz, O., Friday, A., Davies, N.: Towards 'ubiquitous' ubiquitous computing: an alliance with 'the grid'. In: First Workshop on System Support for Ubiquitous Computing Workshop (Ubisys 2003) in association with Fifth International Conference on Ubquitous Computing, Seattle, Washington, U.S (October 2003)
9. Parashar, M., Pierson, J.-M.: When the Grid becomes Pervasive: A Vision on Pervasive Grids
10. Bo-hu, L., Xu-dong, C., Bao-cun, H., et al.: Research and Application on CoSim (Collaborative Simulation) Grid. In: Proceedings of MS-MTSA2006, Alberta, Canada (2006)
11. Want, R., Hopper, A., Falcao, V., Gibbons, J.: The Active Badge Location System. ACM Transactions on Information Systems 10(1), 91–102 (1992)
12. Chen, H.: An Intelligent Broker Architecture for Pervasive Context-Aware Systems, PhD thesis, University of Maryland (2004)
13. Gu, T., Pung, H.K., Zhang, D.Q.: A Service-Oriented Middleware for Building Context-Aware Services. Elsevier Journal of Network and Computer Applications (2004)
14. Dey, A.K.: Providing Architectural Support for Building Context-Aware Applications, PhD thesis, Georgia Institute of Technology (2001)
15. Suri, N., Bradshaw, J.M., Breedy, M.R., et al.: NOMADS: Mobility Support for the CoABS Grid
16. De-gan, Z., Guo-cheng, Y., Yuan-chum, S., et al.: Strategy of seamless transfer for pervasive computing. Control and Decision 20(I) (2005)
17. Wooldridge, M., Jennings, N.R.: Intelligent Agents: Theory and Practice. Knowledge Engineering Review 10(2), 1–67 (1995)

A MCS Based Neural Network Approach to Extract Network Approximate Reliability Function

Wei-Chang Yeh, Chien-Hsing Lin, and Yi-Cheng Lin

Department of Industrial Engineering and Engineering Management
National Tsing Hua University
yeh@ieee.org

Abstract. Simulations have been applied extensively to solve complex problems in real-world. They provide reference results and support the decision candidates in quantitative attributes. This paper combines ANN with Monte Carlo Simulation (MCS) to provide a reference model of predicting reliability of a network. It suggests reduced BBD design to select the input training data and opens the black box of neural networks through constructing the limited space reliability function from ANN parameters. Besides, this paper applies a practical problem that considers both cost and reliability to evaluate the performance of the ANN based reliability function.

Keywords: Reliability, Cost, Simulation, Artificial Neural Network.

1 Introduction

Reliability optimization has been a popular area of research and received significant attention during the past four decades [1-11] due to reliability's critical importance in various kinds of systems. To obtain the network reliability of a complex system analytically is a NP-hard problem [2]. Several authors [12-15] have adopted MCS to measuring the reliability for conveniences and accuracies. Yeh [15] combined statistical and simulation techniques to provide a good estimate to the reliability function in the complex network and consider the cost effect to minimize the total cost. After all, MCS is used to provide an estimator for response under specific parameter settings of a complex system. The ANN approach can be regarded as a statistical method. The feature hid within the designed experiment can be learned by ANN based on the cumulative historical data[16].This paper open the black box of neural networks through taking the network structure and weights as the limited space reliability function of complex system networks. The purposed approach can evaluate network reliability precisely using applicable amount of data. In this paper we also compare different input data selection methods and show their consequences.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 287–297, 2007.

2 Notation, Nomenclature, and Assumptions

Notation

$G(V, E, C)$: a network with the node set $V=\{1,2,\ldots,n\}$, the arc set E, the node cost function C, and nodes 1 and n denote the source node and sink node, respectively.

r_i, $\mathbf{r}$: the reliability of node i and the reliability vector $\mathbf{r}=(r_1, r_2,\ldots, r_n)$, respectively.

$P(\bullet),E[\bullet],Var[\bullet]$: the probability, expected value and variance of $\bullet$, respectively.

p, c : the number of nodes in the shortest path and min-cut between nodes 1 and n of G(V, E, C), respectively.

N, M : the number of replications and the number of simulation runs in each replication of the proposed MCS.

$C(r_i)$, $C(\mathbf{r})$: the cost of the node I under r_i and of G(V, E, C) under $\mathbf{r}$, respectively.

$\mathbf{r}^c$, R^c : the lower-bounds of the reliability vector and the total reliability, respectively.

$R(\mathbf{r}),R^m(\mathbf{r}),R^A(\mathbf{r})$: the exact reliability function, the estimator of $R(\mathbf{r})$ obtained from MCS and ANN, respectively, under $\mathbf{r}$.

$\varphi_i(\ )$: the activation function of neuron i.

$\mathbf{W}$, $\mathbf{W}^k$: the weight matrix of the system network in the ANN model, the weight matrix in kth layer, $\mathbf{W}= (\mathbf{W}^1, \mathbf{W}^2,\ldots,\mathbf{W}^k)$

$R_i(\mathbf{r})$: the expect system reliability of when removing i distinct nodes arbitrary (i distinct nodes are failed) under $\mathbf{r}$.

$R_i^m(\mathbf{r})$: the jth simulation replication of $R^m(\mathbf{r})$ under $\mathbf{r}$.

$\beta_{i,j}$: $\begin{cases} 1, & \text{if there is a path from nodes 1 to } n \text{ after removing} \\ & i \text{ distinct nodes at the } j\text{th simulation replication} \\ 0, & \text{otherwise} \end{cases}$

F_i : the probability of i distinct nodes are failed.

3 Model Formulation

The bi-criterion problems of cost and reliability can be simplified into two types of models without setting them as multi-objective decision analysis (MODA) problems. The first type sets the system reliability as the only goal to maximize with components' cost limitations; the second type is dual problems, where the total cost is minimized under the system and components' reliability constraints. This paper continues the problem proposed by Yeh[15], which is formulated as the following:

$$\text{Minimize} \quad C(\mathbf{r})$$

$$\text{s.t.} \quad R(\mathbf{r}) \geq R^c \tag{1}$$

$$\mathbf{r}=(r_1, r_2, \ldots, r_n) \geq \mathbf{r}^c=(r_1^c, r_2^c, \ldots, r_n^c)$$

$R(\mathbf{r})$ is the system reliability function under the specific system configuration **r**. The reliability function of a system constructed by only simple series or parallel components is easy to obtained, but it is very difficult to find in real-life [2]. In order to acquire the reliability in a complex system, it must overcome three NP-hard problems [2] such as implicating all MCs/MPs candidates; verifying all MCs/MPs candidates whether it is a real MC/MP; and finding reliability through combinatorial principles like inclusion-exclusion, sum of disjoint products, and state space enumeration. All of these procedures possess a computational difficulty that, in the worse case, grows exponentially with network size. Owing to this inherent characteristic of reliability, instead of practically large network systems, the reliability function of only a moderate size network can be solved using the above procedures methods instead of a practical, large network system [1, 6-11]. The MCS based Neural approach presented here could provide a way to extract the approximate function instead of MCS-RSM in previous approach.

4 The Monte Carlo Simulation

Simulations are used widely to provide reference results for arduous and sophisticated problems, and Monte Carol simulation methods is one of the most common simulation techniques that researchers often take. To the complicated problems, e.g. network reliability problems, is suitable to address simulation techniques to receive a close solution. Numerous Monte Carlo simulation methods have been developed for reliability analysis. Most of them focused only on how to design sampling plans to reduce the variance in terms of the known MPs/MCs [12-14]. Yeh introduced a MCS procedure for an elaborate system network without consuming the task of knowing MPs/MCs [15]. The MCS adopted in this paper follows the same procedure in [15].

5 The Artificial Neural Network Model

ANN is a versatile model, and it can be applied for forecasting/mapping problems, clustering problem, and classification problems. Problems for different intentions adopt distinct ANN model. In this study, the problem is to estimate the relationship between the response variable $R(\mathbf{r})$ and the factors r_1, r_2, ..., r_n then derive the total cost of the system from cost functions. The selected ANN model is a multilayer feedforward network (MLFFN) with non-polynomial activation functions as Leshno [17] showed that MLFFN approximated any continuous function in any degree of accuracy. In this study, back-propagation was adopted to train the MLFFN structure. It usually can be divided into several layers, which are the input layer, the hidden layer(s), and the output layer; also, each layer contains several neurons. A Neuron has an activation function that can transfer various input agents into an output signal, and then these signals are transmitted from the input layer to the output layer through the hidden layer(s). The activation function provides a relationship between the signals transmitted in and transmitted out. The relationship between inputs a_j and output y_i of neuron i can be expressed as the following:

$$y_i = \varphi_i\left(\sum_{j=1}^{n} w_{ij} a_j + b_i\right) \tag{2}$$

Where w_{ij} is the weight between neuron j and neuron i, b_i is a constant bias, and activation function $\varphi_i(\ )$ is a continuous, differentiable, and monotonically non-decreasing function such as log-sigmoid, hyperbolic tangent . Fig.1 shows a general MLFFN with one hidden layer and related variables. For examples, $\boldsymbol{W}^k$ is the weight matrix that connects different neurons in the kth layer; $\mathbf{r}=(r_1, r_2, \ldots, r_n)$ represents the reliability of each component that constituted the network system; $R^m(\mathbf{r})=Y_1$ shows the desired target. The weight matrix $\boldsymbol{W}$ changes within the iterative training process, and the estimated response $R^A(\mathbf{r})$ close to $R^m(\mathbf{r})$ with $\boldsymbol{W}$ changes. After training, elements in the weight matrix $\boldsymbol{W}$ and the network structure would provide an approximate reliability function that precisely estimates the overall reliability of the system network.

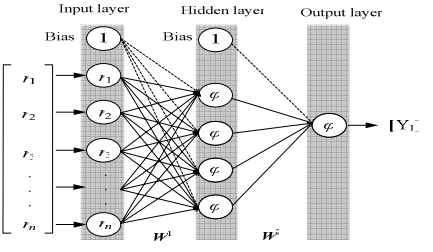


Fig. 1. The simple MLFFN with one hidden layer

While taking ANN as a methodology solves various problems, there are some variables that effect ANN model fitting the problem descriptions. First, variables related to ANN structure needed to be determined, including the number of layers in the ANN model, the amount of nodes in each layer, and connectivity between each layer. The number of layers is determined by the problem complexity; complex problems applied sophisticated ANN structures. However, one hidden layer is usually enough. Moreover, the number of nodes in hidden layer is also identified by problem requirements; a few nodes would lead poor fitting results but train in short time. Finally, connectivity is the degree that nodes communicate with each other. The general ANN we often used is full connected to maintain the interactions between each node.

Second, variables associated to learning efficiency: initial weights, learning rate, and momentum are typical parameter concerning about the learning speed and robustness. Different learning types apply heterogeneous learning parameters. Last, the stopping criteria of ANN, numbers of epochs, MSE of testing, are related to when to stop the training process. These variable that effect ANN model consumed a lot time and effort to determine the optimal parameter. Therefore, this paper studies the selection of training pattern and how to obtain the approximate reliability function through MCS based ANN models. The topics about how to determine the best parameter in ANN model are not included in it.

6 Example and Comparison

In this section, the paper took one benchmark example to explain how to induce the proposed method and showed experimental results. The Benchmark example network shown in Fig. 2 is the modified ARPANET, which is the most frequently cited example [19-23]. This network contains 7 nodes and 9 arcs and two arcs are bidirectional. Before feeding ANN with input training sets, it needs MCS to provide reference responses under the selected input network settings. The ANN take MCS results as training targets and the iteratively training process drives predicting output of ANN close to the training targets. After training, an approximate reliability function is provided by the network weight matrix and structure under limited range reliabilities. With the cost functions listed in Table 7, the nonlinear programming technique solves the cost minimization problem in Eq. 1.

In this example, the network component modules with reduced BBD were adopted as the training sets showed in Eq. 3. The reason that reduces BBD matrix is the original one repeat center points and relative axial points to produce response surface, and these repetitions neither increase the accuracy of neural model nor save time and cost collecting simulation experiment data. Each represents 2^3 different combinations of r_i from Row 1 to row 15 of original BBD matrix, Row 1 to Row 3 and Row 13 to Row 15 are the same ,and the last row represents ten replications of the center point. The reduced BBD matrix saved 33 simulation experiments from original BBD matrix, which is unnecessary in ANN approach.

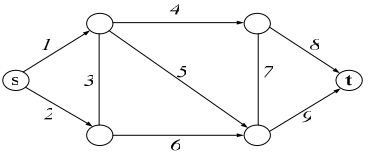


Fig. 2. The modified ARPA network

Table 1. The training data sets and the corresponding approximate reliability using MCS after total 50000 replications

I	r_{i1}	r_{i2}	r_{i3}	r_{i4}	r_{i5}	r_{i6}	r_{i7}	r_{i8}	r_{i9}	R_i^{m}(Std. Error)
1	0.8	0.9	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.96574(0.00248)
2	0.99	0.9	0.9	0.8	0.9	0.9	0.8	0.9	0.9	0.98322(0.00144)
:	:	:	:	:	:	:	:			:
:	:	:	:	:	:	:	:			:
95	0.9	0.9	0.8	0.9	0.99	0.9	0.99	0.9	0.9	0.97776(0.00230)

The training sets are selected from the reduced BBD matrix and The corresponding coded level of the reliability of node i in ith row is explained in Eq.(4), where d_{ij} is in the ith row and jth column of Eq.(3).Follow the procedures of MCS presented in part 4 could obtain the estimator $R^{\mathrm{m}}(\mathbf{r})$ of the system reliability $R(\mathbf{r})$. Table 2 shows the

simulation results of each component module in reduced BBD matrix after M=5000 simulation runs and N=10 replications. All of these 97 data sets are used to training the ANN model, modifying the connection weights between each neuron, and generalizing the ANN model to real network reliability function.

$$\textit{Reduced Box-behnken Design}=\begin{bmatrix} \pm1 & 0 & 0 & \pm1 & 0 & 0 & \pm1 & 0 & 0 \\ 0 & \pm1 & 0 & 0 & \pm1 & 0 & 0 & \pm1 & 0 \\ 0 & 0 & \pm1 & 0 & 0 & \pm1 & 0 & 0 & \pm1 \\ \pm1 & \pm1 & \pm1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \pm1 & \pm1 & \pm1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \pm1 & \pm1 & \pm1 \\ \pm1 & 0 & 0 & 0 & \pm1 & 0 & 0 & 0 & \pm1 \\ 0 & 0 & \pm1 & \pm1 & 0 & 0 & 0 & \pm1 & 0 \\ 0 & \pm1 & 0 & 0 & 0 & \pm1 & \pm1 & 0 & 0 \\ \pm1 & 0 & 0 & 0 & 0 & \pm1 & 0 & \pm1 & 0 \\ 0 & \pm1 & 0 & \pm1 & 0 & 0 & 0 & 0 & \pm1 \\ 0 & 0 & \pm1 & 0 & \pm1 & 0 & 0 & \pm1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{3}$$

$$r_{ij} = \begin{cases} 0.99, & \text{if } d_{ij} = 1 \\ 0.9 \ , & \text{if } d_{ij} = 0 \\ 0.8 \ , & \text{if } d_{ij} = -1 \end{cases} \tag{4}$$

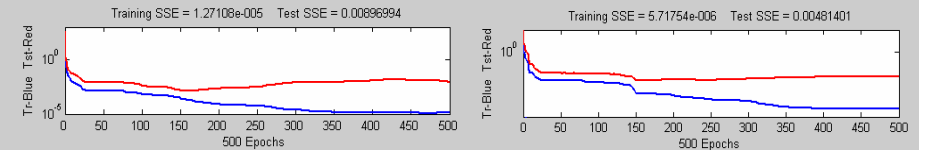


Fig. 3. Training result of BBD designed training sets and 97 random training sets, respectively

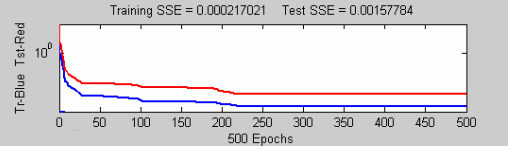


Fig. 4. Training result of reduced BBD training sets

Table 2. The basic configurations related to ANN model in ARPANET example

Number of hidden layer	Single hidden layer
Network structure	9-9-1
Training function	Bayesian regularization
Adaptive learning function	Gradient descent with momentum
Performance evaluation	Sum of squared error
Transfer function	Log-sigmoid function
Connectivity	Full connected
Training epochs	500

The experimental results for different input data selection techniques are shown in Figure 3-5. The testing sets are 500 random selected data sets with exact reliability; they are used to testing the generalization capability. The exact reliability function of this benchmark example is presented in Eq.(5) which is derived from all MPs/MCs with SDP in [21].

$$\begin{aligned} R(r) = {} & r_1 r_4 r_8 + r_2 r_5 r_8 (\overline{r_1} + r_1 \overline{r_4}) + r_2 r_6 r_9 (\overline{r_1}\,\overline{r_5} + \overline{r_1} r_5 \overline{r_8} + r_1 \overline{r_4}\,\overline{r_5} + r_1 \overline{r_4} r_5 \overline{r_8} + \\ & r_1 r_4 \overline{r_8}) + r_1 r_3 r_5 r_8 (\overline{r_4}\,\overline{r_2}) + r_1 r_3 r_6 r_9 (\overline{r_4}\,\overline{r_2}\,\overline{r_5} + \overline{r_4} r_2 r_5 \overline{r_8} + r_4 \overline{r_8}\,\overline{r_2}) + \\ & r_1 r_4 r_7 r_9 (\overline{r_8}\,\overline{r_2}\,\overline{r_3} + \overline{r_8}\,\overline{r_2} r_3 \overline{r_5} + \overline{r_8} r_2 \overline{r_5}) + r_2 r_6 r_7 r_8 (\overline{r_1}\,\overline{r_5}\,\overline{r_9}\,\overline{r_3} + \overline{r_1}\,\overline{r_5}\,\overline{r_9} r_3 \overline{r_4} \\ & r_1 \overline{r_4}\,\overline{r_5}\,\overline{r_9}) + r_1 r_3 r_5 r_7 r_9 (\overline{r_4}\,\overline{r_2}\,\overline{r_8}\,\overline{r_6}) + r_1 r_3 r_6 r_7 r_8 (\overline{r_4}\,\overline{r_2}\,\overline{r_5}\,\overline{r_9}) + \\ & r_1 r_3 r_5 r_6 r_9 (\overline{r_8}\,\overline{r_2}\,\overline{r_3}\,\overline{r_7}) + r_2 r_3 r_7 r_4 r_9 (\overline{r_1}\,\overline{r_5}\,\overline{r_6}\,\overline{r_8}) \end{aligned} \tag{5}$$

Table 2 shows the ANN model configuration in ARPANET example. All of the configuration and training parameters comparing different selection methods are the same, including the initial weights in the weight matrix of system network. Both Figure 3 and Figure 4 show the ANN model closely fitting the training sets and the testing sets to minimum SSE about 150 epochs then slightly increasing. The SSE of testing data sets reveals generalized quality of ANN models. The Bayesian regularization training, which is not easily overtraining and has a good generalization quality, shows stability. The goal of input selection is to find a suitable training design minimize the gap between ANN model and real network reliability model; thus, the experiment results showed the reduced BBD is suitable training sets design for the example.

The ANN model adopted in ARPANET example belongs to the model in Fig.1 which contains 9 input nodes, 9 hidden nodes, and single output neuron in the network. Through the relationship expressed in section 5, it can extract the system reliability function via connection weights and ANN structure. Table 3 shows the trained connecting weights of reduced BBD ANN model. The MCS based Neural approximate reliability function can be expressed as Eq.(6):

$$R^{A}(\mathbf{r}) = 1/\,(1+e^{-ON}) \tag{6}$$

Table 3. The connection weights of ANN model in ARPANET example

Input layer to hidden layer									Bias to hidden layer
5.04	-0.55	-0.47	-0.25	-0.87	-1.52	-0.13	5.83	-1.67	-6.62
-1.19	6.45	-1.05	-0.87	-0.60	-0.61	0.03	2.97	1.68	-7.69
0.19	-0.06	-0.31	0.09	0.11	0.17	-0.55	-0.01	-0.04	0.23
5.14	4.27	-0.92	-2.37	-2.31	-2.35	-0.82	-1.50	-1.68	0.28
0.05	-0.11	-0.40	0.20	0.78	-1.50	2.82	-0.20	0.02	0.11
4.34	-0.13	0.20	-1.39	-0.93	0.13	0.45	-1.31	5.53	-7.80
0.10	-0.06	-0.22	0.03	0.05	0.08	-0.37	-0.03	-0.05	0.12
0.40	0.01	-0.39	0.20	0.11	0.50	-1.02	0.10	0.02	0.42
0.76	1.51	1.34	3.26	2.65	1.65	-0.08	-4.14	-3.91	-0.94
Hidden layer to output layer									Bias to output layer
6.69	7.56	-0.84	-5.50	-2.16	7.22	-0.59	-1.31	5.17	-0.30

ON is the summation of output signals and bias obtained from table 4. The ANN reliability function closely fit the network systems; to illustrate, the mean squared error in the ARPANET example of 500 random testing sets is about 0.0000031568, and it is close to the square of standard errors in simulations. Table 5 and Figure 6 showed the deviation of estimators and the exact reliabilities. Each point in figure 6 represents the average of sorted 10 random testing data sets. From that, it is obviously the ANN model has a good predicting ability.

Table 4. Input expression of each node in the ANN model

Neuron	Inputs
Hidden node1	HN1=5.0403 r_1-0.55464 r_2 -0.46713 r_3 -0.25066 r_4 -0.87329 r_5 -1.5203 r_6 -013181 r_7+5.8272 r_8 -1.6702 r_9 -6.624
Hidden node2	HN2= -1.1898 r_1+6.4542 r_2-1.045 r_3-0.87005 r_4-0.59772 r_5-0.60661 r_6 +0.028639 r_7 +2.9694 r_8 +1.6813 r_9-7.6873
Hidden node3	HN3= 0.18577 r_1 -0.061609 r_2 -0.30566 r_3 +0.091155 r_4 +0.11051 r_5 +0.17123 r_6 -0.55271 r_7 -0.010489 r_8 -0.044932 r_9 +0.22899
Hidden node4	HN4=5.1411 r_1 +4.2741 r_2-0.91676 r_3-2.3727 r_4-2.312 r_5-2.3534 r_6-0.82342 r_7-1.4997 r_8 -1.682 r_9 +0.28273
Hidden node5	HN5=0.049276 r_1 -0.11067 r_2 -0.39787 r_3 +0.20034 r_4 +0.77596 r_5 -1.4998 r_6 +2.8248 r_7 -0.19873 r_8 +0.024043 r_9 +0.11386
Hidden node6	HN6=4.3383 r_1 -0.13379 r_2 +0.19856 r_3 -1.3915 r_4-0.93059 r_5 +0.13354 r_6 +0.45347 r_7 -1.3121 r_8 +5.5303 r_9 -7.8025
Hidden node7	HN7=0.0979 r_1 -0.062434 r_2 -0.22269 r_3 +0.033144 r_4 +0.053194 r_5 +0.076184 r_6 -0.371 r_7 -0.026324 r_8 -0.046117 r_9 +0.1185
Hidden node8	HN8= 0.40117 r_1 +0.0082036 r_2 -0.39113 r_3 +0.19536 r_4 +0.11311 r_5 +0.50021 r_6-1.0205 r_7 +0.098826 r_8 +0.015803 r_9 +0.41894
Hidden node9	HN9= 0.76026 r_1 +1.5147 r_2 + 1.3353 r_3 +3.2614 r_4 +2.6542 r_5 +1.646 r_6-0.084369 r_7 -4.1423 r_8-3.9056 r_9 -0.93626
Output node	ON=6.6911/(1+e^{-HN1})+ 7.5628/(1+e^{-HN2})-0.83881/(1+e^{-HN3}) -5.4959/(1+e^{-HN4})-2.1628/(1+e^{-HN5}) +7.2165/(1+e^{-HN6})- 0.59012/(1+e^{-HN7})- 1.3105/(1+e^{-HN8})+ 5.174/(1+e^{-HN9}) -0.30263

Table 5. The comparison table of deviation of $R^{m}(\mathbf{r})$ and $R^{A}(\mathbf{r})$

Item (500 testing data)	Max Deviation	Min Deviation	Mean Deviation	MAD	MSE
MCS	0.0024	1.571E-06	-0.0042	0.0006	4.99E-07
ANN	0.0082	5.46E-06	0.0003	0.0013	3.15E-06

The cost function of each component is showed in Table 6. It satisfies the general property of cost functions; that is, cost is a monotonically increasing function of component reliability. The object of the formulated model presented in Eq. 1 is minimizing the total cost of the network system; hence, the total cost of the network system is assumed to be the summation of the cost of all components. Table 7 showed the final results of the model considered both cost and reliability constraints; R^{c} and r_i^{c} of Table 7 listed in the 1st row are the constraints of reliability. To evaluate the

efficiency of ANN approximate reliability function relies on the difference of final results that acquired through the exact reliability function and the ANN one. MATLAB optimization toolbox is adopted to solve the nonlinear programming model. From Table 7, it shows the ANN approximate reliability function performed a good result as the exact reliability function did.

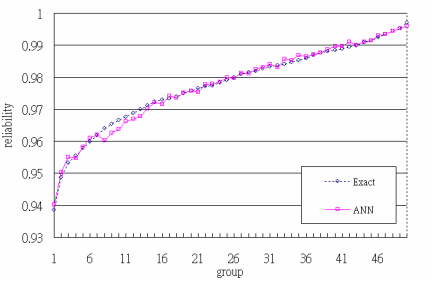


Fig. 5. Comparison of ANN estimate and exact reliability

Table 6. The cost function of each component

node i	α_i	β_i	$C(r_i)=\alpha_i-\beta_i\times\ln(1-r_i)$	node i	α_i	β_i	$C(r_i)=\alpha_i-\beta_i\times\ln(1-r_i)$
1	120	14.7	$120-14.7\times\ln(1-r_1)$	6	100	9.9	$100-9.9\times\ln(1-r_6)$
2	120	14.7	$120-14.7\times\ln(1-r_2)$	7	90	8.75	$90-8.75\times\ln(1-r_7)$
3	90	8.75	$90-8.75\times\ln(1-r_3)$	8	160	12	$160-12\times\ln(1-r_8)$
4	100	9.9	$100-9.9\times\ln(1-r_4)$	9	160	12	$160-12\times\ln(1-r_9)$
5	65	5.64	$65-5.64\times\ln(1-r_5)$				

Table 7. The results solved by MATLAB

	Result	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	$R^A(\mathbf{r})$	$R(\mathbf{r})$	$C(\mathbf{r})$
R^c=.9, r_i^c=.8	ANN	0.800	0.800	0.800	0.800	0.976	0.800	0.800	0.800	0.800	0.900	0.908	1172.10
	Exact	0.800	0.800	0.800	0.800	0.800	0.800	0.800	0.800	0.800	0.874	0.900	1160.05
R^c=.95, r_i^c=.8	ANN	0.800	0.876	0.800	0.800	0.912	0.861	0.800	0.824	0.934	0.950	0.953	1190.08
	Exact	0.800	0.901	0.800	0.800	0.828	0.800	0.800	0.919	0.800	0.946	0.950	1182.01
R^c=.99, r_i^c=.8	ANN	0.968	0.943	0.857	0.840	0.973	0.983	0.829	0.903	0.938	0.990	0.991	1270.65
	Exact	0.800	0.980	0.800	0.800	0.967	0.800	0.800	0.984	0.800	0.984	0.990	1234.65

7 Conclusions

This paper suggests a reduced BBD way to select the input training data and Bayesian regularization training method to train data; in addition, it shows that the suggested method has a better generalization ability and stability. After training, the approximation

reliability function is constructed from the ANN weight matrix, the transfer functions, and the ANN structure. The advantage that ANN could closely predict the corresponding outputs is the reliability lies between the range of 0 and 1, and it is needless to execute data normalization before inducing ANN. It avoids the distortion of real data in the information normalization process and sustains the good generalization results. Therefore, the MCS based ANN approximate reliability function is a good estimator to the real reliability function.

References

1. Aven, T.: Some considerations on reliability theory and its applications. Reliability engineering & systems safety 21, 215–223 (1988)
2. Hoyland, A., Rausand, M.: System Reliability Theory: Models and Statistical Methods. John Wiley and Sons, Chichester (1994)
3. Huang, H.Z.: Fuzzy multi-objective optimization decision-making of reliability of series system. Microelectronics and Reliability 37, 447–449 (1997)
4. Lin, J.S.: Reliability evaluation of capacitated-flow networks with budget constraints. Lie Transactions 30, 1175–1180 (1998)
5. Majety, S.R.V., Dawande, M., Rajgopal, J.: Optimal reliability allocation with discrete cost-reliability data for components. Operations Research 47, 899–906 (1999)
6. Yeh, W.C.: A revised layered-network algorithm to search for all d-minpaths of a limited-flow acyclic network. IEEE Transactions on Reliability 47, 436–442 (1998)
7. Yeh, W.C.: A simple algorithm to search for all d-MPs with unreliable nodes. Reliability Engineering & System Safety 73, 49–54 (2001)
8. Yeh, W.C.: Search for MC in modified networks. Computers & Operations Research 28, 177–184 (2001)
9. Yeh, W.C.: A simple approach to search for all d-MCs of a limited-flow network. Reliability Engineering & System Safety 71, 15–19 (2001)
10. Yeh, W.C.: Search for all d-Mincuts of a limited-flow network. Computers & Operations Research 29, 1843–1858 (2002)
11. Yeh, W.C.: A simple method to verify all d-minimal path candidates of a limited-flow network and its reliability. International Journal of Advanced Manufacturing Technology 20, 77–81 (2002)
12. Easton, M.C., Wong, C.K.: Sequential destruction method for Monte-Carlo evaluation of system reliability. IEEE Transactions on Reliability 29, 27–32 (1980)
13. Landers, T.L., Taha, H.A., King, C.L.: A reliability simulation approach for use in the design process. IEEE Transactions on Reliability 40, 177–181 (1991)
14. Kubat, P.: Estimation of reliability for communication computer-networks - simulation analytic approach. IEEE Transactions on Communications 37, 927–933 (1989)
15. Yeh, W.C.: A MCS-RSM approach for network reliability to minimise the total cost. International Journal of Advanced Manufacturing Technology 22, 681–688 (2003)
16. Stern, H.S.: Neural networks in applied statistics. Technometrics 38, 205–214 (1996)
17. Leshno, M., Lin, V.Y., Pinkus, A., Schocken, S.: Muitilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Networks 6, 861–867 (1993)
18. Fausett, L.: Fundamentals of neural networks: architectures, algorithms, and applications. Prentice-Hall, Inc. Upper Saddle River, NJ, USA (1994)

19. Allan, R.N., Rondiris, I.L., Fryer, D.M.: An efficient computational technique for evaluating the cut-tie sets and sets and common-cause failures of complex systems. IEEE Transactions on Reliability 30, 101–109 (1981)
20. Chatelet, E., Dutuit, Y., Rauzy, A., Bouhoufani, T.: An optimized procedure to generate sums of disjoint products. Reliability Engineering & System Safety 65, 289–294 (1999)
21. Soh, S.T., Rai, S.: Experimental results on preprocessing of path cut terms in sum of disjoint products technique. IEEE Transactions on Reliability 42, 24–33 (1993)
22. Soh, S.T., Rai, S.E.: CAREL - computer-aided reliability evaluator for distributed computing networks. IEEE Transactions on Parallel and Distributed Systems 2, 199–213 (1991)
23. Veeraraghavan, M., Trivedi, K.S.: An improved algorithm for symbolic reliability-analysis. IEEE Transactions on Reliability 40, 347–358 (1991)

An Empirical Study on the Power of the Overlapping Serial Test

Xiaoke Xu[1] and Wai Wan Tsang[2]

[1] Department of Computer Science, The University of Hong Kong
xkxu@cs.hku.hk
[2] Department of Computer Science, The University of Hong Kong
tsang@cs.hku.hk

Abstract. Random number generators (RNGs) are commonly used in simulations. The overlapping serial test is an important test that examines the randomness of RNGs. Its theory has been well-developed but its true ability for rejecting poor RNGs is not known. In this paper, we study the power of the test against the RNGs included in the widely spread *GNU Scientific Library*. By systematically varying the choices of parameters of the test, we find a fine-tuned version that rejects 29 RNGs out of the total of 57 in the library. We will like to warn users not to use these 29 RNGs.

1 Introduction

Nowadays, random number generators (RNGs) are routinely used in simulations and key generation in cryptographic applications. As poor generators will lead to biased results in simulations and successful guesses on cryptographic keys, the quality assurance of generators has drawn much attention ever since computers began to thrive. A basic requirement of a good generator is that it passes all commonly known statistical tests of randomness. The *overlapping serial test* is an important test of randomness developed by Good in 1951 [1]. It is recommended by Knuth [2] and Marsaglia [3] [4]. It is also called the *overlapping m-tuple test* by Marsaglia.

Consider an experiment whose outcome is $0, 1, ..., d-1$ with probability $P(0)$, $P(1), ..., P(d-1)$, respectively. For the efficiency in representing samples and in generating samples in our experiments, d is restricted to be powers of 2 and $k = \log_2 d$ is an integer. The experiment is repeated n times and the outcomes are $Y_1, Y_2, ..., Y_n$. The overlapping serial test treats the outcomes as a cyclic string and examines the distributions of two kinds of overlapped samples. The first kind of samples consist of t outcomes. They are $S_1 = (Y_1, Y_2, ..., Y_t)$, $S_2 = (Y_2, Y_3, ..., Y_{t+1})$, ..., and $S_n = (Y_n, Y_1, ..., Y_{t-1})$. The second kind of samples consist of $t-1$ outcomes. They are $S'_1 = (Y_1, Y_2, ..., Y_{t-1})$, $S'_2 = (Y_2, Y_3, ..., Y_t)$, ..., and $S'_n = (Y_n, Y_1, ..., Y_{t-2})$. The number of samples in both kinds are equal to n. The test statistic is the difference between two *Pearson* forms [5] [6].

$$V = \sum_{|\alpha|=t} \frac{(N(\alpha) - nP(\alpha))^2}{nP(\alpha)} - \sum_{|\alpha|=t-1} \frac{(N(\alpha) - nP(\alpha))^2}{nP(\alpha)} \quad (1)$$

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 298–306, 2006.

In the first summation of the formula, $\alpha = a_1 a_2 \dots a_t$ with $0 \leq a_i < d$. $N(\alpha)$ is the number of times that $S_i = \alpha$, for $i = 1, \dots, n$. $P(\alpha) = P(a_1)P(a_2) \dots P(a_t)$ is the probability that a particular sample, say, S_1, is equal to α. The significances of the symbols appearing in the second summation are similar, except that each sample only consists of $t-1$ outcomes. Asymptotically, V follows the chi-square distribution of $d^t - d^{t-1}$ degrees of freedom. Note that when $t = 1$, the test degenerates to the standard chi-square test.

The most important feature of a test of randomness is its power, i.e., its ability to reject poor RNGs. However, because of the difficulty in theory, the powers of these tests are seldom studied analytically. In this paper, we take an empirical approach to explore the power of the overlapping serial test. We first choose a pool consisting of RNGs of different types and of various quality. We then define the stringency of a test as the number of RNGs it fails in the pool. By systematically varying the choices of k and t of the overlapping serial test, we fine-tune the test for maximum stringency. We also take the same approach to compare the stringency of the test with the gorilla test. The gorilla test is chosen because it is one of the most powerful tests of randomness [7]. Moreover, it examines the Y_i's in a way very similar to the overlapping serial test.

Our study reveals that the overlapping serial test is most stringent when $k = 1$. Moreover, t shall be set to the largest possible value, subject to a constraint imposed by the size of the main memory in the platform computer used for testing. Finally, we find that the power of the fine-tuned test is at par with the gorilla test. The overlapping serial test rejects 29 out of the 57 RNGs in the pool while the gorilla test rejects 28. The two sets of rejected RNGs are mostly overlapping. We would like to alert users not to use these RNGs in simulations.

2 The GSL Pool for Gauging Stringency

We would like to have a pool of RNGs such that a powerful test will fail more RNGs in the pool than a weaker one. The pool shall contain RNGs of different types and of various levels of randomness. One adequate candidate is the RNGs contained in the *GNU Scientific Library (GSL)*. GSL is a free and widely used library for scientific applications. It contains 57 RNGs suggested by experts in random number generation, including 17 *linear congruential generators* (LCG) and 18 *linear feedback shift-register generators* (LFSR). Table 1 lists the names of these generators and the numbers of bits in the words they generate.

The quality of the RNGs in GSL varies in a large range. A good RNG passes all known tests whereas a poor one passes only a few. This phenomenon can be used to distinguish powerful tests from weaker ones. Namely, a weak test can reject only few RNGs in GSL whereas a powerful test can reject many. For examples, the serial correlation test in [2] fails only 1 RNG in GSL but the birthday spacing test [2] [7] fails up to 31.

Let us define the *stringency* of a test as the number of GSL RNGs failed by the test. Suppose the stringency of Test A is higher than Test B and the RNGs failed by Test A is a superset of the RNGs failed by Test B, we conclude that

Table 1. Names and word lengths of the RNGs in GSL

num	name	bits	num	name	bits	num	name	bits	num	name	bits
1	borosh13	32	16	ran1	31	31	random64-glibc2	31	46	ranlxs1	24
2	coveyou	32	17	ran2	31	32	random64-libc5	31	47	ranlxs2	24
3	cmrg	31	18	ran3	29	33	random8-bsd	31	48	ranmar	24
4	fishman18	31	19	rand	31	34	random8-glibc2	31	49	slatec	22
5	fishman20	31	20	rand48	32	35	random8-libc5	31	50	taus	32
6	fishman2x	31	21	random128-bsd	31	36	random-bsd	31	51	transputer	32
7	gfsr4	32	22	random128-glibc2	31	37	random-glibc2	31	52	tt800	32
8	knuthran	30	23	random128-libc5	31	38	random-libc5	31	53	uni	15
9	knuthran2	31	24	random256-bsd	31	39	randu	31	54	uni32	31
10	lecuyer21	31	25	random256-glibc2	31	40	ranf	32	55	vax	32
11	minstd	31	26	random256-libc5	31	41	ranlux	24	56	waterman14	32
12	mrg	31	27	random32-bsd	31	42	ranlux389	24	57	zuf	24
13	mt19937	32	28	random32-glibc2	31	43	ranlxd1	32			
14	r250	32	29	random32-libc5	31	44	ranlxd2	32			
15	ran0	31	30	random64-bsd	31	45	ranlxs0	24			

Test A is more powerful than Test B. The superset condition does not necessarily hold between any two tests as they may check different deficiencies of RNGs. For example, the GSL RNGs rejected by the gorilla test does not have substantial overlap with that of the birthday spacing test. However, in the cases where two tests inspect similar features of samples, the superset condition often holds. Take, for example, the overlapping serial test and the gorilla test. These two tests extract same samples from RNG outcomes but compute different statistics. The RNGs rejected by the two tests are mostly overlapped.

As test outcomes are probabilistic in nature, the superset condition may not hold even if Test A is truely more powerful than Test B. Thus, we will only use this condition as an auxiliary reference instead of a necessary requirement in determining whether a test is more powerful than another.

The GSL pool is a useful tool for tuning the parameters of a test because different versions of the test inspect samples in the same way. The RNGs failed by a low stringency version are likely a subset of the RNGs failed by a higher stringency version.

3 Two-Phase Testing

In this section, we describe the procedure of examining the outcomes of an RNG using the overlapping serial test. First, we describe how to examine a bit sequence using the test. Second, we explain how to extract bit sequences from the RNG outcomes and how to combine the results from testing the bit sequences into a single statistic, and eventually, how to determine pass or fail.

Consider applying the overlapping serial test on a bit sequence $b_1 b_2 b_3 ... b_m$. m is a multiple of k, and $n = m/k$. Each Y_i in Formula (1) is mapped into k bits. S_i consists of kt bits while S'_i consists of $k(t-1)$ bits. As an example, suppose we examine a bit sequence of length $m = 2^{24}$ with the test $[k = 4, t = 6]$. The

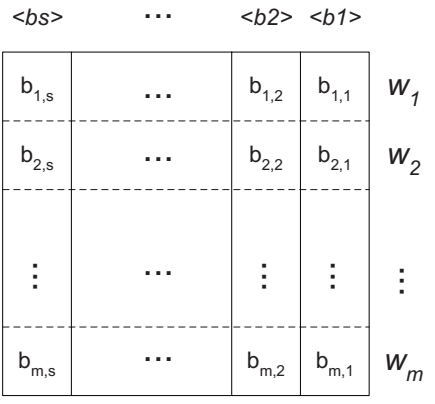


Fig. 1. The bit sequences extracted from the outcomes of an RNG

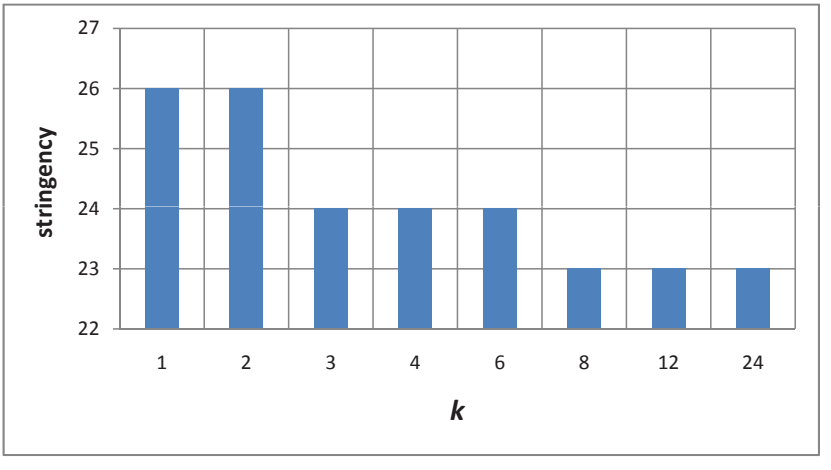


Fig. 2. The stringency of the overlapping serial test against k for $m = 2^{24}$, $t = 24/k$

sample size then is $n = 2^{22}$, while outcomes are $Y_1 = b_1b_2b_3b_4$, $Y_2 = b_5b_6b_7b_8$, ... and $S_1 = (Y_1, Y_2, Y_3, Y_4, Y_5, Y_6)$, $S_2 = (Y_2, Y_3, Y_4, Y_5, Y_6, Y_7)$, ..., etc. V is computed according to Formula (1) and its value is substituted in the chi-square distribution of $d^t - d^{t-1}$ degrees of freedom. If the bit sequence is truly random, the resulting value, $p = Chisq(V, d^t - d^{t-1})$, is a uniform random number in [0, 1). If the samples, S_i's and S_i's, are too evenly distributed, p will tend to be close to 0. If the samples are too unevenly distributed, p will tend to be close to 1.

Now, we are ready to describe a two-phase procedure for examining an RNG. Suppose $w_1, w_2, ..., w_m$ is a sequence of words generated by the RNG and each word consists of s bits. With reference to Figure 1, b_{ij} is the j^{th} bit of w_i. $< bi >$, $1 \leq i \leq s$, is the bit sequence formed by concatenating the i^{th} bits of each word (bits in the i^{th} column).

Table 2. Tables of the p_i's obtained from testing the *lecuyer21* RNG using the overlapping serial test of $[k = 1, t = 24]$ and $[k = 12, t = 2]$

[k = 12, t = 2]		p-value = 0.148					
p1 - p7	0.315	0.012	0.046	0.227	0.524	0.779	0.726
p8 - p14	0.200	0.559	0.425	0.372	0.554	0.068	0.232
p15 - p21	0.506	0.178	0.835	0.352	0.344	0.541	0.717
p22 - p28	0.124	0.034	0.761	0.451	0.425	0.383	0.645
p29 - p31	0.741	0.146	0.770				
[k = 1, t = 24]		p-value = 0.000					
p1 - p7	0.418	0.251	0.026	0.975	0.048	0.071	0.428
p8 - p14	0.409	0.927	0.655	0.041	0.063	0.269	0.417
p15 - p21	***0.005***	0.041	0.177	***0.000***	***0.002***	0.023	0.024
p22 - p28	0.151	0.095	0.381	0.117	0.181	***0.000***	0.033
p29 - p31	***0.007***	0.025	0.110				

In Phase 1, we apply the test to each $< bi >$ to obtain p_i. In Phase 2, the *Anderson-Darling Goodness-of-fit Test (AD Test)* [8] [9], is applied on the p's to check whether they are truly uniformly distributed. We reject the RNG if the resulting p-value is less than the significance level of 0.01.

The upper portion of Table 2 shows the values of the 31 p's obtained in testing the *lecuyer21* RNG using the overlapping serial test of $[k = 12, t = 2]$ in Phase 1. The p-value returned from the AD test on these 31 p_i's in Phase 2 is 0.148. The generator passes this version of the test. The lower portion of the figure shows the p_i's when we test the same generator with the test of $[k = 1, t = 24]$. Note that five of the p_i's are zero or close to zero. The p-value returned by the AD test is 0.000. The generator is rejected.

4 Tuning the Parameters for Maximum Stringency

Consider examining m words generated by an RNG. We want to determine the values of k and t such that the overlapping serial test reaches its maximum stringency. As the test needs to use a large amount of main memory, $2^{kt} + 2^{k(t-1)}$ integer variables, for keeping the numbers of the occurrences of all possible samples of lengths kt and $k(t-1)$, the memory size of the platform computer imposes a key constraint on the choices of k and t. As an example, the test of $[k = 4, t = 6]$ uses $2^{24}+2^{20}$ integer variables. Assuming an integer variable is represented using 4 bytes, the memory needed is 68Mbytes.

4.1 Tuning *k*

Given m, we want to find a choice for k such that the test reaches its maximum stringency. Since the main memory is a key resource, we only consider those versions of the test that use approximately 2^{kt} integer variables. Considering

the cases of $m = 2^{24}$, we gauged the stringencies of the versions of $[k = 1, t = 24], [k = 2, t = 12], [k = 3, t = 8], [k = 4, t = 6], [k = 8, t = 3], [k = 12, t = 2]$, and $[k = 24, t = 1]$. Each of these versions examines all m words being tested. The amount of main memory used lies in $[2^{24}, 2^{24} + 2^{23}]$. The results are shown in Figure 2. There is a clear trend that the stringency decreases as k increases. The version of $k = 1$ rejects most RNGs. It rejects all the generators failed by the version of $k = 12$ and three more: *lecuyer21*, a multiplicative RNG (pp.106-108 of [2]); *minst*, Park and Miller's "minimal standard" RNG; and *ran0*, another *minst* RNG with modified seeding procedure.

The experiment of tuning k was repeated for other values of m. In all cases, the version of $k = 1$ always prevails. We conclude that k shall always be set to 1. In such cases, the amount of overlapping bits between adjacent samples, say, S_i and S_{i+1}, reaches maximum.

4.2 Tuning t

Given m and $k = 1$, we want to find a choice for t such that the test reaches its maximum stringency, subject to the constraint imposed by the memory size of the platform computer. Consider the case where $m = 2^{26}$ and the platform computer has up to 512M bytes free memory. The size of memory imposes a constraint of $t \leq 26$ on the test. Thus, we gauged the stringencies of the versions: $[k = 1, t = 2], [k = 1, t = 6], ...,$ and $[k = 1, t = 26]$. The results are shown in Figure 3. The stringency of the test rises progressively as t increases. Moreover, an RNG rejected by a version of low stringency is also rejected by the more stringent versions. Table 3 shows eight generators in the GSL pool which pass the test when $t = 6$ but are gradually rejected by the test when t increases to 26. The same trend was observed in the experiments with other m values. We conclude that t shall be set to the largest value, subject to the constraint imposed by the memory of the platform computer.

Table 3. Eight GSL generators which pass (P) the test when $t = 6$ but gradually fail (F) when $t = 26$

	t=2	t=6	t=10	t=14	t=18	t=22	t=26
fishman18	P	P	F	F	F	F	F
fishman20	P	P	F	F	F	F	F
lecuyer21	P	P	P	F	F	F	F
minstd	P	P	P	F	F	F	F
ran0	P	P	P	F	F	F	F
random256-bsd	P	P	P	P	P	F	F
uni	P	P	P	P	P	P	F
uni32	P	P	P	P	P	P	F

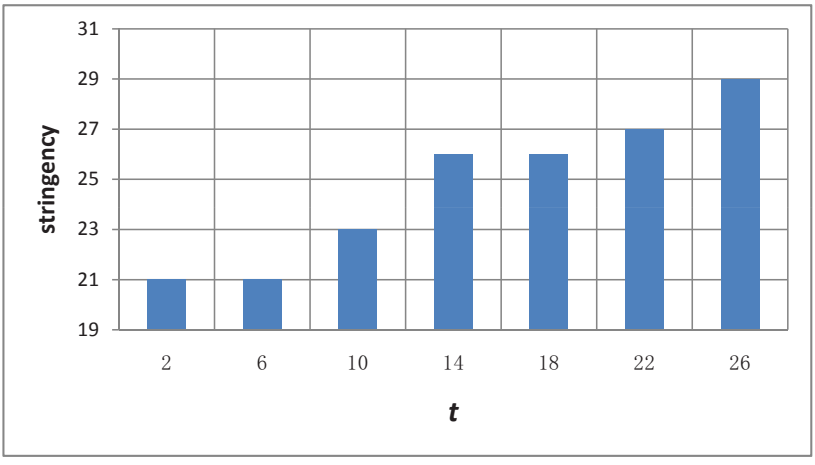


Fig. 3. The stringency of the overlapping serial test against t for $m = 2^{26}$ and $k = 1$

5 Comparison with the Gorilla Test

The gorilla test is the most powerful version of the monkey test and is among the most difficult-to-pass tests for RNGs [7] [10]. It extracts S_i's from $< bi >$ in the same way as the overlapping serial test, but counts the number of all possible S_i's which are missing. Comparing the overlapping serial test with the gorilla test gives us an idea of how powerful the latter really is.

The gorilla test given in [7] examines $m = 2^{26}$ words of an RNG. To be fair, we compare it with the overlapping serial test of $[k = 1, t = 26]$, which examines same number of words. We apply both tests to check the RNGs in the GSL pool. The results are shown in Table 4. The gorilla test rejects 28 out of the 57 RNGs in the pool while the overlapping serial test rejects 29. 27 RNGs are rejected by both tests. Some of these RNGs are widely used in standard C library, e.g., *rand* and *rand48*. The only RNG that fails the gorilla test but passes the overlapping serial test is *knuthran2*, a second-order multiple recursive generator suggested by Knuth. The two RNGs that pass the gorilla test but fail the overlapping serial test are *ran3* and *random256-bsd*. As the two sets of RNGs rejected by the two tests are mostly overlapped, we conclude that the power of the overlapping serial test is at par with the gorilla test.

The overlapping serial test uses an integer variable (4 bytes) for keeping the count of the occurrence of a possible sample, whereas the gorilla test uses only 1 bit for keeping track whether the sample has occurred. Thus, the gorilla test uses far less memory. On the other hand, the gorilla test suffers from a major drawback. The cumulative distribution function (CDF) of its statistic cannot be derived analytically. The p-value of the test is computed from an empirical cumulative distribution obtained via extensive simulation.

Table 4. The results of testing the GSL generators using the overlapping serial test and the gorilla test

Num	Name	OS	Gorilla	Num	Name	OS	Gorilla	Num	Name	OS	Gorilla
1	borosh13	F	F	21	random128-bsd	P	P	41	ranlux	P	P
2	coveyou	F	F	22	random128-glibc2	P	P	42	ranlux389	P	P
3	cmrg	P	P	23	random128-libc5	P	P	43	ranlxd1	P	P
4	fishman18	F	F	24	random256-bsd	***F***	***P***	44	ranlxd2	P	P
5	fishman20	F	F	25	random256-glibc2	P	P	45	ranlxs0	P	P
6	fishman2x	P	P	26	random256-libc5	P	P	46	ranlxs1	P	P
7	gfsr4	P	P	27	random32-bsd	F	F	47	ranlxs2	P	P
8	knuthran	P	P	28	random32-glibc2	F	F	48	ranmar	P	P
9	knuthran2	***P***	***F***	29	random32-libc5	F	F	49	slatec	F	F
10	lecuyer21	F	F	30	random64-bsd	F	F	50	taus	P	P
11	minstd	F	F	31	random64-glibc2	F	F	51	transputer	F	F
12	mrg	P	P	32	random64-libc5	F	F	52	tt800	P	P
13	mt19937	P	P	33	random8-bsd	F	F	53	uni	F	F
14	r250	F	F	34	random8-glibc2	F	F	54	uni32	F	F
15	ran0	F	F	35	random8-libc5	F	F	55	vax	F	F
16	ran1	P	P	36	random-bsd	P	P	56	waterman14	F	F
17	ran2	P	P	37	random-glibc2	P	P	57	zuf	P	P
18	ran3	***F***	***P***	38	random-libc5	P	P				
19	rand	F	F	39	randu	F	F				
20	rand48	F	F	40	ranf	F	F				

6 Future Work

In deriving the CDF of the V defined in Formula (1), the $N(\alpha)$'s are assumed following a joint normal distribution. This assumption is severely violated in the extreme sparse versions of the test, i.e., when m is small and t is very large. So far, we have not yet encountered any noticeable discrepancy between the CDF derived and the true CDF. However, we shall be aware of this possible hazard in practice, and find out how sparse we can go before the theory breaks down.

Acknowledgment

This research is supported by the Hong Kong RGC HKU7143/04E. The authors would like to express their sincere gratitude to Chi-yin Pang for his invaluable comments.

References

1. Good, I.J.: The serial test for sampling numbers and other tests for randomness. In: Proceedings of Cambridge Philosophical Society, vol. 49, pp. 276–284 (1953)
2. Knuth, D.: The Art of Computer Programming, 3rd edn. vol. 2. Addison-Wesley, Reading (1998)
3. Marsaglia, G.: A current view of random number generators. In: Computer Science and Statistics: Proceedings of the 16th Symposium on the Interface, Atlanta, Georgia, pp. 151–158 (1984)

4. Marsaglia, G.: Monkeying with the Goodness-of-Fit Test. Journal of Statistical Software (2005)
5. Pearson, K.: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine 50(5), 157–175 (1900)
6. Tsang, W., Pang, C.Y.: The Mathematics of the Overlapping Chi-square Test. Technical Report TR-2006-12, Department of Computer Science, The University of Hong Kong (September 2006)
7. Marsaglia, G., Tsang, W.: Some difficult-to-pass tests of randomness. Journal of Statistical Software 7(3), 1–8 (2002)
8. Anderson, T., Darling, D.: Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes. The Annals of Mathematical Statistics 23(2), 193–212 (1952)
9. Marsaglia, G., Marsaglia, J.: Evaluating the Anderson-Darling distribution. Journal of Statistical Software 9(2), 1–5 (2004)
10. Marsaglia, G., Zaman, A.: Monkey tests for random number generators. Computers and Mathematics with Applications 26(9), 1–10 (1993)

Idle Resource Supplement Model and Validity Time Designation Model with Reliability Measurement in Grid Computing*

Da Hye Park, Sung Ho Jang, Chang Hyeon Noh, and Jong Sik Lee

School of Computer Science and Engineering
Inha University
Incheon 402-751, South Korea
parkdh@inhaian.net, ho7809@hanmail.net, cromirak@hanmail.net,
jslee@inha.ac.kr

Abstract. Grid computing provides high performance like a super computer through sharing and using distributed heterogeneous computing resources. Grid computing processing time and cost are widely varied since a grid user or a grid middleware can select a variety of distributed heterogeneous resources. Therefore, grid computing absolutely needs a grid resource management method and model. In this paper, we propose two types of resource management model with resource reliability. The first model is the idle resource supplement model. The point of the idle resource supplement model is that adds to idle resources when existing resources can not process jobs. The second model is the validity time designation model which considers grid users. This model processes jobs during validity time that is decided by grid users. This paper evaluates system performance such as utilization, job-loss rate and average turn-around time. And, we estimate experiment results of our models in comparison with those of existing models such as a random model and a round-robin model. The experiment results demonstrate that the two models based on resource reliability measurement improve resource utilization and provide reliable job processing. And, we expect that our proposed models improve grid computing QoS.

1 Introduction

The computer technology has been rapidly developing according to the Moore's Law and has been introducing new technologies. Especially, the internet is one of the most important technologies in the computer technology. The beginning internet was a communication concept between persons in remote site. Recently, it has introduced data sharing and data processing between remote sites. However, it produces large-scaled data which are not solved by a computer, and grid computing [1], [2], [3] has been appeared for solving large-scaled data.

* This research was supported by the IITPA (Incheon Information Technology Industry Promotion Agency), Korea, under the Complex terminal test & service technique development for air distribution RFID equipment.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 307–314, 2007.

The grid computing is the new generation computing which connects geographically dispersed heterogeneous resources such as large-scaled storage, database, and high-technology experiment equipments. That is, the grid computing can solve large-scaled computing problems by geographically dispersed heterogeneous resources. The grid computing has many advantages such as stability, low-in-cost, and flexibility so it has been used to various fields. However, gathering heterogeneous resources produces many difficult problems. Resource unreliability is one of the most critical problems.

Resource reliability is that grid resources can operate without a hitch in the grid computing environment. If it does not guarantee resource reliability, the availability of grid application degrades and performance of distributed computing goes down in the grid computing. Also, in order to improve processing rate of job, the grid computing needs to evaluate performance of resource. Performance of resources can be evaluated by resource reliability. But, the grid computing is difficult to evaluate resource reliability because it consists of distributed heterogeneous resources. So, we propose resource management models based on resource reliability measurement.

This paper is organized as follows. Section 2 presents related work. Section 3 presents the resource management models with resource reliability measurement. Section 4 presents analysis results of job-loss rate of models before actual experiment. And, Section 5 demonstrates efficiency of our models in comparison with existing models. The conclusion is in Section 6.

2 Related Work

2.1 Grid QoS

In the grid computing, to guarantee Grid Quality of Service (QoS) [4], [5] is necessary for application. It can increase the grid computing utilization of users, if it guarantees Grid QoS. The grid computing classifies Grid QoS into four parameters.

Security. It is a parameter about security level in the grid computing. The Resource provider has the right to access service according to security management strategy. And, it provides a message cryptograph and access control.

Performance. It is a parameter to measure utilization and latency. In order to guarantee its parameter, resources have to provide high utilization and low latency.

Integrity. It is a parameter about accuracy of interaction. Appropriate execution of transaction guarantees interaction. Grid resources need to assure success of transaction.

Availability. It is a parameter about possibility that grid resources can process job.

In order to assure these parameters of Grid QoS, the grid computing is needs appropriate resource management method. We propose a grid resource management method with reliability measurement.

2.2 Grid Reliability

The grid computing is dynamically virtual organization and is concept which solves problems through resource sharing. And, processing time and cost is widely changed to choose distributed heterogeneous resources. Therefore, the grid computing absolutely needs a grid resource management method and model. In order to solve grid resource management problem, grid researches are generally utilizing grid model applying economic theory. Typical grid models applying economic theory [6] are the commodity market model, the bargaining model, and the auction model. But, these grid models applying economic theory do not include resource reliability policy for efficient resource allocation. Recently, many grid researches focus on the resource reliability [7], [8], [9].

The grid reliability represents accuracy and precision of model and implies conceptions such as stability, consistency, and predictability. In the grid computing, resource reliability assures resources which can quickly process job for grid users. Namely, the grid resource reliability represents that resources can process job rapidly and can allocate job stably. Thus, if it guarantees the grid resource reliability, it can improve computing performance and increase utilization of the grid application. In the grid environment, to guarantee reliability of distributed heterogeneous resources is very complex and difficult because resources in the grid computing have different policies and are distributed. So, a mechanism which can apply to all grid resources needs for to guarantee grid resource reliability.

Currently, researches for grid resource reliability focus on a trust mechanism [10], [11]. The trust mechanism gives the credits to resource. It uses the credits which reflect the resource reliability during job execution. But, problem of the trust mechanism is that the resource provider may cheat others to increase its trust. So, we evaluate resource reliability based on historical resource status data and apply it to the resource management.

3 Resource Management Model

In this section, we introduce an idle resource supplement model and a validity time designation model based on resource reliability measurement.

The first model is the idle resource supplement model. This model utilizes that input of resources is open in the grid computing. The idle resource supplement model is constructed as shown in Fig. 1. The idle resource supplement model consists of five components. Grid user sends jobs to coordinator. A coordinator calculates resource reliability with resource status data. Then, a coordinator allocates jobs to resource which can process jobs quickly. Resources can process a job at one time. If resource is processing job, it puts job into queue. And, if Queue Size (QS) is changed, resource sends QS to its analyzer. Analyzer measures resource status with QS. And, if resource status is changed, analyzer sends measured resource status duration time to a coordinator. However, if resource that is selected by resource reliability evaluation can not process jobs, this model appends idle resource (IR) and processes jobs. IR has own analyzer like existing resources and also sends its status data to coordinator.

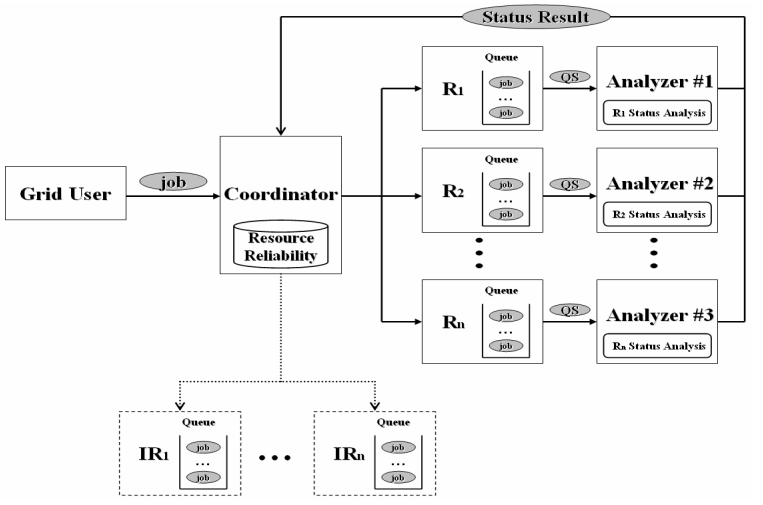


Fig. 1. Architecture of the Idle Resource Supplement Model

The second model is the validity time designation model which considers grid users. This model processes jobs during validity time that is decided by grid users. The validity time designation model is composed of four types of components as shown in Fig. 2. Grid user sends job designating validity time to a coordinator. A coordinator calculates resource reliability according to resource status data and selects a proper resource and sends job which has validity time. If resource is passive, it can process jobs. But, if resource is busy, it put received job into queue. And, jobs in queue are disappeared over validity time. Resources send QS to its analyzer and analyzer measures resource status with QS. If resource status is changed, analyzer sends measured resource status duration time to a coordinator. However, if it is impossible for existing resources to process job, a coordinator does not send jobs to resources and stores jobs into its own queue. A coordinator sends stored jobs when resources can process jobs.

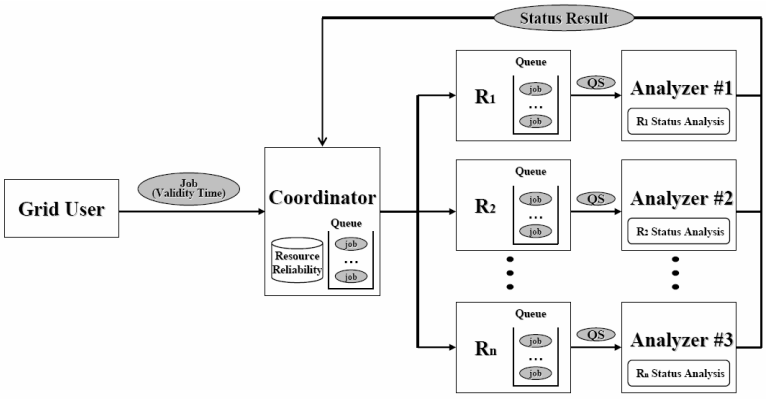


Fig. 2. Architecture of the Validity Time Designation Model

4 Analysis of Job-Loss Rate

We analyzed the job-loss rate of the models before the actual experiment. In order to analysis, we have assumed that all resources are identical with the queue size and the processing time. And, if status of resource is the low performance (if resource exceeds the queue size), it produces job loss. Table 1 shows analysis results of job-loss rate of the idle resource supplement model, the validity time designation model, the round-robin model, and random model.

Table 1. Analysis results of job-loss rate of the models (N; total number of produced job, M; total resource number, L; queue size of resources, P; probability that resource will be the low performance)

Idle resource supplement model	Validity time designation model	Round-robin model	Random model
$\frac{N-(M\times L)}{N}\times P$	0	$\frac{N-(M\times L)}{N}$	P

In above equation, N is total number of produced job and M is total resource number. L indicates queue size of resources and P indicates probability that resource will be the low performance. Job-loss rate of the idle resource supplement model is $\{(N-ML)/N\}*P$, because job-loss rate is produced when all resources are low performance. Job-loss rate of the validity time designation model is 0, because coordinator does not send jobs to resources and stores jobs into own queue in case it is impossible for existing resources to process job. In the round-robin model, job-loss rate is $\{(N-ML)/N\}$, because this model does not select resource measuring reliability of resources. In the random model, job-loss rate is P, because the random model produces job-loss when selected resource has low performance. To this analysis results, we have expected that the validity time designation model will show 0, the idle resource supplement model will show much lower job-loss rate than the random model and the round-robin model.

5 Experiments

In order to demonstrate efficiency of the proposed resource management models using resource reliability, we implemented two models which are the idle resource supplement model and the validity time designation model on the DEVSJAVA modeling and simulation [12] environment. If existing resources can not process jobs, the idle resource supplement model appends idle resources. This model utilizes fluidity of resources in the grid computing environment. The validity time designation model assigns validity time to jobs that are generated by grid user. And, we compared performance of our models to those of the round-robin model and the random model. We implement many resources because the grid computing is composed of variously heterogeneous resources. Also, each of the resources has different processing time and QS. Namely, experiment environment is composed of various resources. We generated about 10000 jobs for reliable experiment and measured performance such as utilization, average turn-around time, and job-loss rate.

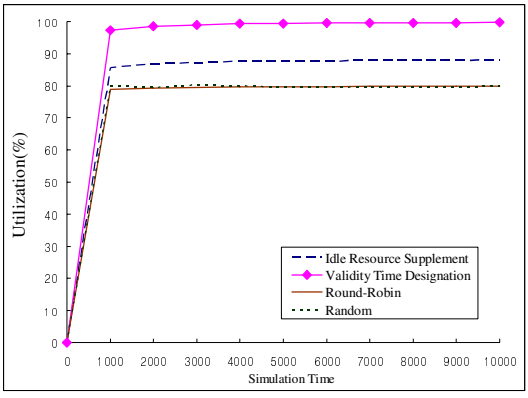


Fig. 3. Comparison of Utilization (Idle Resource Supplement Model vs. Validity Time Designation Model vs. Round-Robin Model vs. Random Model)

The first experiment is to measure utilization of models. Fig. 3 illustrates utilization of four models by simulation time. The validity time designation model shows the highest utilization of four models. And, the idle resource supplement model shows lower utilization than the validity time designation model, but higher utilization than the random model and the round-robin model. This result demonstrates that the resource management model based on resource reliability measurement improves overall system performance.

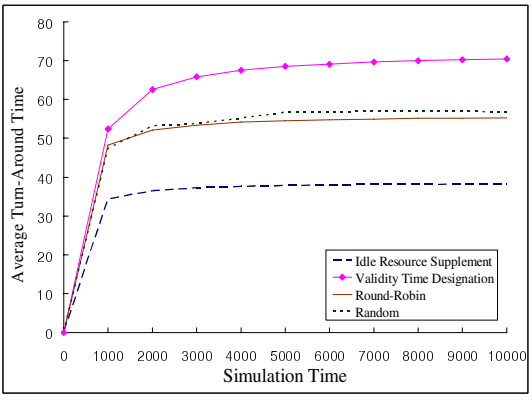


Fig. 4. Comparison of Average Turn-Around Time (Idle Resource Supplement Model vs. Validity Time Designation Model vs. Round-Robin Model vs. Random Model)

The second experiment is to measure average turn-around time. Fig. 4 shows variations of average turn-around time by simulation time. The idle resource supplement model shows the lowest average turn-around time of four models. That is, the idle resource supplement model provides reduced processing time compared with the random model and the round-robin model. However, the validity time designation

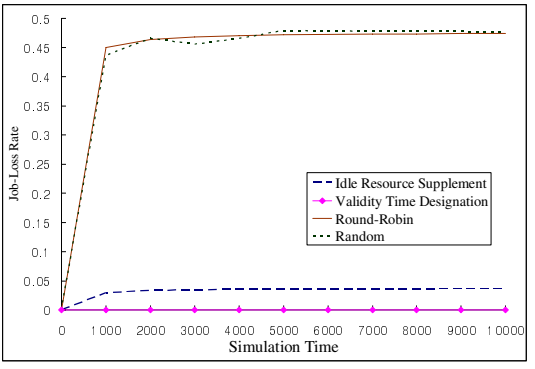


Fig. 5. Comparison of Job-Loss Rate (Idle Resource Supplement Model vs. Validity Time Designation Model vs. Round-Robin Model vs. Random Model)

model shows the highest average turn-around time of the models. Because the validity time designation model processes a lot of jobs compared with the random model and the round-robin model, average turn-around time of the validity time designation model is prolonged.

The third experiment is to measure job-loss rate of models. Fig. 5 illustrates utilization of four models by simulation time. The idle resource supplement model recorded the 3.52% job-loss and the validity time designation model recorded the 0% job-loss while the round-robin model recorded the 47.14% job-loss and the random model recorded the 47.86% job-loss when simulation time is 5000. This result demonstrates that the idle resource supplement model and the validity time designation model are stable and provide trustworthy job process.

6 Conclusion

In order to guarantee efficient job processing in grid computing, this paper proposes two resource management models based on resource reliability measurement. We utilize historical status data of resources in order to measure grid resource reliability. Status of resource is divided into three parts. And, we calculate resource reliability by resource status duration time. We construct two models using the resource management model. The first model is the idle resource supplement model. The point of the idle resource supplement model is that adds to idle resources when existing resources do not process jobs. The second model is the validity time designation model. The validity time designation model processes jobs during validity time that is decided by grid users.

For performance evaluation, we constructed two models on the DEVSJAVA modeling and simulation. And, we estimated our models in comparison with the random model and the round-robin model. These experiment results showed that proposed models provide efficient resource allocation and stable job processing in comparison with the random model and the round-robin model. The resource management based on the resource reliability measurement provides effective job processing and stable resource allocation in the grid computing environment.

References

1. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
2. Foster, I.: What is the Grid? A three point checklist. Argonne National Lab and University of Chicago(2002), http://www-fp.mcs.anl.gov/ foster/Articles/WhatIsTheGrid.pdf
3. Berman, F., Fox, G., Hey, T.: Grid computing: making the global infrastructure a reality. J. Wiley, New York (2003)
4. Al-Ali, R.J., ShaikhAli, A., Rana, O.F., Walker, D.W.: Supporting QoS-based Discovery In Service-oriented Grids. In: Proceedings of 17th International Parallel and Distributed Processing Symposium (IPDPS 2003) (2003)
5. Foster, I., et al.: End-to-End Quality of Service for High-End Applications. Computer Communications, 1375–1388 (2004)
6. Buyya, R.: Economic-based Distributed Resource Management and Scheduling for Grid Computing (April 2002), available at http://www.buyya.com/thesis/
7. Shi, X., Jin, H., Qiang, W., Zou, D.: Reliability Analysis for Grid Computing. In: Jin, H., Pan, Y., Xiao, N., Sun, J. (eds.) GCC 2004. LNCS, vol. 3251, pp. 787–790. Springer, Heidelberg (2004)
8. Li, C., Xiao, N., Yang, X.: Application availability measurement in computational grid. In: Li, M., Sun, X.-H., Deng, Q.-n., Ni, J. (eds.) GCC 2003. LNCS, vol. 3032, pp. 151–154. Springer, Heidelberg (2004)
9. Coit, D.W.: System reliability prediction prioritization strategy. In: Proceedings of Annual Reliability and Maintainability Symposium (2000)
10. Luo, J., Ji, P., Wang, X., Zhu, Y.: A Novel Method of QoS Based Resource Management and Trust Based Task Scheduling. In: Shen, W.-m., Lin, Z., Barthès, J.-P.A., Li, T. (eds.) CSCWD 2004. LNCS, vol. 3168, pp. 21–32. Springer, Heidelberg (2005)
11. Kamvar, S.: EigenRep: Reputation Management in P2P Networks. Technical Report, SCCM-02-16, Stanford University (2002)
12. Zeigler, B.P., et al.: The DEVS Environment for High-Performance Modeling and Simulation. IEEE CS & E 4(3), 61–71 (1997)

Implementation of a Tactic Manager for the Simulation of a Target Motion Analysis between a Submarine and a Surface Ship Warfare

Myeong-Jo Son[1], Doo-Yeoun Cho[2], Ju-Hwan Cha[1], Kyu-Yeul Lee[3], Tae-Wan Kim[3,*], and Jun-Kyu Park[4]

[1] Dept. of Naval Architecture & Ocean Engineering, Seoul National University, Korea
{prokper3,jhcha79}@snu.ac.kr
[2] Dept. of Naval Architecture & Marine Engineering, Mokpo National University, Korea
dycho@mokpo.ac.kr
[3] Dept. of Naval Architecture & Ocean Engineering and Research Institute of Marine System Engineering, Seoul National University, Korea
{kylee,taewan}@snu.ac.kr
[4] Agency for Defense Development, Korea
junkyu@add.re.kr

Abstract. A tactic manager which can change the behavior of a simulation model according to the tactics defined outside of the model has been studied and implemented. Based on DEVS(discrete event system specification) formalism, we generated a simulation model which is equipped with the interface to the tactic manager. To demonstrate the effectiveness of the tactic manager, a target motion analysis in the warfare between a submarine and a surface ship is simulated.

Keywords: Modeling & Simulation, Tactic Manager, Target Motion Analysis, DEVS, Warfare between a Submarine and a Surface Ship.

1 Introduction

It needs tremendous budget and period in developing underwater vehicles such as submarines and torpedoes. So it is necessary to predict the performance of those accurately using modeling and simulation(M&S) for minimizing the developing risk which is caused by trials and errors occurred in design process and construction process and for optimizing the design. M&S is also necessary to predict the effect according to the various tactics in operating process of underwater vehicles.

1.1 The Concept and the Necessity of Tactic Manager

In real world, a submarine is operated by the captain's decision. In simulation, however, the submarine model is controlled by decision data or tactic data which were

* Corresponding author.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 315–326, 2007.

defined by users previously. Tactic Manager takes a role supporting the model to determine the manner of action dynamically according to the predefined tactics.

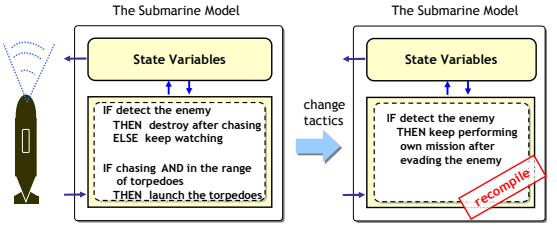


Fig. 1. Tactics implemented inside a model

In case of the simulation model without Tactic Manager, there're problems such as rewriting and modifying the model every time when tactics are changed because tactics should be defined statically in the model (Fig. 1). On the other hand, Tactic Manager takes tactics as input from outside of the model through files and controls actions of the model analyzing those inputs. In other words, Tactic Manager has a benefit such as simulating the various scenarios without modifying the model when tactics are changed, but with modifying tactic definition files only (Fig. 2).

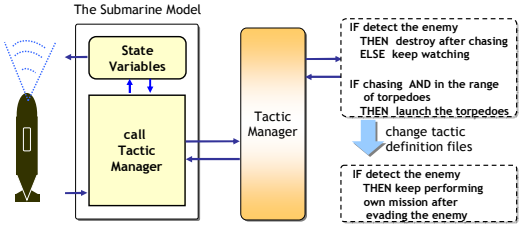


Fig. 2. Separation of tactics from a model

1.2 Paper Structure

In this research, we studied the concept and the definition of Tactic Manager. The simulation model is constructed with DEVS(Discrete EVent system Specification) formalism [1, 5, 6] which has been used widely in discrete event simulation. In addition the interface for passing the parameter between Tactic Manager and the model is added. Table based tactic definition methods were used to simplify implementing of Tactic Manager. To demonstrate the effectiveness of Tactic Manager, a target motion analysis in the warfare between a submarine and a surface ship is simulated.

The structure of this paper is as followed. In chapter 2, the scenario of the warfare between a submarine and a surface ship which will be simulated using Tactic Manager is described. In chapter 3, the concept and the algorithm of Tactic Manager is presented. In chapter 4, simulation results are presented and its applications are discussed. Finally, we derive conclusions and suggest future works in chapter 5.

2 The Scenario Description of the Warfare between a Submarine and a Surface Ship

The own submarine is diesel submarine of type 209, and performs the coastal barrier mission usually, for relatively short cruise range and its peculiar calmness. The barrier mission consists of submarine's moving with fixed patterns in the predefined area and detecting the enemy ship passed across that area, and attacking it after approaching. The enemy surface ship(battleship, supplies transport ship) passes across barrier area of the submarine with high speed. When the submarine detects the enemy surface ship during its barrier mission, it starts TMA(Target Motion Analysis) procedures to estimate the kinematic state(range, bearing, course, velocity and etc) of the target and approaches to the target at the same time. The bearing is the target's direction of relevance to the observer which is measured from true North regarded as 000°. After having gotten the kinematic state of the enemy through TMA, it is examined that the target exists within maximum range of the torpedo. And then, the submarine launches torpedoes when the target exists within a half of the torpedo's maximum range for increasing hitting possibility of torpedoes. Else, the submarine continues the approach process to the target and then launches.

2.1 Target Motion Analysis(TMA) and a Leg

2.1.1 Target Motion Analysis(TMA)

TMA is estimation methods for getting the kinematic state of the target on surface or in underwater [3, 4]. The kinematic state of the target is estimated easily when the active sonar which emits sound waves and analyzes reflection sound waves is used. But TMA is necessary procedure to attack the target after detection for the submarine which should be in stealth and so, use only the passive sonar which detects the enemy with noises radiated from the target. TMA procedure consists of more than two legs. The manner of straight moving in constant course and speed is called a leg [2, 4].

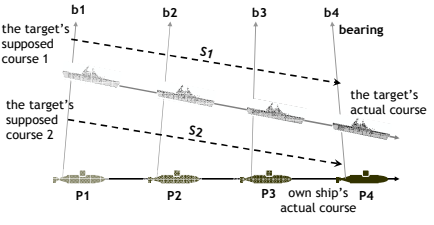


Fig. 3. First leg for the target motion analysis

When the observer moves from P1 to P4 as shown in Fig. 3, the target moves from b1 to b4. After first leg, the bearing shown as gray arrows and the bearing rate presented as dot arrows and distance ratio are known. Both of S1 which represents the target with slow speed and being far away and S2 which represents the target with fast speed and being close to the observer can be possible TMA solution. In other words, there's tremendous TMA solution with single leg. Because of this reason, the unique TMA solution

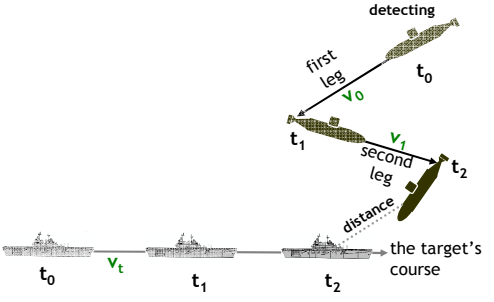


Fig. 4. Two legs for the unique solution of the target motion analysis

can be estimated by performing another leg different in course and speed from first leg (Fig.4). In general, best TMA solution can be estimated when the observer performs second leg changing the course more than 50° from previous leg [2].

2.1.2 Leg Methods

Legs used in TMA procedure are as followed (Fig. 5). Tactics with respect to this simulation can be generated by composing these legs in various manners.

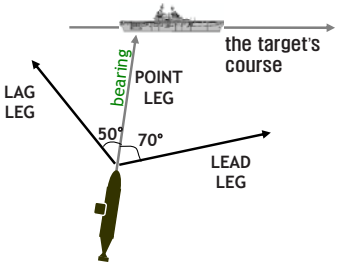


Fig. 5. Three types of leg for the target motion analysis

① POINT leg

POINT leg is the leg that moves straight in the same course to the bearing got from the passive sonar. In other words, it is the leg that moves to the position of the target at that time. It is used usually as first leg of TMA.

② LEAD leg

LEAD leg is the leg that changes the course by adding some angle to the bearing with same course of the target, in other words, moves ahead of the target. The added angle is usually 70° and leg time is 10 minutes. LEAD leg has a benefit that can reduce the distance between the observer and the target. However, when the distance between the observer and the target is close and the speed of the observer is slower than that of the target, there's a defect that the observer cannot approach the target for being crossed each other and being fell behind.

③ LAG leg

LAG leg is the leg that changes the course by adding some angle to the bearing withopposite course of the target, in other words, moves behind of the target. The added angle is usually 50° and leg time is 10 minutes. LAG leg has a benefit that can open the attack range for reaching a favorable attack position when the distance between the observer and the target is very close. On the other hands, during the LAG leg, if the speed of the target is much faster than that of the observer, the observer cannot approach the target for being fell behind.

2.2 Submerged Approach Region(SAR)

If the submarine which has slower speed than the enemy surface ship intends to launch the torpedoes and then destroy the enemy, it should remain in SAR(Submerged Approach Region). SAR is defined as followed. TDZ(Torpedo Danger Zone) is defined as circle whose diameter is the maximum range of torpedoes [2]. In addition, LLOA(Limiting Line Of Approach) is defined as tangent line to TDZ with the angle $\psi = \sin^{-1}(S_o / S_t)$ (S_o=velocity of own submarine, S_t= velocity of the target) from the course of the target. The open region which is closed with two LLOA and one TDZ is called SAR (Fig. 6).

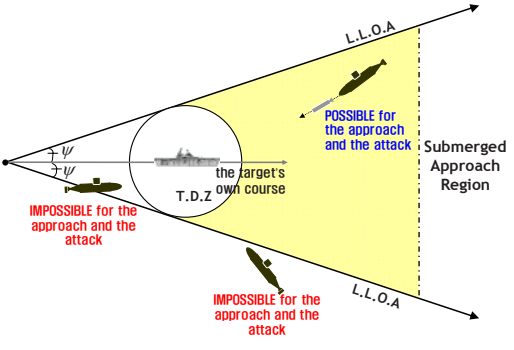


Fig. 6. Definition of Submerged Approach Region

If the submarine remains within SAR, it can attack enemy surface ship after appropriate legs. Else it cannot attack the enemy without changing its velocity.

2.3 The Scenario of Target Motion Analysis Simulation

When simulation begins, own submarine starts at fixed position(x, y) and performs the barrier mission with predefined movement patterns. And its velocity is fixed velocity chosen randomly between 2~8 knots. Enemy surface ship moves straight in the course of 000 ° and its velocity is fixed velocity chosen randomly between 10~24 knots. Its starting poison is varied according to chosen velocity. If the submarine detects the enemy with the passive sonar, it approaches to the enemy for attacking in advance, and at same time, it performs TMA procedure to estimate the kinematic state of the enemy.

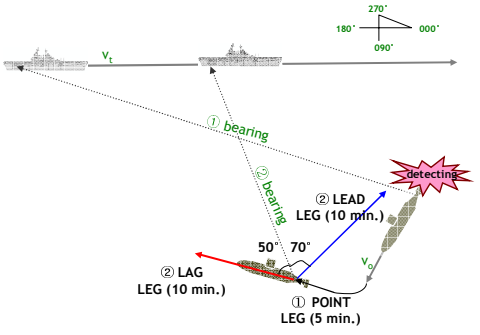


Fig. 7. Detection of a target ship and the target motion analysis procedure

In this simulation, TMA procedure consists of three different legs. When third leg has finished, the submarine is able to attack because it can estimate range and velocity, course of the enemy. First leg of TMA is POINT leg for 5 minutes. After this, bearing rate of the enemy is computed. Second leg is either LEAD leg or LAG leg and for 10 minutes (Fig. 7). Third leg is either POINT leg or different leg from second leg. In other words, if second leg is LEAD leg, then third leg will be LAG leg and if second leg is LAG leg, then third leg will be LEAD leg (Fig. 8). After the end of TMA procedure, it is examined that the submarine remains within SAR.

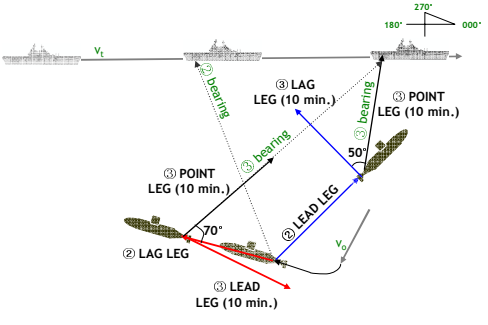


Fig. 8. Second & third leg of the target motion analysis procedure

If the submarine is out of SAR and the enemy is out of the maximum range of torpedoes, the submarine can't attack the enemy because the submarine is slower than the enemy so it cannot approach closer at all as time goes by. So in this situation, the simulation is over. On the other hands, if the submarine remains within SAR, it should prepare the attack procedure. Fist, if the enemy is within a half of the maximum range of torpedoes, torpedoes are launched. Else, it is examined whether the submarine can approach the enemy closer or not, considering the battery remains of the submarine. If it is possible, the approach procedure continues until the enemy

exists within a half of the maximum range of torpedoes, else torpedoes are launched immediately (Fig.9).

Torpedoes are launched to the estimated impact position to where they move straight with 45 knots speed. When they reach that position, they acquire the enemy with the passive sonar. If the acquisition is succeeded and torpedoes reach to the enemy, the enemy is destroyed. Else, the enemy is missed.

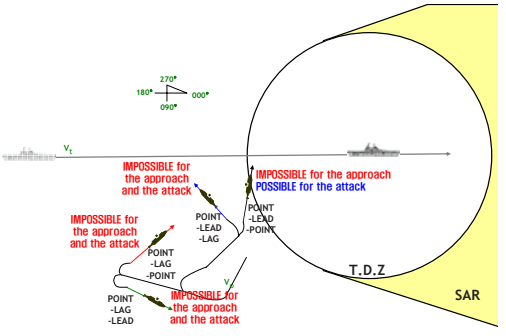


Fig. 9. Attack procedure after the target motion analysis

The submarine which launched torpedoes has tactical option either retreating or keeping approaching to the enemy. If it retreats, the simulation is over with the results whether attack is success or not. On the other hands, when it keeps approaching to the enemy, if the attack is succeeded, the simulation is over. Else, torpedoes remained and the submarine is still within SAR, additional torpedoes are launched.

3 Tactic Manager

3.1 The Concept of Tactic Manager

Tactic Manager has simple structure as "IF(decision condition) Then (Tactic 1), Else (Tactic 2)". In other words, Tactic Manager is action/tactic determining engine which can choose tactic 1 if the model satisfies specific decision condition, or tactic 2 if not.

3.2 The Description of Tactics

3.2.1 Phase Tactics

It consists of three different phase – default phase, detection and approach phase and attack phase. Whether detect or not and range are used as decision condition for these.

```
If (detection = = TRUE)
Then { IF( range ! = NULL)
       Then (attack phase)
       Else (detection and approach phase)}
Else(default phase)
```

3.2.2 Detail Tactics

① Default Phase

The default phase is that the submarine patrols in barrier with predefined movement patterns. And there's no decision conditions.

② Detection and Approach Phase

Current LEG number and previous LEG type are used as decision conditions.

```
If (current LEG number == 1)
Then (100%: LEG Type= 'POINT')
Else If (current LEG number ==2)
Then (70%: LEG Type= 'LEAD'
      30%: LEG Type= 'LAG')
Else {If(previous LEG Type= 'LEAD')
       Then(50%: LEG Type= 'POINT'
           50%: LEG Type= 'LAG')
       Else((50%: LEG Type= 'POINT'
            50%: LEG Type= 'LEAD')}
```

The detail tactic which satisfies specific decision condition can be chosen by the stochastic manner. The MOE(Measure Of Effectiveness) analysis simulation is possible with these stochastic based tactic selection.

③ Attack Phase

Range, number of torpedoes and battery remains are used as decision conditions.

```
If( range < 0.5* the maximum range of torpedoes)
Then ( Launch torpedoes )
Else {IF(number of torpedoes>0)
       If (battery remains> 40%)
       {Then (Approach)
       Else If (battery remains >20%)
       Then(50%: Approach
            50%: Retreat)
       Else(Retreat)
     Else(Retreat)}
```

3.3 Tactic Description File

Tactic description files are necessary to link tactics defined previous section from Tactic Manger to the model. Text file(.txt) is used to write tactic description files. Fig. 10 shows tactic description file which has whole tactics of the simulation.

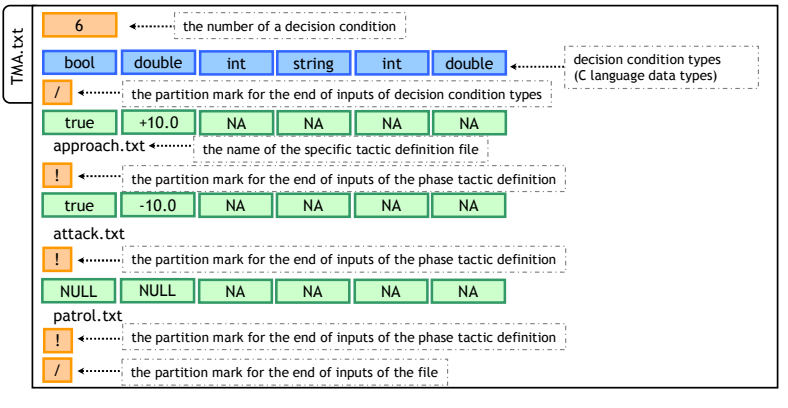


Fig. 10. Structure of a tactic definition file

All decision conditions which are used in tactic description files are written on the top of files. First the number of decision conditions is written and then the types of each decision conditions are described in sequence. C language based data types are used for decision condition types. The example of decision conditions for this simulation is as followed (Fig. 11).

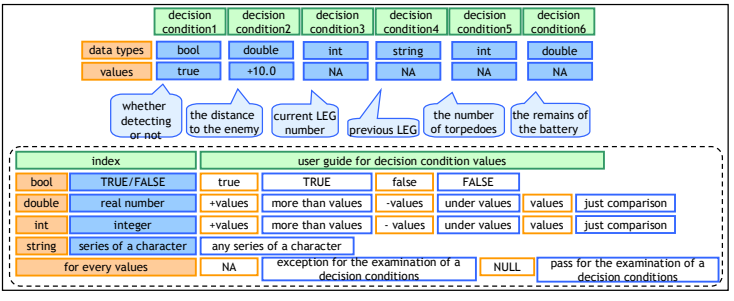


Fig. 11. Structure of a decision condition

Decision conditions which are described in above figure mean that the submarine detects the enemy and range is more than 10 nautical miles. In addition, decision conditions of the value "NA" mean that these are excluded in the examination of the decision conditions. Also, decision conditions of the value "NULL" mean that these are passed all the way in the decision condition examination. In case of the third decision conditions of Fig. 10, four values are excluded in the examination for being

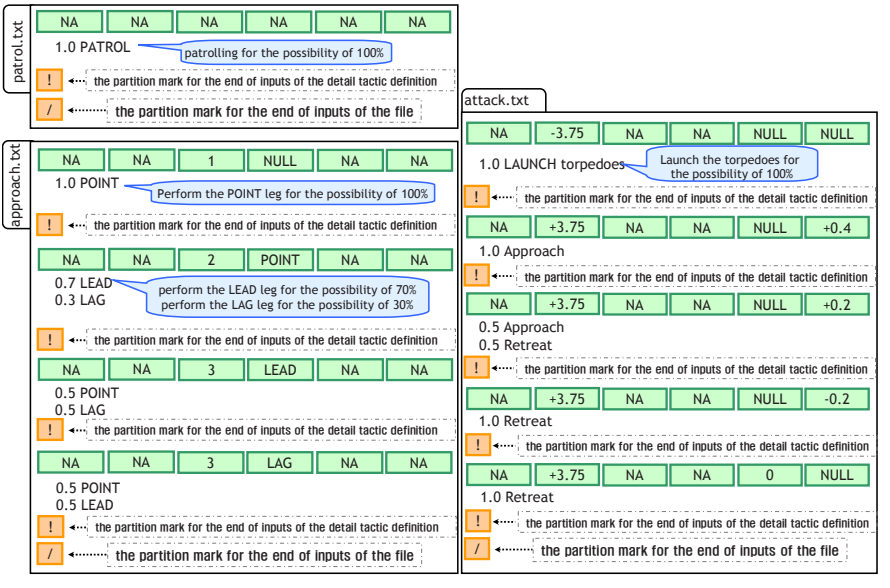


Fig. 12. Example of a tactic definition file

"NA" and rest values are "NULL". So, when selecting the tactic, "PATROL.TXT" is selected as phase tactic description file name always. As Tactic Manager has sequential searching pattern, the tactic which has to be selected always as the default tactic like above example should be written on the bottom of the tactic description file.

Fig. 12 shows detail tactic description files which are used in this simulation. Tactic Manger examines the decision condition, and phase tactic description file name which is passed the examination is returned to Tactic Manager. Then, Tactic Manager reads detail tactic description file of which name is same as returned file name and examines again for selecting the appropriate detail tactic. Finally, Tactic Manager passes selected tactic to the model. This passed tactic is defined as string data type, and the manner of action or tactics, parameters for the model are defined already in the model according to various tactics.

3.4 The Interface between the Model and Tactic Manager

Interface function between the model and Tactic Manager is shown in Fig. 13. The submarine model calls Tactic Manager when it needs to be determined appropriate tactic and passes own states as parameters. Then Tactic Manager determines the tactic with respect to those states and returns it to the model. The submarine model makes an action according to returned tactic.

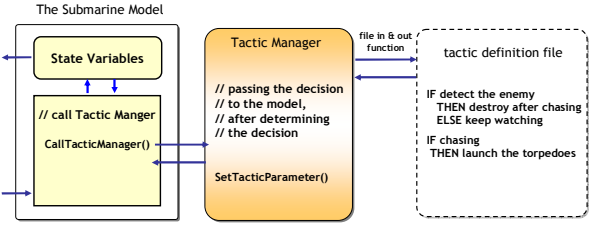


Fig. 13. Interface function between a model and a tactic manager

```
int TacticManager::CallTacticManager( char* id_string, ... ) {
  // make the list of state variables
  // through parsing the parameters from AtomicModel
     CObArray StateVariables;
     for(p=id_string; *p; p++) {
              switch(*++p){
                case 'd':  ival = va_arg(ap, int);
                           StateVariables.Add( new Integer(ival) );
                           break;
                case 'f':  dval = va_arg(ap, double);
                           StateVariables.Add( new Float(dval) );
                           break;
                     .... }   }
    // Apply the tactic which is determined
    //from the list of state variables  and tactics to the model
      Cstring result = SearchTactic( StateVariables, m_pModel );
      m_pModel->SetTacticParameters( result ); }
```

Fig. 14. Pseudo code of CallTacticManager() interface function

These process is made by interface function between CallTacticManager() in Tactic Manager Class and SetTacticParameters() in atomic model class. Among these, pseudo code of CallTacticManager() which passes the states of the model to Tactic Manger is described in Fig. 14.

As Tactic Manger should be implemented in general for applying to every model, not depending on the specific model, parameters number of interface function that is for passing states of the model has to be flexible. For this reason, parameter changeable function in C/C++ is used.

4 Simulation Results

Regardless of TMA tactics that own submarine detects the enemy surface ship and approaches to it, the possibility for the attack success is analyzed as changing the velocity of the submarine from 2 to 8 knots and changing the velocity of the enemy surface ship from 10 to 24 knots (Fig. 15-(a)).

As the submarine's velocity grow faster, the possibility for the attack success increases. And as the enemy surface ship's velocity grow faster, the possibility for the attack success decrease. It is striking that the possibility for the attack becomes maximum when the submarine's velocity is relatively slow as 2~4 knots and the velocity of the enemy is 14 knots. This shows that if the velocity of the enemy is not fast enough, accurate TMA solution cannot be estimated and if the enemy moves with high speed, the submarine can't approach the enemy at all.

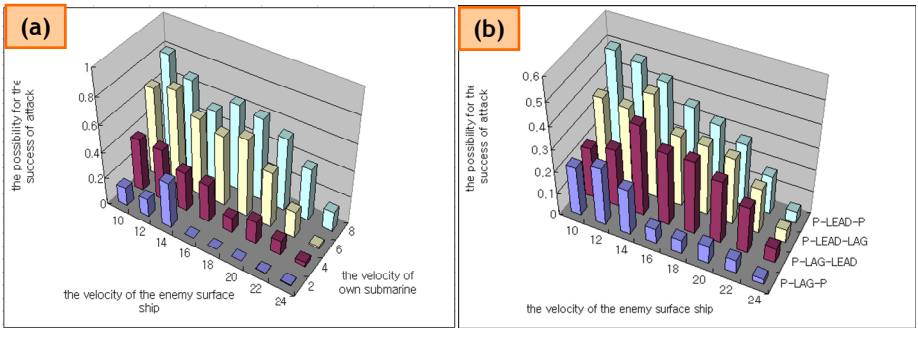


Fig. 15. Simulation results - (a) Possibility of attack success according to the speed of the target and the submarine, (b) Possibility of attack success according to tactics

Four different TMA tactics(POINT-LEAD-POINT, POINT-LEAD-LAG, POINT-LAG-POINT, POINT-LAG-POINT) are simulated. This means that the simulation is possible without modifying the model, through changing the tactic description files only.

Fig. 15-(b) shows the possibility of attack success according to four different TMA legs tactics as the velocity of the enemy surface ship increase from 10 to 24 knots and the submarine's velocity is fixed being chosen between 2 ~ 8 knots uniformly.

The knowledge which can support the captain of the submarine to make a decision among various TMA tactics in the real warfare can be generated with simulating the various tactics in the manner described above.

5 Conclusion and Future Works

In this research, the concept and the algorithm of Tactic Manager has been studied for predicting the effect of various tactics through underwater vehicle simulation. According to these, Tactic Manager which deals with table based tactic description files has been implemented and a target motion analysis in the warfare between a submarine and a surface ship is simulated to demonstrate the effectiveness of this. If tactics are changed, we can get simulation results without modifying the model, through changing the tactic description files only when we use implemented Tactic Manager.

For future works, we want to expand Tactic Manager to deal with the tactic description files which is not only table based but also script based that can be written more easily by users. And for the manner of selecting the tactic, AI is supposed to be adapted in Tactic Manager to make a decision through considering simulation results, not just to behave in deterministic or stochastic manners.

Acknowledgements

This research is performed as a part of those research projects below and supported by institutes followed.

a) Underwater Vehicle Research Center of Agency for Defense Development, Korea, SM-11 research project; "A study on the architecture of network-based virtual composition system for the function and performance simulation of underwater vehicles",
b) Research Institute of Marine System Engineering, Seoul National University, Korea
c) BK 21 project, Marine Technology Education and Research Center, Seoul National University, Korea

And we would like to appreciate these supports.

References

1. Bang, K.-W.: Combined Discrete Event and Discrete Time Simulation Framework for Shipbuilding Process Planning, Master Thesis, Seoul National University (2006)
2. Bakos, G.K.: Submarine Approach and Attack Tactics – Simulation and Analysis, Master's Thesis, Naval Postgraduate School, Monterey California (1995)
3. Cunningham, A., Thomas, A.: Target Motion Analysis Visualization. In: Asia Pacific Symposium on Information Visualisation (APVIS 2005). Conference in Research and Practice in Information Technology, Sydney, Australia, vol. 45 (2005)
4. Nardone, S.C., Graham, M.L.: A Closed-Form solution to Bearings-Only Target Motion Analysis. IEEE Journal of Oceanic Engineering 22(1), 1–11 (1997)
5. Zeigler, B.P.: Object-Oriented Simulation with Hierarchical, Modular Models. Academic press, London (1990)
6. Zeigler, B.P., Praehofer, H., Kim, T.G.: Theory of Modeling and Simulation, 2nd edn. Academic Press, London (2000)

Simulation Study on E-commerce Recommender System Based on a Customer-Product Purchase-Matrix*

Chi-myung Kwon and Seong-yeon Kim**

Hadan-Dong 840, Saha-Gu, Division of MIS, Dong-A University
Pusan, Korea 604-714
{cmkwon,sykim1}@dau.ac.kr

Abstract. This paper investigates the efficiencies of CF method and SVD-based recommender system for producing useful recommendations to customers when large-scale customer-product purchase data are available. Simulation experiments on synthetic transaction data show SVD-based recommender system yields a better performance than the CF method. Reduced product dimensionality from SVD may be more effective in generating a reliable neighborhood than CF method, and thereby it may improve the efficiency of recommendation performance. In applying SVD-based recommender system, the recommendation quality increases as the size of the neighborhood increase up to a certain point, but after that point, the improvement gains diminish. Our simulation results also show that an appropriate number of products for recommendation would be 10 in term of the error of false positives since around this point, the recall is not small, and both precision and F1 metric appear to be maximal. Even though the recommendation quality depends upon the dimension and structure of transaction data set, we consider such information may be useful in applying recommender system

1 Introduction

The large E-commerce sites offer lots of products for sale. Choosing among so many options is challenging for customers. Recommender systems have emerged to resolve this problem. A recommender system receives information from customers about which products they are interested and recommends products they likely need. While a customer is at the E-commerce site, these systems find the customer's behavior, develop a model of behavior and apply this model to recommend products to the customers. Collaborative filtering (CF) is the most successful system technology to date, and it is used in many of recommender system on the Web [15]. CF systems build a database of preferences for products by customers and then recommend products to a target customer base on the opinions of other customers.

In large E-commerce sites, customer-product purchase data present that even active customers may have purchased well under 1% of the products. With association rules it is common to find rules having support and confidence higher than a user-defined

* This paper was supported by Dong-A research fund in 2004.
** Corresponding author.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 327–336, 2007.

minimum. When the frequency of a customer's purchased-product is low, even we may find a strong association rule for product recommendation, rules with very low support are often uninteresting since they do not describe sufficiently large populations. (From simulation results applying E-miner version 9.5 of SAS to synthetic product-purchase data, the minimum support is 0.4%). In applying CF, the problem space can be formulated as a matrix of users versus items, with each cell representing user's rating on a specific item. This matrix is generally very sparse since each user rates only a small percentage of the total number of items. In such a case, it may be one way to use a customer-product purchase-matrix instead of rating matrix for recommendation [12].

The tremendous growth of customers and products poses several key challenges for a successful recommender system in the E-commerce domain. These include producing high quality of recommendation and lessening system response time for recommendation. Therefore we need to improve the procedure of identifying the neighbors, which require much computational efforts, and to achieve high coverage in the face of sparseness of customer-product purchase-matrix. The structure of customer-product purchase-matrix is quite similar to that of term-document matrix used in information retrieval community. Latent semantic indexing (LSI) is a widely used dimensionality reduction technique in information retrieval, which uses SVD and derives most important latent structure in the association of terms and documents [5]. Similarly to the approach of LSI, we try to capture the customer's characteristics of purchasing products in reduced product-space through SVD, and use this information to achieve the high quality of recommendation and to reducing efforts in neighborhood generation.

This paper focuses on SVD-based algorithm that has potential to improve scalability and quality of recommender system for a case that the customer-product purchase-matrix is available. To this end, we generate synthetic transaction data in retailing environment, and analyze the effectiveness of considered recommender system based on synthetic data.

2 Collaborative Filtering Algorithms

In this section, we briefly outline the main ideas of CF algorithm. CF systems recommend products to a target customer based on the opinion of other customers. These systems employ statistical techniques to find a neighborhood of like-minded customers. Customers in the neighborhood tend to buy similar products with a target customer. Suppose we have a database of customer ratings for items, where the customer u indicates his/her interest in an item i on a numeric scale $r_{u,i}$. It is now possible to define similarity measures between two customers, a and u, where a customer profile simply consists of a vector of numeric ratings. Usually the neighborhood formation scheme uses Pearson correlation as a measure of proximity:

$$\rho_{a,u} = \sum_{i=1}^{m} (r_{a,i} - \bar{r}_a)(r_{u,i} - \bar{r}_u) / \sqrt{\sum_{i=1}^{m} (r_{a,i} - \bar{r}_a)^2 (r_{u,i} - \bar{r}_u)^2}. \quad (1)$$

Here $\bar{r}_u$ and $\bar{r}_a$ are the mean ratings of the customers a and u, respectively.

The most important step in CF-based recommender system is to form a proximity-based neighborhood between a target customer and a number of like-minded customers. The neighborhood formation process is to find an ordered list of l customers whose similarity is greater than other customers. Once the size l of neighborhood of target customer a is formed, CF system uses several algorithms to produce recommendation [7]. Group Lens [10] first introduced automated CF system using a neighborhood-based algorithm. Group Lens used Pearson correlations to weight user similarity, used all available correlated neighbors, and computed a final prediction by performing a weighted average of deviations from the neighbor's mean:

$$p_{a,i} = \bar{r}_a + \sum_{u=1}^{l} (r_{u,i} - \bar{r}_u)\rho_{a,u} / \sum_{u=1}^{l} \rho_{a,u} . \quad (2)$$

Here $p_{a,i}$ represents the prediction for the target customer a for item i. The Ringo music recommender [14] and the Bellcore Video Recommender [8] expanded upon the original GroupLens algorithm. Breese et al. [4] performed an empirical study of several variants of neighborhood-based CF algorithm.

While these correlation-based prediction schemes were shown to perform well, they suffer from several limitations. Fist, correlation between two customers can be only be computed over those items that both customers have rated. If customers can choose among thousands of items to rate, it is likely that overlap of rated items between two customers will be small in many cases. Therefore, many of the computed correlations are based on a few observations and thus the correlations may not be reliable measure of similarity. Second, nearest neighbor algorithm requires computation that grows with both the number of customers and the number of products. With a lot of customers and products, CF algorithm may suffer serious scalability problems. Finally, the fact that two customers do not rate the same items does not necessary mean that they are not like-minded. In real life scenario, different product names can refer the similar objects. Correlation-based algorithm can not find this latent association and treat them differently. These weakness of the correlation-based CF algorithm leads to explore alternative methods. In the context of information retrieval, there exists a similar problem that we are facing in CF algorithm. One of these approaches, latent semantic indexing (LSI) is based on dimensionality reduction of the initial data through the singular value decomposition (SVD). In next section, we will address how this technique can be used for recommendation tasks in E-commerce site. (For a more detailed information about CF algorithm, see the references [3] and [7]).

3 SVD-Based Recommender System

LSI is a dimensionality reduction technique that has been widely used in information retrieval to solve the synonymy [2]. Given a term-document matrix, LSI is used to capture most of the important underlying structure in the associations between terms and documents with reduced dimensionality. With the same spirit, we are trying to

capture relationship among customers based on their purchasing records of products by reducing the dimensionality of the product space. If we can increase density of sparse matrix and thereby find more information about purchased-products, it may improve the accuracy of recommender system. Also discovery of latent relationship from input data may potentially solve the synonymy problem in recommender system. We expect that SVD technique nicely maps into recommendation tasks in E-commerce site. The rest of section presents the construction of SVD-based recommender algorithm for the purpose of generating top-N recommendations. We first present SVD technique to produce a low-dimensional representation of a given matrix, and then compute neighborhood in the reduced space. Finally we present a procedure to generate a list of top-N product recommendation for customers.

3.1 SVD on Customer-Product Purchase Matrix

In E-commerce sites, transaction data is a collection of historical purchasing records of n customers on m products. It can be represented as an (n x m) customer-product purchase-matrix A where a_{ij} is 1 if the ith customer has purchased the jth product, and 0, otherwise. This matrix is a large-scale which is very sparse because even active customers may purchase only a small subset of all products. We suppose that $rank(A) = r$ and $n \geq m$. Then the SVD on A is defined as

$$A = U\Sigma V^T . \tag{3}$$

Here U and V are orthogonal matrices, $\Sigma = diag(\sigma_1, \ldots, \sigma_m)$, $\sigma_i > 0$ $(1 \leq i \leq r)$, and $\sigma_i = 0$ $(i \geq r+1)$. The first r columns of U and V define the orthogonal eigenvectors associated with the r nonzero eigenvalues of AA^T and $A^T A$, respectively. The SVD on A derives a latent customer-product purchase relationship model from the orthogonal matrices U and V, and the diagonal matrix Σ. These matrices reflect a breakdown of the original relationships into linearly independent vectors. In equation (3), if $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m$, the use of k-largest singular triplets is equivalent to approximating the original customer-product purchase matrix A by A_k defined as

$$A_k = \sum_{i=1}^{k} u_i \sigma_i v_i^T , \tag{4}$$

Here u_i and v_i^T are the column vectors of U and V, respectively. The A_k obtained from the truncated SVD captures most of the important underlying structure in the association between customers and products. Intuitively, the number of k is much smaller than the number of products m, thus A_k is a reduced k-dimensional representation closest to the original customer-product purchase matrix A having product-space of m. We now discuss to form a neighborhood of target customer in the reduced space based on A_k (see the reference [2] for a more detailed description of SVD).

3.2 Neighborhood Formation in the Reduced Space

The process to find a neighborhood of target customer focuses on the problem how to identify the customers as neighbor, which is the most important part in recommender system. We note that A_k shows purchasing trends of *n* customers in *k*-dimensional transformed product space. To form a neighborhood of target customer *a*, we need to transform the original product-purchase vector q_a of customer *a*, into *k*-meta product space [5]:

$$\hat{q}_a = q_a^T U_k \Sigma_k^{-1} \tag{5}$$

Here $U_k = (u_1, \ldots, u_k)$ and $\Sigma_k = diag(\sigma_1, \ldots, \sigma_k)$. The proximity between two customers is usually measured using either the correlation or the cosine measure [12]. We compute the cosine angle between the *k*-dimensional two vectors $\hat{q}_a$ and $\tilde{q}_b$, which is given by

$$\cos(\hat{q}_a, \tilde{q}_b) = (\hat{q}_a \cdot \tilde{q}_b) / (\| \hat{q}_a \| * \| \tilde{q}_b \|) \tag{6}$$

After computing the proximity between a target customer and every customer *b* in data file, we select the *l* nearest customers to a target customer and form the neighborhood of size *l* for a target customer *a*.

3.3 Top-N Recommendation Generation

The final step of SVD-based recommender system is to derive the top-N recommendation to a target customer from his/her neighborhood. Sarwar et al. [12] presented most-frequent item recommendation and association rule-based recommendation procedures for top-N recommendation. Most-frequent item recommendation looks into the neighborhood and each neighbor scans through his/her purchase data and perform a frequency count of the products. After all neighbors are accounted for, the system sorts the products according to their frequency count and simply returns the N most frequent products as recommendation that have not yet been purchased by the target customer. Association rule-based recommendation uses association rule for only the *l* nearest neighbors and select the N highest ranked products as the recommended set. We here use the most-frequent item recommendation rule to produce the top-N recommendation.

4 Simulation Experiment

We generate synthetic transaction data to evaluate the performance of the algorithms over a large range of data characteristics. These transactions mimic the transactions in the retailing environment. Based on these data, we investigate how much the recall, precision and F1 measure of the algorithms change according to the size of neighborhood, the number of recommended products and correlation level of large item sets. Also we analyze the performance of SVD-based recommender system.

4.1 Generation of Customer–Product Purchase Matrix

In real world, people tend to buy set of items together. Each such set is potentially a maximal large item set. A transaction may contain more than one large item set (LIS). Transaction sizes are typically clustered around a mean and a few transactions have many items. Typically the sizes of LIS also are clustered around a mean and a few large item sets having a number of items.

We first determine the size of each transaction. Note that each item is chosen with the same probability *p*, and there are *m* items. Then the expected number of items in a transaction is given by a binomial distribution with mean of *mp*, and is approximated by Poisson distribution with mean *mp*. Thus we pick the size of transaction from Poisson distribution with mean *mp*. For a convenience of simulation experiments, we set the numbers of products and customers to be 500, and 1000 respectively, and generate the transaction data of training set and the same size of test set. We suppose that the average number of products to purchase for each customer is 1.4% of products, which is equal to 7.

We then assign items to the transaction. Each transaction is assigned a series of potentially large item sets. LIS are chosen from a set of *T*, which consists of item sets. The number of item sets in *T* is set to 1000. An item set in *T* is generated from first picking the size of the LIS from Poisson distribution with mean 2. Items in the first item set are randomly selected. To model the phenomenon that LSI often has common items, some fraction of items in a subsequent item sets are chosen from previous item set generated. We select three levels of fraction, 0.25, 0.5 and 0.75 to investigate the performance of recommendation algorithms according to correlation levels of LSI. The remaining items are picked randomly.

Each item set in *T* has a weight associated with it, which corresponds to the probability that this item set is picked. The weight is picked from uniform distribution U(0,1) and is then normalized so that the sum of the weights for all the item sets in *T* is 1.

To model the phenomenon that all the items in LSI are not always bought together, we assign a corruption level c to the selected LSI. When adding the LSI to a transaction, we keep dropping an item from LSI if a random number from U(0,1) is less than c, which is equal to 0.5. Thus for an item size s, we will add s items to the transaction (1-c) of the time, (s-1) items c(1-c) of the time, etc.

Finally we generate the transaction for each customer to have Poisson distribution with mean 7. (For a more detailed description generate synthetic data of transactions, see the reference [1]).

4.2 Evaluation Metrics

As we mentioned in section 4.1, we divided data sets into two parts - the training set and test set. We apply the recommender systems to the training set and recommend the top-N products to a target customers consisting of test set. Then we explore the test set and match products with our top-N set. Products appear in both sets are

members of the hit set. The ratios of hit set size to both the test size and the top-N set size would be measures to evaluate the performance of recommender systems. We now define recall and precision as the following [14]:

$$\text{Recall} = (\text{size of hit set}) / (\text{size of test set}). \tag{7}$$

$$\text{Precision} = (\text{size of hit set}) / N. \tag{8}$$

In equations (7) and (8), we see that increasing the number N tends to increase recall but decreases precision. The fact that two measures are critical for the quality of recommendation leads us to use the standard F1 metric that gives equal weight to them and is defined as follows:

$$Fl = (2 \times \text{recall} \times \text{precision})/(\text{recall} + \text{precision}). \tag{9}$$

We compute the recall, precision and F1 for each customer and calculate their average as our metrics.

4.3 Experimental Results

We apply the CF algorithm and SVD-based recommender system to synthetic transaction data, and compute the evaluation measures of recall, precision and F1 metric. We explore how much the size of neighborhood, number of products recommended to a target customer and correlation level to generate LSI for transaction data affect these three metrics. To determine the effect of neighborhood size, we performed experiments with increase its size from 10 to 50 by 10. We also varied the recommendation products size from 5 to 25 by 5 to find its effect to three metrics. Table 1 presents the experimental results for three different fraction levels of 0.25, 0.5 and 0.75 to generate LIS.

As we see in Table 1, the performances of SVD-based recommender system are better than those of the CF algorithm. In applying the SVD-based recommender system, the size of neighborhood has a significant impact on the recommendation quality. In general, the recommendation quality increases as the size of neighborhood increases. However, after a certain point, the improvement gain diminishes and the quality becomes worse (see the figure 1). These results are coincident with those obtained from work of Sarwar et al [12]. On the other hand, the performances of the CF algorithm decrease as the size of neighborhood increases in our case.

Figure 2 presents the changes of three performance metrics of recall, precision and F1, respectively, from applying SVD algorithm for various sizes of recommended product and neighborhood. As we increase the size of recommended products, the recall increases regardless of the neighborhood size, but the precision and F1 metric appear maximally at a point around 10 of recommended products. After 10, these two measures decrease. As Sarwar et al. [14] pointed out, it seems that they depend on the dimension and characteristics of data set. We also find that given the same size of recommended products, three metrics do not show much difference if the neighborhood size is over 20. Finally we observed that the effect of correlation ratio in generating LSI to three metrics is negligible.

Table 1. Evaluation Metrics according to the size of neighborhood(c), the number of recommenaded products (N) and fraction level generating LIS (in each cell, recall, precision and F1 are orderedly given from above)

C	Method	Fraction level=0.25 N 5 10 15 20 25	Fraction level=0.50 N 5 10 15 20 25	Fraction level=0.75 N 5 10 15 20 25
10	SVD	2.59 4.94 5.32 5.54 5.75 0.52 0.49 0.36 0.28 0.23 0.86 0.90 0.66 0.53 0.44	2.73 5.00 5.34 5.57 5.77 0.55 0.50 0.36 0.28 0.23 0.91 0.91 0.67 0.53 0.44	2.79 5.05 5.38 5.61 5.81 0.56 0.51 0.36 0.28 0.23 0.93 0.92 0.67 0.53 0.45
	CF	2.54 3.31 3.88 4.20 4.34 0.51 0.33 0.26 0.21 0.17 0.85 0.60 0.49 0.40 0.33	2.58 3.40 3.94 4.27 4.42 0.52 0.34 0.26 0.21 0.18 0.86 0.62 0.49 0.41 0.34	2.60 3.40 3.95 4.25 4.38 0.52 0.34 0.26 0.21 0.18 0.87 0.62 0.49 0.41 0.34
20	SVD	2.51 5.45 6.48 6.70 6.74 0.50 0.55 0.43 0.34 0.27 0.84 0.99 0.81 0.64 0.52	2.58 3.40 3.94 4.27 4.42 0.52 0.34 0.26 0.21 0.18 0.86 0.62 0.49 0.41 0.34	2.64 5.74 6.56 6.70 6.74 0.53 0.56 0.44 0.34 0.27 0.88 1.03 0.82 0.64 0.52
	CF	2.47 3.17 3.72 4.20 4.51 0.49 0.32 0.25 0.21 0.18 0.82 0.58 0.47 0.40 035	2.47 3.23 3.84 4.31 4.58 0.50 0.32 0.26 0.22 0.18 0.82 0.59 0.48 0.41 0.35	2.53 3.22 3.86 4.31 4.58 0.51 0.32 0.26 0.22 0.18 0.84 0.59 0.48 0.41 0.35
30	SVD	2.32 5.57 6.75 6.90 6.98 0.46 0.58 0.45 0.35 0.28 0.77 1.01 0.84 0.66 0.54	2.43 5.81 6.79 6.93 6.98 0.49 0.58 0.45 0.35 0.28 0.81 1.06 0.85 0.66 0.54	2.52 5.92 6.79 6.92 6.98 0.50 0.59 0.45 0.35 0.28 0.84 1.08 0.85 0.66 0.54
	CF	2.41 3.06 3.58 4.06 4.43 0.48 0.31 0.24 0.20 0.18 0.80 0.56 0.45 0.39 0.34	2.43 3.10 3.64 4.15 4.52 0.41 0.31 0.24 0.21 0.18 0.81 0.56 0.46 0.40 0.35	2.47 3.10 3.67 4.16 4.56 0.49 0.31 0.25 0.21 0.18 0.82 0.56 0.46 0.40 0.35
40	SVD	2.16 5.55 6.79 6.96 7.00 0.43 0.56 0.45 0.35 0.28 0.72 1.01 0.85 0.66 0.54	2.30 5.82 6.84 6.96 7.00 0.46 0.58 0.46 0.35 0.28 0.77 1.06 0.86 0.66 0.54	2.40 5.92 6.86 6.97 7.01 0.48 0.59 0.46 0.35 0.28 0.80 1.08 0.86 0.66 0.54
	CF	2.35 2.94 3.49 3.96 4.36 0.47 0.29 0.23 0.20 0.17 0.78 0.54 0.44 0.38 0.34	2.39 3.01 3.55 4.06 4.46 0.48 0.30 0.24 0.20 0.18 0.80 0.55 0.44 0.39 0.34	2.42 2.98 3.56 4.07 4.47 0.48 0.30 0.24 0.20 0.18 0.81 0.54 0.45 0.39 0.34
50	SVD	2.00 5.46 6.81 7.00 6.99 0.40 0.55 0.45 0.28 0.28 0.67 0.99 0.85 0.54 0.54	2.15 5.75 6.85 6.96 6.99 0.43 0.58 0.46 0.35 0.28 0.78 1.05 0.86 0.66 0.54	2.27 5.87 6.88 6.96 7.00 0.46 0.59 0.46 0.35 0.28 0.76 1.07 0.86 0.66 0.54
	CF	2.32 2.85 3.41 4.36 4.24 0.46 0.29 0.23 0.17 0.17 0.77 0.52 0.43 0.34 0.33	2.35 2.93 3.47 3.97 4.33 0.47 0.29 0.23 0.20 0.17 0.78 0.53 0.43 0.38 0.33	2.35 2.89 3.50 3.97 4.37 0.47 0.29 0.23 0.20 0.18 0.78 0.53 0.44 0.38 0.34

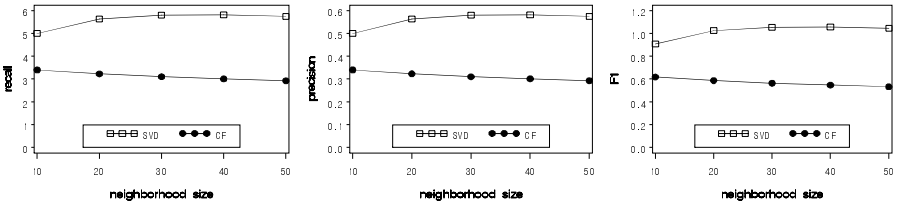


Fig. 1. Impact of neighborhood size on recall, precision and F1 metrics for CF and SVD-based recommender system when number of products recommended is 10

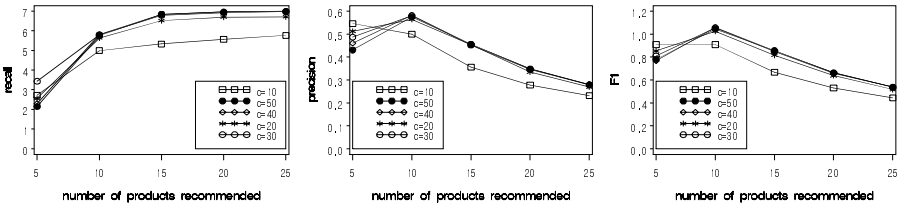


Fig. 2. Impact of number of products recommended on recall, precision and F1 metrics with different neighborhood size for SVD-based recommender system

5 Conclusion

This research investigates the efficiencies of the classical CF method and SVD-based recommender system for the purpose of producing useful recommendations to customers when large-scale customer-product purchase data are available. We experimentally evaluate SVD-based recommender system as well as the CF recommender system over a large range of input data. Simulation experiments on synthetic transaction data show SVD-based recommender system yields a better performance than the CF method with respect to the recall, precision and F1 metrics. Our experimental results show that reduced product-dimensionality by SVD may yield more reliable similarity between two customers, which is critical in forming neighborhood, and thereby it may give a performance gain in recommendation compared to CF system.

In applying SVD-based recommender system, the recommendation quality increases as the size of the neighborhood increases up to a certain point, but after this point, the improvement gains diminish. Our simulation results also show that an appropriate number of products for recommendation would be 10 in term of the error of false positives since around this point, the recall is not small, and both precision and F1 metric appears to be maximal. Even though the recommendation quality depends upon the dimension and structure of transaction data set, we consider such information may be useful in constructing recommender system.

References

1. Agawal, R., Srikant, R.: Fast Algorithms for Association Rules. In: Proceedings of the 20th VLDB Conference, pp. 487–499 (1994)
2. Berry, M., Dumais, S., O'Brian, G.: Using Linear Algebra for Intelligent Information Retrieval. SIAM Review 37, 573–595 (1995)
3. Billsus, D., Pazzani, M.: Learning Collaborative Information Filters. In: Proceedings of ICML, pp. 46–53 (1998)
4. Breese, J., Heckerman, D., Kadie, C.: Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, pp. 43–52 (1998)
5. Deerwester, S., Dumais, S., Furnas, G., Landauer, T.: Indexing by Latent Semantic Analysis. J. of the American Society for Information Science 41(6), 391–407 (1990)

6. Good, N., Schafer, B., Konstan, L., Borchers, A., Sarwar, B., Herlocker, J., Riedl, J.: Combining Collaborative Filtering with Personal Agents for Better Recommendations. In: Proceedings of the AAAI Conference, pp. 439–446 (1999)
7. Herlocker, J.H., Joseph, A., Borchers, A., Riedl, J.: An Algorithm Framework for Performing Collaborative Filtering. In: Proceedings of the Conference on Research and Development in Information Retrieval (1999)
8. Hill, H., Stead, L., Rosenstein, M., Furnas, G.: Recommending and Evaluating Choices in a Virtual Community of Use. In: Proceedings of ACM, pp. 194–201 (1999)
9. Kautz, H., Selman, B., Shah, M.: Combining Social Networks and Collaborative Filtering. Communications of the ACM 40(3), 63–65 (1997)
10. Konstan, J., Miller, B., Martz, D., Herlocker, J., Gordon, L., Riedl, J.: GroupLens: Applying Collaborative Filtering to Usenet News. Communication of the ACM 40(3), 77–87 (1997)
11. Resnick, P., Varian, H.: Recommender Systems. Special Issue of Communications of the ACM 40(3) (1997)
12. Sarwar, B., Karypis, G., Konran, J., Riedl, J.: Analysis of Recommendation Algorithm for E-Commerce. In: Proceedings of the 2nd ACM Conference on Electronic Commerce (2000)
13. Sarwar, B., Karypis, G., Konran, J., Riedl, J.: Application of Dimensionality Reduction in Recommender System – A Case Study. In: ACM WebKDD Workshop (2000)
14. Shadanand, U., Maes, P.: Social Information Filtering: Algorithm for Automating Word of Mouth. In: Proceedings of ACM, pp. 210–217 (1995)
15. Shafer, J., Konstan, J., Riedl, J.: Recommender Systems in E-Commerce. In: Proceedings of ACM E-Commerce Conference (1999)

A New Performance Evaluation Method for Automatic Target Recognition Based on Forward Cloud

Fang Wang, Yanpeng Li, and Xiang Li

School of Electronic Science and Engineering, NUDT, Changsha 410073, China
susana0718@yahoo.com.cn

Abstract. As the development of Automatic Target Recognition (ATR) technology, the performance evaluation for it becomes more and more important. In this paper, we introduced the cloud theory into the fuzzy comprehensive evaluation, in order to resolve the problem that the evaluation results are impressible to the values of index weights. At last, we compared our evaluation method with the general fuzzy comprehensive evaluation through the simulations with field test data. Then the simulation results proved the availability of the method we supposed.

1 Introduction

Since the middle period of last century, radars have been used in the field of Automatic Target Recognition (ATR). Until now, there have been many different kinds of ATR algorithms and systems based on different techniques and theories. As the development of ATR algorithms and systems, the performance evaluation of them becomes more and more important.[1,2] Effective performance evaluation can provide a scientific guidance of the ATR system improvements for the researchers, at the same time, offer a rule of selection for the users. In one hand, during the investigating process, the researchers can comparatively easily find the problem existing in their ATR system via the performance evaluation of each module and stage, and the former results can be used as the rules of improvements. In the other hand, during the using process, owing to the complexity of battlefields and the variety of martial aims and the performance limitations of ATR systems, the users need to choose an appreciate ATR system which can realize the certain martial aim under the certain battlefield, and the performance evaluation and prediction of every possible ATR system under different battlefield situation naturally become the best warrants. As a result, we can save research fare, curtail research cycle, and upgrade the system performance. From the literatures, we know that compared with the research on ATR system, the one on performance evaluation is much younger. Up to now, in the aspect of theory system, there have been several performance evaluation frameworks[3,4]; in the aspect of evaluation method, there also have been a number of methods, such as the evaluation method based on fuzzy comprehensive evaluation[3], the evaluation method base on fuzzy integral[3], the evaluation method based on fuzzy cluster analysis[3], confusion matrix method[5], the evaluation method based on ROC performance curve[6], decision analysis[7] etc. Specially, fuzzy comprehensive evaluation has already played a very 1important role in complex system multi-factor comprehensive evaluation, fuzzy

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 337–345, 2007.

judgment, weather forecast, agriculture production, and so on.[8,9] It is a better one among the existing evaluation methods because that it comprehensively considers the effect of various system performance indexes on the system unitary performance, and the effect of various uncertain factors together.

In the course of fuzzy comprehensive evaluation, the weight values of the performance indexes are significant to the final evaluation result. However, it is very difficult to choose the reasonable values for them in practice. In matter of fact, we can only decide the relative importance of the system performance indexes, for example important or very important. Therefore, in this paper, we introduced a model called forward cloud in order to solve the problem above, which can realize the transformation from a qualitative concept to a quantified value. Then the validity of our modification was testified through the comparison between the general fuzzy comprehensive evaluation and our method.

In Section II we introduce some basic knowledge of the cloud theory, including the concept of cloud and the algorithm of the normal forward cloud. Then we present the performance evaluation method for ATR system based on the forward cloud in Section III. The simulation and comparison results which can validate our approach are given in Section IV. Finally we arrive at a conclusion and outline shortcomings of our method and areas of future work in Section V.

2 Basic Knowledge of Cloud Theory

The cloud model is a transformation model between a qualitative concept and a quantified value, which is presented based on the traditional Fuzzy Set theory and Probability Statistics theory by Li Deyi. In other words, it can realize the transformation between a qualitative concept described in words and its numerical representation.

2.1 Definitions [11]

Consider a domain U which is represented by precise values and can be arbitrary number of dimensions. And C is a qualitative concept on it. If a quantitative value $x \in U$ and x is a stochastic realization of C, the distribution of x on domain U comes to be known as a *cloud model* (and *cloud* for short) and each x is a cloud drop. The numerical characters of cloud are Ex, En and He, respectively called the expectation, entropy and super-entropy of cloud. They reflect the quantitative characters of concept C: Ex is the most typical sample of C; En reflects the area which can acceptant by lingual value C in numerical space; He is the entropy of En, which reflects the uncertainty of En and is determined by both the randomness and fuzziness of En.[11] For instance, with the data attained in an large-scope statistical test about the fuzzy concept "youngster" done by Zhang Nanlun et al, we can generate a one dimension normal cloud to describe "youngster". The result of the test and the cloud we generated are respectively shown in Fig. 1 and Fig. 2. Here we generated 1000 cloud drops and set $Ex = 24$, $En = 5$, $He = 0.5$ respectively.

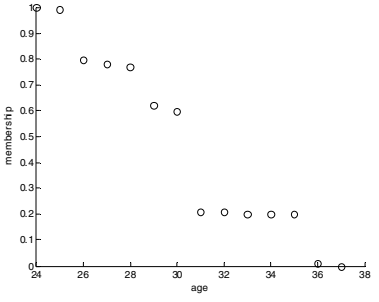


Fig. 1. The membership degree of "young ster" gotten from the test done by Zhang Nanlun etc

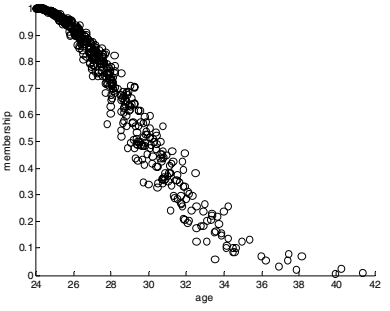


Fig. 2. The description of "youngster" using normal forward cloud

The most important cloud model is the normal one because of its universality which is based on the universality of normal distribution and campanula membership function.[10]

2.2 Forward Cloud Generator

Forward cloud generator is a forward transformation from a qualitative concept to a quantitative value, that is, the generator will obtain the range and the distribution of a quantitative value from the qualitative information expressed by a lingual value. An algorithm of M-dimension normal forward cloud generator is as follows:

Input: $Ex = [Ex_1, Ex_2, ..., Ex_M]^T$ expectation array

$En = [En_1, En_2, ..., En_M]^T$ entropy array

$He = [He_1, He_2, ..., He_M]^T$ super-entropy array

N the number of cloud drops to be generated

Output: N M-dimension cloud drops

Step1: according to En and He, generate an M-dimension normal random number $\sigma = [\sigma_{1j}, \sigma_{2j}, \cdots, \sigma_{Mj}]$ which is has an expectation of $[En_1, En_2, ..., En_M]$ and a variance of $[He_1^2, He_2^2, ..., He_M^2]$;

Step2: generate another M-dimension normal random number $X = [x_{1j}, x_{2j}, \cdots, x_{Mj}]$ whose expectation and variance are $[Ex_1, Ex_2, ..., Ex_M]$ and $[\sigma_{1j}^2, \sigma_{2j}^2, \cdots, \sigma_{Mj}^2]$ respectively;

Step3: calculate $\mu_j = \exp\left[-\sum_{i=1}^{M} \frac{(x_{ij} - Ex_i)^2}{2\sigma_{ij}^2}\right]$, and then $(x_{1j}, x_{2j}, ..., x_{Mj}, \mu_j)$ is a drop composing the cloud aimed at;

Step4: repeat Step1-Step3 until generating N cloud drops.

In fact, the drop generated by the normal forward cloud generator above is a random variable whose expectation and variance are Ex and $En^2 + He^2$ respectively.

3 Performance Evaluation Based on Forward Cloud for ATR

Combining with the forward cloud, the evaluation method base on fuzzy comprehensive evaluation can be modified as the following. Then it is not only extensively applicable for various kinds of ATR system, but also overcoming the shortcoming that the evaluation result is sensitive to the index weights. The detailed algorithm is as follows:

(1) Determine the set of target kind, the set of target kind weight, the set of performance index, the set of index weight and the set of remark.

The set of target kind: $T = \{t_1, ..., t_S\}$, and S is the number of target kind;

The set of target kind weight: $WT = \{wt_1, ..., wt_S\}, \sum_{i=1}^{S} wt_i = 1$;

The set of performance index: $U = \{u_1, ..., u_K\}$, and K is the number of performance index;

The set of index weight: $W = \{w_1, ..., w_K\}$;

The set of remark: $V = \{v_1, ..., v_F\}$, and F is the number of evaluation grade;

Where w_j is the weight of index u_j. It may be either a proper numerical value or a lingual atom denoting the importance of u_j such as important, general important. It is very difficult to obtain the exact value for each index weight in practice, so adopting a lingual atom like above to denote the importance of index in respect to the whole performance is more scientific and reasonable.

(2) Determine the index intervals corresponding to each remark for the i th $(i = 1, ..., S)$ target kind.

For example, when the set of remark is {Better, Good, Middle, Bad}, the index intervals are shown in Table 1.

Table 1. Indexes intervals corresponding to the remarks for the i th target kind

	Better	Good	Middle	Bad
Index 1	$> d_{11}^i$	$(d_{12}^i, d_{11}^i]$	$(d_{13}^i, d_{12}^i]$	$\le d_{13}^i$
Index 2	$> d_{21}^i$	$(d_{22}^i, d_{21}^i]$	$(d_{23}^i, d_{22}^i]$	$\le d_{23}^i$
……	……	……	……	……
Index S	$> d_{S1}^i$	$(d_{S2}^i, d_{S1}^i]$	$(d_{S3}^i, d_{S2}^i]$	$\le d_{S3}^i$

(3) Describe each weight of evaluation index with 1-dimension normal cloud.

The weight cloud should be generated according to the important degree and its relative numerical domain of each index ascertained by experts. For each middle grade, there is a bi-restriction $[C_{\min}, C_{\max}]$, so we can use a cloud like symmetric cloud to describe it, and contrarily, a half-cloud should be adopted. Here, we make the numerical domain aforementioned on $[0,1]$ and play the middle value of restriction as the expectation of cloud, then

$$\begin{cases} Ex = (C_{\min} + C_{\max})/2 \\ En = (C_{\max} - C_{\min})/6 \\ He = k \end{cases} \tag{1}$$

Here k is a constant that indicates the fuzzy degree of the index weight. It reflects the discrepancy for different person about the lingual atom which has been used to describe the importance of index. An example for the cloud of index weight is illuminated in Fig. 3. The set of remark is {general important, important, very important} and their relative intervals are [0,0.3], [0.3,0.7], [0.7,1.0] in turn; the set of index has three elements and their weights are "general important", "important" and "very important" respectively.

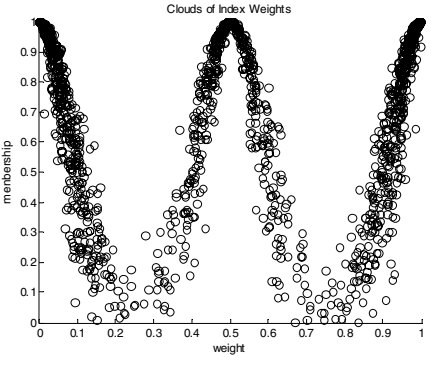


Fig. 3. An example of the clouds of index weights

(4) Calculate the values of each performance index according to the experiment data, and then obtain the performance evaluating matrix of the system under evaluation to the i th target kind.

Suppose the system under evaluation has carried through H_i groups target recognition experiment for the i th kind of target, then from each group we can attain a value of each index. Remark the time of index j dropping into each interval as $h_{j1}^i, \ldots, h_{jF}^i$, then the single factor judgment vector is

$$\left[\frac{h_{j1}^{i}}{H_i},...,\frac{h_{jF}^{i}}{H_i}\right]=\frac{1}{H_i}\left[h_{j1}^{i},...,h_{jF}^{i}\right]$$

In a similar way, the judgment matrix is

$$R_i=\left[r_{jn}\right]_{K\times F}=\frac{1}{H_i}\begin{bmatrix}h_{11}^{i} & h_{12}^{i} & \cdots & h_{1F}^{i}\\ \vdots & \vdots & \vdots & \vdots\\ h_{K1}^{i} & h_{K2}^{i} & \cdots & h_{KF}^{i}\end{bmatrix}$$

(5) Combined with the cloud of index weight, arrive at the result of fuzzy comprehensive evaluation.

On the assumption that the weight cloud relative to index j is WC_j, the drops whose membership degrees are higher than a fixed threshold λ are selected and then the vector W_j of index j is composed by the values that relative to them. If the number of drops selected is G, it is obvious that the dimension of W_j is G, too. Similarly we can get the weight vectors for the rest of indexes we concerned, and compose the weight matrix $W_{matrix}=\left[W_1^T,\cdots,W_K^T\right]$ of all weight vectors. Then uniform all the row of W_{matrix} and remark the unitary matrix as W_{matrix}'. Thereby, the formula of fuzzy comprehensive evaluation is

$$B_{i0}=\left(b_{gn}\right)_{G\times F}=W_{matrix}'\circ R \tag{2}$$

Where $b_{gn}=\bigvee_{j=1}^{K}\left(w_{gj}\wedge r_{jn}\right), g=1,\cdots,G, n=1,\cdots,F$, then

$$B_i=\left(b_1^i,\cdots,b_F^i\right)=\left(\bigvee_{g=1}^{G}b_{gn}\right)_{1\times F} \tag{3}$$

(6) Repeat (2)~(5) until all kinds of target are computed completely.
(7) Combined with the weights of target kind, all the evaluation results gotten above are synthesized to obtain a final evaluation result.

If the matrix of all evaluation results gotten above is $B=\left(b_1,\cdots,b_F\right)$, then the final evaluation result is

$$b_n=\frac{1}{Sum}\sum_{i=1}^{S}wt_i b_n^i,\quad n=1,\cdots,F \tag{4}$$

Here $Sum=\sum_{i=1}^{S}\sum_{n=1}^{F}wt_i b_n^i$.

4 Simulation Results

Referring to the evaluation framework and the indicator system proposed in [3], the simulation conditions are set as follows:

The set of target kind is $T = \{t_1, t_2\}$; the set of target kind weight is $WT = \{0.3, 0.7\}$; the set of index is $U =$ {the mean of MRR (the measurement of recognition rate), the variance of MRR}; the set of index weight is $W =$ {extremely important, very important}; the set of remark is $V =$ {Good, Middle, Bad}

In matter of fact, the mean of MRR reflects the capability of correct recognition of an ATR system, while the variance of MRR reflects the stability of the former capability. In addition, the importance of index is divided into three grades in this simulation: extremely important, very important, and important. The intervals relative to remarks are shown in Table 2.

Table 2. Intervals relative to remarks

Index	Good	Middle	Bad
Mean of MRR	>0.8	(0.7,0.8]	<=0.7
Variance of MRR	<0.05	[0.05,0.2)	>=0.2

The data we used in the simulation are supplied from a field test data set. Our method has been done for many times in cases that the weight clouds correspond to different intervals and the result of each time are shown in Fig. 4. In Fig. 5, they are the results of the general fuzzy comprehensive evaluation proposed in [3], which are obtained under different index weights with the same data.

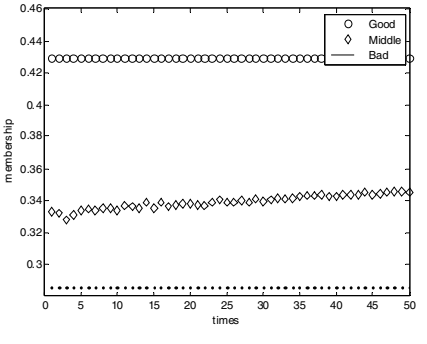


Fig. 4. The evaluation results using the method presented in this paper when the weight clouds are corresponding to various range

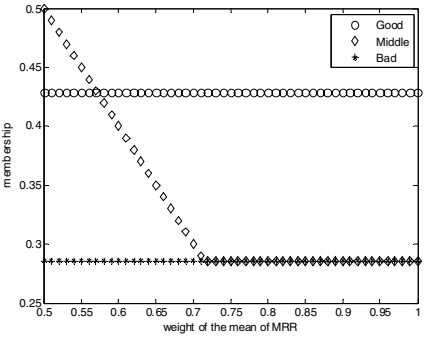


Fig. 5. The evaluation results using the method presented in literature [3] with various index weight of the mean of MRR

Fig. 5 illuminates that the general fuzzy comprehensive evaluation is sensitive to the changes of index weights. The evaluation result changes at the point of MRR=0.57 in Fig. 5, that is, if the weight of MRR is near that value, then a very small change of the weight will lead to an opposite evaluation result. Whereas, from Fig. 4, we can get the conclusion that the changes of the intervals relative to weight clouds will not affect the evaluation result as long as the lingual atoms which represent the importance of indexes are not changed.

5 Conclusion

In this paper, we introduce the forward cloud model into the ATR performance evaluation in order to solve the problems existing in fuzzy comprehensive evaluation. The simulation results given in Section IV illuminate the validity and advantage of our method comparing with other ones. However, the following need pointing out:

i. the comprehensive model using in this paper has been set up on the base of multiplex operator $\vee - \wedge$, in fact this operator only considers the effect of important indexes and ignores other relative unimportant indexes. Although the processing is very simple, too much information is wasted during it. Therefore the users can select other operators, such as weighted mean operator or all-round restrict operator or balanceable mean operator, to set up the comprehensive model in practice.
ii. Our method is modified only for the effect of index weight. So how to avoid the effects of other man-made numerical values is an area of future work.
iii. Lastly, the application of backward cloud in ATR performance evaluation and the performance evaluation in the case of small samples are also areas of future work.

References

1. Wenxian, Y.: Intelligentized Recognition Method and Application in Radar Target Recognition System on vessel[D], School of Electronic Science and Engineering, NUDT, Changsha, China (1992)
2. Arnold, C.W.: Improve ATR Performance Evaluation Via Mode Seeking. In: Signal processing, sensor fusion, and target recognition IV Conference. SPIE Proceedings, Orlando, FL, Bellingham, WA, vol. 2484, pp. 574–582 (1995)
3. Yanpeng, L.: Performance Evaluation in Automatic Target Recognition-Foundation, Theoretic System and Related Research[D], School of Electronic Science and Engineering, NUDT, Changsha, China (2004)
4. Fries, S., Klausmann, P., Jager, U., Saur, G., Willersinn, D., Hofele, G., Thonnessen, U.: Evaluation Framework for ATR Algorithms. In: SPIE, vol. 3718, pp. 438–448 (1999)
5. Brian, C.: Bassham: Automatic target recognition classification system evaluation methodology: [D] (September 2002)
6. Alsing, S.G., Blasch, E.P., Bauer Jr., K.W.: Tree-dimensional(3-D) receiver operating characteristic(ROC) trajectory concepts for the evaluation of target recognition algorithms faced with the unknown target detection problem. In: SPIE, vol. 3718, pp. 449–458 (1999)

7. Bassham, C.B., Klimack, W.K., Bauer Jr., K.W.: ATR evaluation through the synthesis of multiple performance measures. In: Proceedings of SPIE, vol. 4729, pp. 112–121 (2002)
8. Zhongxiong, H.: Fuzzy Mathematics and Applications[M], pp. 128–188. Tian Jin Science and Technology press, Tian Jin (1982)
9. Hongxing, L., Qun, W., Qinzhi, D.: Engineering Fuzzy Mathematics and Applications[M], pp. 132–136. Tian Jin Science and Technology press, Tian Jin (1993)
10. Deye, L., Haijun, M., Xuemei, S.: Membership Cloud and Membership Cloud Generator [J]. Journal of Computer Research and Development 32(6), 15–20 (1995)
11. Deyi, L., Yi, D.: Artificial Intelligence with Uncertainty. National Defense Industry Press (2005)

Secure Multicast Tree Structure Generation Method for Directed Diffusion Using A* Algorithms*

Jin Myoung Kim, Hae Young Lee, and Tae Ho Cho

School of Information and Communication Engineering, Sungkyunkwan University
300 Cheoncheon-dong, Jangan-gu, Suwon 440-746, Korea
{kjm77,sofware,taecho}@ece.skku.ac.kr

Abstract. The application of wireless sensor networks to areas such as combat field surveillance, terrorist tracking, and highway traffic monitoring requires secure communication among the sensor nodes within the networks. Logical key hierarchy (LKH) is a tree based key management model which provides secure group communication. When a sensor node is added or evicted from the communication group, LKH updates the group key in order to ensure the security of the communications. In order to efficiently update the group key in directed diffusion, we propose a method for secure multicast tree structure generation, an extension to LKH that reduces the number of re-keying messages by considering the addition and eviction ratios of the history data. For the generation of the proposed key tree structure the A* algorithm is applied, in which the branching factor at each level can take on different value. The experiment results demonstrate the efficiency of the proposed key tree structure against the existing key tree structures of fixed branching factors.

1 Introduction

Wireless sensor networks (WSNs) have emerged as an innovative class of networked embedded systems due to the union of ever smaller, less costly embedded processors and wireless interfaced with micro-sensors based on micro-mechanical systems (MEMS) technology [1,2]. WSNs are composed of small autonomous devices or sensor nodes that are networked together [3]. The applications of WSNs range from indoor applications such as smart homes and health monitoring in a hospital to outdoor applications such as highway traffic monitoring, combat field surveillance, security and disaster management [4].

Some of the most important outdoor applications, including military surveillance and the tracking of enemy ship movements and terrorist threats involve privacy and security issues [5,6]. However, sensor networks are highly vulnerable to security attacks, since the individual sensor nodes are anonymous and the communication among them is made via wireless links [7,8]. One way to implement secure wireless

* This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Advancement). (IITA-2006-C1090-0603-0028).

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 346–354, 2007.

communication in WSNs is through the use of message encryption [3]. The sensor nodes in the network share a secret encryption key(s) for encrypting the messages exchanged among them.

Sensor nodes are evicted from the network if they are compromised and exhibit malicious behavior in the network [9]. When this eviction occurs, all of the keys known to the sensor node must be changed and new keys must be securely delivered to the remaining sensor nodes, in order to prevent the evicted node from forging false messages. When the addition of the sensor nodes occurs due to the reorganization of the network or the installation of complementary nodes, it is necessary to update all known keys, so that the newly added nodes cannot understand the encrypted messages formed by the current keys [3,6]. This key update is typically accomplished by broadcasting encrypted messages, called re-keying messages, containing the new key(s).

The symmetric key technique is mostly adopted in WSNs. LEAP [10] is a key management protocol for sensor networks. It establishes four types of keys for each sensor node. Di Pitro *et al*. enhanced the LKH (Logical key hierarchy) to create a directed diffusion based LKH. This technique provides mechanisms for nodes joining and leaving groups where the key hierarchy is used to effectively re-key all nodes within the leaving node's hierarchy [1]. LKH is a tree based key management model. The LKH model was introduced in [11] to deal with secure multicasting. This model allows for the efficient update of the group key when a sensor node is added or evicted from the group. In this paper, we derive an extension to LKH.

Our central goal is to generate a secure multicast structure for the management of the group key. Generally, when a sensor transmits 1 bit of data, the energy consumption of the sensor is higher than that required for computation [12]. In this paper, we focus on the generation of the key tree for the purpose of reducing the number of re-keying messages. The eviction and addition ratios of the sensor nodes vary depending on where the sensor nodes are deployed. We generate the secure multicast tree structure by considering the eviction and addition ratios of the sensor nodes. We apply the A* algorithm to generate the structure, in order to reduce the number of re-keying messages. The experimental result is performed against the LKH. The result shows that the proposed method is more efficient in terms of the number of re-keying messages.

2 Related Works

There are many types of routing protocols in WSNs that requires key management to ensure their security. The key management procedure is an essential constituent of network security.

2.1 Directed Diffusion and Propagation Limiting Method

Directed diffusion, one of these flat network routing is a data-centric routing algorithm, where the BS sends queries to specific regions and waits for data from the nodes in the specific regions [13]. An interest message is a query which specifies what a user wants. In order to create a query, an interest message is defined using a list of attribute-value pairs such as object type, interval, selected region, and etc. After the interest message is propagated throughout the sensor field, gradients are setup to

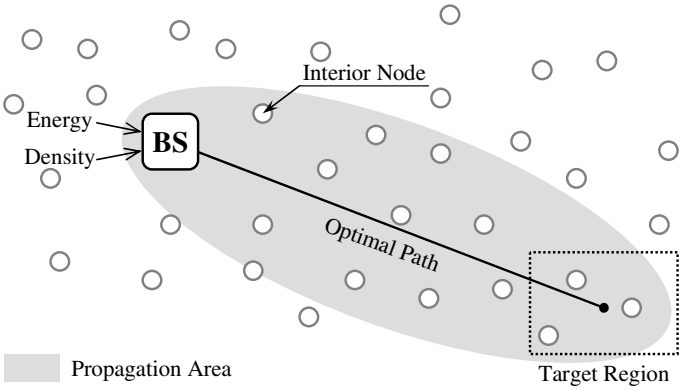


Fig. 1. The propagation limiting method (PLM)

draw data satisfying the query towards the requesting node. The gradient specifies an attribute value and a direction of data flow [13,14].

The propagation limiting method (PLM) [15] in which the data transmission area is limited according to a threshold value for reducing the energy consumption in the network. A fuzzy rule-based system is exploited to determine the threshold value by considering the energy and density of all the deployed nodes (Fig. 1). Compare to the classical flooding, PLM consumes the limited energy more efficiently since only part of the nodes participates in the propagation process. Thus, the nodes that do not participate in the propagation can save their energy, which results in the prolong network lifetime.

2.2 Logical Key Hierarchy

Wallner *et al*. proposed a hierarchical keying model called LKH. In this model, the key distribution center (KDC) maintains a key tree which is used for the update and

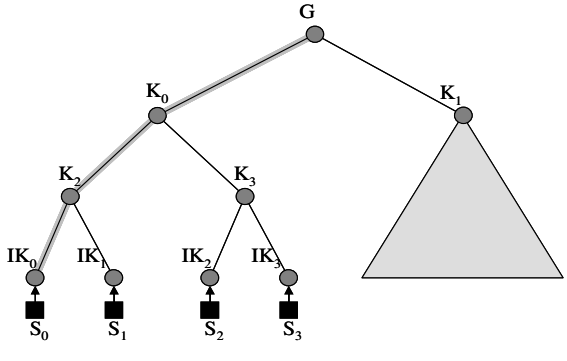


Fig. 2. Logical Key Hierarchy

distribution of the group key. Each internal node in the tree represents a cryptographic symmetric key. The center associates each group member with one terminal node of the tree. In the sensor network, the hierarchical technique can be applied in support of other keying protocols such as pre-deployed keying, to provide a mechanism for maintaining the freshness of the shared deployed cryptographic keys [7,11].

The binary key tree structure in Fig. 2, a series of cryptographic symmetric keys, e.g., G, K_0, $\cdots$, K_3, IK_0, $\cdots$, IK_3, are logically distributed in a tree rooted at the KDC. The terminal nodes in the tree correspond to the sensor nodes, e.g., S_0, $\cdots$, S_3, in Fig. 2. Each terminal node knows all the keys from its leaf node up to the root. For example, S_0 stores keys IK_0, K_2, K_0, and G. G is the group key that is shared by all of the sensor nodes in the group. The group key is used to encrypt all group communication traffic. IK_0 is the unique individual key of S_0. K_0, $\cdots$, K_3 are called key encrypted keys (KEKs). KEKs serve the sole purpose of encrypting new keys during re-keying.

2.3 Wang *et al.*'s Re-keying Scheme

Our re-keying scheme is based on Wang *et al.*'s key-oriented re-keying scheme [16]. Accordingly, for the tree in Fig. 2, when the addition and eviction of sensor nodes occur in the group, the numbers of re-keying messages needed for addition and eviction are as follows:

$$
\begin{aligned}
&Add\_MSG = h \\
&Evict\_MSG = (d-1)(h-1) \\
&where \\
&\quad \text{Add\_}MSG : \text{Number of messages needed for the addition of sensor nodes} \\
&\quad \text{Evict\_}MSG : \text{Number of messages needed for the eviction of sensor nodes}
\end{aligned}
\tag{1}
$$

In this equation, h and d are the height and degree of the tree, respectively. For example, in Fig. 2, h and d are 4 and 2, respectively, and the number of re-keying messages needed for the addition and eviction of nodes are 4 and 3, respectively.

3 Dynamic Group

In directed diffusion based sensor networks, initial key tree structure is organized by considering the history ratio of the addition and eviction operation. In this section, we describe an operation for addition and eviction of sensor nodes.

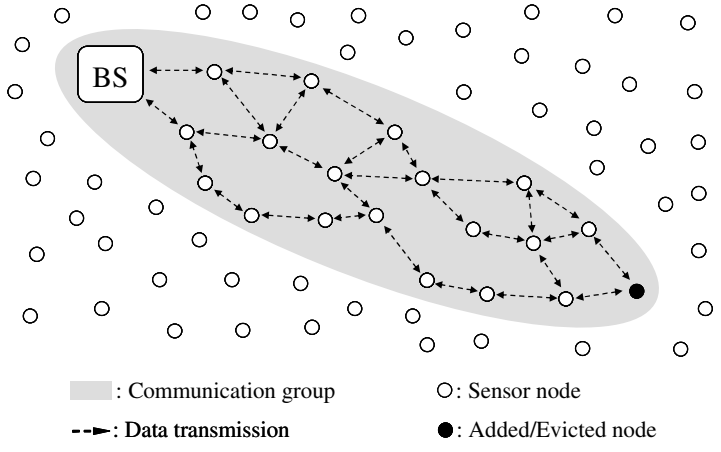


Fig. 3. Communication group in directed diffusion

3.1 Group Initialization

To protect the sensor networks from the potential attacks, confidential and authenticated group communication has to be established among the nodes and BS. Fig. 3 shows a communication among sensor nodes and BS within a propagation area. A shade area represents the propagation area. The area corresponds to a communication group. In the group, each node has a same group key, key encrypted keys and individual key. We apply the BS and nodes play the role of the multicast group members and KDC respectively. We use a group initialization mechanism that is introduced in [1].

In this paper, we assume that BS can know the history ratio of the event of the added and evicted node within the territory. These ratio values are available for the networks.

3.2 Addition Operation

An addition operation is an event that a node is joined to a communication group. The addition operation of the node occurs due to the reorganization of the communication for the collection of the sensed data or the installation of complementary nodes. In Fig. 3, for the transmission of the sensed data from a source node to BS, the source node sends interests message to its neighbor nodes. If BS receives the message from the node then BS generates a new group key or key encrypted keys, and sends to each node. The keys are sent to the nodes by Wang *et al.*'s key-encrypted re-keying scheme.

3.3 Eviction Operation

An eviction operation is an event that a node is evicted from a communication group. When a node is evicted the group, it can either be due to spontaneous or forced eviction. Spontaneous eviction is the result of insufficient energy, change of the node's sensing interest region and self-awareness of the process of being compromised. On the other hand, forced eviction is the result of intrusion detection that decides the node in question

is no longer trustable. In Fig. 2, if an evicted node is evicted from the secure channel then BS generates a new group key and KEKs and sends to the nodes in the secure channel except for the evicted node. The keys also are sent to the nodes by Wang et al.'s key-encrypted re-keying scheme.

4 Proposed Key Tree Structure

In order to reduce the number of the transmissions needed to refresh the keys that are stored in sensor nodes when addition eviction operation occurs, we use a key tree structure that can have various branching factors in each level of the key tree. The collection of branching factors is represented by *vT*, as shown in the sequence below.

$$vT = \{d_0, d_1, ..., d_h\}$$

In *vT*, d_h represents the branching factor at level *h* of the tree. The branching of the nodes at a particular level is identical, i.e., the branching factor for all of the nodes at level 1 is d_1.

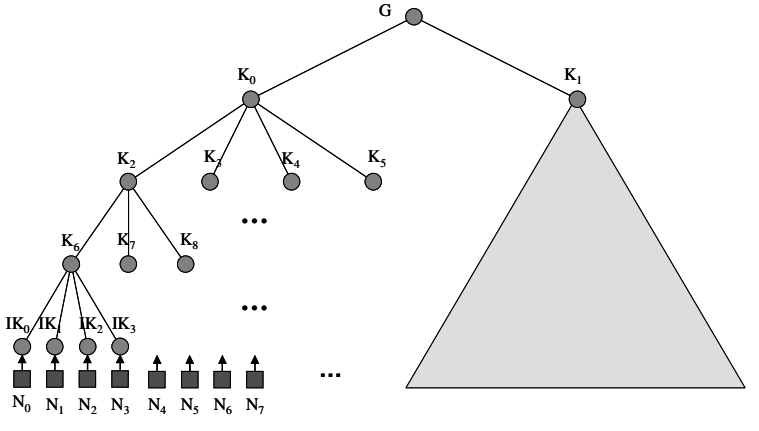


Fig. 4. The proposed key tree structure

Fig. 4 shows the example of the proposed key tree structure. The key tree structure in the figure can be represented by {2, 3, 4, 3}. As shown Section 2.2, *G* and K_i (*i*=0, 1, ⋯) are a group key and KEKs respectively in the figure. For directed diffusion based sensor networks, we apply A* algorithm to generate the key tree structure by considering the ratio of the addition and eviction operation in the networks.

5 Application of A* to the Key Tree Structure Generation

In directed diffusion that is applied PLM, confidential and authenticated communication is established among sensor nodes within a propagation area in order to ensure the security of the communication group. The nodes within a propagation area correspond to

the communication group members. We use a tree based key management technique and apply A* algorithm to generate the key tree structure for efficient key refresh.

The number of re-keying messages in the proposed structure and heuristic evaluation function that is employed in A* algorithm can be found in [17]

6 Experimental Result

An experiment was performed to compare the proposed structure with the existing key structure of the binary and ternary tree structures. The performance criterion is the average number of re-keying messages needed when addition/eviction events occur. The average number of re-keying messages is classified by the various numbers of sensor nodes and the ratio of the addition and eviction operation within a communication group. We measured the average number of re-keying messages required according to the various history data of the addition and eviction operations. The A* algorithm based key structures are constructed for various ratios of addition and eviction operations to measure the average number of re-keying messages in order to show how effective the proposed method is.

Table 1. Average number of re-keying messages for various network sizes and history data

Ratio / Size	Add = 0.1 Evict = 0.9	Add = 0.3 Evict = 0.7	Add = 0.5 Evict = 0.5	Add = 0.7 Evict = 0.3	Add = 0.9 Evict = 0.1
20	5.0/5.1/5.8	5.0/5.3/5.4	5.0/5.5/5.0	4.6/5.7/6.7	3.6/5.9/4.2
40	6.0/6.1/7.7	5.9/6.3/7.1	6.0/6.5/6.5	5.2/6.7/5.9	3.9/6.9/5.3
60	6.1/6.1/7.7	6.3/6.3/7.1	6.5/6.5/6.5	5.5/6.7/5.9	4.1/6.9/5.3
80	7.0/7.1/9.6	7.0/7.3/8.8	7.0/7.5/8.0	6.2/7.7/7.2	4.5/7.9/6.4
100	7.1/7.1/9.6	7.3/7.3/8.8	7.0/7.5/8.0	6.2/7.7/7.2	4.8/7.9/6.4
120	7.1/7.1/9.6	7.3/7.3/8.8	7.5/7.5/8.0	6.5/7.7/7.2	4.9/7.9/6.4
140	7.9/8.1/9.6	7.7/8.3/8.8	7.5/8.5/8.0	6.5/8.7/7.2	5.0/8.9/6.4

In Table 1, the size and ratio represent the number of sensor nodes and the ratio of addition/eviction events in the networks, respectively. The resultant value, e.g., 5.0/5.1/5.8, indicates the average number of re-keying messages required for the addition/eviction events in the A* algorithm based key structure, binary tree key structure and ternary tree key structure, respectively. Accordingly, Table 1 shows that the A* algorithm based key structure needs an average number of re-keying messages which is less than or equal to that of the other key tree structures.

The simulation result in Fig. 5 shows the measurement result of the remaining energy level of sensor nodes after refreshing group key for events. The events consist of the addition and eviction of sensor nodes and randomly occur in the network. In this simulation, 500 sensor nodes are randomly placed in 400m x 400m territory. We assumed that the energy level of each sensor node is 0.25J. The energy consumption caused by a message transmission and reception is 16.25μJ per meter and 12.5μJ. Each node has the transmission range of 40 meters.

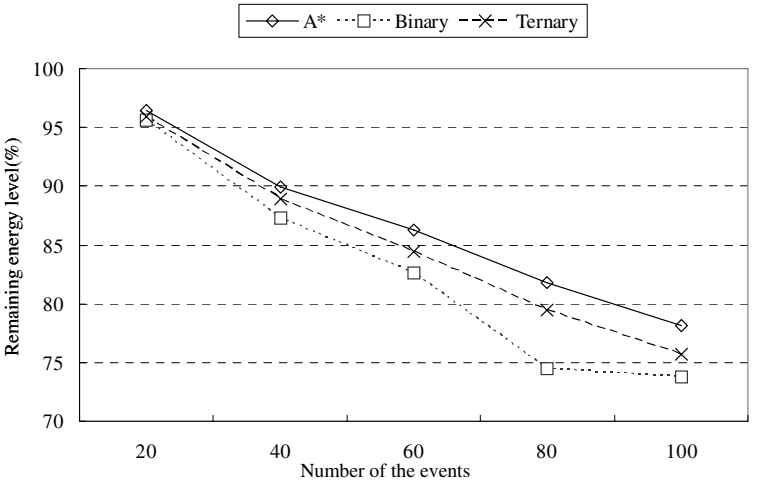


Fig. 5. Remaining energy level of all sensor nodes in the networks

We can know that the remaining energy level in the key tree that is generated by the proposed method is higher then the binary and ternary structure in all range when the messages

7 Conclusion

In various applications, WSNs enable the monitoring of the target system or area. Especially in areas such as military, commercial, and privacy applications, ensuring security is the most important issue. In this paper, we showed how a secure multicast key tree structure can be generated by exploiting the A* algorithm in directed diffusion. The algorithm generates the key tree structure by considering the history data of addition and eviction of sensor nodes in the territory. The generated key tree structure is effective, when frequent addition and eviction occurs, in that the number of re-keying messages required for managing the key system is reduced compared to that of the existing key tree structures.

Future works will improve the proposed method by considering the re-keying message update cycles and propose a multi addition and eviction event of sensor nodes. The storage issue involved in managing the hierarchical key structure will also be considered.

References

1. Pietro, R.D., Mancini, L.V., Jajodia, S.: Providing secrecy in key management protocols for large wireless sensors networks. AdHoc Network 1, 455–468 (2003)
2. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless Sensor Networks: A Survey. Computer Networks Journal 38, 393–422 (2002)

3. Eltoweissy, M., Wadaa, A., Olariu, S., Wilson, L.: Group Key management scheme for large-scale sensor networks. Ad Hoc Network 3, 668–688 (2005)
4. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A Suvey on Sensor Networks. IEEE Communications Magazine 40, 102–114 (2002)
5. Sahoo, P.K., Chen, J.J.-R., Sun, P.-T.: Efficient Security Mechanisms for the Distributed Wireless Sensor Networks. ICITA'50 (2005)
6. Eltoweissy, M., Younis, M., Ghumman, K.: Lightweight Key Management for Wireless Sensor Networks. In: IEEE International Conference Performance on Computing, and Communications, pp. 813–818 (2004)
7. Carman, D., Kruus, P., Matt, B.: Constraints and approaches for distributed sensor networks security. NAI Technical Report (2000)
8. Jolly, G., Kuscu, M., Kokate, P.: A hierarchical key management method for low-energy wireless sensor networks. UMBC Online Document (2002)
9. Chan, H., Perrig, A., Song, D.: Random key pre-distribution schemes for sensor networks. In: IEEE 2003 Symposium on Security and Privacy (2003)
10. Zhu, S., Setia, S., Jajodia, S.: LEAP: efficient security mechanisms for large-scale distributed sensor networks. ACM CCS, 27–31 (2003)
11. Wallner, D., Harder, E., Agee, R.: Key management for multicast: Issues and architectures. IETF RFC 2627 (1999)
12. Pottie, G.J., Kaiser, W.J.: Wireless Integrated Network Sensors. Communications of the ACM 43, 51–58 (2000)
13. Intanagonwiwat, C., Govindan, R., Estrin, D., Heidemann, J., Silva, F.: Directed diffusion for wireless sensor networking. IEEE ACM T. Network 11(1), 2–16 (2003)
14. Heideman, J., Silva, F., Intanagonwiwat, C., Govindan, R., Estrin, D., Ganesan, D.: Building Efficient Wireless Sensor Networks with Low-Level Naming. In: Proc. of SOSP, pp. 146–159 (2001)
15. Chi, S.H., Cho, T.H.: Fuzzy Logic based Propagation Limiting Method for Message Routing in Wireless Sensor Networks. In: Gavrilova, M., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganà, A., Mun, Y., Choo, H. (eds.) ICCSA 2006. LNCS, vol. 3983, pp. 58–67. Springer, Heidelberg (2006)
16. Wong, C., Gouda, M., Lam, S.: Secure group communications using key graphs. IEEE/ACM Transactions on Networking 8, 16–30 (2000)
17. Kim, J.M., Cho, T.H.: A* based Key Tree Structure Generation for Group Key Management in Wireless Sensor Networks. Computer Communications, Elsevier (submitted, 2007)

Simulation of Soft Deformable Objects for Virtual Reality Medical Applications

Kup-Sze Choi

Centre for Integrative Digital Health, School of Nursing
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
hskschoi@inet.polyu.edu.hk

Abstract. Interactive simulation of deformable objects is a key component in various virtual reality medical applications. While many methods based on mass-spring models have been proposed, this paper presents an alternative that considers deformation as a result of forces propagating sequentially from one mass point to another, where matrix formulation is not required. The simulation speed is controllable by varying the scope of localized deformation. Detailed study on the propagation sequence and penetration depth is performed. Simulation speed is improved with the aid of a lookup table created by pre-computation. Results on deformable simulation and timing performance are presented. The approach is feasible for developing medical applications involving interactive simulation of soft deformable objects.

Keywords: Virtual reality, surgical simulation, deformable simulation, mass-spring models.

1 Introduction

Virtual reality (VR) based computer simulation plays an increasingly important role in heath care and medicine. In medical education, training of medical procedures has employed VR technologies to improve the learning curve, from grasping, cutting to suturing in minimally invasive surgery [1], for instance. Besides, VR has also been incorporated in diagnosis and treatment. Virtual colonoscopy has emerged as an alternative diagnostic tool that avoids the discomfort and risk of perforation in conventional approach. A recent example is the monitoring of fetus development by using real time interactive visualization technique. In psychotherapy, virtual scenarios are reproduced in the cyberspace to help patients to overcome phobia [2]. VR is also helpful to rehabilitation in developing therapeutic tools with variable difficulty levels and quantitative evaluation of the recovery progress [3].

Many of these VR applications involve interactive simulation of soft deformable objects. The virtual objects should change their shapes autonomously in response to user's manipulation. Ideally, such changes should be realistic and take place in real time. Realism, however, usually comes at the expense of simulation speed. Physics-based approaches have been proposed in attempt to achieve this goal and the two major streams are the mass-spring model (MSM) and the finite element method

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 355–364, 2007.

(FEM). Generally speaking, the former is easier to implement, computationally more efficient but less accurate than the latter. In this paper, an alternative mass-spring model, called Force Propagation Model (FPM), is presented, where deformation of soft objects is modeled as a sequential propagation of forces from one mass point to another. The simulation speed is controllable by the scope of propagation. Detailed study on the propagation sequence and the effect of penetration depth on deformable simulation is performed. To expedite the simulation, pre-computation is performed to eliminate run-time determination of the propagation sequence. The rest of the paper is organized as follows. Section 2 gives an overview of MSM and FEM. Section 3 introduces the proposed FPM and gives details of the method. The pre-computation of propagation sequence is discussed in Section 4. Deformation of virtual objects using FPM and the timing results are presented in Section 5 and 6 respectively. A summary is finally given in Section 7.

2 Simulation of Deformable Objects

Many methods have been proposed for simulating deformable objects. A comprehensive review can be found in reference [4]. Among these methods, MSM and FEM represent two main schools of physics-based techniques. In MSM, a deformable object is discretized into a system of mass points (or nodes) connected by springs. Deformation is then obtained by formulating the particle dynamics using the Newton's Law and the Hooke's Law. It requires the formulation of stiffness matrices, and the solution to a system of linear equations that involves inverse matrix computation using numerical methods. MSM has been used in various areas, including endoscopic surgery training [1], facial tissue deformation and craniofacial surgery [5].

In FEM, the formulation begins by decomposing the object into a number of finite elements. Global deformation is then determined by computing the nodal displacement of each constituting element, in which continuum mechanics, i.e. the generalized 3D Hooke's Law, is applied to model the stress-strain relationship. Like MSM, formulation of FEM involves matrix algebra and the differential equations obtained from the Newton's Law are solved for the dynamics of the nodes. Details are available in standard texts [6]. FEM is mathematically more rigorous and accurate, but it is in general computationally intensive and not suitable for interactive applications unless the numerical computations are accelerated. FEM has also been widely used in VR medical applications, in particular, virtual surgery [7], to obtain more accurate simulation results.

While the deformable model presented in this paper is a variation of MSM, it employs a different mathematical treatment that does not require any matrix formulation in solving the equations of the particle dynamics, thus avoiding the matrix computations involved in conventional MSM. Interactions due to the forces acting on the surface nodes of deformable objects propagate to the inner nodes one by one, until reaching a pre-defined boundary. The mechanism is shown schematically in Fig. 1. Instead of considering deformation globally, this model effectively focuses computation resources on the interacting nodes in the localized regions. The simulation is thus scalable by specifying the degree of localization.

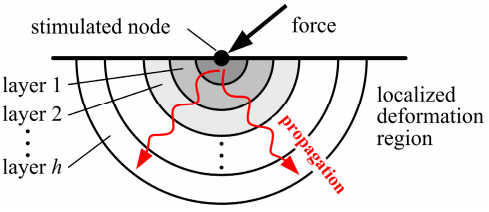


Fig. 1. Schematic diagram showing the mechanism of FPM

3 The Force Propagation Model

The basic idea of FPM is introduced in the section, with detailed discussions on the propagation sequence and the penetration depth.

3.1 Dynamics of Mass Points

In FPM, when an external force (referred to as stimulus hereafter) is applied to a node in the system, the force propagates from the point of contact, namely the stimulated node, to its neighboring nodes via the interconnecting springs. The process proceeds in an ordered manner from the nearest neighbors to the farthest ones. The dynamics of the nodes is governed by the Newton's law of motion. The nodal displacement of the i-th node, $\mathbf{u}_i$, due to an external force $\mathbf{F}_i$, is given by

$$m_i \ddot{\mathbf{u}}_i + d_i \dot{\mathbf{u}}_i + \sum_j \frac{k_{ij}(|\mathbf{r}_{ij}| - l_{ij})}{|\mathbf{r}_{ij}|} \mathbf{r}_{ij} = \mathbf{F}_i . \quad (1)$$

where m_i and d_i are respectively the mass and damping constant of the node i, $\mathbf{r}_{ij}$ is the vector distance between node i and node j, l_{ij} and k_{ij} are respectively the rest length and the stiffness. Equation (1) can be solved by using classical numerical integration techniques. In the current implementation, solvers based on the central difference method and the Runge-Kutta method are developed to allow for flexibility in performing numerical integration depending on the required speed of convergence and accuracy.

3.2 Sequence of Propagation

To simulate the propagation process, it is necessary to determine the sequence of nodes that forces would propagate. Deformation is simulated by evaluating sequentially the new nodal position on per-node basis according to the priority given by the sequence. The sequence is investigated under three situations, namely, single stimulus, intra-layer propagation, and multiple stimuli.

Single Stimulus. When only one mass point is subjected to external force, the propagation order is similar to that of breadth-first traversal, which begins from the contact point to all its immediate neighbors, followed by the neighbors of neighbors, until reaching the maximum depth of traversal – the *penetration depth* that limits the

size of the deformation region. For a mass-spring network, penetration depth is also referred to as the *topological distance* from the contact point to the farthest node.

Intralayer Propagation. In principle, the forces due to the nodes in one layer would exert on all the nodes in the next layer simultaneously via the springs. This would revert to the formulation of stiffness matrix in conventional MSM. In FPM, deformable simulation is reduced to the evaluation of the position of individual nodes and hence no matrix formulation is required. Note that the order of propagation for the nodes in the same layer, i.e. the intralayer propagation sequence, is neglected, however. This is justified experimentally by performing deformable simulation with four different intralayer propagation sequences as defined below.

- *D* – the default sequence, following the order of breadth-first traversal
- *R* – the reverse of the normal sequence *D*
- *Rnd1*, *Rnd2* – two randomized sequences

In this experiment, forces are applied to a cluster of 21 surface nodes of a regular mass-spring lattice. The effect of the propagation sequence on deformation is studied by comparing the nodal position evaluated by two different sequences. The normalized difference in nodal position Δ is given by

$$\Delta\ (\%) = \frac{1}{l} | \mathbf{p}_i(\xi) - \mathbf{p}_i(\zeta) | \times 100\% \cdot \tag{2}$$

where $\mathbf{p}_i(\xi)$ and $\mathbf{p}_i(\zeta)$ are the simulated position vectors of the node i obtained by using the propagation sequence ξ and ζ respectively, l is natural length of the spring. The experimental result is given in Table 1. It indicates that the maximum difference in nodal position due to deformable simulations using the sequence *D* and *R* (i.e. the worst case) is 0.77% of the spring's natural length, which should be visual imperceptible. When the results of the two randomized sequences are compared, the maximum difference is only 0.31%. Hence, it is concluded that the intralayer propagation sequence only plays a minor role in FPM and can be neglected. In the current implementation, the sequence directly follows the order of breadth-first traversal.

Table 1. Effect of intralayer propagation sequence

Δ (%)	*D* vs *R*	*D* vs *Rnd1*	*D* vs *Rnd2*	*Rnd1* vs *Rnd2*
Mean	0.10	0.04	0.05	0.02
Max	0.77	0.43	0.44	0.31
Min	0.22×10^{-4}	1.08×10^{-4}	0.69×10^{-4}	0.38×10^{-4}

Multiple Stimuli. The scenario is complicated by the presence of multiple localized deformation regions which may interact or overlap with each other. Care should be taken to ensure the synchronization of force propagation in these regions. This is modeled by performing depth synchronization such that the layers of nodes at the same depth but belonging to different regions have the same priority in the overall propagation process.

3.3 Penetration Depth

Penetration depth determines how far the forces would propagate and controls the number of nodes involved in deformable simulation. It can be adjusted to vary the simulation speed against the degree of localization. To estimate the optimum depth, it is required to calibrate the model by performing deformable simulation at various penetration depth. By monitoring the incremental change in the simulated profile of a deformed object at increasing depth, it is anticipated that no noticeable change would be observed as the penetration is further increased beyond a certain value, which is the optimum depth (see Fig. 2). The estimation process is automated by setting a threshold of perceptible change in the simulated profile. As the optimum depth is model-dependent, it means that an extra pre-computation step is required to estimate the optimum depth for each model to be used before simulation can take place. The effect of the direction of applied force and spring stiffness on the optimum depth has been studied by using this estimation approach.

Direction of External Force. The effect of the direction of external force on the optimum depth is studied by applying a force of constant magnitude to an isotropic lattice, but at different angles to the surface normal. Fig. 2 shows the simulated profiles of the deformed lattice when the applied force is at 0°, 30°, 45° and 60° to the surface normal. It is found that the difference in the estimated depths is not

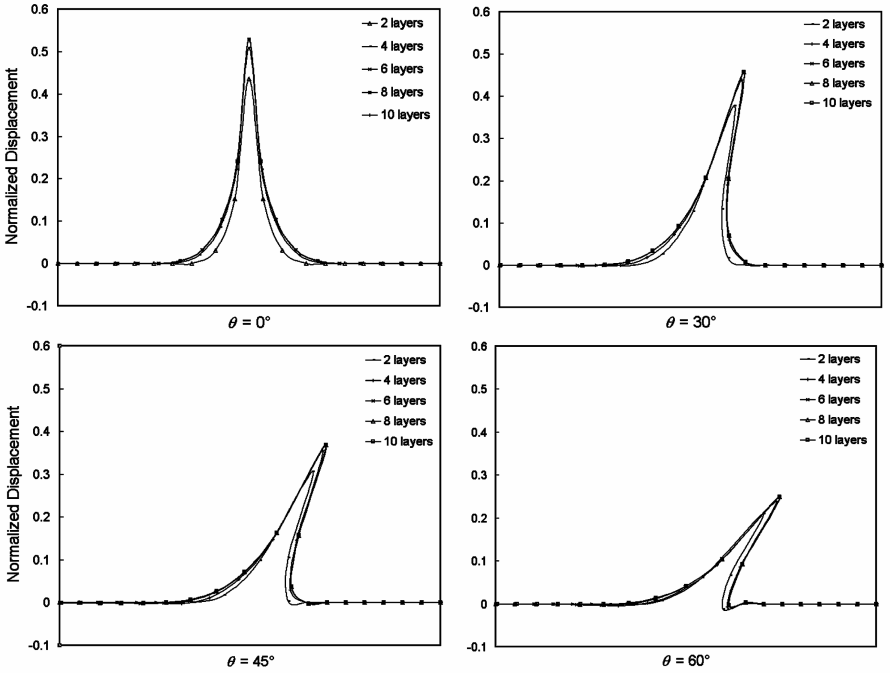


Fig. 2. Simulated profiles obtained by pulling a surface node of an isotropic lattice

remarkable. The optimum depth is 6 layers in all the four cases, regardless of the force direction. Further increasing the penetration beyond 6 layers does not produce significant difference. The result shows that the optimum penetration depth is not sensitive to the direction of the applied force.

Stiffness. Investigation has been made by estimating the optimum depth of isotropic lattices, each with different spring stiffness. It is found that that the greater the stiffness, the deeper the required penetration depth. This can be explained by the fact that, under the same external force, it is more difficult to deform a hard object and only results in deformation in a highly localized region. Hence, a small penetration depth is sufficient for simulating hard objects.

4 Pre-computation of Propagation Sequence

Determining propagation sequence by conducting breadth-first search at run time can reduce the simulation speed, especially for large penetration. To improve the timing performance, the sequence is pre-computed offline and stored in a lookup table (see Fig. 3). For each surface node i on the object, the table stores the neighbors of node i at different penetration depth, from the group of nearest neighbors at layer 1, $g(i,1)$, to farthest group at the optimum depth n, $g(i,n)$. When a force is applied to node p, suppose the specified depth is h ($\leq n$), the sequence can be obtained simply by making a table lookup, i.e. the successive retrieval of $g(p,1)$, $g(p,2)$ up to $g(p,h)$. When multiple surface nodes are subjected to external forces, to ensure depth synchronization, neighbors of each node at the same depth are retrieved and grouped together first before those at the next layer are appended. For example, if two surface nodes p and q are moved by a user, the corresponding propagation sequence is

$$\{ \{g(p,1), g(q,1) \}, \{g(p,2), g(q,2) \}, ..., \{g(p,h), g(q,h) \} \} . \quad (3)$$

Note that common neighbors may exist between $g(p,j)$ and $g(q,j)$ at some common layer j (i.e. $g(p,j) \cap g(p,j) \neq \varnothing$ for some j) if the localized deformation regions due to

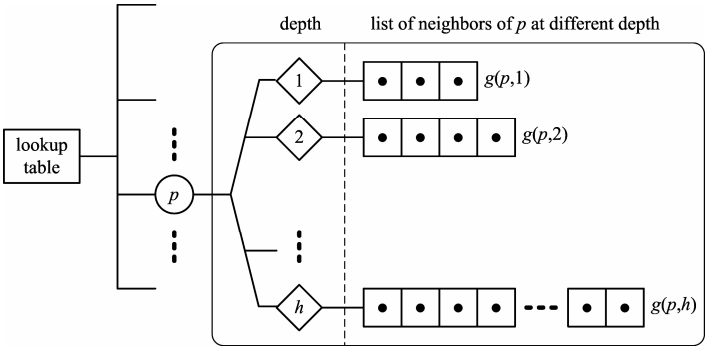


Fig. 3. Schematic diagram of the lookup for storing the propagation sequences

node p and q overlap. This can occurs when p is close to q, or when the scope of propagation is large. It is therefore necessary to preclude repeated nodes appearing in the propagation sequence. After pre-computation is performed, the appropriate sequence can be obtained promptly by table lookup without performing breadth-first search on the fly.

5 Deformable Simulation

FPM has been applied to simulate both volumetric and surface models. Fig. 4 and Fig. 5 show the deformation of a lattice containing 20×4×10 mass points, with two stimuli applied at the opposite sides (left) and at the same side (right) respectively. The penetration depth is set to 10 layers to demonstrate FPM's ability in handling multiple stimuli and overlapping deformation regions. Next, the deformation of models created by triangulated surface meshes is shown in Fig. 6. In addition, an interactive system is built on a multithreaded platform with PHANToM (from SensAble Technology Inc.) as the haptic user interface. The endpoint of the PHANToM's stylus is modeled as the probing object and point-based haptic rendering technique is applied [8]. The probe-polygon collisions are handled by using the haptic software toolkit. Fig. 7 shows the interactive deformation of a surface-based heart model. A virtual probe is used to represent the end effector of the PHANToM arm.

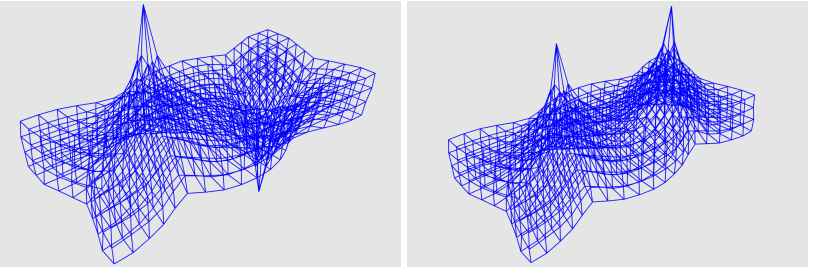


Fig. 4. Lattice deformation represented with wireframe

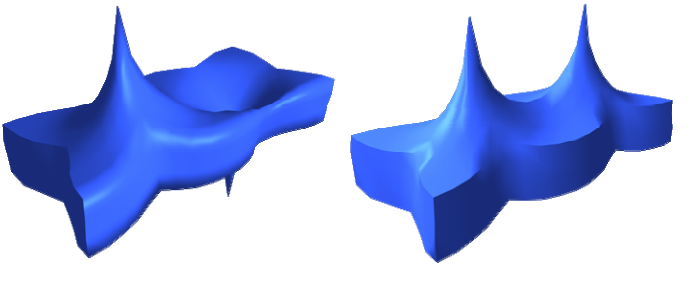


Fig. 5. Surface rendering of the deformed lattice shown in Fig. 4

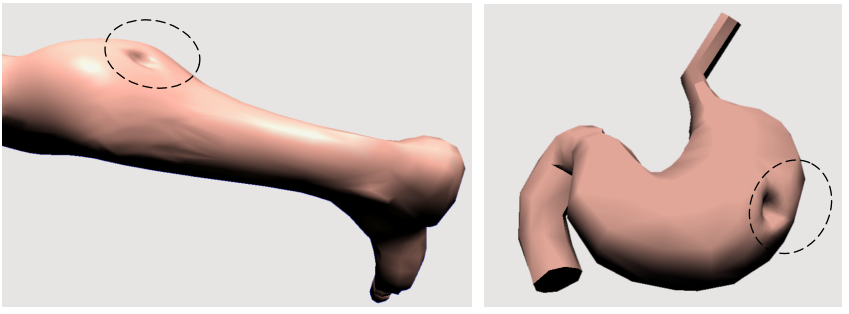


Fig. 6. Simulation using surface-based models. Deformed regions marked with dashed lines.

Fig. 7. Interactive deformable simulation of a virtual heart model (3372 nodes, 6476 triangles)

6 Timing Performance

FPM is scalable simply by adjusting the penetration depth. Its time complexity is O(N) where N is the number of nodes in the propagation sequence. In terms of the penetration depth m, the complexity for surface and volumetric model are respectively O(m^2) and O(m^3). By limiting the extent of deformation, computation can be expedited to facilitate real-time interactive applications. The efficiency of FPM is studied by measuring the time spent on calculating the new nodal positions. On an Intel Pentium IV 2.26 GHz computer, experiment is carried out by applying a single

Table 2. Computation time for nodal position update

Penetration depth	Number of nodes involved	Time (msec/time-step)
2	19	0.16
4	85	0.36
6	231	0.86
8	489	2.45
10	891	5.83
12	1469	12.46
14	2255	26.10

stimulus to the surface of a lattice containing 40×20×40 nodes. Time elapsed for 100 time-steps is measured at different penetration depth. The data are tabulated in Table 2. It is found that is the time taken to update the nodal position is $O(m^{2.83})$ which agrees with the scalability of volumetric model as discussed above.

7 Summary

Comprehensive analysis and evaluation of the proposed FPM are presented in this paper. Simulation is performed on per-node basis to avoid the matrix formulations in conventional MSM. Algorithms are developed to determine the sequences of force propagation and to synchronize the process in overlapping or multiple regions. Simulation speed of FPM is scalable against accuracy simply by tuning the penetration depth. While the optimum depth depends on material stiffness, it is insensitive to the angle of the applied forces. Run-time breadth-first search is eliminated by creating a lookup table to store the sequences at various penetration depths. FPM has been used to simulate interactive cutting of deformable objects [9]. To further exemplify FPM's ability for VR medical simulation, an integrated application will be developed to simulate the real scenarios in the training of a specific medical procedure. In addition, FPM is being exploited to serve as general technique for simulating collaborative tasks through multi-user interactions over the network [10]. This is attributed FPM's scalability where simulation speed can be adjusted dynamically by varying the penetration depth to adapt to network traffic.

Acknowledgement. The work was supported in part by the Research Grants Council of the Hong Kong SAR (No. PolyU 5145/05E and No. PolyU 5147/06E). The author would like to thank F.L. Chung and K. Sylvain for their support to this work.

References

1. Kühnapfel, U., Çakmak, H.K., Maab, H.: Endoscopic Surgery Training Using Virtual Reality and Deformable Tissue Simulation. Computers and Graphics 24, 671–682 (2000)
2. Banos, R.M., Botella, C., Perpina, C., Alcaniz, M., Lozano, J.A., Osma, J., Gallardo, M.: Virtual Reality Treatment of Flying Phobia. IEEE Transactions on Information Technology in Biomedicine 6, 206–212 (2002)
3. Deutsch, J.E., Lewis, J.A., Burdea, G.: Virtual Reality-Integrated Telerehabilitation System: Patient and Technical Performance. In: International Workshop on Virtual Rehabilitation, pp. 140–144 (2006)
4. Nelean, A., Müller, M., Keiser, R., Boxerman, E., Carlson, M.: Physically Based Deformable Models in Computer Graphics. Computer Graphics Forum 25, 809–836 (2006)
5. Mollemans, W., Schutyser, F., Van Cleynenbreugel, J., Suetens, P.: Fast Soft Tissue Deformation with Tetrahedral Mass Spring Model for Maxillofacial Surgery Planning Systems. In: Barillot, C., Haynor, D.R., Hellier, P. (eds.) MICCAI 2004. LNCS, vol. 3217, pp. 371–379. Springer, Heidelberg (2004)
6. Zienkiewicz, O.C., Taylor, R.L.: The Finite Element Method. 5th edn. Butterworth Heinemann, Boston (2000)

7. Bro-Nielsen, M.: Finite Element Modeling in Surgery Simulation. Proceedings of the IEEE 86, 490–503 (1998)
8. Zilles, C.B., Salisbury, J.K.: A Constraint-based God-object Method for Haptic Display. In: IEEE International Conference on Intelligent Robots and Systems, vol. 3, pp. 146–151. IEEE Press, New York (1995)
9. Choi, K.S.: A Digital Design Approach to Interactive Cutting of Deformable Objects with Adaptive Mesh Refinements. Computers and Graphics 30, 233–243 (2006)
10. Qin, J., Choi, K.S., Heng, P.A.: An Adaptive Framework using Cluster-based Hybrid Architecture for Enhancing Collaboration in Surgical Simulation. In: Medicine Meets Virtual Reality Conference, pp. 367–372. IOP Press, Amsterdam (2007)

Applying Open Source Game Engine for Building Visual Simulation Training System of Fire Fighting

Diping Yuan, Xuesheng Jin, Jin Zhang, and Dong Han

Chinese People's Armed Police Force Academy
Langfang, Hebei province, 065000, P.R. China
dpyuan2002@yahoo.com.cn

Abstract. There's a growing need for fire departments to adopt a safe and fair method of training to ensure that the firefighting commander is in a position to manage a fire incident. Visual simulation training systems, with their ability to replicate and interact with virtual fire scenarios through the use of computer graphics or VR, become an effective and efficient method for fire ground education. This paper describes the system architecture and functions of a visual simulated training system of fire fighting on oil storage, which adopting Delat3D, a open source game and simulation engine, to provide realistic 3D views. It presents that using open source technology provides not only the commercial-level 3D effects but also a great reduction of cost.

1 Introduction

Fire is a kind of unexpected incident that imperils people's security and property. In recent years, with the great appearance of urban high-rise buildings, special constructions such as subway, nuclear power plant and modern petrochemical enterprise, a great deal of special fires occur, with flowing, explosive and destructive features. This has put forward a significant task for fire departments: how to take corresponding extinguishing measures and technology to improve the extinguishing efficiency [6].

Fire fighting training in the fire departments is fundamental to the safety of personnel at a fire incident as well as the success in dealing with this emergency in an efficient and effective manner [6]. A failure to offer adequate training has resulted in deaths and injuries to fire department personnel.

Traditional training way gives priority to perform physical and skill training with the aid of simulating equipment, while the commander's tactical training on purpose of ensuring the success of fire extinguishing and rescue, as well as safeguarding firefighters' safety is fulfilled through live exercise, tabletop exercise and paper-based study[6]. Live exercise achieves remarkable success in improving commander's mental quality and commanding experience, it is necessary, but not all the time. It's too expensive and exist safety problem. Tabletop exercise lacks of real time interactive and evaluating objectivity. It's also time-consuming in constructing process. Paper-based study is less viewable, and could only offer assumption and simulation of simple fire scene. In recent years, the application of 3-D animation technology achieves the CAI system, which furthers the tabletop exercise and has the same problem. For traditional tactical training method, there are the problems of time, resourced and evaluating objectivity [3].

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 365–374, 2007.

The rapid progress of computer graphics, artificial intelligence, visual simulation and virtual reality, and their successful applying in military training[4] and entertainment, make fire departments across the world realize these technology's benefits for firefighting training and assessment. Computer based visual simulation training system is fast becoming the fire service's first choice.

In General, a visual simulation fire fighting tactical training system should include the following basic modules:

Training parameters initializing. Set a wider variety of fire situations, fire fighting personnel, equipments, surrounding environments, and weather condition etc.

Firefighting tactical managing. Providing an interface for the trainees to analysis, decide, make and releasing a fire fighting action order or plan relying on the related tactics.

Management evaluating. Objectively analysis and assess the tactical plan and disposition put forward by trainees. Give trainees a reasonable and detail explanation as well as induce the next step of fire situation.

Fire scenarios display. Providing a 2-D and 3-D virtual fire situation panorama or scenarios, which can make trainees a virtual fire feelings so that they can immerse and interact, really "learning while doing".

Fire scenarios and action process replaying. Record whole simulation training process and replay so as to comment on it after the incident.

Compared firefighting visual simulation training systems with military training applications, they all need applying AI technology in sensing prior and current status, making decision, finding path and evaluating results. In visualization aspect, they all need to provide a user interface allows intuitive interaction and rapidly increases situational awareness. However, firefighting training systems have their different features:

It concerns personal training or group training under a LAN environment. Large and complicated distributed simulation environment is not needed.

It emphasizes more on small scale area and local detail visual simulation than large scale range. Such as internal of a building, a fired oil tank, a fired plane etc.. Large scale terrain simulation is not cared.

It focuses on smoke and fire's dynamic or time varying fidelity.

It focuses on human behavior mode in the emergence condition.

Investment in developing and maintain firefighting visual simulation training systems is far less than that of in military training system.

Features list above decide it is very important that building a visual simulation training system of firefighting based on a cheap, high effective and real-time rendering engine.

In order to validate the mentioned thought above, we developed a fire fighting command visual simulation training system prototype, based on Delta 3D, an open source game engine. The mission of this system is to improve and enhance commanders' tactical command ability and quality in collaborative dealing with oil tank fire. Through implementing this prototype, we hope to summarize and form a standard

framework for developing similar training systems, such as high rise building fire, petrochemical fire, airplane fire etc[5][7][8].

2 System Architecture

2.1 Hardware Architecture

A basic training unit consists of four sections. That is: Instructor Operation station, Incident Command Station, Cooling and Extinguish Crew Command Station and Logistics and Support Crew Command Station, linked by a network. These four sections play different roles and perform various functions(Figure 1).

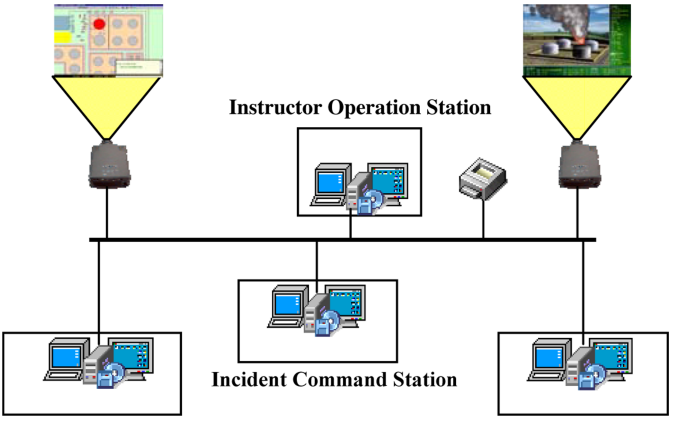


Fig. 1. System hardware architecture

The Instructor Operation station controls the training process in total. It provides such functions as:

Setting oil tanks information in the oil storage, including fired oil tank and fired-adjacent oil tank.

Setting oil tank's fire types.

Setting available fire fighting resources, including personnel and equipments.

Making action reference plan automatically, depending on the tactic knowledge base. In here, we used AI expert system technology.

Judging trainee's action plan, making the evolution of fire scenarios.

Training process replaying.

The other three commander sections, as the trainees, provide tactics action plans and deployment plans collaborative, according the current fire situation, and report all the plans to the Instructor Operation station to be judged. Each station consists of a PC with 2 display adapters, showing respective 2-D fire ground geography information, text information and 3-D interactive virtual fire scene. All these information can

also be projected to the large screen to meeting the teaching and viewing needs. In a network environment, multi basic training units can make up of a fire fighting command training group, so as to meeting various types of oil tank fire' training needs.

2.2 Software Architecture

The software architecture is the key to implement the system function(Figure 2). The system bases on client/server mode, adopts Microsoft COM/DCOM, gives full play to the functions of information transmission, 2-D information display in oil tank storage, 3-D animation playing, etc. under the network environment. A great deal of oil tanks geography data, 3-D animation data of simulation fire scene, knowledge base data of providing automatic assessment for computer are all maintained in the back-end database server. Through the interface and service offered by application program server, it could regulate and storage data from back-end database.

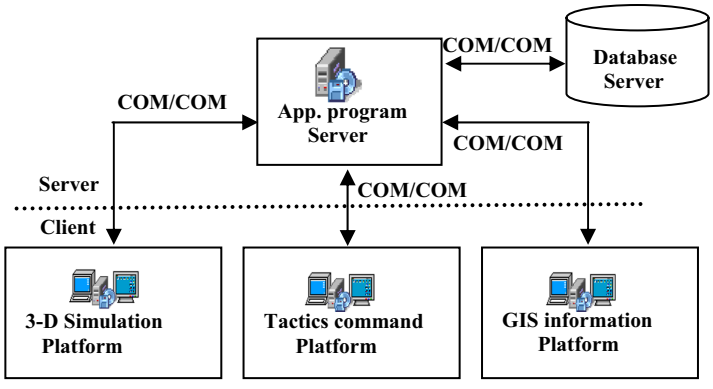


Fig. 2. Systems software architecture

3 System Function

Accordingly, System software consists of four parts: the Instructor Operation System, Incident Commander System and two Crew Commanders Systems. Listed below is the function structure graph of the four systems.

Instructor Operation System totally control training process. Incident Commander System realizes such functions as making and releasing fire fighting tactical plan, deploying teams and equipments for the other 2 crew commander systems. Cooling and Extinguish and Logistics and Support crew commander systems, according respective duty, perform orders from the IC commander system and deploy the fire-fighting resource on the 2-D map. All results are showed on the 3-D interactive fire scenarios.

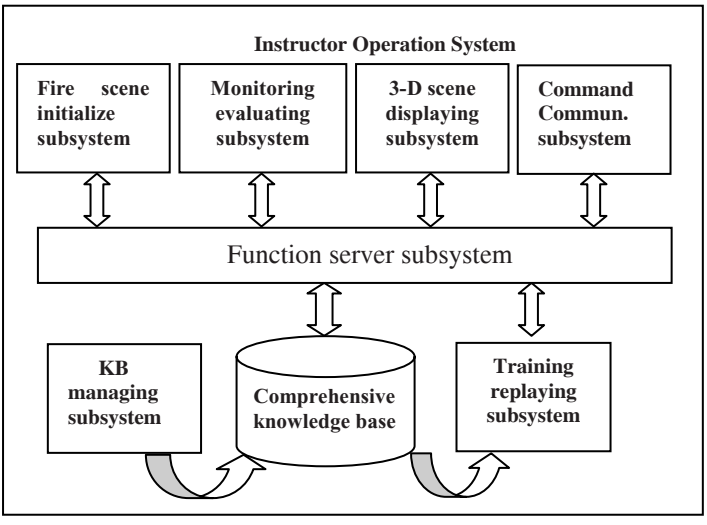


Fig. 3. Instructor operation system function graph

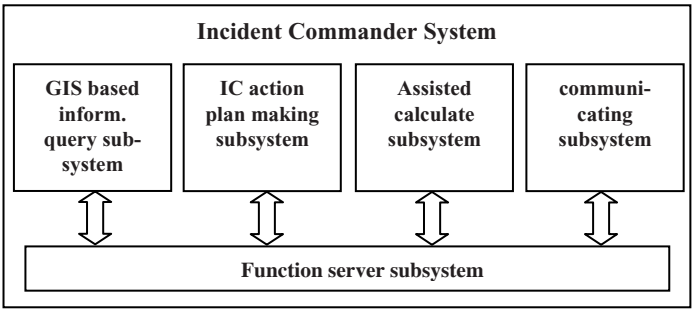


Fig. 4. Incident Commander System function graph

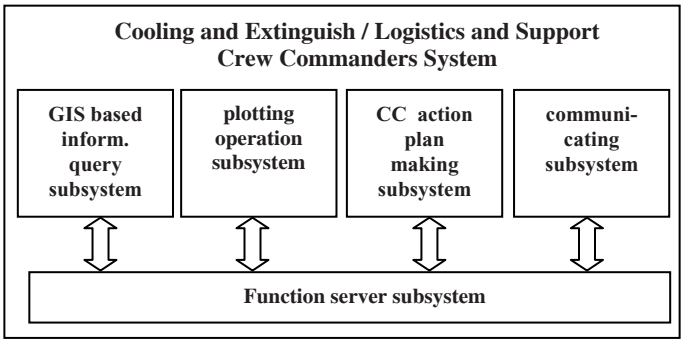


Fig. 5. 2 Crew Commanders System function graph

3.1 Monitoring and Evaluating Training Subsystem

This subsystem is the key of monitoring and directing training process. It has the functions:

In accordance with various fires derived from initial scene parameters, depending on the tactical management and forces disposition inferring knowledge base and making use of logical reasoning mechanism, the computer automatically generates fire extinguishing plan for reference.

In accordance with current fire extinguishing command process, this system provides an integrated analyzing and decision-making mechanism combing auto evaluating of machine and that of specialist, to perform intellectual assessment on tactical application during whole fire extinguishing command process. In light of comprehensive judging and explaining results, it automatically guides the fire situation, and directs the trainees, according to the changeable fire scene conditions, to input new fire extinguishing managing plan, then enter next process.

Based on the comprehensive knowledge base, it could randomly change current fire situation to directly interfere in the fire scene state, then make the training content more complicated and close to the actual situation.

It is in the form of PDU to transmit and maintain all states and data information of current training stations, and to replay and react training.

3.2 Comprehensive Knowledge Base and Its Managing Subsystem

From the considerations of quantitative and qualitative integration, we have developed a set of standard and dynamic comprehensive knowledge decision-making system, rule base expert system, which contains requisite fire professional knowledge of extinguishing oil tank fire. It is the foundation of auto judging tactical plan and auto directing fire scene, meanwhile, whether it is perfect or not, which determines the performance and work efficiency of the whole system. The comprehensive knowledge base is formed of fire extinguishing command library, random case managing command library, operating plan judging base, random case managing plan judging base, fire scene development judging base, random case fire scene development base, tactical principle base and other databank, and it adopts regular knowledge denoting method and object-oriented technology, containing over 1000 articles of rules, in addition, it offers open interface to expand and maintain the content of knowledge base timely.

3.3 Fire Extinguishing Plan-Making Subsystem

Incident Commander's command and each Crew Commander's command plans are composed of many articles of commands. They are classified and maintained in the knowledge base in terms of hierarchy and commands of two levels, totally 9 categories and more than 100 articles. In line with the comprehensive knowledge base, the system provides operating interfaces of two levels commanding plans. The commanders at all levels, according to the current fire situation, judge the fire type and developing state, then make a tactical operating plan to fulfill real time commanding process driven by command controlling and evaluating decision.

3.4 The Plotting Operation Subsystem on the Basis of GIS Vector Object

The system provides vector graphs of various traveling fire protection facilities, such as vehicles, squirt guns and so on. The forces disposition of each crew commander station bases on the increasing, deleting, revising bearing and other operation on the planar figure of MapX, a 2-D GIS plug in released by the Mapinfo. Each vector graph not only provides various information related to the illustration to realize ichnography information query, deflation and magnification without distortion. but All 2-D vector objects are correspondent with 3-D object, and the operation results are reflected real timely in the 3-D virtual fire scenarios. More importantly, it lays a foundation for monitoring and assessing consumption and loss of calculating power in subsystem, then fulfilling the auto-judge and analysis of forces disposition.

3.5 Function Server Subsystem

On purpose of mutual sharing code function, the system adopts object-oriented technological design and performance. The objects composing of each subsystem are functionally classified as substantial object, controlling object, functional object and calculating object. Among them, functional object and calculating object are in the function server subsystem (application program server).

By regulating different controlling objects, and controlling objects in charge of regulating the functional and calculating objects in function server subsystem, each subsystem performs specific functional operation in subsystem, as well as the data exchange and transfer among subsystems and so on.

4 Visualization

While 2D views of the fire ground are sufficient for this type of simulation systems, we believe the ability to browse full size fire areas, observe fire and smoke's dynamic shapes, and see human's behavior will give the trainees much better judge how the most serious threat might be positioned and what they should do firstly. Therefore, 3-D scene displaying subsystem is one of the important parts of the whole system.

Since there are less historical video materials to record typical oil tank fire cases, we should choose graphics based rendering to simulate fire scene and apply particle combustion modeling for fire and smoke.

As we have mentioned in the introduction, It is very important to select a proper rendering engine, which should not only guarantee to fulfill the 3-D real time interactive fire ground, but make the correspondence between 3-D full size fire scenarios and the 2-D plane map of disposing forces.

Beginning we consider utilize MultiGen Vega and DI-Guy. Vega is an API that is an extension of Vega Scene Graph, which replaces Performer. It provides a high-performance cross-platform software environment for real time visual simulation applications. As a successful commercial rendering engine, Vega is widely used in military simulation training systems. Vega's functions are enough strong to support our system's demands. However, it's too expensive and will make against to popularize the application among fire departments.

We also consider use UnReal game engine developed by the Epic Games and ran into the same licensing and updating issues. Moreover, the underlying engine is not open, which make the developers have no way to modify if it did not meet the specific system demands[2]. This make us shift to consider the open source game engine and select the Delta3D ultimately.

4.1 Delta3D

Delta3D is an open source game engine designed specifically for games, simulations, or other graphical applications. It is being developed at the MOVES Institute of the Naval Post Graduate School. The primary goal of Delta3D is to provide a single, flexible API with the basic elements needed by all visualization applications, based on an open source architecture. Its design Philosophy[1] follows:

Keep everything open to avoid lock-ins and increase flexibility.

Make it modular so we can swap anything out as technologies mature at different rates.

Make it multi-genre since we never know what type of application it will have to support next.

Build a community so the user doesn't have to pay all the bills.

Delta3D integrates many other well known Open Source projects which form the different modules of the engine(software architecture see also Figure 1. in [1]). For example, it use Open Scene Graph (OSG) and OpenGL for rendering, Open Dynamics Engine(ODE) for physics, Open AL for audio, Character Animation Library 3D (CAL3D) for animating characters, as well as projects such as Trolltech's Qt, Crazy Eddie's GUI (CEGUI), Xerces-C, Producer, InterSense Tracker Drivers, HawkNL, and the Game Networking Engine (GNE). Rather than bury the underlying modules, Delta3D integrates them together in an easy-to-use API which allows direct access to important underlying behavior when needed. This also means that as these other source products improve, Delta3D also benefits from those improvements, without modifying the application code entirely.

In addition to the underlying components, Delta3D provides a variety of tools such as the Simulation, Training, and Game Editor (STAGE), the particle editor, a stand-alone model viewer etc.

Using STAGE, the developer can input and position all the model types that OSG supports in the world, make them move, insert triggers, and incorporate game logic. This makes it easy for both professionals and novices to build advanced levels for Delta3D applications.

The particle system editor allows developers to use graphical tools to change the properties of a particle system and see the effects immediately in real time. This greatly speeds the development process by eliminating the need to run the application to see the effects of changing a particle system's properties.

The model viewer allows developers to load a model OSG supports quickly and view it from all angles without having to write an application to do this.

Further, Delta3D offers high performance rendering of environmental features such as the sky, clouds, etc. The developer can merely input a time and weather conditions

(clear, partially cloudy, overcast, etc.) and Delta3D will procedurally generate the clouds and position the sun to match.

4.2 3-D Scene Displaying Subsystem

Above Delta3D's features make it particularly well suited for developing our 3-D scene displaying subsystem.

In fact, we applied 3DS Max and MultiGen Creator to achieve the 3-D modeling, using OSGConv and OSGExp convert OpenFlight and 3DS Studio Max file into Open Scene Graph format so as to be used in Delta3D. The subsystem is developed by using VC++ call Delta3D API as a low level rendering engine. Rendering is on the basis of fire scene development inferring mechanism and implement the dynamic simulation of training process in different phases(Figure 6).

Fig. 6. Crew 3-D and 2-D fire scenarios

Based on the oil tank fire theory and historical cases, 10 combustion types are classified as basic combustion types. The 3-D interactive dynamic picture of oil tank fire is consists of 7 components:

Oil tank storage scene.
The fire form of oil tank combustion.
The smoke form of oil tank combustion.
The effect on fire, smoke and forces disposition by wind direction and wind force.
Fire protection measures with forces disposition.
Character animation.
Audio effects.

5 Conclusion

Visual simulation training system for firefighting can effectively solve some problems affecting training quality, caused by cost, field, time, weather, security and other

factors. It can not only be applied to the demonstrating and researching implementation of fire tactical operation and training in fire protection forces or first responders, but also meet the teaching needs in the fire educational institutions and training bases.

In the process of developing, we realize that 3-D immersive and interactive real time fire scenarios is a important parts, which can arouse trainees' interesting and put them into a active training behavior. Open source game engine, such as Delta3D, makes a great platform for fire departments which wish to build a computer based realistic firefight training application without a big budget.

References

1. McDowell, P., Darken, R., Sullivan, J., Johnson, E.: Delta3D: A Complete Open Source Game and Simulation Engine for Building Military Training Systems. The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology 3(3) (2006)
2. Doris, K., Larkin, M., Silvia, D., McDowell, P.: Applying Gaming Technology to Tomahawk Mission Planning and Training, Report ID 05F-SIW-004. In: SISO 2005 FALL SIW Workshop. Orlando (2005)
3. Ross, F.: Put the AI in training, Fire Chief, United Kingdom (2002)
4. Xinping, Z.: Distributed Virtual Battlefield Environment. system simulation, vol. 13 Suppl. Beijing (2001)
5. Stam, J.: Interacting with smoke and fire in real time. Communication of ACM 43(7), 76–83 (2000)
6. Heyuan, W.: Fire fighting tactical training, the Press of Shanghai science and technology, ShangHai (1999)
7. Bukowski, R., Sequin, C.: Interactive simulation of fire in virtual building environments. In: Proceedings of SIGGRAPH 1997, Los Angeles (1997)
8. Balducelli, C.: Computer Simulation of Emergency Scenarios for Multi-Agents Cooperative Training. In: The Proceedings of the 1994 SCS Simulation Multiconference, San Diego (1994)

A Modified Numerical Integration Method for Deformable Object Animation

Xinrong Hu[1,2], Lan Wei[1], and Dehua Li[2]

[1] Dept. of Computer Science&Technology,WuHan University of Science&Engineering,WuHan,P.R.China,430074
[2] Institute for Pattern Recognition and Artificial Intelligence, State Education Commission Laboratory for Image Processing and Intelligence Control,HuaZhong Univ. of Science&Technology,WuHan 430074
hxr@wuse.edu.cn

Abstract. A satisfied deformable object simulation should be general, accurate, efficient and stable. Explicit, implicit and semi-implicit numerical integration methods have contributed to large performance enhancements in the field of deformable simulation. Cloth is the most representative deformable object. In this paper, we propose an improved embedded Runge-Kutta integration method to solve the deformable simulation that based on classical spring-mass model. Traditional embedded Runge-Kutta methods generally solve ordinary differential equations of deformable object simulation. Most of them tend to concentrate on the efficiency of the simulation process, and not the fidelity of the simulation result. We investigate and determine the extent to which the overall quality must be compromised in order for the stable conditions to be satisfied. The improved Runge-Kutta method proposed in our paper incorporates the fixed time step and adaptive time step in solving cloth motion equations to achieve a controllable error evaluation. Compared with the other Runge-Kutta methods, the proposed method has some advantages for cloth simulation: controllable error evaluation without extra computations, excellent efficiency, good stability and satisfied precision. Experiment demonstrates that this method results in simulation efficiency improvements and is considerable practicable.

Keywords: deformable object, animation, numerical integration, embedded Runge-Kutta, ordinary differential equation, spring-mass model, time step.

1 Introduction

The simulation of deformable object is a hard but challengeable work. It attracts more and more interests of graphics researchers. Cloth is the most representative deformable object and cloth simulation can be used to solve this complex problem. Garment designers, textile engineers and computer graphics researchers are all interested in predicting the motion and static drape of cloth. It is an important integral component of virtual character animation and can be used in future film, games and virtual reality applications. A general, accurate, efficient and stable technique is a challenge for solving 3D deformable simulation. However, the textile and garment industries have only recently acquired the proper physical models and computational

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 375–383, 2007.

strategies to successfully simulate cloth drape and manipulation. In order to improve the fidelity of 3D cloth simulation, maintaining a reasonable computation time and error, a deeper study of the cloth model and the identification of cloth behaviors at different levels are necessary.

The study is not intended to integrate another more precise physical model of cloth behavior, but rather focus on the real-time constraints for the simulation and the visual cloth motion features to which an observer is sensitive. Most of the existing approaches use a general-purpose simulation method using discrete simulation model for the cloth [3-4]. Unfortunately, simulations that simply calculate all potentially colliding vertices may generate a realistic result, but do not provide a guaranteed frame time and error. An effective method should be implemented that avoids heavy computation of the spring-mass model wherever possible.

This problem has already received more and more attentions from mechanical and textile engineers, as well as computer scientists. In fact, spring-mass model [6] is the most frequently considered discrete cloth representations, and tends to prefer discrete ordinary differential equation (ODE) model, typically formulated through Lagrangian or Newtonian dynamics, formally simpler and computationally less expensive.

In this paper, we aim to develop an implicit integration method--embedded Runge-Kutta (RK) method solving the deformable motion equations, which is exemplified by cloth simulation. While simulating cloth, the improved method can reduce the computational costs, adjust the size of time step automatically and control error estimate availably, without losing accuracy or generality.

2 Related Works

2.1 Spring-Mass Model

The simplest and intuitive way of designing a cloth simulation system is to consider the cloth as being discrete into a set of masses that interact with each other through springs [10]. The discrete cloth structure is called spring-mass model. The technique was first introduced in 1988 by Hamann and Parent, and was developed further in 1995 by Provot [7,11]. Since then, spring-mass model is the most common modeling tool for cloth simulation. The spring-mass model considers a triangular mesh where the vertices are masses and the edges are springs with constant rigidity and optional viscosity, as shown in Fig.1 and Fig.2.

This representation simulates cloth by modeling the low level structure of the fabrics and the anisotropic behavior due to different warp and weft properties. This model yields visually plausible deformable cloth surface, which can be solved through adding cross-springs shown in figure 1. Bending springs are connected between two mass points across each edge. Tension springs are connected between two nearest neighbor mass points along horizontal and vertical direction. Shearing springs are connected between two diagonal mass points. The three types of springs have different mathematical and physical formulations to describe the forces applied on the cloth. The combination of the mass points and spring forces is integrated with respect to time to provide a new acceleration for each point. Through the spring-mass model, the motion and static drape of cloth can be computed with several of numeric integration methods.

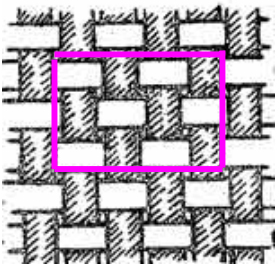


Fig. 1. Close view of cloth

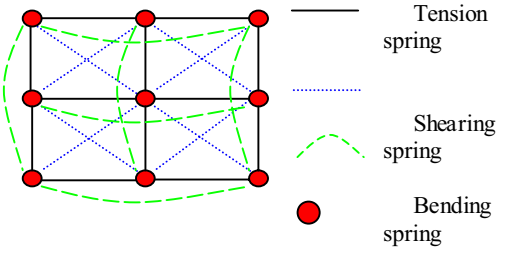


Fig. 2. Corresponding simulation model

2.2 Numeric Integration Methods for Cloth Simulation

Numerical integration is the essence of spring-mass based cloth simulation and is the most important factor in the design of cloth simulation systems [11]. With numerical integration methods, an ordinary differential equation is solved to compute the time-varying state of the masses.

Semi-implicit integration scheme was used to solve the motion differential equation of the particle system [8], which including explicit space domain and implicit time domain. It offers first-order accuracy and a better stability than a pure explicit integration scheme. It does not require the resolution of a large sparse linear system for each of iterations. Several semi-implicit integration methods have been discussed in terms of stability and accuracy [5]. However, large time step may be used, most of the time without any instability issue. When time steps are infinite, it is equivalent to iterations of the Newton resolution method.

Since the pioneering work of Baraff and Wikin[15], implicit integration scheme is currently widely used in applications including cloth simulation, from real time animation in Virtual Reality applications to accurate garment simulation for design and prototyping applications. It has good simulation stability. The computation may be performed more quickly than semi-implicit scheme because of adopting large simulation time steps. Many implicit integration methods are now available in cloth simulation. Detail review is described by Hairer [14].

Simulation errors, which usually break the stability of explicit integration methods, only appear in the implicit scheme as a form of “numerical damping” that does not convergence to stability completely. This may alter the motion of cloth, which results in decreasing the fidelity of simulation result and increasing the size of time step. This may be a problem when expecting an accurate cloth motion and simulation.

In order to alleviate the disadvantage of implicit integration, a various different methods have proposed for cloth simulation, such as Euler method, Midpoint method, higher-order RK method, and so on. However, the accuracy or simulation error of these methods is difficult to control.

Whether explicit integration schemes or implicit ones, they all have a badly limitation that a fixed time steps must be given in advance. If the time step is too large, the cloth simulation will exceed the stability area and cannot achieve stability. Else if the time step is too small, the cloth simulation will increase computational time. An improved method to provide more accuracy, efficiency and stability is expected. The scheme proposed in this paper is based on the embedded RK method.

3 Embedded Runge-Kutta Method

3.1 Motion Differential Equation of Cloth Simulation

As mentioned above, the surface of motion cloth is represented with spring-mass model. That is, the cloth structure is discretied into a mesh of point masses. The dynamics of the system is then simulated in time according to Newton's second law:

$$F_{\text{int}\,ernal} + F_{external} = m\frac{d^2x}{dt^2} + C_{damping}\frac{dx}{dt} \tag{1}$$

Here, $F_{\text{int}\,ernal}$ and $F_{external}$ are the internal forces and external forces applied on the mass points respectively. $x(t)$ is the location of mass point at time t. $C_{damping}$ is the damping coefficient. For cloth simulation, the motion of masses is always described with two order or higher order differential equation of time t. When simulating the dynamic cloth, the initial state is known. Solving equations of the cloth motion is a solving initial conditions of the ordinary differential equations.

In these schemes of solving equations, fourth-order and five-order RK scheme are applied widely in deformation simulations because of their excellent performances. However, a better solution to solving ordinary differential equations of the cloth behavior should be that time steps are adaptive and errors are predictable.

3.2 Fourth-Order Runge-Kutta Scheme

To animate such a spring-mass system, the following fourth-order RK scheme can be used:

$$v_{i+1} = v_i + h(f_1 + 2f_2 + 2f_3 + f_4)/6 \tag{2}$$

$$x_{i+1} = x_i + v_{i+1}dt \tag{3}$$

Here, $v_{i+1} = v_i + h$

$f_1 = f(t_i, v_i)$;

$f_2 = f(t_i + \frac{h}{2}, v_i + \frac{h}{2}f_1)$;

$f_3 = f(t_i + \frac{h}{2}, v_i + \frac{h}{2}f_2)$;

$f_4 = f(t_i + h, v_i + hf_2)$;

Where f is the acceleration of the mass point at time t; v_i is velocity of the mass point at time t_i; h is a fixed time step of the simulation system;

The final overall errors of RK scheme achieves to $O(h^4)$ and partial errors achieves to $O(h^5)$. That is, the higher the orders of RK scheme is, the more accuracy the simulation system is., and the more complicated the computation is. At the same time, since we often have to handle collisions (which give rise to discontinuities in the motion during simulation), this scheme is not appropriate. The scheme is a recurrent process. The computation may not a crucial problem, whereas the difficult problem is the choice of time step h. The simulation system can diverge rapidly since assuming the force as constant over too large a time step may induce a wild change in position.

In practice, we effectively notice a stable behavior of the system only for a small time step. Therefore, RK scheme in cloth simulation is often unsatisfied in practice. We can adjust the time steps to reduce the computation times: small time steps are adopted in some parts of the simulated cloth, while large time steps are adopted in other parts. That is, according to the cloth animation requirements, time steps are adaptive to the simulation and the system is stable. In the following section, we perform a study of RK method relatively to embedding parameters that control the time steps and errors.

3.3 Embedded Runge-Kutta Scheme

An alternative time step adjustment method is based on the embedded RK formulas, originally invented by Fehlberg. An interesting fact about RK formulas is that for orders M higher than four, (M+1) function evaluations are required.

Fehlberg[11,13] proposed a fifth order method with six function evaluations where another combination of the six functions gives a fourth-order method. The difference between the two estimates can then be used as an estimate of the truncation error to adjust the time step size. Since Fehlberg's original formula, several other embedded RK formulas have been found.

The general form of a fifth-order Runge- Kutta formula is:

$$\begin{cases} f_1 = hf(t_i, v_i) \\ f_2 = hf(t_i + a_2 h, v_i + b_{21} h f_1) \\ f_3 = hf(t_i + a_3 h, v_i + b_{31} f_1 + b_{32} f_2) \\ f_4 = hf(t_i + a_4 h, v_i + b_{41} f_1 + b_{42} f_2 + b_{43} f_3) \\ f_5 = hf(t_i + a_5 h, v_i + b_{51} f_1 + b_{52} f_2 + b_{53} f_3 + b_{54} f_4) \\ f_6 = hf(t_i + a_6 h, v_i + b_{61} f_1 + b_{62} f_2 + b_{63} f_3 + b_{64} f_4 + b_{65} f_5) \\ v_{i+1} = v_i + c_1 f_1 + c_2 f_2 + c_3 f_3 + c_4 f_4 + c_5 f_5 + O(h^5) \end{cases} \tag{4}$$

The embedded sixth-order formula is:

$$v'_{i+1} = v_i + c'_1 f_1 + c'_2 f_2 + c'_3 f_3 + c'_4 f_4 + c'_5 f_5 + c'_6 f_6 + O(h^6) \tag{5}$$

From the equation (4) and (5), the error estimate:

$$error = v'_{i+1} - v_{i+1} \tag{6}$$

Here, $error$ is the controllable tolerance of specified error. Compared the error with the allowable accuracy, the system can decide if time step is adjusted automatically.

Runge-Kutta-Cash-Karp[11,13] is an another method of solving ordinary differential equations. This method is essentially the same as the Runge-Kutta-Fehlberg method. The motion equations of solving initial conditions with Cash-Karp method are similar with Fehlberg method. The difference between these two methods is the coefficients. Whereas the former is said to give a more efficient method. In these two methods, we can determine the new time step h_{new} with the help of the accuracy, the old time step h, and the error $|v'_{i+1} - v_{i+1}|$, the expression is:

$$h_{new} = h \left| \frac{accuracy}{error} \right|^{1/5} \tag{7}$$

Cash-Karp method is the most accurate integration technique available on the simulator, having a sixth order error term.

3.4 Problems with Embedded Runge-Kutta Scheme

Embedded RK scheme offers a range of controllable errors, and can achieve a higher accuracy and an adaptive time step. When the motion function is smooth, the embedded RK scheme can speed the computations with large time step. However, the time step in this scheme is adaptive but not controllable efficiently, especially in cloth simulation. We assumed two frame display methods in cloth simulation: one is called dynamic display; the other is called fixed display. The former display a frame as soon as the state of this frame is computed. The latter display a computed frame under the governing of program timer.

Here, we discuss the problems related with dynamic display. As known, the time difference between two frames is not distinct with either method of solving motion equations. It is assumed that the time T used to compute a frame state is 0.1ms(millisecond). Every frame is completed computing and displayed within 0.1ms in order to improve the fidelity of simulation. However, in previous embedded RK methods, time step is hard to control availably. If time step is equal to computation time of frame state, the simulation is satisfied. While time step is far bigger than computation time, for example, time step is 1.0ms and computation time is 0.1ms. Then the problem is raised: the frame achieved after 1.0ms is displayed at time 0.1ms. So the simulation speeds up 10 times (as fast play in films) and results a distortion badly. Similarly, when function is not smooth and time step maybe smaller than computation time, the simulation slows down as slow play in films. We call this behavior as “discordant simulation”.

4 Improved Embedded Runge-Kutta Method

In this paper, we proposed an improved embedded RK method□adaptive computation of fixed time step. Is fixed time step contrary with adaptive computation? The answer is no. It utilizes the human vision resorting effect: human eyes do not detect discontinuousness if 24 frames display in one second. For computing conveniently, it plays 25 frames per second in our method. That is, the fixed time step is assumed 0.04s(second) that is used for timing display. An adaptive time step adjustment algorithm is implemented in the fixed time step.

It is assumed that current frame state, next frame state and fixed time step are known. We can specify an error threshold value. Based on RK scheme, we define the improved embedded RK method as follows:

```
1   procedure improved ERK
2   time of achieving next frame = fixed time step
3   current time = 0
4   testing time step = time step of next frame = fixed time step/4
5   while ( current time < time of achieving next frame)
6      testing time step = time step of next frame
7      Implement the procedure of general ERK method to derive error
```

8 if(error < error threshold value)
9 time step of next frame = Factor0* testing time step*(error/ESP)$^{1/5}$
10 go to line 12
11 testing time step = Factor1* testing time step*(error/ESP)$^{1/5}$
12 current time + = testing time step
13 current frame status = next frame status
14 testing time step = current time - time of achieving next frame
15 Implement the procedure of general ERK method to derive error

In the improved method described as above, fixed time step is adapted to achieve a considerable error estimate form line 5 to 13 of the procedure. Line 15 ensures adaptive time step and simulation precision.

5 Experiments and Conclusion

5.1 Simulation Results of Runge-Kutta Scheme Family

We animate the cloth motion with general fourth order Runge-Kutta (RK4) method, Runge-Kutta-Cash-Karp (RKCK) method, Runge-Kutta-Felhberg(RKF) method and improved embedded Runge-Kutta-Fehlberg (MERK) method respectively. The following simple procedure provides the final part of our experiments.

Fig.3 to Fig.6 show cloth simulation results with these four methods:

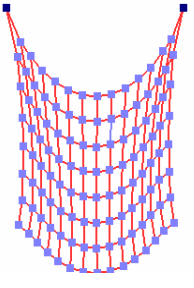


Fig. 3. RK4 method(t=300ms)

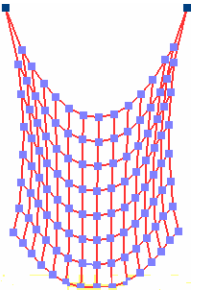


Fig. 4. RKF method(t=300ms)

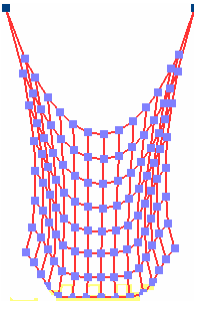


Fig. 5. RKCK method(t=300ms)

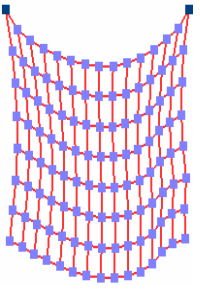


Fig. 6. MERK method(t=300ms)

5.2 Conclusions

The methods introduced in all of the previous sections usually allow us to take sizeable time steps forward, without loss of stability. Even so, there are still times when the time step size must be reduced to avoid divergence. There are a large number of methods for altering the size of a time step, for both explicit and implicit integrators, but these methods tend to concentrate on the efficiency of the simulation, not the fidelity and controllable error estimate. Our goal is animation, thus visually pleasing results, meaning a numerically stable solution is the deciding voice. The improved method that proposed in our paper provides a satisfied efficiency and a good fidelity without any extra computational cost.

As seen from 5.1, our experiments show that a good locality behavior and a scalable error strategy are crucial for an efficient solution of ODEs of cloth simulation using embedded RK methods. With the improved embedded RK method proposed in our paper However, due to the large range of different types and characters of cloth simulation systems, hardware and memory architectures, compilers and optimization techniques, we cannot give a general recommendation on which implementation should be used.

Table 1 shows the data comparisons of these four methods in cloth simulation on HP graphics workstation XW6000.

Table 1. Experiment data comparison

Method	RK4	RKF	RKCK	MERK
cloth points	120	120	120	120
Time step (ms)	0.01(Constant)	Adaptive	Adaptive	Constant +Adaptive
Time to stability(ms)	280	296	300	300
Error	bigger	small	smaller	controllable
Precision	good	better	better	better
Stability	good	better	better	best
Implement	easy	easier	easier	easy

For a simple simulation system that emphasized particularly on the computation time, RK4 maybe a better choice to solve the cloth motion equations. However, it may result the precision decrease and a unsatisfied stability. For a cloth simulation system that emphasized particularly on precision of system, embedded RK methods are the better choice to achieve satisfied simulation results. If simulation fidelity is expected at the same time, an improved embedded RK method must be implemented.

Considering the controllable error and stability, improved embedded RK is currently successful in ODEs solutions of our cloth simulation system. However, the careful selection of the more efficient implementation for the solution of a specific cloth simulation system is usually worth the effort as it can save a large percentage of computation time. Further improvements of the performance of improved embedded RK solvers may be obtained by a specialization in a particular RK method.

Acknowledgments. Funding for this work was provided by the HuBei Educational Council, China (2004D005) and Hubei Digital Textile Equipment Key laboratory (DTL200702), it is greatly appreciated.

References

1. Parks, D., Forsyth, D.A.: Improved Integration for Cloth Simulation. Eurographics short presentations (2002)
2. Volino, P., Magnenat-Thalmann, N.: Accurate Garment Prototyping and imulation, Computer Aided Design and Applications. CAD Solutions 2(5), 645–654 (2005)
3. Choi, K.-J., Ko, H.-S.: Stable but responsive cloth. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques, pp. 604–611. ACM Press, New York (2002)
4. Oshita, M., Makinouchi, A.: Real-time Cloth Simulation with Sparse Particles and Curved Faces. In: Proceedings of Computer Animation 2001, Seoul, Korea (November 2001)
5. Volino, P., Magnenat-Thalmann, N.: Comparing efficiency of integration methods for cloth simulation. In: Computer Graphics International 2001, pp. 265–272 (2001)
6. Eberhardt, B., Weber, A., Straßer, W.: A fast, flexible particle-system model for cloth draping. IEEE Computer Graphics and Applications 16(5), 52–59 (1996)
7. Eischen, J., May-Plumlee, T., Kenkare, N., Pandurangan, P.: Accurate 3d Virtual Drape Simulations: A Developmental Method. In: Proceedings of International Textile and Apparel Association 2003 annual conference (2003)
8. Mathews, J.H., Fink, K.D.: Numerical Methods Using MATLAB, 4th edn. Pearson Education, Inc. (2004)
9. Prieto, M., Santiago, R., Espadas, D., Llorente, I.M., Tirado, F.: Parallel Multigrid for Anisotropic Elliptic Equations. Journal of Parallel and Distributed Computing 61(1), 96–114 (2001)
10. Etzmuss, O., Gross, J., Straßer, W.: Deriving a Particle System from Continuum Mechanics for the Animation of Deformable Objects. IEEE Transaction on Visualization and Computer Graphics 9(4), 538–550 (2003)
11. Hauth, M., Etzmuss, O., Straßer, W.: Analysis of numerical methods for the simulation of deformable models. The Visual Computer (2002)
12. Parks, D., Forsyth, D.A.: Improved integration for cloth simulation. In: EUROGRAPHICS Short Presentation (2002)
13. Foshee, J.W.: Resolution Independent Curved Seam In Clothing Animation Using a regular particle grid, master thesis (2004)
14. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I: Nonstiff Problems. Springer, Berlin (1993)

Performance Monitoring Based on UML Performance Profile

Dong Kwan Kim[1], Chul Jin Kim[2], and Eun Sook Cho[3]

[1] Virginia Tech, Department of Computer Science
Blacksburg, Virginia 24061, USA
ikek70@vt.edu
[2] Architecture Group, Samsung Electronics Co. Ltd.
Gangnam-Gu, Seoul 135-524, Korea
chuljin777.kim@samsung.com
[3] Seoil College, Department of Software
Jungnang-Gu, Seoul 131-702, Korea
escho@seoil.ac.kr

Abstract. In this paper we propose a way of measuring software performance metrics such as response time, throughput, and resource utilization. It is obvious that performance-related Quality of Service (QoS) is one of the important factors which are satisfied for users' needs. The proposed approach uses UML performance profile for the performance specification and aspect-oriented paradigm for the performance measurement. Code instrumentation in AOP is a mechanism to insert source code for performance measurement into business logic code. We used AspectJ, an aspect-oriented extension to the Java. AspectJ code for performance measurement is separated from Java code for functional requirements. Both AspectJ and Java code can be woven together for the performance measurement. The key component of the proposed approach is an AspectJ code generator. It creates AspectJ code for the performance measurement from the UML [1] models containing performance profile.

Keywords: Performance, UML performance profile, Aspect-Oriented, AspectJ.

1 Introduction

Due to the inherent complexity of enterprise applications, it isn't trivial to guarantee a specific Quality of Service (QoS). QoS requirements are timeliness, scalability, high availability [2]. Enterprise business applications are required to satisfy QoS requirements so as to meet users' needs. In this paper the QoS is limited to performance issues such as response time, throughput, and resource utilization. Performance degradation is often encountered during systems operation in the real world. Therefore it is a first step to monitor the performance of the systems. After that, the appropriate remedies can be conducted in order to improve the performance.

Software performance is a cross-cutting concern as well as a non-functional requirement. The cross-cutting concern is not confined only a few modules, but often spread over multiple functional and non-functional elements. With performance often

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 384–393, 2007.

cross-cutting requirements, it makes sense to use Aspect-Oriented Programming (AOP) to represent this aspect of requirements, since that is the purported strength of this technique. AOP is a promising software engineering technique for expressing cross-cutting characteristics of software systems. We define and localize performance as an aspect of the software using AOP. This brings together performance requirements over the software components of the system into an artifact that can be instrumented for performance measurement, yet remain consistent with the relevant models. We believe that this could increase the maintainability and understandability of the system.

We introduce an alternative AOP-based approach for performance evaluation to aid software engineers in evaluating the performance of software systems. A key strength of our approach is the use of a code instrumentation mechanism of AOP. AspectJ code for the performance evaluation is generated from UML models and instrumented between the flows of business logic code. The AspectJ code is then used to check if the performance requirements are satisfied. Based on our preliminary research, we anticipate improved performance verification as well as development cost and time savings for performance evaluation using the proposed approach.

The paper is organized as follows. Section 2 outlines some background information on UML performance profiles, AOP, and AspectJ. In Sections 3 and 4, our proposed approach for the performance monitoring is discussed along with a case study. Section 5 discusses related work. We then conclude the paper with key observations.

2 Background

2.1 UML Performance Profile

The Object Management Group (OMG) has published UML profiles for schedulability, performance, and time specification [3] that defines a set of quantitative performance annotations to be added to UML models. The performance profile facilitates capturing performance requirements, associating performance-related Quality of Service (QoS) characteristics, specifying execution parameters, and presenting performance results computed by modeling tools or found in testing [3] [4]. The performance analysis domain model describes the basic abstractions used in performance analysis, including scenarios, steps, resources and workload. A performance context in the performance analysis domain model specifies one or more scenarios that are used to explore various dynamic situations involving a specific set of resources [3]. Performance represents a dynamic aspect of a system in the specific time. Therefore it is necessary to extract and define some interesting scenarios from an entire execution of the system. Scenarios define a sequence of execution paths and can have QoS requirements such as response times, utilization, or throughputs. They consist of one or more steps. A step may use resources to perform its function and its granularity depends on the level of abstraction. Resources are another basic abstraction and can be realized in active or passive. They participate in one or more scenarios in the specific performance context. A workload is an abstraction of the intensity of demand for the execution of a specific scenario. It can be realized in open or close workload.

2.2 Aspect-Oriented Programming

Aspect-Oriented Programming (AOP) is a programming paradigm that promotes the principle of separation of concerns to enable programmers to focus on aspects [5]. It may be a promising technology that goes beyond object-oriented programming to support the entire software development life cycle from requirements analysis to testing. When developing software systems, we should consider non-functional requirements such as logging, performance, persistence, concurrency, and security. These non-functional requirements are considered as cross-cutting concerns in AOP. An aspect in AOP is a suitable software artifact which represents and localizes the cross-cutting concern. AOP supports many benefits in managing non-functional requirements: the support of high-level abstraction, localization of cross-cutting concern, better modularization, maintainability, and reusability, separation implementation of logic and non-functional requirements, code instrumentation. Especially, code instrumentation is a way to change the behaviors of existing applications without touching their code.

AspectJ [5] [6] is an aspect-oriented extension to the Java programming language that provides convenient constructs for supporting separation of concern concepts like *aspect*, *join point*, *pointcut*, *advice*, and *weaving*. *Join point* is a well-defined point in the program flow. It can be method calls, method executions, object instantiations, constructor executions, field references and handler executions. *Pointcut* picks out certain join points and values at those points [7] [8]. Boolean operations or wildcards allow programmers to flexibly pick key join points. *Advice* is executed when a join point is reached. It extends or modifies the original joint point function by defining additional behaviors. The code fragment can be executed before, after, or around a given pointcut [9]. *Pointcut* and *advice* are useful constructs to dynamically affect execution flow of Java program. The main work of AspectJ implementation is to ensure that aspect and non-aspect code (normally, Java code) run together in a properly coordinated fashion. This coordination process is called aspect weaving and involves making sure that applicable advice runs at the appropriate join points [6].

3 AOP-Based Performance Monitoring

To measure the performance data of a running system is a way in knowing the status of the system. It is an easy way to hardcode the performance measurement code between business logic so as to measure performance metrics such as response time or throughput. However the hard-coding approach has many disadvantages-it could make software systems more complicated because code for performance evaluation is tangled with code for functionalities. It could also result in bad maintainability for software systems. For better modularization, the performance measuring code is needed to be separated from business logic code. In this sense, AOP is a suitable alternative. Code instrumentation mechanism of AOP facilitates capturing runtime data for the performance measurement. Code for performance measurement is separated from logic code and can be reused. The logic code isn't needed to be touched for the performance measurement. As mentioned in previous sections, the proposed approach generates AspectJ code so as to allow users easily to evaluate performance. The

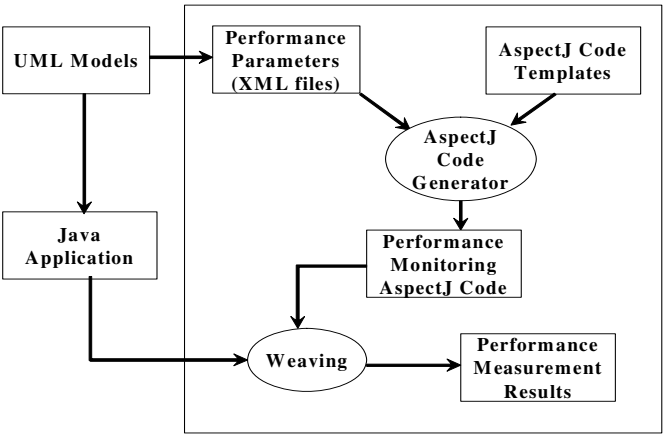


Fig. 1. Flow of AspectJ code generation

AspectJ code generator is not a general-purpose AspectJ code generator. It is used for generating AspectJ code for only performance measurement. The AspectJ code can be generated from UML models annotated with performance profiles.

Fig. 1 depicts detailed procedures for generating AspectJ code. The biggest box in Fig. 1 indicates the AspectJ code generator. Performance requirements can be obtained from class or sequence diagrams. These diagrams should contain performance profiles that are defined on the performance domain analysis model. The performance parameters in Fig. 1 are made from the performance requirements. They in XML format are provided to the AspectJ code generator. Another major component for code generation is code templates. It is common code for concrete performance aspects and is implemented based on performance requirements defined on the performance analysis domain model. Let's suppose a performance aspect for response time. Common parts of the aspect for response time are placed in the code template and specific ones come from performance requirements. AspectJ code generator constructs an executable AspectJ model from code templates and performance requirements. Java applications are made based on UML models. At the weaving step, the generated AspectJ code is inserted at the appropriate spot of Java code and intercepts the execution flow of Java code.

The classes and their relationships are illustrated in the performance analysis domain model in Fig. 2. Each class has its own attributes-e.g. *Workload* class has a *responseTime* attribute.

The attributes are important to measure the performance of a system. They can be a performance metric. Table 1 shows the mapping of the attributes of the classes to aspects. The first column has some of classes in the performance analysis domain model. This work covers some of whole classes that are considered as a performance metric. The second column shows the attributes of the classes in the first column. Some of all attributes are considered and mapped to the aspects. The last column in Table 1 has the performance aspects-e.g. response time aspect, throughput aspect, and

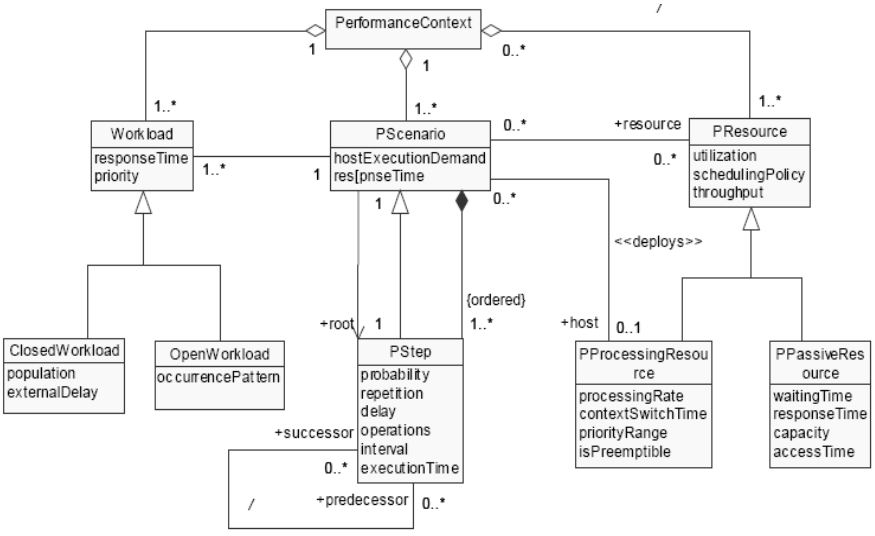


Fig. 2. Performance analysis domain model [3]

utilization aspect. Each aspect has a corresponding attribute. Those aspects are generated by the AspectJ code generator.

Suppose that the *Workload::responseTime* attribute is used in a sequence diagram. The AspectJ code generator creates an aspect that corresponds to the response time attribute. This response time attribute can be realized in the *ResponseTimeAJ* during implementation phase.

Table 1. Domain attribute to aspects mapping

Class name	Attribute name	Aspect
Workload	Workload::responseTime	ResponsetTimeAJ
CloasedWorkload	ClosedWorkload::population	PopulationAJ
	ClosedWorkload::externalDelay	ExtDelayAJ
Step	Step::hostExecutionDemand	ExecutionDemandAJ
	Step::responseTime	ResponsetTimeAJ
	Step::delay	DelayAJ
	Step::interval	IntervalAJ
Resource	Resource::utilization	UtilizationAJ
	Resource::throughput	ThroughputAJ
PassiveResource	PassiveResource::accessTime	AccessTimeAJ
	PassiveResource::responseTime	ResponsetTimeAJ
	PassiveResource::waitTime	WaitTimeAJ

4 Case Study

Duke's Bookstore example [10] is used to demonstrate the proposed approach. The Duke's Bookstore is a simple online bookstore application: its services are browsing a book, putting or removing a book into a shopping cart, payment, and sending an acknowledgement. It is implemented in Java Servlets and JDBC.

You can see a requirement for response time in Fig. 3 illustrates a response time requirement and indicates a probability for the response time that the delay in showing detailed book information will not take longer than half a second in 95%. If the elapsed time for `selectBook` method exceeds 500 milliseconds the response time constraint isn't satisfied. `ResponseTimeAJ` aspect has one pointcut and two pieces of advice. `Before` advice measures a system time before the method is executed. `After` advice measures a system time right after it finishes. The difference between measured times is the execution time of the `selectBook` method. Source code for the `ResponseTimeAJ` can be generated from the performance aspect code templates.

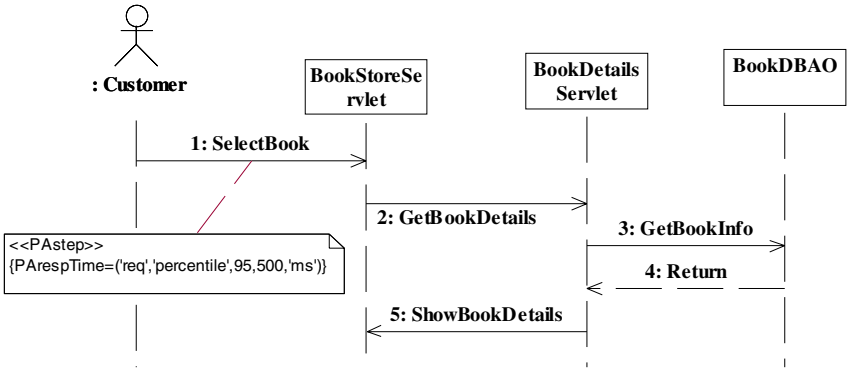


Fig. 3. Sequence diagram for requesting detailed book information

The AspectJ code for the response time aspect is as follows:

```
public aspect ResponseTimeAJ{
  long begin, end, elapsedTime;
  long timelimit=500;
  pointcut selectBookMethod():execution(void BookStore-
Servlet. selectBook ());
  before() : selectBookMethod () {
    begin=System.currentTimeMillis();
  }
  after() returning : selectBookMethod () {
```

```
      end=System.currentTimeMillis();
      elapsedTime=end-begin;
      if(elapsedTime > timelimit)
         //Do something.
    }
  }
```

The Duke's Bookstore application stores book data into a database. The utilization of the database means how frequently the database has been used during the unit time. We need to know the throughput of the database to get the utilization. The service time and the number of completed jobs should be measured for the utilization. .The service time of database can be obtained by measuring the connection and disconnection time of database. To get a connection time of DB, we have `dbConnection` pointcut. It measures the time right after users connect to DB. This time indicates the database begins the service. Another pointcut is `dbDisconnection`. It measures the time right before users disconnect to the database. Next step is to figure out how many query transactions have been done during the service time. `statementExec` pointcut is used for that. Whenever a SQL query executes, the number of jobs increases by one. When we get the service time and the number of completed query transactions, we can get the throughput of the database. And then, we can calculate the utilization of the database. In this case study, we assume the initial value of the service rate is 500.

The AspectJ code for the database utilization aspect is as follows:

```
public aspect UtilizationAJ{
  long begin, end, elapsedTime, beginCon, endCon, con-
nectionTime; long throughput=0; long job=0; long
unit=60; float utilization=0; long servicerate=500;
  pointcut dbConnection():
    call(Connection+ DataSource.getConnection(..)) ||
    call(Connection+ Driver.connect(..)) ||
    call(Connection+ DriverManager.getConnection(..));
  after() returning : dbConnection() {
    beginCon=System.currentTimeMillis();}
  pointcut dbDisconnection():call(void Connec-
tion.close());
  before() : dbDisconnection() {
    endCon=System.currentTimeMillis();
    connectionTime=endCon-beginCon;
    connectionTime/=1000;
    throughput = job*unit/connectionTime;
    utilization=(new Long(throughput).floatValue()) /
(new Long(servicerate).floatValue())* 100;
  }
  pointcut statementExec():call(*
java.sql..*.execute*(..));
  after() returning : statementExec() { job++; }
}
```

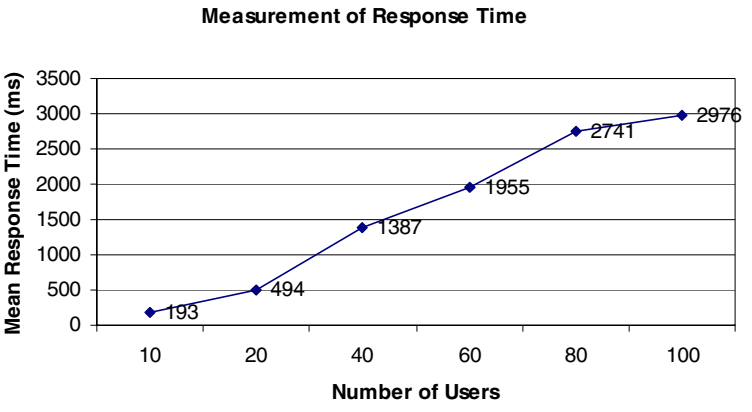


Fig. 4. Measurement of response time using the response time aspect

In this case study, we used Apache JMeter [11] as a workload generator and Tomcat [12] as a Servlet container. MySQL [13] was used as a database management system. The response times of the scenario for displaying detailed information are shown in Fig. 4. The sequence diagram for the scenario is illustrated in Fig. 3. The response time aspect which was generated by the AspectJ code generator captured those response times. The average response time indicates how long it takes to display the detailed information of a selected book. The number of users in Fig. 4 means how many people can concurrently visit the Duke's Bookstore web site. We measured the response time of 6 cases: 10, 20, 40, 60, 80, and 100. As the number of users increases, the average response time goes up. The response time will take longer as the concurrent users grow.

5 Related Work

Considerable research has been conducted regarding performance model such as queuing network, extended queuing network, layered queuing network, stochastic process algebra, and generalized stochastic petri net. Along with these approaches, research based on the UML performance analysis domain model has emerged where a performance model can be generated from UML models using performance profiles. In [14], Arief and Speirs introduced a tool that allows system developers to design a new system using the UML class and sequence diagram notations. The tool automatically generates simulation programs from those diagrams along with some random and statistics information. Park and Kang [15] explained how AspectJ code can be constructed from UML-based performance models and used for simulating performance prior to coding the software. Ho et al. [16] proposed a UML All pUrpose Transformer (UMLAUT) framework as methodological support for building and manipulating UML models with aspects. Copper et al. [17] presented a UML based approach to model and analysis performance aspect. They used Rapide ADL and Armani and modeled response time performance aspect in a Domain Name Service example.

Bodkin [18] addressed how to use AspectJ in order to monitor performance of applications. Practically useful examples for aspect-oriented performance monitoring system were mentioned in the article-Servlet monitoring, database request monitoring, Servlet request tracking, and Java Database Connectivity (JDBC) monitoring. Pearce et al. [19] introduced an experimental aspect-oriented Java profiler, called DJProf that employs AspectJ to insert the necessary instrumentation for profiling and can be used to profile Java programs without modification and does not require users to have any knowledge of AspectJ. DJProf also supports several different modes of profiling e.g. heap usage, object lifetime, wasted time, and time spent. Ng et al. [20] showed how to facilitate the process of static or dynamic program comprehension using code instrumentation of AOP. Friend, Eclipse, and Compress were analyzed to demonstrate the benefits of AOP.

6 Conclusions

In this research we proposed an approach for a performance monitoring using UML performance profile and aspect-oriented programming. The goal of the work is to measure performance metrics that are specified in UML models. To reduce the gaps between design and implementation, the code generator helps developers make AspectJ code for performance monitoring. In UML, performance specification is based on the performance analysis domain model which defines classes and their attributes for the performance metrics such as response time, throughput, and resource utilization. For the code generation, it is a first step to define the relationship between the performance analysis domain model and performance aspects. The code generator creates AspectJ code based on the relationship information. Performance metrics as cross-cutting concerns can be separated from the primary models and implemented as aspects. The current code generator supports some of classes in the performance analysis domain model. We plan to implement the performance AspectJ code templates which are intended to fully support the performance analysis domain model. We anticipate considerable effort in providing a seamless performance monitoring framework from design to implementation phase. Another related interesting topic is performance prediction. A mathematical or formal performance model is necessary for the performance prediction. The proposed approach can be extended as adding modules for the performance prediction.

References

1. Unified Modeling Language (UML) Tutorial (2007), http://pigseye.kennesaw.edu/dbraun/csis4650/A&D/UML_tutorial/index.htm
2. Lodi, G., Panzieri, F., Rossi, D., Turrini, E.: SLA-Driven Clustering of QoS-Aware Application Servers. IEEE Transactions on Software Engineering 33(3), 186–197 (2007)
3. Object Management Group (OMG): UML profile for Schedulability, Performance, and Time Specification (2005)
4. Bennett, A., Field, A.: Performance Engineering with the UML Profile for Schedulability, Performance and Time:a Case Study. In: 12th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems (2004)

5. Kiczales, G., Lamping, J., Mendhekar, A., Maeda, C., Lopes, C., Loingtier, J., Irwin, J.: Aspect-Oriented Programming. In: Proc. of the 11th European Conf. on Object-Oriented Programming, pp. 220–242 (1997)
6. Kiczales, G., Hilsdale, E., Hugunin, J., Kersten, M., Palm, J., Griswold, W.G.: An Overview of AspectJ. In: Proc. of the 15th European Conf. on Object-Oriented Programming, pp. 327–353 (2001)
7. AspectJ (2007), http://www.eclipse.org/aspectj
8. Miles, R.: AspectJ Cookbook, O'REILLY (2005)
9. Kandé, M.M., Kienzle, J., Strohmeier, A.: From AOP to UML-A Bottom-Up Approach. In: AOSD'2002 Workshop on Aspect-Oriented Modeling with UML (2002)
10. Armstrong, E., Ball, J., Bodoff, S., Carson, D.B., Evans, I., Green, D., Haase, K., Jendrock, E.: The J2EE 1.4 tutorial (2005)
11. Apache JMeter (2007), http://jakarta.apache.org/jmeter/
12. Apache Tomcat (2007), http://tomcat.apache.org/
13. MySQL (2007), http://www.mysql.com/
14. Arief, L.B., Speirs, N.A.: A UML Tool for an Automatic Generation of Simulation Programs. In: Proc. 2nd Int'l Workshop on Software and Performance, pp. 71–76 (2000)
15. Park, D., Kang, S.: Design Phase Analysis of Software Performance Using Aspect-Oriented Programming. In: Proc. 5th Aspect-oriented Modeling Workshop in conjunction with UML 2004 (2004)
16. Ho, W., Pennaneac'h, F., Plouzeau, N.: UMLAUT: A Framework for Weaving UML-based Aspect-Oriented Designs. In: Proc. of the Technology of Object-Oriented Languages and Systems (TOOLS 33), pp. 324–334. IEEE Computer Society, Los Alamitos (2000)
17. Cooper, K., Dai, L., Deng, Y.: Modeling Performance as an Aspect: a UML Based Approach. In: The 4th AOSD Modeling With UML Workshop (2003)
18. Bodkin, R.: AOP@Work: Performance monitoring with AspectJ (2005), http://www-128.ibm.com/developerworks/java/library/j-aopwork10/
19. Pearce, D.J., Webster, M., Berry, R., Kelly, P.H.J.: Profiling with AspectJ, Software: Practice and Experience (2006)
20. Ng, D., Kaeli, D., Kojarski, S., Lorenz, D.H.: Program Comprehension Using Aspects. In: ICSE 2004 Workshop on Directions in Software Engineering Environments (WoDiSEE 2004) (2004)

Analysis of Variable Slit Structure on Pneumatic System

Chongho Youn[1], Heuangkhamsene Chanthalang[2], Kenji Kawashima[1], and Toshiharu Kagawa[1]

[1] Precision and Intelligence Laboratory, Tokyo Institute of Technology, R2-45 4259 Nagatuta, Midori-Ku, Yokohama, 226-8503, Japan
{youn,kkawashi,kagawa}@pi.titech.ac.jp
[2] Department of Mechano-Micro Engineering, Tokyo Institute of Technology, R2-45 4259 Nagatuta, Midori-Ku, Yokohama, 226-8503, Japan
heuangkhamsene.c.aa@m.titech.ac.jp

Abstract. Pressure reducing valves are widely used to maintain the pressure of gas reservoirs to specific values. In a normal valve, supply pressure is depressurized with an orifice plate. When pressurized air passes through the orifice plate, considerable noise occurs at the downstream side. To solve this problem, we have developed the slit structure that can maintain a laminar flow and can reduce the noise. At the former slit structure, the height of the slit was fixed. In this study, we proposed an improved slit structure that the height of slit is variable. Coned disk springs are installed between the slit structures and the height of them is precisely controlled. The flow characteristics of the slit are investigated theoretically and experimentally. The experimental results indicated that the noise level decreased by approximately 40dB.

Keywords: Theoretical analysis, Slit structure, Laminar flow, Noise.

1 Introduction

Pneumatic systems are widely used in industrial fields from the viewpoint of low cost and safety. Breather valves are used to maintain the pressure of gas reservoirs to specific values as shown Fig. 1. In a normal valve, supply pressure is depressurized with an orifice plate. When pressurized air passes through the orifice plate, a considerable noise and pressure fluctuation occur at the downstream. Therefore, reduction of noise and pressure fluctuation originating from the valve are required.

These phenomena have been investigated both analytically and experimentally. In addition, several methods, such as the use of diffusers [1], wrapping pipe with sound-damping materials [2], changing the plug structure [3], and the use of silencers [4] have been developed to reduce noise and pressure fluctuation. However, with these methods, when the flow rate increases, the flow might become turbulent. In some cases, sonic flow occurs even when the pressure ratio is lower than 0.528. Turbulent and sonic flow can generate considerable noise and shock waves.

J.-W. Park, T.-G. Kim, and Y.-B. Kim (Eds.): AsiaSim 2007, CCIS 5, pp. 394–402, 2007.

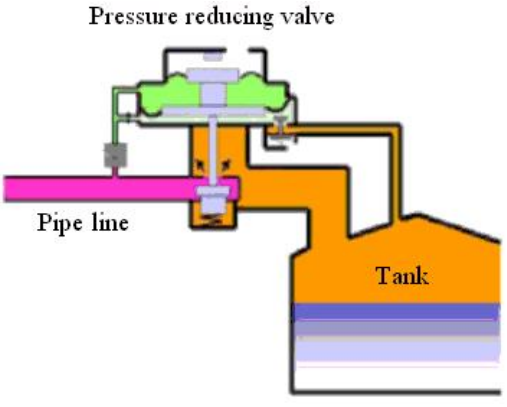


Fig. 1. Pressure reducing valve which connect between pipe line and gas tank

Therefore, the authors have proposed the slit structure that can reduce noise and pressure fluctuation. This reduces the noise by changing orifice into the slit structure. The flow of the slit structure reduces the noise by suppressing the generation of turbulence and shock wave. In former research, it was found that the slit structure of height 0.05mm has the silencing effect more than about 40dB in comparison with the orifice [5], and the height of the slit was fixed.

In this paper, we proposed an improved slit structure that the height of it is variable by using coned disk springs. The structure of variable slit structure is contained in Sec. 2, and the flow characteristics of the slit are investigated theoretically and experimentally in Sec. 3, 4. Finally, the effect of the noise reduction on the slit is investigated experimentally.

2 Variable Slit Structure

Figure 2 shows a schematic drawing of the variable slit structure. The slit structure consists of three elements, a disk, a guiding bar and coned disk springs. The upper part of Fig. 2 shows the cross section of the structure. The upper disk consists of a flow inlet and an upper surface of the radial slit. The inner diameter of upper disk is 8.5mm. The guiding bar guides the movement of the disk. Coned disk springs are installed between the disk structures and the height of them are precisely controlled.

The compressed air enters from the center of the upper disk and is exhausted outward through the radial slit. The lower part of Fig. 2 shows the top view of the lower disk. The lower disk consists of a lower surface of the radial slit and a groove for holding the disc springs. The outer diameter of upper disk is 50mm, and the depth of groove is 0.5mm.

The structure of the coned disk spring is shown in Fig. 3. The material of it is SUS304, and the spring constant is 619.5N/mm. The diameter of the coned disk spring is 8mm. The thickness is 0.3mm, and height is 0.55mm. It is located on groove of disk. The height difference of groove and coned disk spring makes the 50μm slit. The height of it is precisely controlled with a control force.

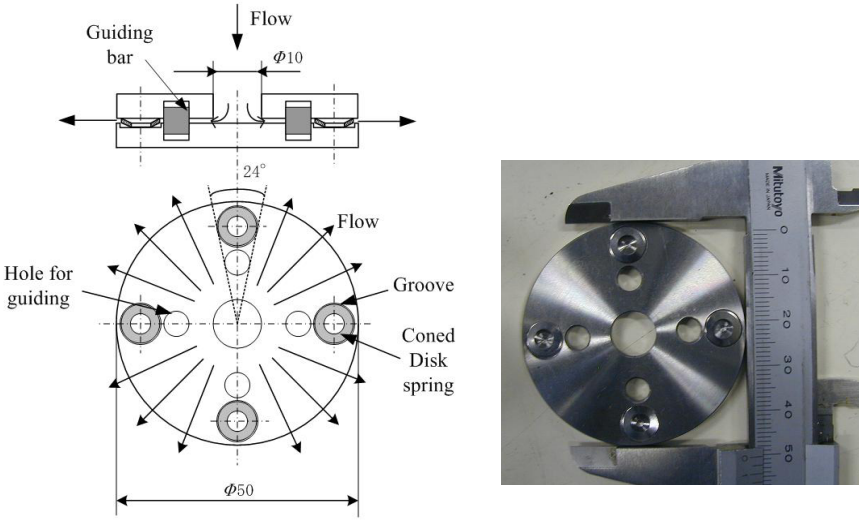


Fig. 2. Schematic of variable slit structure

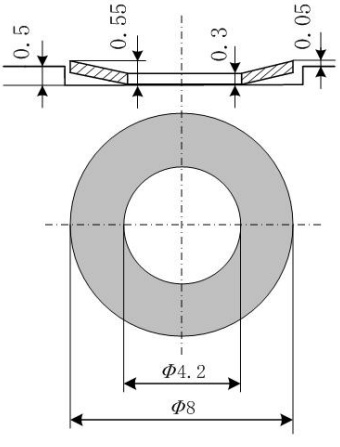


Fig. 3. Structure of coned disk spring

The picture and the structure of new valve are shown in Fig. 4. In this study, we produced a new valve which has one layer of slit as shown in the right side of Fig. 4. The compressed air with supply pressure P_s enters from the center of the disks and is exhausted to atmosphere pressure P_a through the slits. The pneumatic cylinder is set up below the disk, and the piston of it is connected with the disk. The control pressure P_c is supplied to the cylinder, and the cylinder moves the disk. Therefore, the height of slits is controlled with the pneumatic cylinder by changing the control pressure. The height of slit is measured by a microscope.

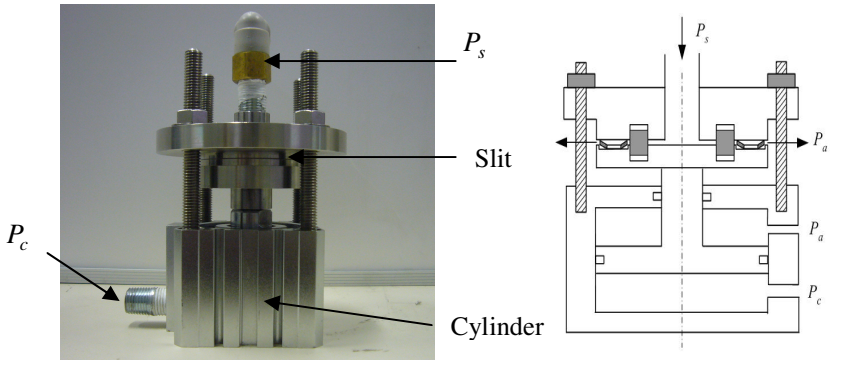


Fig. 4. Picture and Structure of new valve

3 Analysis of the Slit Structure

The flow rate characteristics of the slit structure are investigated theoretically under the assumptions that the flow is laminar and passes through the slit under isothermal conditions.

The relationship between the pressure drop dP and the average flow velocity $\bar{u}_r$ is given by the next equation

$$dP_r = -\frac{12\mu\bar{u}_r}{h^2}dr \tag{1}$$

which comes from the Navier-Stokes equation of steady state laminar flow between parallel plates . In Eq. (1), μ stands for the viscosity of the air and h is the height of the parallel slit. The following equation is obtained from the continuity equation:

$$(\bar{u}A\rho)_{r_2} = (\bar{u}A\rho)_r \quad r_1 \le r \le r_2 \tag{2}$$

where $\bar{u}$, A, and ρ are the velocity, the cross-sectional area, and density with respect to radial r_2 and r accordingly. Using the state equation of gases

$$P = \rho R\theta \tag{3}$$

where R and θ are the gas constant and average temperature of air, respectively. The following equation is obtained From Eq. (2) and Eq. (3).

$$\bar{u}_r = \frac{A_{r_2}\rho_{r_2}}{A_r\rho_r}u_{r_2} = \frac{A_{r_2}P_{r_2}}{A_rP_r}u_{r_2} = \frac{r_2P_{r_2}}{rP_r}u_{r_2} \tag{4}$$

The effective area of the slit A_r at radius r is given by the following equation:

$$A_r = 2\pi rh\frac{264^\circ}{360^\circ}. \tag{5}$$

It is assumed that the degree of cross-sectional area is 264°as shown in fig. 2. Substituting Eq. (4) into Eq. (1) yields the following equation:

$$dP_r = -\frac{12\mu r_2 P_{r_2} \bar{u}_{r_2}}{h^2 r P_r} dr \tag{6}$$

When Eq. (6) is integrated, and the boundary condition of $P_{r_2} = p_a$ is substituted, the following equation is obtained.

$$P_r = \sqrt{\frac{24\mu \bar{u}_2 r_2 P_a}{h^2} \ln \frac{r_2}{r} + P_a{}^2} \tag{7}$$

Substituting $\bar{u}_2 = Q / A_{r_2}$ into Eq. (7), the following equation is obtained.

$$P_r = \sqrt{\frac{24\mu Q r_2 P_a}{h^2 A_{r_2}} \ln \frac{r_2}{r} + P_a{}^2} \tag{8}$$

If the flow rate Q is given, the pressure P_r at $r = r_1$ is obtained from Eq. (8).

The extra pressure drop ΔP_i must be considered in the inlet region [6]. This is evaluated using the following equation:

$$\Delta P_i = \xi \frac{\rho \bar{u}_{r_1}^2}{2} \tag{9}$$

where ξ is the friction coefficient.

If the flow rate Q is given, the inlet average velocity $\bar{u}_{r_1}$ is obtained from the equation $\rho_a Q = (\rho \bar{u} A)_{r_1}$. As a result, the supply pressure P_s is given by adding the losses in the slit P_{r_1} and the losses in the inlet region ΔP_i.

$$P_s = P_{r_1} + \Delta P_i \tag{10}$$

4 Characteristics of the Slit Structure

4.1 Flow Characteristics

The flow characteristics were measured experimentally using the experimental apparatus shown in Fig. 5. Buffer tank and two regulators are set up on the upstream side. Supply and control pressures were regulated by each regulator. The pressure and the flow rate were measured using a bourdon tube pressure gauge and a float-type area flow meter, respectively.

The measured flow characteristic is shown in Fig. 6. The vertical axis is flow rate and the horizontal axis is a radius. The triangular symbols show the experimental result, and the solid line is the theoretical result. The theoretical result is calculated by Eq. 10 using the parameters listed in Table 1. The friction coefficient was chosen to approximately match the experimental results. The supply pressure is set constant at 500kPa. The control pressure was varied from atmospheric pressure to 580kPa, and the height of slit is varied from 50μm to 0μm. It is thought that from the fact that theoretical analysis and experimental result agree well the flow by the viscosity is dominant.

The pressure distribution along r direction is examined with Eq. 8 and Eq. 9. The result is shown in Fig. 7. The vertical axis is pressure distribution and the horizontal axis is a radius. The pressure loss is almost linear line when the height of slit is 10μm. However, the pressure loss in the inlet region is getting large when the height of slit increases. It means the inlet length effect is greater when the height of slit increases.

Table 1. Parameters of the slit structure

r_1 (m)	r_2 (m)	h (μ m)	ξ
0.05	0.025	0 ~ 50	0.45

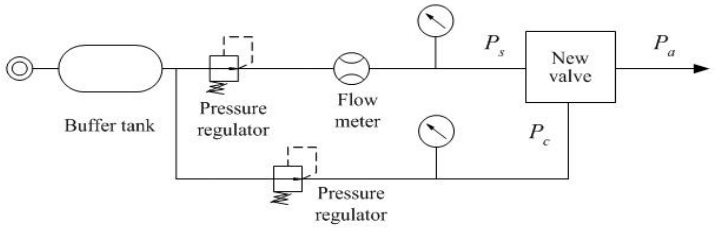


Fig. 5. Experimental apparatus to measure flow characteristics

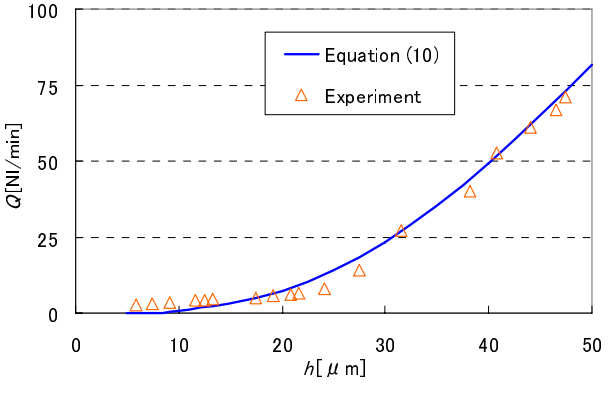


Fig. 6. Flow characteristics of the slit structure

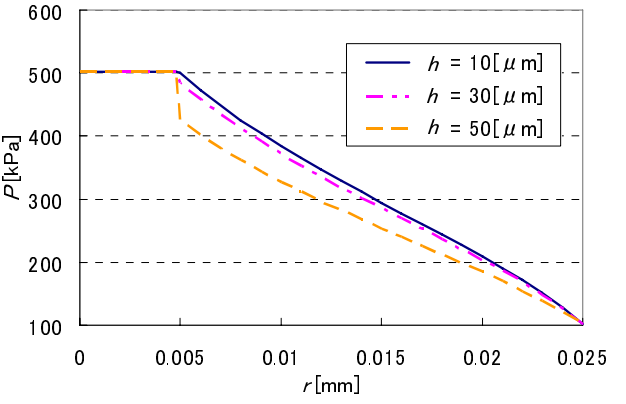


Fig. 7. Pressure distribution along *r* direction

4.2 Reynolds Number

The Reynolds number is examined from the result by the theoretical analysis. The Reynolds number is calculated by Eq. (11).

$$\mathrm{Re} = \frac{\rho_r \bar{u}_r D_h}{\mu} = \frac{\rho_a Q D_h}{\mu A_r} \tag{11}$$

Here,

$$D_h = 4\frac{2\pi rh}{2(2\pi r + h)} \cong 2h \tag{12}$$

Reynolds number along r direction is shown in Fig. 8. The vertical axis is Reynolds number and the horizontal axis is a radius. When the height of the slit rises, the Reynolds number rises because flow rate increases. Because the Reynolds number is a

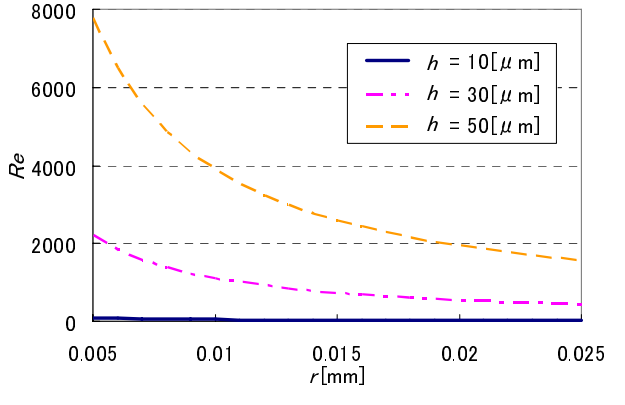


Fig. 8. Reynolds number along *r* direction

function of the cross-sectional area, the flow on the radial shape has the effect of reducing the Reynolds number. Because the area ratio of the inlet for the outlet is 0.2, the Reynolds number of the outlet becomes 20% of the inlet. The Reynolds number at the outlet is 2000 or less as a result.

5 Noise Level

The noise level of the valve was measured using a microphone sensor (Custom corp. ; SL-1370), according to the Japanese Industrial Standards (JIS) [7]. The background noise level of the room is at 30dB. Microphone was placed in the room at an angle of 45 degree from the center axis of the valve as shown in Fig.9. The distance from the valve to the microphone was 1.0m.

The noise levels of the orifice and the newly developed valve with the variable slit structure were compared. The experimental results are shown in Fig.10. The horizontal axis is flow rate, and the vertical axis is the noise level. We confirmed in advance that the flow characteristics of the orifice and the new valve were approximately the same. The results indicated that the noise level decreased approximately 40dB in the new valve.

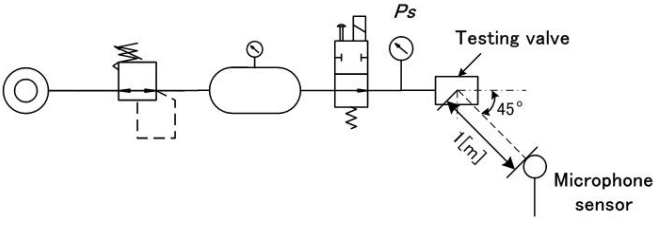


Fig. 9. Pneumatic circuit and noise measuring system

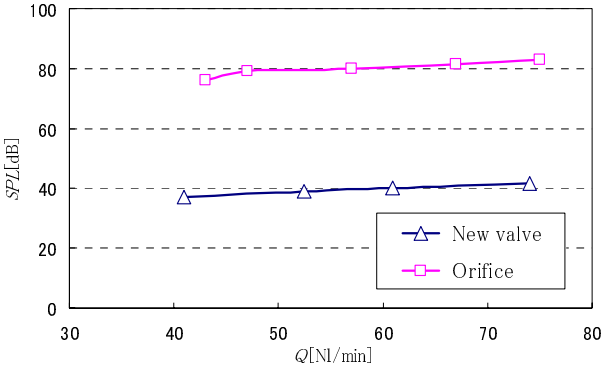


Fig. 10. Noise level of orifice and new valve

6 Conclusion

This paper proposed variable slit structures instead of orifice structures. The height of slit is variable by using coned disk springs. The coned disk springs, which diameter is 8mm, are installed between the slit structures and the height of them is precisely controlled with control pressure.

The flow characteristics of the slit are investigated theoretically and experimentally. The pressure distribution and Reynolds number is examined. The experimental results indicated that the noise level decreased by approximately 40dB.

References

1. Boger, H.W.: Designing Valves and Downstream Devices as Low Noise Packages, Heat./Piping/Air Cond (1971)
2. Bell, L.H.: Industrial Noise Control. Mech. Eng (Am. Soc. Mech. Eng.) 88, 417–426 (1993)
3. Amini, A., Owen, I.: A Practical Solution to the Problem of Noise and Fluctuation in a Pressure-Reducing Valve. Exp. Therm. Fluid Sci. 10, 136–141 (1995)
4. Davies, P.O.L.A., Harrison, M.F., Collins, H.J.: Acoustic Modeling of Multiple Path with Experimental Validations. J. Sound Vib. 200(2), 195–225 (1997)
5. Youn, C., Kawashima, K., Kagawa, T.: Fundamental Analysis of Super Low Noise Control Restriction for Compressible Fluid. In: The 18th I.C.H.P., pp. 387–394 (2003)
6. Stone, C.R., Wright, S.D.: Nonlinear and Unsteady Flow Analysis of Flow in a Viscous Flowmeter. Trans. Inst. Meas. Control(London) 16(3), 128–141 (1994)
7. JIS B8379, The pneumatic silencer (In Japanese) pp. 653–657 (1995)

Author Index