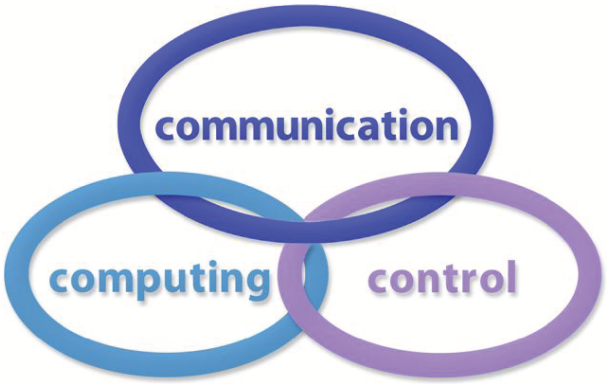
INTERNATIONAL JOURNAL

of

COMPUTERS, COMMUNICATIONS & CONTROL

With Emphasis on the Integration of Three Technologies

IJCCC



Year: 2012 Volume: 7 Number: 4 (November)

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).

Agora University Editing House

CCC Publications

www.journal.univagora.ro

International Journal of Computers, Communications & Control

EDITORIAL ADDRESS:
R&D Agora Ltd. / S.C. Cercetare Dezvoltare Agora S.R.L.
Piaţa Tineretului 8, Oradea, jud. Bihor, Romania, Zip Code 410526
Tel./ Fax: +40 359101032
E-mail: ijccc@univagora.ro, rd.agora@univagora.ro, ccc.journal@gmail.com
Journal website: www.journal.univagora.ro

International Journal of Computers, Communications & Control

Mario de J. Pérez Jiménez
Dept. of CS and Artificial Intelligence
University of Seville, Sevilla,
Avda. Reina Mercedes s/n, 41012, Spain
marper@us.es

Dana Petcu
Computer Science Department
Western University of Timisoara
V.Parvan 4, 300223 Timisoara, Romania
petcu@info.uvt.ro

Radu Popescu-Zeletin
Fraunhofer Institute for Open
Communication Systems
Technical University Berlin, Germany
rpz@cs.tu-berlin.de

Imre J. Rudas
Institute of Intelligent Engineering Systems
Budapest Tech
Budapest, Bécsi út 96/B, H-1034, Hungary
rudas@bmf.hu

Yong Shi
Research Center on Fictitious Economy
& Data Science
Chinese Academy of Sciences
Beijing 100190, China
yshi@gucas.ac.cn
and
College of Information Science & Technology
University of Nebraska at Omaha
Omaha, NE 68182, USA
yshi@unomaha.edu

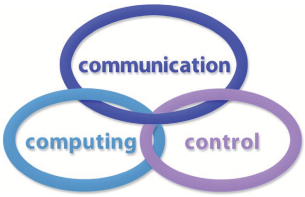
Athanasios D. Styliadis
Alexander Institute of Technology
Agiou Panteleimona 24, 551 33
Thessaloniki, Greece
styl@it.teithe.gr

Gheorghe Tecuci
Learning Agents Center
George Mason University, USA
University Drive 4440, Fairfax VA 22030-4444
tecuci@gmu.edu

Horia-Nicolai Teodorescu
Faculty of Electronics and Telecommunications
Technical University "Gh. Asachi" Iasi
Iasi, Bd. Carol I 11, 700506, Romania
hteodor@etc.tuiasi.ro

Dan Tufiş
Research Institute for Artificial Intelligence
of the Romanian Academy
Bucharest, "13 Septembrie" 13, 050711, Romania
tufis@racai.ro

Lotfi A. Zadeh
Professor,
Graduate School,
Director,
Berkeley Initiative in Soft Computing (BISC)
Computer Science Division
Department of Electrical Engineering
& Computer Sciences
University of California Berkeley,
Berkeley, CA 94720-1776, USA
zadeh@eecs.berkeley.edu

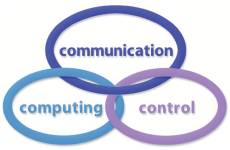


DATA FOR SUBSCRIBERS
Supplier: Cercetare Dezvoltare Agora Srl (Research & Development Agora Ltd.)
Fiscal code: 24747462
Headquarter: Oradea, Piata Tineretului Nr.8, Bihor, Romania, Zip code 410526
Bank: MILLENNIUM BANK, Bank address: Piata Unirii, str. Primariei, 2, Oradea, Romania
IBAN Account for EURO: RO73MILB0000000000932235
SWIFT CODE (eq.BIC): MILBROBU

International Journal of Computers, Communications & Control

Short Description of IJCCC

Title of journal: International Journal of Computers, Communications & Control
Acronym: IJCCC
Abbreviated Journal Title: INT J COMPUT COMMUN
International Standard Serial Number: ISSN 1841-9836, E-ISSN 1841-9844
Publisher: CCC Publications - Agora University
Starting year of IJCCC: 2006
Founders of IJCCC: Ioan Dzitac, Florin Gheorghe Filip and Mişu-Jan Manolescu
Logo:



Number of issues/year: IJCCC has 4 issues/odd year (March, June, September, December) and 5 issues/even year (March, June, September, November, December). Every even year IJCCC will publish a supplementary issue with selected papers from the International Conference on Computers, Communications and Control.
Coverage:

- Beginning with Vol. 1 (2006), Supplementary issue: S, IJCCC is covered by Thomson Reuters - SCI Expanded and is indexed in ISI Web of Science.
- Journal Citation Reports(JCR)/Science Edition:
  - Impact factor (IF): JCR2009, IF = 0.373; JCR2010, IF = 0.650; JCR2011, IF = 0.438.
- Beginning with Vol. 2 (2007), No.1, IJCCC is covered in EBSCO.
- Beginning with Vol. 3 (2008), No.1, IJCCC, is covered in Scopus.

Scope: IJCCC is directed to the international communities of scientific researchers in universities, research units and industry. IJCCC publishes original and recent scientific contributions in the following fields: Computing & Computational Mathematics; Information Technology & Communications; Computer-based Control.
Unique features distinguishing IJCCC: To differentiate from other similar journals, the editorial policy of IJCCC encourages especially the publishing of scientific papers that focus on the convergence of the 3 "C" (Computing, Communication, Control).
Policy: The articles submitted to IJCCC must be original and previously unpublished in other journals. The submissions will be revised independently by at least two reviewers and will be published only after completion of the editorial workflow.

International Journal of Computers, Communications & Control,
ISSN 1841-9836, Vol. VII (2012), No. 4 (November).
Abbreviated Journal Title in ISI Web of Science: INT J COMPUT COMMUN.
This journal is a member of, and subscribes to the principles of, the Committee on Publication Ethics (COPE): http://publicationethics.org.

Contents

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 594-605

Control Schemes for a Quadruple Tank Process

A. Abdullah, M. Zribi

Ali Abdullah, Mohamed Zribi
Electrical Engineering Department
Kuwait University
P. O. Box 5969, Safat-13060, Kuwait
E-mail: ali.abdullah@ku.edu.kw
E-mail: mohamed.zribi@ku.edu.kw

Abstract: This paper deals with the control of a quadruple tank process. A gain scheduling controller, a linear parameter varying controller and an input-output feedback linearization controller are proposed for the quadruple tank process. The derivation of the three control schemes is presented in details. Moreover, the proposed control schemes are implemented using an experimental setup. The experimental results indicate that the developed control schemes work well and are able to regulate the output of the process to its desired value. Additionally, the implementation results demonstrate that the input-output feedback linearization controller gave the best performance.
Keywords: quadruple tank process, gain scheduling control, linear parameter varying control, input-output feedback linearization control.

1 Introduction

The quadruple tank process is a highly nonlinear system which has been used to test different multivariable control schemes. Several controllers were designed for this process. For example, a decentralized proportional integral (PI) controller [1, 2], a decentralized PI controller with sliding mode features [3], a decoupled proportional integral and derivative (PID) controller [4], an internal model controller [5], a model predictive controller [6, 7], a quantitative feedback controller [8] and an H_∞ controller [9] were proposed for the control of the quadruple tank process. These control schemes were designed using the linearized model of the quadruple tank process around different operating points. Therefore, these controllers can not guarantee good performances of the controlled system over the whole operating range of the quadruple tank process because of the inherent nonlinearities of the quadruple tank process.
In order to achieve good performances over the whole operating range of the quadruple tank process, other control techniques were reported in the literature. Nonlinear model predictive controllers where designed in [10, 11] for the process. In [12], a sliding mode controller was designed and implemented on the process. However, it should be noted that a singularity is encountered when using this controller. The singularity occurs when one of the four tanks is empty. Hence, the proposed controller can not be implemented in such case. In [13], a linear decentralized PI controller was designed based on the approximated nonlinear model of the process over a selected range of operation of the process. The simulation results show a good tracking behavior over the selected operating range. On the other hand, the work in [14] dealt with the design of a gain scheduling PI controller for the process; the gain scheduling design is done according to the operating input voltages.
In this paper, three well-known controllers consisting of a gain scheduling control [15-17], a linear parameter varying control [18-21] and an input-output feedback linearization control [22, 23] are designed and implemented to control the water levels in the quadruple tank process over the whole range of operation of the process. Moreover, an integral action is included in the three control schemes in order to achieve good tracking performances [22]. The implementation results

are presented to show the effectiveness of the proposed control schemes.
The paper is organized as follows. The dynamic model of the process and the control objective of the paper are presented in sections 2 and 3 respectively. A gain scheduling controller, a linear parameter varying controller, and an input-output feedback linearization controller are designed in sections 4, 5 and 6 respectively. The experimental results are presented and discussed in section 7. Finally, some concluding remarks are given in section 8.

2 The dynamic model of the quadruple tank process

A schematic diagram and a picture of the quadruple tank process are shown in Fig. 1. The system consists of four cylindrical tanks and two pumps; these pumps are connected to valves for water distribution. Pump 1 is used to distribute water from the water reservoir to tanks 1 and 4, while pump 2 is used to distribute water to tanks 2 and 3. Four pressure sensors which are located at the bottom of each tank are used to measure the water levels in the tanks.

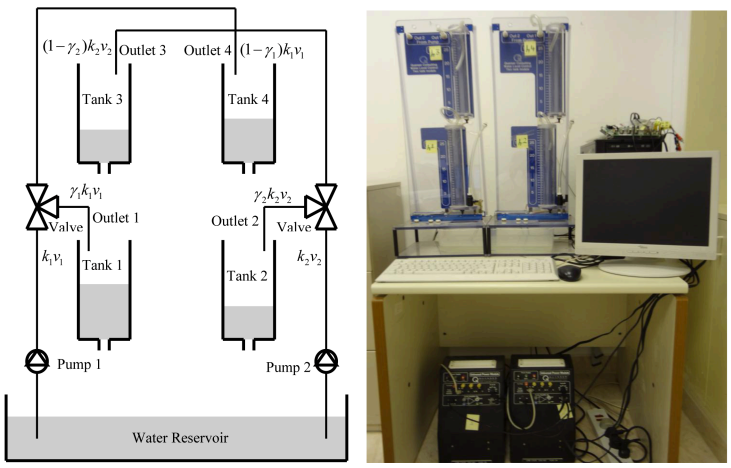


Figure 1: A schematic diagram (left) and a picture (right) of the quadruple tank process

The dynamic model of the quadruple tank process can be written as [1]:

$$\begin{bmatrix} \dot{h}_1 \\ \dot{h}_2 \\ \dot{h}_3 \\ \dot{h}_4 \end{bmatrix} = \underbrace{\begin{bmatrix} -p_1\sqrt{h_1} + p_2\sqrt{h_3} \\ -p_4\sqrt{h_2} + p_5\sqrt{h_4} \\ -p_7\sqrt{h_3} \\ -p_9\sqrt{h_4} \end{bmatrix}}_{f(h)} + \underbrace{\begin{bmatrix} p_3 & 0 \\ 0 & p_6 \\ 0 & p_8 \\ p_{10} & 0 \end{bmatrix}}_{B} \underbrace{\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}}_{v}, \quad y = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}}_{C} \underbrace{\begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix}}_{h} \tag{1}$$

where $p_1 = a_1\sqrt{2g}/A_1$, $p_2 = a_3\sqrt{2g}/A_1$, $p_3 = \gamma_1 k_1/A_1$, $p_4 = a_2\sqrt{2g}/A_2$, $p_5 = a_4\sqrt{2g}/A_2$, $p_6 = \gamma_2 k_2/A_2$, $p_7 = a_3\sqrt{2g}/A_3$, $p_8 = (1-\gamma_2)k_2/A_3$, $p_9 = a_4\sqrt{2g}/A_4$, and $p_{10} = (1-\gamma_1)k_1/A_4$. The variables and the parameters of the process are the water level h_i, the cross-section area A_i, the outlet cross-section area a_i of tank i ($i = 1, 2, \ldots, 4$), the voltage v_j applied to pump j, the constant gain k_j of pump j, the constant of the valve γ_j connected to pump j ($j = 1, 2$). The output of the system is y and the gravitational acceleration is g.

3 The control objective of the paper

The objective of the paper is to design control schemes such that the outputs of the process, i.e. h_1 and h_2, asymptotically converge to the desired levels h_1^o and h_2^o. The steady state value of the applied voltage vector $v^o = [v_1^o, v_2^o]^T$ which can maintain the water level vector at $h^o = [h_1^o, h_2^o, h_3^o, h_4^o]^T$ must satisfy the following equilibrium equations:

$$-p_1\sqrt{h_1^o}+p_2\sqrt{h_3^o}+p_3v_1^o=0,\ -p_4\sqrt{h_2^o}+p_5\sqrt{h_4^o}+p_6v_2^o=0,\ -p_7\sqrt{h_3^o}+p_8v_2^o=0,\ -p_9\sqrt{h_4^o}+p_{10}v_1^o=0 \quad (2)$$

Clearly, equations (2) imply that one can only select the values of two water levels. For instance, if we select the values of h_1^o and h_2^o (since they represent the desired outputs of the system) then it can be shown that the steady state values of h_3^o and h_4^o must satisfy the following matrix equation:

$$\begin{bmatrix} \sqrt{h_3^o} \\ \sqrt{h_4^o} \end{bmatrix} = \begin{bmatrix} p_2/p_1 & (p_3p_9)/(p_1p_{10}) \\ (p_6p_7)/(p_4p_8) & p_5/p_4 \end{bmatrix}^{-1} \begin{bmatrix} \sqrt{h_1^o} \\ \sqrt{h_2^o} \end{bmatrix} \quad (3)$$

It should be noted that the inverse of the matrix in (3) exists when $p_2p_5p_8p_{10} \neq p_3p_6p_7p_9$, which is equivalent to $\gamma_1 + \gamma_2 \neq 1$.

4 Design of a gain scheduling controller

In this section, a gain scheduling controller is designed using the classical approach of gain scheduling design [15]. At first, the nonlinear dynamic model of the process under a linear controller is linearized around several operating points. Then, a linear controller is designed at each operating point to meet the required specifications. Finally, the resulting linear controllers are interpolated according to the water levels h_1 and h_2 to produce a single gain scheduling controller. The obtained controller is used to regulate the output of the process from one operating point to another operating point.

Consider the following state feedback integral controller:

$$v = -K_h h - K_\sigma \sigma - K_e e, \quad \dot{\sigma} = e = r - y \quad (4)$$

where $r = [r_1, r_2]^T$ is the reference vector and K_h, K_σ, and K_e are the gains of the controller. The closed loop system when using the controller (4) into the model of the process given by (1) is such that:

$$\dot{h} = f(h) - B[(K_h - K_eC)h + K_\sigma\sigma + K_er], \quad \dot{\sigma} = r - Ch, \quad y = Ch \quad (5)$$

When $r = [h_1^o, h_2^o]^T$, the closed loop system (5) has an equilibrium point at (h^o, σ^o) where h^o satisfies (3), $e = 0$, and $\sigma^o = -K_\sigma^{-1}[K_hh^o + v^o]$ provided that the matrix $K_\sigma \in R^{2\times 2}$ is nonsingular. To obtain a linear system, we linearize the closed loop system (5) about (h^o,σ^o) to yield:

$$\dot{x} = \left(\begin{bmatrix} A & 0_{4\times 2} \\ -C & 0_{2\times 2} \end{bmatrix} - \begin{bmatrix} B \\ 0_{2\times 2} \end{bmatrix} \begin{bmatrix} K_h - K_eC & K_\sigma \end{bmatrix}\right) \underbrace{\begin{bmatrix} h-h^o \\ \sigma-\sigma^o \end{bmatrix}}_{x}; \quad A = 0.5 \begin{bmatrix} -p_1/\sqrt{h_1^o} & 0 & p_2/\sqrt{h_3^o} & 0 \\ 0 & -p_4/\sqrt{h_2^o} & 0 & p_5/\sqrt{h_4^o} \\ 0 & 0 & -p_7/\sqrt{h_3^o} & 0 \\ 0 & 0 & 0 & -p_9/\sqrt{h_4^o} \end{bmatrix} \quad (6)$$

where $A = \partial f(h)/\partial h|_{h=h^o}$. The gains K_h, K_σ, and K_e are designed using the linearized feedback system (6) such that all closed loop poles lie inside a prescribed region shown in Fig. 2. Note

that by placing the closed loop poles inside the shaded region, we ensure that good responses are obtained. This is the case because by placing the closed loop poles inside this region results in i) a minimum decay rate τ, ii) a minimum damping ratio $\zeta = \cos(\varphi)$, and iii) acceptable control gains penalized by ρ.
Now, assume that the output of the process needs to be regulated to the ith operating point $(h_1^{o^i}, h_2^{o^i})$ and then to the $(i+1)$th operating point $(h_1^{o^{i+1}}, h_2^{o^{i+1}})$. To achieve this task, the linearized feedback system given by (6) is used to obtain a set of gains $\{K_h^l, K_\sigma^l, K_e^l\}$, $(l = 1, \ldots, 4)$, which are designed to meet the above mentioned specifications at each of the following operating points: $(h_1^{o^i}, h_2^{o^i})$, $(h_1^{o^i}, h_2^{o^{i+1}})$, $(h_1^{o^{i+1}}, h_2^{o^i})$ and $(h_1^{o^{i+1}}, h_2^{o^{i+1}})$, respectively. The corresponding set of controllers are given by $v^l = -K_h^l h - K_\sigma^l \sigma - K_e^l e$ for $l = 1, 2, \ldots, 4$. Using the bilinear interpolating method [16], these control outputs are interpolated according to the water levels h_1 and h_2 to produce the following gain scheduling controller:

$$v = \eta_1 v^1 + \eta_2 v^2 + \eta_3 v^3 + \eta_4 v^4 \tag{7}$$

where $\eta_1 = (h_1^{o^{i+1}} - h_1)(h_2^{o^{i+1}} - h_2) / ((h_1^{o^{i+1}} - h_1^{o^i})(h_2^{o^{i+1}} - h_2^{o^i}))$, $\eta_2 = -(h_1^{o^{i+1}} - h_1)(h_2^{o^i} - h_2) / ((h_1^{o^{i+1}} - h_1^{o^i})(h_2^{o^{i+1}} - h_2^{o^i}))$, $\eta_3 = -(h_1^{o^i} - h_1)(h_2^{o^{i+1}} - h_2) / ((h_1^{o^{i+1}} - h_1^{o^i})(h_2^{o^{i+1}} - h_2^{o^i}))$ and $\eta_4 = (h_1^{o^i} - h_1)(h_2^{o^i} - h_2) / ((h_1^{o^{i+1}} - h_1^{o^i})(h_2^{o^{i+1}} - h_2^{o^i}))$.
The controller given by (7) is applied to the quadruple tank process to regulate (h_1, h_2) to the operating point $(h_1^{o^i}, h_2^{o^i})$ and then to the next operating point $(h_1^{o^{i+1}}, h_2^{o^{i+1}})$. It should be noted that the controller (7) does not guarantee the stability and the performance of the system over the whole range of operation of process; this drawback can be overcome by using the LPV approach developed in the next section.

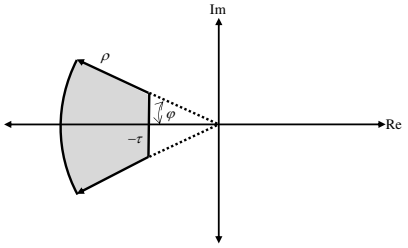


Figure 2: Pole-placement region

5 Design of a linear parameter varying controller

This section presents the design of a linear parameter varying (LPV) controller for the quadruple tank process. In order to design an LPV controller, the process model (1) needs to be written as a quasi-LPV model. To achieve this goal, a standard polynomial fitting technique [26] is used to approximate the nonlinear terms $\sqrt{h_i}$ with ${}_i h_i$ for $0 \leq h_i \leq \overline{h}_i = 30\ cm$, where ${}_i$ is obtained as ${}_i = 0.583 - 4.036 \times 10^{-2} h_i + 1.73 \times 10^{-3} h_i^2 - 3.659 \times 10^{-5} h_i^3 + 2.981 \times 10^{-7} h_i^4$ for $i = 1, \ldots, 4$. It can be shown that the parameters ${}_i$ are bounded such that $0.1 = {}_i \leq {}_i \leq {}_i = 0.6$. Notice that the parameter vector $= [{}_1, {}_2, {}_3, {}_4]^T$ is varying inside a hyper-rectangle region with 2^4 vertices defined as $\Lambda_j \in \{(\upsilon_{1,j}, \ldots, \upsilon_{4,j}) | \, \upsilon_{i,j} \in \{{}_i, {}_i\}\}$ for $j = 1, \ldots, 2^4$, where $\upsilon_{i,j} \in R$ is the ith element of $\Lambda_j \in R^4$.

Therefore, the process model (1) can be written in the following quasi-LPV form:

$$\begin{bmatrix}\dot{h_1}\\ \dot{h_2}\\ \dot{h_3}\\ \dot{h_4}\end{bmatrix}=\underbrace{\begin{bmatrix}-p_{11} & 0 & p_{23} & 0\\ 0 & -p_{42} & 0 & p_{54}\\ 0 & 0 & -p_{73} & 0\\ 0 & 0 & 0 & -p_{94}\end{bmatrix}}_{A()}\underbrace{\begin{bmatrix}h_1\\ h_2\\ h_3\\ h_4\end{bmatrix}}_{h}+\underbrace{\begin{bmatrix}p_3 & 0\\ 0 & p_6\\ 0 & p_8\\ p_{10} & 0\end{bmatrix}}_{B}\underbrace{\begin{bmatrix}v_1\\ v_2\end{bmatrix}}_{v},\ y=\underbrace{\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\end{bmatrix}}_{C}\underbrace{\begin{bmatrix}h_1\\ h_2\\ h_3\\ h_4\end{bmatrix}}_{h} \tag{8}$$

Let the LPV state feedback integral controller be such that:

$$v=-\underbrace{(\bar{K}_{h_0}+\sum_{i=1}^{4}{}_i\bar{K}_{hi})}_{\bar{K}_h()}h-\bar{K}_\sigma\sigma-\bar{K}_e e,\quad \dot{\sigma}=e=r-y \tag{9}$$

where the controller gains $\bar{K}_{h_l}$ ($l=0,1,\ldots,4$), $\bar{K}_\sigma$, and $\bar{K}_e$ are designed such that the required specifications are met for all admissible values of the parameter vector .
The closed loop dynamic model of the process when using (8) and (9) is such that:

$$\dot{\overline{x}}=(\underbrace{\begin{bmatrix}A() & 0_{4\times2}\\ -C & 0_{2\times2}\end{bmatrix}}_{\bar{A}()}-\underbrace{\begin{bmatrix}B\\ 0_{2\times2}\end{bmatrix}}_{\bar{B}}\underbrace{\begin{bmatrix}\bar{K}_{he}() & \bar{K}_\sigma\end{bmatrix}}_{\bar{K}()})\underbrace{\begin{bmatrix}h\\ \sigma\end{bmatrix}}_{\overline{x}}+\underbrace{\begin{bmatrix}-B\bar{K}_e\\ I_{2\times2}\end{bmatrix}}_{B_c}r,\quad y=\underbrace{\begin{bmatrix}C & 0_{2\times2}\end{bmatrix}}_{C_c}\overline{x} \tag{10}$$

The gain $\bar{K}()=[\bar{K}_{he}()\ \ \bar{K}_\sigma]$ is designed to guarantee the stability of the closed loop system for any possible trajectory . To obtain a good performance of the closed loop system, the closed loop poles of the system (10) (at the 2^4 vertices) are forced to lie in the left half of the complex plane and inside the region shown in Fig. 2. This objective is achieved by considering the LPV gain $L()$ such that $L()=L_0+\sum_{i=1}^{4}{}_iL_i=\bar{K}()P$, where P is a positive definite matrix. Then, the following set of LMIs [20] are solved for P and L_l ($l=0,\ 1,\ldots,\ 4$):

$$\begin{gathered}M(\Lambda_j)+M^T(\Lambda_j)+2\tau P<0,\\ \begin{bmatrix}-\rho P & M(\Lambda_j)\\ M^T(\Lambda_j) & -\rho P\end{bmatrix}<0,\begin{bmatrix}\sin(\varphi)(M(\Lambda_j)+M^T(\Lambda_j)) & \cos(\varphi)(M(\Lambda_j)-M^T(\Lambda_j))\\ \cos(\varphi)(M^T(\Lambda_j)-M(\Lambda_j)) & \sin(\varphi)(M(\Lambda_j)+M^T(\Lambda_j))\end{bmatrix}<0\end{gathered} \tag{11}$$

where $M(\Lambda_j)=\bar{A}(\Lambda_j)P-\bar{B}L(\Lambda_j)$ for $j=1,2,3,\ldots,2^4$. Once the matrices P and L_l ($l=0,\ 1,\ldots,\ 4$) are obtained using any available software such as the LMI Control Toolbox [25], the controller gains $\bar{K}_{h_0}$, $\bar{K}_{h_l}$ and $\bar{K}_\sigma$ are calculated using the matrix equations $\begin{bmatrix}\bar{K}_{h_0}-\bar{K}_eC & \bar{K}_\sigma\end{bmatrix}=L_0P^{-1}$ and $\begin{bmatrix}\bar{K}_{hl} & 0_{2\times2}\end{bmatrix}=L_lP^{-1}$ for $l=1,\ldots,4$, where the gain $\bar{K}_e$ is designed to improve the closed-loop performance of the system.
The following proposition gives the main result of this section.

Proposition 1. *The LPV controller (9) with gains obtained using (11) guarantees the stability of the closed loop system (10) and the regulation of the system output y to its desired value over the whole range of operation of the process. Furthermore, the closed loop poles at each vertex of the scheduling variable are located inside the region shown in Fig. 2.*

Proof: Consider the matrix equation $\bar{K}()=L()P^{-1}$. Let $V=\overline{x}^TP^{-1}\overline{x}$ be a Lyapunov function candidate for the system (10), then the time derivative of V along the trajectories of the system while assuming that $r=0$ is given by $\dot{V}=(P^{-1}\overline{x})^T[\overline{A}()P-\overline{B}\,L()+P\overline{A}^T()-L()^T\overline{B}^T](P^{-1}\overline{x})=$

$(P^{-1}\overline{x})^T[M() + M^T()](P^{-1}\overline{x})$. Given that $M(\Lambda_j) + M^T(\Lambda_j) < -2\tau P < 0$ at each vertex Λ_j. then $M() + M^T() < -2\tau P < 0$ for all admissible values of which implies that $\dot{V} < 0$. Therefore, the LPV controller (9) guarantees the stability of the closed loop system (10). Furthermore, the integral term in the LPV controller (9) ensures the regulation of the output to its desired level. Moreover, the LMIs (11) ensure that the closed loop poles at each vertex Λ_j are located inside the region shown in Fig. 2 (see [20]). □

The LPV control design approach is computationally intensive because the LMIs (11) need to be solved at the 2^4 vertices. Therefore, the following section presents a controller which is less computationally intensive than the proposed LPV controller.

6 Design of an input-output feedback linearization controller

This section deals with the design of an input-output feedback linearization controller for the quadruple tank process.
Let the input-output feedback linearization controller be such that:

$$v_1 = (1/p_3)[\kappa_1 e_1 + \kappa_2 \int_0^t e_1 d\acute{t} + p_1\sqrt{h_1} - p_2\sqrt{h_3}],\ v_2 = (1/p_6)[\kappa_3 e_2 + \kappa_4 \int_0^t e_2 d\acute{t} + p_4\sqrt{h_2} - p_5\sqrt{h_4}] \quad (12)$$

where $e_1 = r_1 - h_1$ and $e_2 = r_2 - h_2$. The positive controller gains κ_1, κ_2, κ_3, and κ_4 are chosen such $\kappa_1 > 2\sqrt{\kappa_2}$ and $\kappa_3 > 2\sqrt{\kappa_4}$.
The following proposition gives the main result of this section.

Proposition 2. *The input-output feedback linearization controller controller (12) when applied to the quadruple tank process (1) guarantees the exponential convergence of the water levels h_1 and h_2 to their desired values h_1^o and h_2^o respectively as t tends to infinity. Moreover, the controller (12) guarantees the boundedness of the water levels $h_3(t)$ and $h_4(t)$ (i.e., $0 \leq h_3(t) \leq q_1$, and $0 \leq h_4(t) \leq q_2$ for some positive constants q_1 and q_2).*

Proof: The application of the controllers v_1 and v_2 given by (12) to the dynamical model of the quadruple tank process (1) results in the following error dynamics:

$$\dot{e}_1 = -\kappa_1 e_1 - \kappa_2 \int_0^t e_1 d\acute{t}, \quad \dot{e}_2 = -\kappa_3 e_2 - \kappa_4 \int_0^t e_2 d\acute{t} \quad (13)$$

The error dynamics (13) can be written as $\ddot{e}_1 + \kappa_1\dot{e}_1 + \kappa_2 e_1 = 0$ and $\ddot{e}_2 + \kappa_3\dot{e}_2 + \kappa_4 e_2 = 0$. By choosing $\kappa_1 > 2\sqrt{\kappa_2}$ and $\kappa_3 > 2\sqrt{\kappa_4}$, we are guaranteed that the characteristic equations $s^2 + \kappa_1 s + \kappa_2 = 0$ and $s^2 + \kappa_3 s + \kappa_4 = 0$ have negative real roots. In this case, the solutions of (13) are given by:

$$e_1(t) = c_1 exp(-\lambda_1 t) + c_2 exp(-\lambda_2 t), \quad e_2(t) = c_3 exp(-\lambda_3 t) + c_4 exp(-\lambda_4 t) \quad (14)$$

where $-\lambda_i$ $(i = 1, \ldots, 4)$ are the roots of the above characteristic equations, and c_i $(i = 1, \ldots, 4)$ are constants which depend on the initial conditions and the values of the λ_i. Therefore, the errors e_1 and e_2 exponentially converge to zero as t tends to infinity (i.e., the water levels h_1 and h_2 exponentially converge to their desired levels h_1^o and h_2^o respectively as t tends to infinity).
Using equations (12)-(14), it can be shown that $\dot{h}_3$ and $\dot{h}_4$ in (1) can be written as follows:

$$\dot{h}_3 = -p_7\sqrt{h_3} - (p_5 p_8/p_6)\sqrt{h_4} + m_1(t), \quad \dot{h}_4 = -(p_2 p_{10}/p_3)\sqrt{h_3} - p_9\sqrt{h_4} + m_2(t) \quad (15)$$

where,

$$m_1(t) = (p_8/p_6)(c_3\lambda_3 exp(-\lambda_3 t) + c_4\lambda_4 exp(-\lambda_4 t) + p_4\sqrt{h_2^o - c_3 exp(-\lambda_3 t) - c_4 exp(-\lambda_4 t)})$$
$$m_2(t) = (p_{10}/p_3)(c_1\lambda_1 exp(-\lambda_1 t) + c_2\lambda_2 exp(-\lambda_2 t) + p_1\sqrt{h_1^o - c_1 exp(-\lambda_1 t) - c_2 exp(-\lambda_2 t)})$$

It can be shown that $m_1(t)$ and $m_2(t)$ are bounded from above for all $t \geq 0$, i.e., $m_1(t) \leq \overline{m}_1$ and $m_2(t) \leq \overline{m}_2$ where $\overline{m}_1$ and $\overline{m}_2$ are some constants. Hence, it can be concluded that $\dot{h}_3 \leq -p_7\sqrt{h_3} + \overline{m}_1$ and $\dot{h}_4 \leq -p_9\sqrt{h_4} + \overline{m}_2$.
Consider the differential equation $\dot{\hat{h}}_3 = -p_7\sqrt{\hat{h}_3} + \overline{m}_1$ with the initial value $\hat{h}_3(0) = h_3(0)$. Let Lyapunov function candidate $V_1(\hat{h}_3) = \hat{h}_3^2$, then $\dot{V}_1 = -2\hat{h}_3(p_7\sqrt{\hat{h}_3} - \overline{m}_1)$. Notice that $\dot{V}_1$ is negative when $\sqrt{\hat{h}_3} > (\overline{m}_1/p_7)$ (which corresponds to $V_1 > (\overline{m}_1/p_7)^4$). This means that all solutions starting such that $V_1(0) > (\overline{m}_1/p_7)^4$ will decrease monotonically but will never go below the line $V_1 = (\overline{m}_1/p_7)^4$, while all solutions starting such that $V_1(0) \leq (\overline{m}_1/p_7)^4$ will increase monotonically but they will never cross the line $V_1 = (\overline{m}_1/p_7)^4$ because $\dot{V}_1$ is negative for $V_1 > (\overline{m}_1/p_7)^4$. Therefore, we can conclude that $V_1 \leq max\{V_1(0), (\overline{m}_1/p_7)^4\}$ or $\hat{h}_3 \leq max\{\hat{h}_3(0), (\overline{m}_1/p_7)^2\}$. The comparison principle [22] leads us to conclude that $h_3(t) \leq \hat{h}_3(t)$. Since $h_3(t) \geq 0$, we can conclude that $0 \leq h_3(t) \leq q_1 := max\{h_3(0), (\overline{m}_1/p_7)^2\}$. Similar arguments can be used to conclude that $0 \leq h_4(t) \leq q_2 := max\{h_4(0), (\overline{m}_2/p_9)^2\}$. □

Therefore, it can be concluded that the proposed controller (12) guarantees the exponential convergence of h_1 and h_2 to their desired values as well as the boundedness of h_3 and h_4.

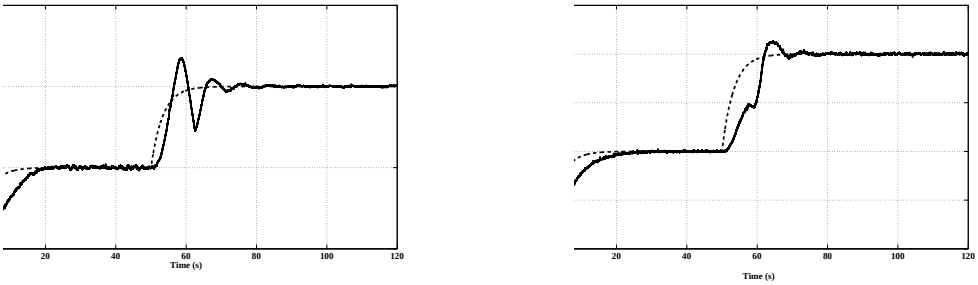


Figure 3: The water levels of tank 1 (solid: left) and tank 2 (solid: right) when using the gain scheduling controller. The references r_1 and r_2 are depicted using the dashed lines

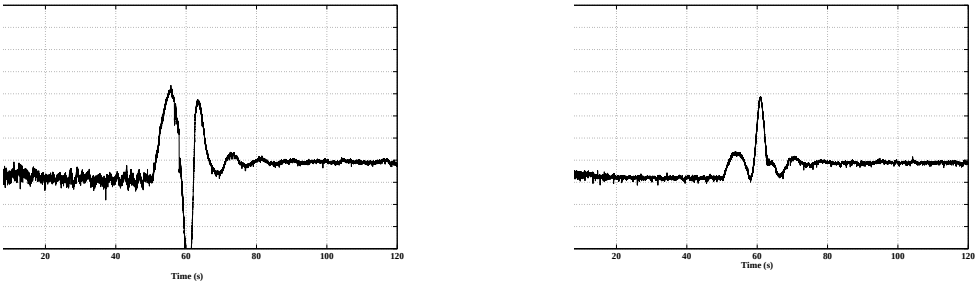


Figure 4: The inputs v_1 (left) and v_2 (right) when using the gain scheduling controller

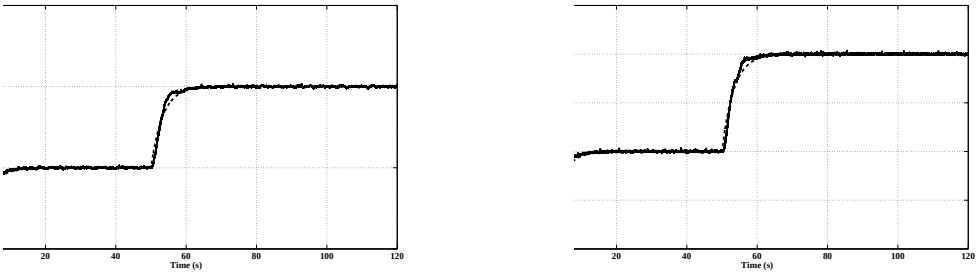


Figure 5: The water levels of tank 1 (solid: left) and tank 2 (solid: right) when using the linear parameter varying controller. The references r_1 and r_2 are depicted using the dashed lines

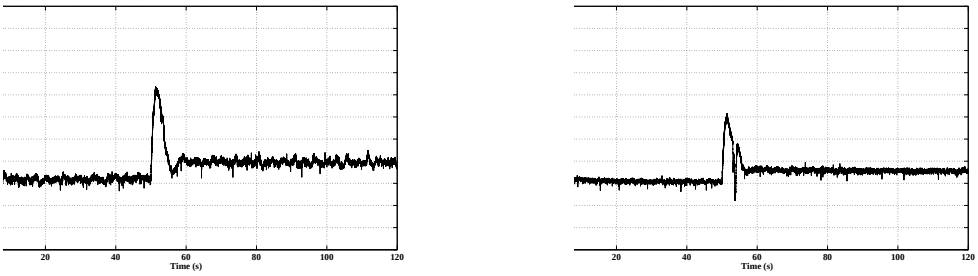


Figure 6: The inputs v_1 (left) and v_2 (right) when using the linear parameter varying controller

7 Experimental results

The three proposed control schemes are implemented using an experimental setup of the quadruple tank process manufactured by Quanser Consulting Inc. [24]. The physical parameters of the quadruple tank system are as follows: $A_1 = A_2 = A_3 = A_4 = 15.5179\ cm^2$, $a_1 = a_2 = a_3 = a_4 = 0.1781\ cm^2$, $g = 981\ cm/s^2$, $k_1 = k_2 = 3.3\ cm^3/Vs$, $\gamma_1 = 0.66$, and $\gamma_2 = 0.75$. The sampling rate of the process is 10^{-3} seconds. The reference signal $r = [r_1,\ r_2]^T$ is chosen such that r_1 changes its amplitude from 5 cm to 10 cm at $t = 50$ sec and r_2 changes its amplitude from 4 cm to 8 cm at $t = 50$ sec.
At first, the gain scheduling controller given by (7) is used to regulate the output y to the operating point $(h_1^{o^i},\ h_2^{o^i}) = (5\ cm,\ 4\ cm)$ and then to the operating point $(h_1^{o^{i+1}},\ h_2^{o^{i+1}}) = (10\ cm,\ 8\ cm)$. The controller gains, at each operating point, are designed such that the closed loop poles lie inside the region shown in Fig. 2 where $\tau = 0.03$, $\zeta = 20^\circ$ and $\rho = 2$. The experimental results are shown in Figs. 3-4. Fig. 3 shows the water levels in tanks 1 and 2, while Fig. 4 shows the input voltages to pump 1 and pump 2. It can be seen from the figures that the water levels h_1 and h_2 track the desired reference signal r. However, the water level h_1 exhibits a percent overshoot of about 20 % and a settling time of about 20 seconds while the water level h_2 exhibits a percent overshoot of about 6 % and a settling time of about 20 seconds. Also, the input voltages stay within reasonable ranges. It should be noted that the performance of the system can be further improved through proper tuning of the parameters of the controller.

Secondly, the linear parameter varying controller given by (9) is designed and implemented such that the system is stable over the whole operating range $0 \leq h_i \leq 30\ cm$; the closed loop poles are located inside the region shown in Fig. 2 where $\tau = 0.03$, $\zeta = 20^\circ$ and $\rho = 2$. The experimental results are shown in Figs. 5-6. Fig. 5 shows the water levels in tanks 1 and 2, while Fig. 6 shows the input voltages to pump 1 and pump 2. It can be seen from the figures that the

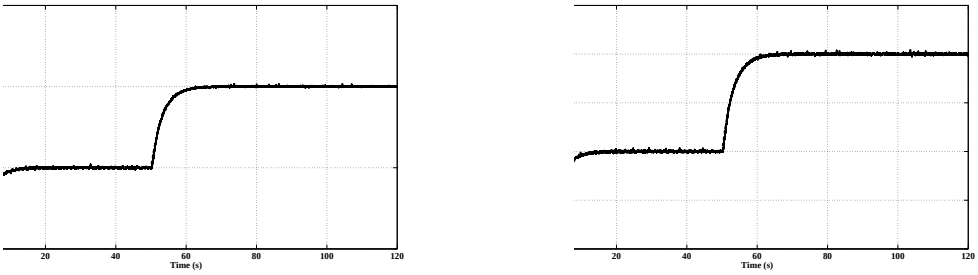


Figure 7: The water levels of tank 1 (solid: left) and tank 2 (solid: right) when using the input-output feedback linearization controller. The references r_1 and r_2 are depicted using the dashed lines

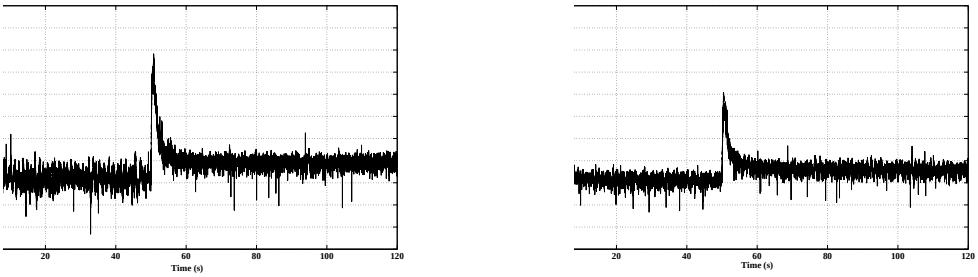


Figure 8: The inputs v_1 (left) and v_2 (right) when using the input-output feedback linearization controller

water levels h_1 and h_2 track the desired reference signal r with no overshoot and a settling time of about 10 seconds for both h_1 and h_2. Also, it can be seen that the input voltages stay within reasonable ranges.

Thirdly, the input-output feedback linearization controller given by (12) is applied to the quadruple tank process. The parameters of the controller are taken to be $\kappa_1 = \kappa_3 = 3$ and $\kappa_2 = \kappa_4 = 1$. The experimental results are shown in Figs. 7-8. Fig. 7 shows the water levels in tanks 1 and 2, while Fig. 8 shows the input voltages to pump 1 and pump 2. It can be seen from the figures that the water levels h_1 and h_2 track the desired reference signal r very well. Also, the input voltages stay within reasonable ranges (but display a bit more chattering).

To compare the performances of the proposed control schemes, the errors $e_1 = r_1 - h_1$ and $e_2 = r_2 - h_2$ are plotted in Fig. 9. It is clear from this figure that all the errors converge to zero. However, the errors for the input-output feedback linearization controller are less than the errors of the other two controllers. In addition, the figures show that the linear parameter varying controller gave better results than the gain scheduling controller. Furthermore, it is noted that the input-output feedback linearization controller can be implemented easily.

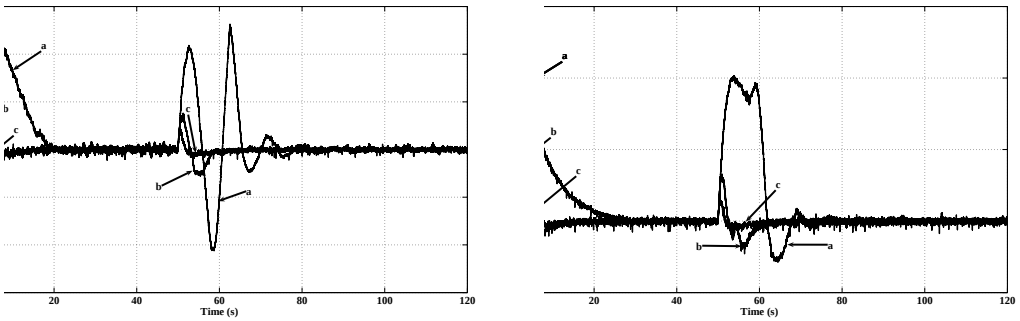


Figure 9: The errors e_1 (left) and e_2 (right) when using the gain scheduling controller (a), the linear parameter varying controller (b) and the input-output feedback linearization controller (c)

8 Conclusion

A gain scheduling controller, a linear parameter varying controller, and an input-output feedback linearization controller are proposed for the quadruple tank process. At first, we propose to use a gain scheduling controller. However, this controller does not ensure the stability and the performance of the closed loop system over the whole operating range. Therefore, a linear parameter varying controller which guarantees the stability and the performance of the system over the desired operating range is proposed to control the process. However, the design of the linear parameter varying controller is quite complicated. Therefore, to reduce the design/implementaion complexity, an input-output feedback linearization controller is derived for the process. Experimental results are presented for the three control schemes. The implementation results indicate that the three proposed control schemes work well and are able to regulate the output of the system to its desired value. However, the implementation results indicate that the input-output feedback linearization controller gave the better performance in comparison with the other two controllers. Future research will address the design of fault-tolerant control schemes for the quadruple tank process.

Bibliography

[1] K.H. Johansson, "The quadruple-tank process: a multivariable laboratory process with an adjustable zero," *IEEE Trans. Control Syst. Technol.*, vol. 8, pp. 456-465, 2008.

[2] B. Moaveni, A. Khaki-Sedigh, "Input-output pairing for nonlinear multivariable systems," *J. Appl. Sci.*, vol. 22, pp. 3492-3498, 2007.

[3] F. Garelli, R.J. Mantz, H. De Battista, "Limiting interactions in decentralized control of MIMO systems," *J. Process Control*, vol. 16, pp. 473-483, 2006.

[4] K.J. Astrom, K.H. Johansson, Q.G. Wang, "Design of decoupled PI controllers for two-by-two systems," *IEE Proc. Control Theory Appl.*, vol. 149, pp. 74-81, 2002.

[5] E.P. Gatzke, E.S. Meadows, C. Wang, F.J. Doyle III, "Model based control of a four-tank system," *Comput. Chem. Eng.*, vol. 24, pp. 1503-1509, 2000.

[6] M. Mercangoz, F.J. Doyle III, "Distributed model predictive control of an experimental four-tank system," *J. Process Control*, vol. 17, pp. 297-308, 2007.

[7] D. Henriksson, A. Cervin, J. Akesson, K. Arzen, "On dynamic real-time scheduling of model predictive controllers," *41st IEEE Conference on Decision and Control*, Las Vegas, pp. 1325-1330, 2002.

[8] S.M. Mahdi Alavi, A. Khaki-Sedigh, B. Labibi, M.J. Hayes, "Improved multivariable quantitative feedback design for tracking error specifications," *IET Control Theory Appl.*, vol. 1, pp. 1046-1053, 2007.

[9] R. Vadigepalli, E.P. Gatzke, F. Doyle III, "Robust control of a multivariable experimental four-tank system," *Ind. Eng. Chem. Res.*, vol. 40, pp. 1916-1927, 2001.

[10] R. Findeisen, F. Allgower, L.T. Biegler, *Assessment and Future Directions of Nonlinear Model Predictive Control, Springer*, Berlin, 2007.

[11] N. Barhoumi, S. HadjSaid, F. M'sahli, "Constrained nonlinear model predictive control of hybrid dynamic systems," *J. Auto. Syst. Eng.*, vol. 3, pp. 46-56, 2009.

[12] P.P. Biswasa, R. Srivastavaa, S. Raya, A.N. Samanta, "Sliding mode control of quadruple tank process," *Mechatronics*, vol. 19, pp. 548-561, 2009.

[13] B. Labibi, H.J. Marquez, T. Chen, "Decentralized robust output feedback control for control affine nonlinear interconnected systems," *J. Process Control*, vol. 19, pp. 865-878, 2009.

[14] F.D. Bianchi, R.J. Mantz, C.F. Christiansen, "Multivariable PID control with set-point weighting via BMI optimisation," *Automatica*, vol. 44, pp. 472-478, 2008.

[15] W.J. Rugh, J.S. Shamma, "Research on gain scheduling," *Automatica*, vol. 36, pp. 1401-1425, 2000.

[16] R.A. Hyde, K. Glover, "The application of scheduled H_∞ controllers to a VSTOL aircraft," *IEEE Trans. Autom. Control*, vol. 38, pp. 1021-1039, 1993.

[17] N. Aouf, D.G. Bates, I. Postlethwaite, B. Boulet, "Scheduling schemes for an integrated flight and propulsion control system," *Control Eng. Pract.*, vol. 10, pp. 685-696, 2002.

[18] J.S. Shamma, M. Athans, "Guaranteed properties of gain scheduled control for linear parameter varying plants," *Automatica*, vol. 3, pp. 559-564, 1991.

[19] P. Apkarian, P. Gahinet, "A convex characterization of gain-scheduled H_∞ controllers," *IEEE Trans. Autom. Control*, vol. 5, pp. 853-864, 1995.

[20] M. Chilali, P. Gahinet, "H_∞ design with pole placement constraints: an LMI approach," *IEEE Trans. Autom. Control*, vol. 41, pp. 358-367, 1996.

[21] P. Apkarian, H.D. Tuan, "Parameterized LMIs in control theory," *SIAMJ. Control Optim.*, vol. 4, pp. 1241-1264, 2000.

[22] H.K. Khalil, *Nonlinear Systems*, Prentice-Hall, New Jersey, 2002.

[23] J.J.E. Slotine, W. Li, *Applied Nonlinear Control*, Prentice-Hall, New Jersey, 1991.

[24] J. Apkarian, *Coupled Water Tank Experiments Manual*, Quanser Consulting Inc., Canada, 1999.

[25] P. Gahinet, A. Nemirovski, A. Laub, M. Chilali, *LMI Control Toolbox User's Guide*, The MathWorks Inc., Natick, MA, 1995.

[26] G.E. Forsythe, M.A. Malcolm, C.B. Moler, *Computer Methods for Mathematical Computations*, Prentice-Hall, New Jersey, 1977.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 606-616

The Role of Visual Rhetoric in Semantic Multimedia: Strategies for Decision Making in Times of Crisis

A.M.P. Brasoveanu, I. Dzitac

Adrian M.P. Brasoveanu
MODUL University Vienna, Department of New Media Technology,
Austria, 1 Am Kahlenberg, 1190 Vienna,
E-mail: adrian.brasoveanu@modul.ac.at

Ioan Dzitac
1. Aurel Vlaicu University of Arad,
Romania, 310330 Arad, 2 Elena Dragoi, and
2. Agora University
Romania, 410526 Oradea, 8 Piata Tineretului
E-mail: rector@univagora.ro

Abstract: As semantic multimedia is approaching mainstream, even the great improvements that can be seen in its classic schools, like the data mining inspired Information Retrieval based on metadata analysis, or Computer Vision, might not be enough. We identify a new group that gains traction in the semantic multimedia community and which uses as starting point developments from psychology and visual communication. For the purposes of this article we restrict our domain to visual rhetoric as we consider it to yield the biggest potential for future developments.
Living in times when the periods between crises seem to be shorter and shorter, we look at how developments in semantic multimedia can be used for predicting and overcoming crises. We analyze at least 2 aspects related to this: using information visualization to understand the evolution of crises and creating multi-layered semantic multimedia technologies that can easily be adapted to use a variety of sources and solve problems from different domains. In both cases we show how techniques inspired by visual rhetoric (information linking, framing, composition) in conjunction with named entity recognition offer a lot of benefits. The section related to multi-layered semantic multimedia technologies also draws on the lessons learned while designing a prototype application aimed at improving tourism decision making process.
The article ends with a discussion on evaluation methods for multi-layered semantic technologies applications. We look at how to evaluate them on both levels: mechanisms (information linking versus raw named entity recognition when generating visuals, for example), and decision making strategies (Do such systems actually solve real problems related to crises, create jobs or at least can they be repurposed to solve other problems than the one with which we have started?).
Keywords: semantic multimedia, visual rhetoric, text to visual matching, interactive documentary, crisis strategies, multimedia storytelling

"The progress of civilization can be read in the invention of visual artifacts, from writing to mathematics, to maps, to printing, to diagrams, to visual computing."
Stuart Card et. al. - *Readings in Information Visualization*

1 Introduction

In today's fast Web there is a need for elegant mechanisms that can help us understand how to process, store, retrieve and present the huge amount of information contained in multimedia files. Semantic multimedia [23] [24], the branch of Semantic Web focused on the analysis of multimedia documents, traditionally employed methods like metadata analysis, feature extraction

or multimodal analysis, and as a result, the core researchers were split in two large groups: a group that had its roots in text processing and data mining, and a group with roots in computer vision [21] [23] [24]. Of course at times one will need to apply both methods to get meaningful information from a multimedia system. Today, new groups are slowly emerging, driven by the advances from fields like visual communication, psychology or biology. The group at which we adhere is still a small one and tries to tie the ideas from the modern visual rhetoric to information visualization, multimedia processing and generation, interactive documentaries and other semantic multimedia areas.

Visual rhetoric [11] is one of the new disciplines taken into account by the semantic multimedia researchers. We consider this cross-pollination natural since both fields deal with the interpretation of visual media (paintings, photographs, movies, games, etc.). The question we would like to address in this paper is how can we design semantic technologies that take into account the findings from visual rhetoric?

The rest of the paper is organized in 4 sections. Section 2 starts with a discussion about the various schools and definitions of visual rhetoric; and identifies several ideas that have a potential for growth in the field of semantic multimedia. It also contains a review of the related work. Section 3 continues with a short analysis of the role that visual methods have in decision making with a special focus on crisis economics. Section 4 presents some ideas about prevention of crises and a case study built around a prototype application. We conclude our paper with a discussion on the various evaluation methods that can be used in order to assess the success of our enterprise, both on the level of the mechanisms described (comparisons between different methods for generating multimedia content) and on the level of decision making strategies.

2 The Role of Visual Rhetoric in Semantic Multimedia and Related Work

Images, video files or graphics of any kind (paintings, infographics, interactive visualizations) always tell more than we would like to admit. Since the biggest processing engine we have is our brain we should pay more attention to how we process and present any information using multimedia channels. Visual rhetoric is one of the modern disciplines that can help us do precisely this, if we take the time to study it and apply it to our representation and interactivity problems [11]. While visual rhetoric is not new, its theoretical treatment and multimedia applications are.

The seeds of this discipline can be found in the articles about art theory, film art and iconology, published in the German Space since the third decade of the 20^{th} century. Probably the most famous exponents of that period were Rudolf Arnheim [2] and Ernst Gombrich [9]. The term visual rhetoric was rarely used at the time, but most of the elements discussed in their essays (from equilibrium to lighting or color, but also space or dynamics) are included in modern treatments of visual rhetorics. The influence of this movement goes well beyond visual communication, and their ideas can be found almost everywhere from architecture to game design. This wave was mostly focused on issues of representation and composition in art.

During the '60s and '70s there was a French wave of visual rhetoric, which was inspired from the film criticism of Cahiers du Cinema, literary criticism and philosophy, its most famous exponents being Roland Barthes [3] and Jacques Bertin [4]. Cinema, photography and charts were the focus of the essays published during this wave.

Today, the dominant current in visual rhetoric theory is Anglo-Saxon, Gunther Kress [17] or Charles Hill [11] being some of its most respected proponents. Eric Kandel [16], belongs probably to both the German and Anglo-Saxon wave (he left Vienna when he was 9 years old in 1938, but he always tried to keep contact with fellow Austrians like Ernst Kris or Ernst Gombrich who were

influential in establishing the grounds on which a scientific theory of visual rhetoric will someday be formed). The current wave is one of solid grounding, multimodality, mathematics, computer science and semiotics being used to connect the dots between the various long running threads. Complex questions are asked (questions like: What is the role of music in a certain scene from a movie? How can certain elements be used in the same scene to enhance its meaning?). By doing this, the current wave starts to deconstruct the authorial intentions in a scientific manner.

While there is no single all-encompassing definition of visual rhetoric to this day, there is a consensus regarding the fact that you can create a visual rhetoric space for any discipline. Some of the fields where visual rhetoric can and should be used are investigated in [11] together with the possible definitions of visual rhetoric as seen from those fields of study. Basically all definitions agree on one aspect: visual rhetoric is a form of communication that uses images to construct meaning or arguments. By extension, visual literacy defines the way we respond to images and it implies that we are already well trained in how to read images.

The beginnings of visual rhetoric were controversial (as proved by the Ernst Haeckel biological images forgery case [8]), but its continuous improvement, especially during the last two decades, has led to its acceptance as one of the leading areas of research in visual communication. It is enough to look up the list of publications from the premier venues for semantic web, multimedia or information visualization publications (Journal of Web Semantics, ACM Multimedia, IEEE Multimedia, IEEE TVVG, ACM TOMCCAP) during the last years (2008 - 2012) to discover that some of the articles that received a lot of attention (Best Paper Awards, quotations, discussions in other papers, even sequels) apply ideas inspired by visual rhetoric like: visual query suggestion [30], narrative visualization [20], affective image classification [18], framing effects [13] and color naming models and their applications [10]. Hullman and Diakopoulos [13] apply their ideas on visualization rhetoric to a class of visualizations identified by Segel and Heer [20] as narrative visualizations.

Narrative visualizations do not just visually present some numbers, but also draw attention to the story behind those numbers, and in doing so they need to deploy an entire arsenal of techniques like provenance rhetoric, mapping rhetoric (visual metaphors, contrast, etc), linguistic or procedural rhetorics [13]. The paper about color naming models [10] is important mainly for library builders (especially JavaScript libraries), while [18] uses features generally used in psychology and biology to create an affective image classification. Some of the metrics used in [18] are color (name, contrast, features), texture (wavelet, Tamura, etc.), composition (depth of field, rule of thirds) or content (human faces, skin).

An interesting problem in our view is that of using images to illustrate arguments. An existing approach towards this problem involves finding the entities from texts (or even a search box) and associating them with images from medical literature (using caption processing, image processing, and topic discovery), and can be found in the Structured Literature Image Finder [1]. The training data used manually annotated images from different subcellular locations. SLIF, developed at CMU, is basically a restricted search engine, one that has as domain medical literature. For medical texts software like this is extremely valuable, as it is often the case that an image presents an entire story.

Identifying images that best represent entities is a challenging task by itself, but identifying images that would best illustrate a point of view is an even more daunting task. It is however the task that anyone involved in visual rhetoric would like to solve, especially people involved in advertising. Bocconi, Hardman and Nack [5] were able to generate matter of opinion documentaries by "framing" into a larger conversation clips from interviews that captured people's reactions to the September 11 terrorist act. After proving how to build the "framing" mechanisms using semantic graphs, they conclude that the role of visuals in providing support for the subject matter in documentaries needs to be further developed. We used this assumption as a

starting point for our investigations and discovered that there are several theories in social semiotics [11] [17] [27] dedicated to supporting verbal meaning with visual artifacts, and we started to build a framework around these theories.

What we noticed by reviewing the literature is that while these articles are hardly related when it comes to subject (apparently there aren't many connections between visual query suggestion, information visualization or color naming models), and some of them do not even mention the term visual rhetoric (except for [13] which refers to "visualization rhetoric"), almost all of them quote some of the pioneers of the field (Rudolf Arnheim [2], Johanes Itten [14], Roland Barthes [3], or Jacques Bertin [4], for example), and some were influenced by the research group led by Alberto del Bimbo [7]. This suggests that a third group is gaining lots of traction in the area of semantic multimedia, and that this group embraces theories from art, psychology or biology. The surveyed papers are built upon the philosophy of the early pioneers of the field, while we base our work on theories developed in the last 15 years.

3 Visual Rhetoric for Decision Making in Times of Crisis

Mainstream economists almost always fail to predict crises, and the example of the current crisis (started in 2008 with the collapse of the American housing market, and at the time of writing - 2012 - still unfinished) is one of the best. There a few economists who are said to have predicted this crisis (Nouriel Roubini, Peter Schiller, Nassim Taleb and others) [34], but there are hardly any graphics that prove their theories. The most interesting theories present dragon-kings (significant or meaningful outliers), black swans (in essence events that are almost impossible to predict) and other models that would lead to accurate predictions. Dragon-kings models proposed by ETH's Didier Sornette [15] [22] [31] do not involve events that can't be predicted, as black swan events suppose. Sornette's group anticipated some short bubbles and published the results sometimes few days before the actual events occurred [31]. Other economists just improved SOM models and run them against datasets related to the current crisis [19]. While they did not predict a new crisis or the length of the current one with a great accuracy, these models are still useful and can help us understand what happens in today's high frequency trading markets.

We think that the worst part when it comes to crisis prediction or visualization is that even some of the best visualizations (take the visualizations from one of the top contests, for example [33]) do not present us with the visual cues (dragon-kings, black swans) that would make us easily understand what happened during the last years. Just plotting some data without highlighting the events that suggest bubbles or other crisis scenarios is not going to help us too much. We have to spend more time thinking about the results of the visualizations and how to present those results in such a way that they are easy to understand. This means more time spent tackling the problems related to manipulation and bias, since graphics can in the same time improve our understanding of current or past events, but they can also be used to manipulate the public opinion. In today's connected world, where most of the governments put their data online (including financial data), we think that this dual nature of visualization rhetoric is something we must carefully address. [13] shows how to "frame" narrative visualizations to present different points of view related to the same event (poll predictions). Another method to insert more information into visualizations would be to combine various visual metaphors as demonstrated in [12].

We think there are multiple reasons why current visualizations do not take into account such phenomena.

First, many visualizations are not done by interdisciplinary teams, so if the researchers have not heard about such events (dragon kings, black swans, etc) they will not be represented in the

end product (this assumption does not apply to bigger outlets with traditions in visualization like The New York Times or Guardian).

Second, most of the visualization rhetoric used today still comes from several sources (usually Jacques Bertin [4], William Cleveland [6], Edward Tufte [25] [26], Leland Willkinson [29], the last one being more recent than the rest) mostly focused on visual presentation of data, but not on framing and the consequences of framing and story selection.

Third, there is no single, uniform, easy to understand and use visual rhetoric for any type of decision making. This means that all visualization designers need to master visual literacy. For today's visualization designers there is an easy path towards mastery of both visual rhetoric and visual literacy: they can start learning online (outlets like: [36]- [41]) and then go on and read the masters (classics: [4], [6], [25], [26], [29], or modern: [42]- [44]).

Just by choosing a story and applying some creative layering techniques over a plot we will not be able to predict and prevent a crisis. If we want to be able to do that, we will have to rely on other mechanisms like job creation, for example. Or we will have to create technologies that work on multiple levels and are easy to adapt for solving different problems. Prevention is not going to be of much help as long as the visualizations we use do not show us the dragon-kings or black swans.

4 Multi-layered Strategies for Overcoming Crisis - Semantic Technologies to Increase Profits

A good case study for our ideas would involve a situation in which we can produce multimedia content in order to leverage some of the advantages that interactive visual environments have over traditional media. Tourism industry, entertainment, or politics could be considered some of the premier venues where image is everything and such applications would help a lot. Take for example this scenario: you have several videos in which speakers tell stories about what they like in a city. If you want to create a short movie from this clips, you will likely want to convince those who watch it that the city being spoken about is really beautiful. This means you will have to replace in many places the images of the speakers with those of the objects or concepts being spoken about. Using visual rhetoric is just one of the best ways to automate this process.

Our approach towards the use of visual rhetoric in semantic multimedia is based on some of the more recent findings, like [11] and [17]. The main goal of our project is to understand how to use visual rhetoric in multimedia environments. Some secondary goals for the project are to use textual to visual matching methods and visual content generation methods. Most of the work presented in Section 2 uses visual rhetoric only for interpretation/explanation purposes (this visualization presents the evolution of a player during the last season, for example), whereas our goals are more inclined towards the possible applications of visual rhetoric in the space of multimedia content generation (automated movie generations, automated footprint generation, automated summary generations).

The case study involves a prototype web application called Interview eXplorer and used to generate visuals for a series of interviews with people (mostly students) regarding their lives in a foreign city (Vienna, the capital of Austria). We selected Vienna as the city that must be explored, not only because we spend a lot of time in the city, but also because:

- It is the city where the whole movement with integration between science and art started around 1900, according to the Nobel Prize laureate Eric Kandel [16];
- It is a city with lots of historical sites which are well represented on any media property, from Twitter to Wikipedia (Belvedere or Stephansdom, for example);
- It is a big city with enough green spaces (Donauinsel, Vienna Woods, etc.);

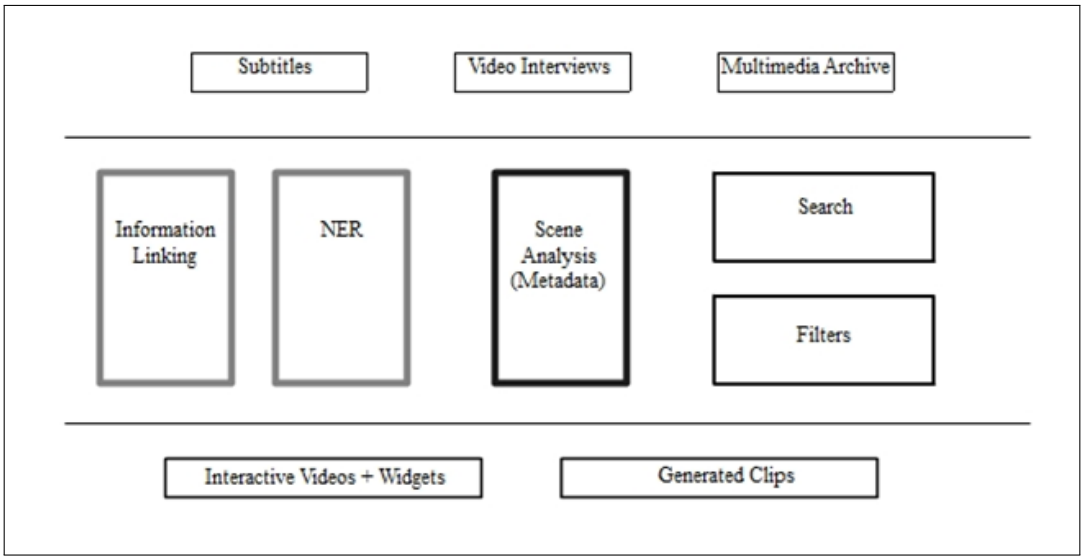


Figure 1: The components of the prototype

• The landscape changes quite fast because there are lots of new construction sites;

• While it does not have so many iconic images like London, Paris or New York, it consistently ranks higher on Mercer's Quality of life tops for few years now [36] (which can make for interesting life questions which should theoretically be harder to illustrate through images).

The generated short documentaries (webisodes) can be used in effective marketing campaigns for attracting students or tourists, for example, but they can also be used as research tools in economy or social sciences in order to understand the needs and the habits of the population from a certain area.

The interviews contain questions related to the parts of Vienna that attract the students (buildings, parks, Danube - easier to illustrate if we use an automated approach based on named entity recognition), but also questions about how well they integrated into the new environment (work, friends, social life - harder to illustrate even if you use a manual approach, and it gets even worse with automated approaches).

The application can function in 2 modes: Search (where we can just see the proposed visuals and some widgets with additional information about the entities mentioned) and Explore (where we can see proposals on how to illustrate episodes with the related identified visuals).

For additional information we integrated widgets that display data from Wikipedia or Twitter, or the maps from Google, for example. Integrating Facebook widgets was problematic, as some of the people interviewed felt that it would be a serious breach of their privacy. They were basically not aware that all you need to find a person on Facebook is a name. People that had Twitter accounts were more likely to agree to display their information on the Twitter widget, because they see this as free publicity.

Currently, all the videos, images and interviews used in the prototype are from the personal collection of the author. A long term goal of the project would be to use any free images or videos available on the web that are related to the topic discussed.

The front end of the prototype was programmed in JavaScript, while the back end uses several other programming languages. The communication between different components is done using the JSON format.

Since this is a prototype, and not a commercial application, we are not able to provide metrics regarding its profitability. The expression from the title of this section should be interpreted as: since semantic technologies are now cheaper than they used to be, they can easily be used to

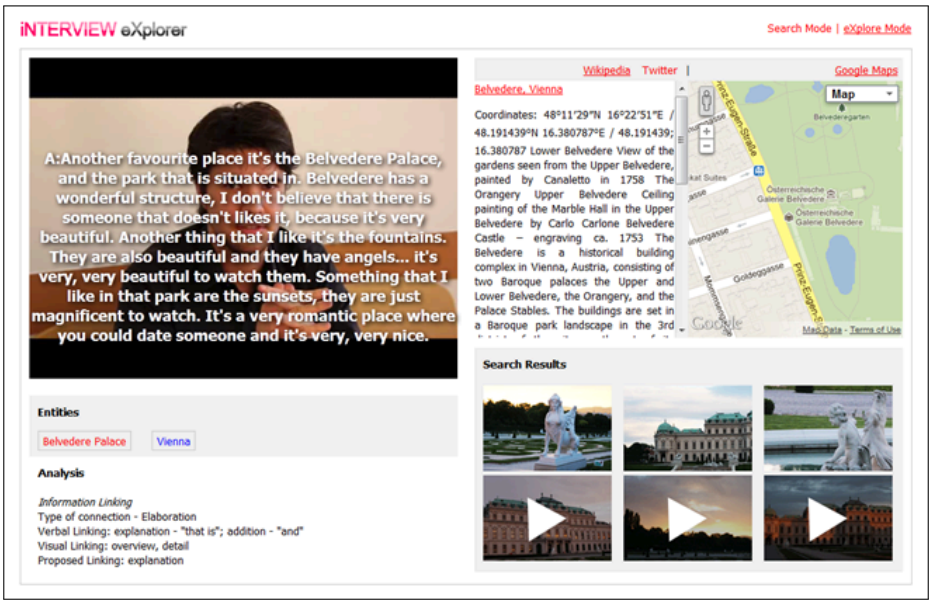


Figure 2: Interview Explorer Prototype - First Iteration

Figure 3: Visual suggestions tab from the eXplore mode. This simple visualization offers only several suggestions for replacing the frames of each video block.

create applications that provide lots of features at an accessible price. Maybe the most interesting aspect related to the development of such applications is the fact that you can create a stack of technologies to solve a specific problem (replace the image of the person that is interviewed with images about the things he is talking about) and easily repurpose them to solve other problems (like use the generated documentaries for marketing purposes, or add an interactive visualization layer on top and provide data that can be useful for government to understand foreign citizens living in Austria, for example). Seen from this perspective, the fact that semantic web is now becoming mainstream should help a lot of companies create the tools that will help them easily navigate through the bad periods they are now confronting.

5 Discussion and Future Work

By looking at the work presented in Section 2, it can be easily seen that there is a new current in semantic multimedia. This current might not yet be on par with Information Retrieval or Computer Vision, but it is certainly developing into something powerful and useful in the same time. It might not be a fully formed school like the ones we mentioned, but given the large number of papers in top venues we think that this school will be important in the next couple of years. Transforming visual rhetoric into science is not something that will happen overnight. It takes time to build the mathematical models and the simulations that are associated with scientific processes, and also needed in order to reproduce the results.

As we have seen in Section 3, many crisis prediction models should also include a visual component (SVM visualization, dragon-king, black swan or something else). We think that visualizations should take this finding into account and such patterns should be discovered as soon as possible. It is hard for us to understand why big companies invest millions or billions of dollars or euros into visualization systems for real-time stock trading, but when it comes to presenting their findings to the public they almost never show the patterns that could lead to crises. Before fractal analysis we could argue that there were no good mathematical models to predict crises [15] [22] [31], but now since we have such models we should use them.

The prototype from Section 4 is going through new iterations. Future work will involve trying to replace images of speakers with images of the concepts being spoken about. We will also add new layers to our architecture: a visualization layer, a sentiment analysis layer, and so on.

We are currently undertaking an evaluation of the system on components level. It is currently the only method to evaluate it since our system uses images and videos that are not from standardized datasets like TRECVID [35].

We also do user evaluations against each component, because it is important to know what the users feel about the end result. First reactions of the interviewed persons were that the system looks good and it is useful. They would even see themselves using it, if it would be open to public, since they consider it has the potential to be great in the space of personalized entertainment. We will perform a more detailed survey to assess the strength and weaknesses of our approach in the eyes of the users. Since today cinema viewers are accustomed to seeing complex narratives, we know that their expectations are high. For example, the ending from one of last year's most critically acclaimed movies: *Tinker, Tailor, Soldier, Spy* mixes scenes that happened during at least 4 periods of time (the distance between each being several years) in only 5 minutes (the focus is on the relationship between two characters, and the ascension of the third in light of the recent events), but still manages to keep us involved and provide a satisfactory and artsy conclusion in the same time. We are aware that reaching such an artistic mastery requires a long commitment from our side, but our first target is not *Tinker, Tailor, Soldier, Spy,* when it comes to generating an engaging narrative (as that movie is also a dramatic adaptation of a novel), but rather *The Autobiography of Nicolae Ceausescu*, a documentary which manages to

tell the dictator's story without any background narration. From our point of view reaching the artistic and information complexity level of this documentary (editing, scoring, narrative comprehension, etc.) will take several years, but the first steps toward this goal have already been made.

Assessing the success of our prototype system on the level of decision making strategies represents a complex process which is also likely to take several years. It involves developing future versions and showing the end product to potential customers in order to draft a commercial version in one day. A commercial version might easily function on multiple levels as we suggested. For example, if we would implement it for a touristic city portal, the system could generate both the ads that could be served to visitors (ads for ski during winter, and for hiking during summers), but also the presentations for various locations from the city and webisodes that show different aspects of living in that community.

Combining such systems with social media monitoring for example would allow us to extract more information from chained events like the Arab Spring. This will offer us an unprecedented level of access to information to real historical events, which is a thing all historians would like. Adding powerful prediction models to such a system would make it the ultimate crises prediction and intervention tool. We are only a few steps away from such systems as they are predicted in movies like *Minority Report*.

Bibliography

[1] A. Ahmed, L.P. Coelho, A. Arnold, J. Kangas, A.B. Sheikh, E. Xing, W. Cohen, R.F. Murphy, *Structured literature image finder: Parsing text and figures in biomedical literature*, Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 8 (2-3), 151-154, 2010

[2] R. Arnheim, *Art and Visual Perception: A Psychology of The Creative Eye*, 2nd Edition, Cambridge, 2004.

[3] R. Barthes, *Image, Music*, Text, Hill and Wang, 1977.

[4] J. Bertin, *Semiology of Graphics: Diagrams, Networks, Graphs*, 2nd Edition, ESRI Press, 2010.

[5] S. Bocconi, F. Nack, L. Hardman, *Automatic generation of matter-of-opinion video documentaries*, Journal of Web Semantics, 6 (2), 139-150, 2008.

[6] W.S. Cleveland, *The Elements of Graphing Data,* 2nd Edition, Hobart Press, 1994

[7] C. Colombo, A. Del Bimbo, P. Pala, *Semantics in Visual Information Retrieval*, IEEE Multimedia, 6(3): 38-53, 1999.

[8] P. Dombrowski, *Ernst Haeckel's controversial visual rhetoric*, Technical Communication Quarterly 12: 303-319, 2003.

[9] E.H.Gombric, *Art and Illusion. A Study in the Psychology of Pictorial Representation*, Millennium Edition, Princeton University Press, 2000.

[10] J. Heer, M. Stone, *Color Naming Models for Color Selection, Image Editing and Pallette Design*, ACM CHI, May 5-10, 2012, Austin, Texas, USA.

[11] C.A. Hill, M. Helmers, *Defining Visual Rhetorics,* Lawrence Erlbaum Associates, London, 2004.

[12] A.Hubmann-Haidvogel, A.M.P. Brasoveanu, A. Scharl, M. Sabou, S. Gindl. *Visualizing Contextual and Dynamic Features of Micropost Streams,* In Proceedings of the WWW'12 Workshopon on 'Making Sense of Microposts'. Lyon, France, April 16, 2012, CEUR Workshop Proceedings Vol-838, 34 - 40, 2012.

[13] J. Hullman, N. Diakopoulos, *Vizualization Rhetoric: Framing Effects in Narrative Visualization,* IEEE TVCG, Vol. 17, No. 12, 2011, 2231-2240.

[14] J. Itten, *The Art of Color: The Subjective Experience and Objective Rationale of Color*, John Wiley, New York, 1973.

[15] Z.-Q. Jiang, W.-X. Zhou, D. Sornette, R. Woodard, K. Bastiaensen, P. Cauwels, *Bubble diagnosis and prediction of the 2005-2007 and 2008-2009 chinese stock market bubbles*, Journal of Economic Behavior & Organization 74 (3), 149-162, 2010.

[16] E. R. Kandel, *The Age of Insight: The Quest to Understand the Unconscious in Art, Mind, and Brain, from Vienna 1900 to the Present*, Random House, 2012

[17] G. Kress, T. van Leeuwen, *Reading Images*, Routledge, London, 2006.

[18] J. Machajdik, A. Hanburry, *Affective Image Classification using Features Inspired by Psychology and Art Theory*, Proceedings of the ACM Multimedia, October 25-29, 2010, Firenze, Italy, 83-92.

[19] P. Sarlin, D. Marghescu, *Visual Predictions of Current Crises: A Comparison of Self-Organizing Maps with Probit Models*, TUCS Techical Report, No 978, June 2010.

[20] E. Segel, J. Heer, *Narrative Visualization: Telling Stories with Data*, IEEE TVCG, Vol. 16, 2010, 2231-2240.

[21] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, *Content Based Image Retrieval at the End of the Early Years*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 22 (12), page 1349-1380, 2000.

[22] D. Sornette, *Dragon-kings, black swans and the prediction of crises*, International Journal of Terraspace Science and Engineering, 2 (1), pp. 1-18, 2009.

[23] S. Staab, A. Scherp, R. Arndt, R. Troncy, M. Gregorzek, C. Saathoff, S. Schenk, L.Hardman., *Semantic multimedia*, Reasoning Web, 4th International Summer School, Venice, Italy. Volume 5224 of LNCS, Springer, 125-170, 2008.

[24] R. Troncy, B. Huet, S. Schenk, *Multimedia Semantics.* Metadata, Analysis and Interaction, Wiley 2011.

[25] E. R. Tufte, *The Visual Display of Quantitative Information*, 2nd Edition, Graphics Press, 2001.

[26] E. R. Tufte, *Beautiful Evidence*, Graphics Press, 2006.

[27] T. van Leeuwen, *Introducing Social Semiotics*, Routledge, 2005.

[28] W.-N. Wang, Y.-L. Yu, *Image Emotional Semantic Query Based On Color Semantic Description*, Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005, p.4571-4576.

[29] L. Wilkinson, *The Grammar of Graphics*, Second Edition, Springer, 2005.

[30] Z.J. Zha, L. Yang, T. Mei, M. Wang, Z. Wang, T.S. Chua, X.S. Hua, *Visual Query Suggestion: Towards Capturing User Intent in Internet Image Search*, ACM TOMCCAP 6, Issue 3, 1-19, 2010.

[31] W. Yan, R. Rebib, R. Woodard, D. Sornette, *Detection of Crashes and Rebounds in Major Equity Markets*, http://adsabs.harvard.edu/abs/2011arXiv1108.0077Y, last accessed: 8.04.2012.

[32] http://mazamascience.com/WorkingWithData/?p=870, last accessed: 08.04.2012.

[33] http://www.informationisbeautifulawards.com/, last accessed: 08.04.2012.

[34] http://www.economicpredictions.org/who-predicted-the-financial-crisis.htm, last accessed: 08.04.2012.

[35] http://trecvid.nist.gov/, last accessed: 08.04.2012.

[36] http://www.mercer.com/articles/quality-of-living-survey-report-2011, last accessed: 12.07.2012.

[37] http://www.visualisingdata.com/, last accessed: 12.07.2012.

[38] http://www.informationisbeautiful.net/, last accessed: 12.07.2012.

[39] http://flowingdata.com/, last accessed: 12.07.2012.

[40] http://www.visualcomplexity.com/vc/, last accessed: 12.07.2012.

[41] http://visual.ly/, last accessed: 12.07.2012.

[42] http://eagereyes.org/, last accessed: 12.07.2012.

[43] http://blog.typekit.com/2012/06/12/designing-data/, last accessed: 12.07.2012.

[44] http://twitter.com/nytgraphics/, last accessed: 12.07.2012.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 617-631

Nondeterministic Algorithm for Breaking Diffie-Hellman Key Exchange using Self-Assembly of DNA Tiles

Zheng Cheng
Zhejiang University of Technology
288 Liuhe Road, Hangzhou, P.R. China
chengzhenhust@gmail.com

Abstract: The computation based on DNA tile self-assembly has been demonstrated to be scalable, which is consider as a promising technique for computation. In this work, I first show how the tile self-assembly process can be used for computing the modular multiplication by mainly constructing three small systems including addition system, subtraction system and comparing system which can also be parallely implemented the discrete logarithm problem in the finite field $GF(p)$. Then the nondeterministic algorithm is successfully performed to break Diffie-Hellman key exchange with the computation time complexity of $\Theta(p)$, and the probability of finding the successful solutions among many parallel executions is proved to be arbitrarily close to 1.
Keywords: Modular multiplication; Discrete logarithm; Nondeterministic; Diffie-Hellman; Key exchange; Self-assembly; DNA tiles

1 Introduction

Nature is a rich source for computing devices design. In recent years, computing models and algorithms inspired by biological systems are deeply investigated, e.g., membrane computing inspired by cells and DNA computing inspired by DNA molecules. Most of such computing models inspired by cells (or DNA molecules inside cells) are proved to be universal and computationally efficient [1, 2]. This work focuses on computing systems based on DNA molecules, especially the self-assembly of DNA tiles. Since Adleman demonstrated that the DNA recombination techniques can be used to solve combinational search problem [3], the field of DNA-based computing has developed very fast. DNA computing potentially provides a degree of parallelism and high density storage far beyond that of conventional silicon-based computers [4].

DNA tile self-assembly is a crucial process, by which the small objects can autonomously assemble into big complexes [5]. Seeman proposed DNA nanotechnology to make self-assembled nanostructures from DNA molecules [6]. Based on this landmark work, Winfree utilized a kind of DNA nanostructure called DX (double crossover) tile to realize a patterned lattice and the complex algorithmic pattern [7]. Winfree et al. demonstrated that two dimensional DNA self-assembly can be capable of Turing-universal computation [8]. Winfree and Eng proved that self-assembly of linear, hairpin and branched DNA molecules can be generated regular, bilinear and context-free languages respectively [9]. DNA tile algorithmic self-assembly can also be used to create crystals with patterns of binary counters [10, 11] and Sierpinski triangles [12] to implement arbitrary circuit [13]. However, those crystals are deterministic. Mao experimentally implemented the first algorithmic DNA tile self-assembly [14], where a logical computation (cumulative XOR) is performed. Brun proposed the application of DNA tile self-assembly in arithmetic [15]. DNA self-assembly is also used to cope with combinational NP-complete problems, such as solving the satisfiability problem [16] by using 2D DNA self-assembly tiles, nondeterministically factoring numbers [17], deciding a system of subset sum problem [18]. DNA self-assembly has potential applications in the cryptography. XOR computation on pairs of bits

can be used to execute a one-time pad cryptosystem, which provides theoretically unbreakable security [19].

To provide modern security features, the algorithms for public-key cryptosystems such as RSA [20], Diffie-Hellman key agreement [21], the digital signature algorithm [22] and systems based on elliptic curve cryptography [23] are widely used. All these algorithms have one thing in common: they all need operate the modular multiplication and exponent multiplication [24, 25]. In 1976, Diffie and Hellman [26] proposed the public-key distribution scheme based on the discrete logarithm problem in a finite field $GF(p)$. Chen [27] gave deterministic algorithm to break Diffie-Hellman key exchange and constructed the basic tiles in which the complex modular multiplication operation in decimal is contained, so it can't be easily executed based on the experiment of simple binary arithmetic and logical operations. Here I mainly propose the nondeterministic algorithm to solve this problem by the addition, subtraction and comparing operations for binary numbers, which can be performed easily in experiments. The computation time complexity of our algorithm is $\Theta(p)$, and the probability of finding the successful solutions among the many parallel executions is proved to be arbitrarily close to 1.

The rest of this paper is structured as follows: Section 2 will describe the tile self-assembly model. Section 3 gives the method of computing modular multiplication using DNA tile self-assembly. Section 4 shows the process of breaking Diffie-Hellman key exchange based on modular multiplication by self-assembling. The conclusion will summarize the contribution of our work.

2 Algorithmic DNA tile self-assembly

The abstract Tile Assembly Model [28] is a formal model of crystal growth which is designed to model self-assembly of molecules such as DNA. Rothemund and Winfree [29] defined the abstract tile assembly model, which provides a rigorous framework for analyzing algorithmic self-assembly. The tile assembly model extends the theory of Wang tilings [30] of the plane by adding a natural mechanism for growth.

For the tile self-assembly model, the assembly complexes take place by starting with the seed tile denoted as the basic tile type set, which can be produced the seed configuration S. For each tile, it can be represented by the binding domains $\sum$ which is a 4-tuple $\{\sigma_N, \sigma_E, \sigma_S, \sigma_W\} \in \sum^4$. Here, N, E, S, W is labeled as the direction of north, east, south and west respectively. The set of directions is a set of four functions from positions to positions denoted as $D = \{N, E, S, W\}$, i.e. $\mathbb{Z}^2$ to $\mathbb{Z}^2$ such that all positions (x, y), $N(x, y) = (x, y+1)$, $E(x, y) = (x+1, y)$, $S(x, y) = (x, y-1)$, $W(x, y) = (x-1, y)$. For a tile t, for $d \in D$, $bd_d(t)$ is used to denoted as the binding domain of tile t on d's side.

Given a tile system $\mathbb{S} = \{T, g, \tau\}$ which is designed preparedly, here the parameter T is a function $\mathbb{Z} \times \mathbb{Z} \to T$, which is a configuration of the tile self-assembly model. g is a strength function $\sum \times \sum \to \mathbb{R}$, which denotes the strength of the binding domains, and $\tau \in \mathbb{N}$ is the temperature. When the growth of process terminates, these can be produced a unique final configuration $\mathbb{S}$ based on the seed configuration S. S is also a function $\mathbb{Z}^2 \to \Gamma$, here Γ is a set of tiles which can be used to design the configuration of the tile self-assembly model. Here, I mainly use the abstract tile assembly model to break Diffie-Hellman key exchange which is shown in Figure 1. Intuitively, the model has tiles or squares that stick or don't stick together based on various binding domain including $\sigma_N, \sigma_E, \sigma_S, \sigma_W$ on their four sides.

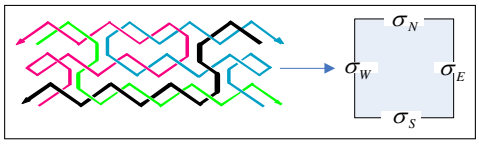


Figure 1: The structure of basic tile unit.

3 The algorithm for computing modular multiplication based on DNA tile self-assembly

Various techniques for speeding up modular multiplication have been reported in literature. Among them, two major approaches stand out: one is based on the interleaved modular multiplication algorithm where the multiplier is processed from the most significant position [31]. The other one is based on the Montgomery algorithm where the multiplier is processed from the least significant position [32]. The key that enables the linking of these two approaches is a new representation of residue classes modulo. The two methods account for most of the complexity in terms of time and resources needed. So this gives reason to search for dedicated solutions which compute the modular multiplications efficiently with minimum resources. Here, I use the method of interleaved modular multiplication based on DNA tile self-assembly. It takes advantage of these techniques and the ones that may eventually be devised to further boost speed.

For the integers X, Y, M with $0 \leq X, Y < M$, in order to get the result of $X*Y \mod M$, I use the algorithm which gives a pseudo code implementation of interleaved modular multiplication as follows:

(1)$P = 0$;
(2) for $(i = n-1;\ i \geq 0;\ i = i-1\ )\{$
(3) $P = 2 * P$;
(4) $P = P + x_i * Y$;
(5) if $(P \geq M)\ P = P - M$;
(6) if $(P \geq M)\ P = P - M;\}$

Here, n is the number of bits of X. x_i is denoted as the i-th bit of X. The idea of interleaved modular multiplication is very simple: the first operand is multiplied with the second operand bitwise and added to the intermediate result. The intermediate result is reduced with respect to the modulus. For this purpose at most two subtractions per iteration are required. The process is to interleave multiplication and reduction such that the intermediate results are kept as short as possible. For every $x_i (0 \leq i \leq n-2)$, there is a left shift, a partial product computation, an addition, and at most two subtractions. The partial product computation and left shift operations are easily performed by using an array of AND gates and wiring respectively. In each iteration of the loop, the estimation of the previous iteration can be added to the intermediate result.

In this section, I will introduce the algorithm for performing modular multiplication through constructing three small systems which are addition system, subtraction system and comparing system. Examples can be given to indicate how the tile assembly model performs the computations.

3.1 Addition system

This system will do addition operations between two positive integers. When the comparison result at the previous step is "<" which means the value of P is smaller than the modulus M, then

this system will make addition operations between two integers P and M. If $x_i = 0(0 \leq i \leq n-2)$, the result is the value of P which should be shifted one bit from right to left, intuitively the sum is the value of $2 * P$. If $x_i = 1(0 \leq i \leq n-2)$, the addition between P and Y is done, here P should be shifted one bit from right to left and Y does not need to shift one bit.

Theorem 3.1. Let the addition system be denoted as $\sum_+$={ab, abcd, a/bcd, a*bcd, a*bc, a*b, a, a*, #, a*bcd/, a/bc; a, b, c, d $\in \{0,1\}$}, and T_+ be the set of tiles as described in Fig.2(c). Let $g_+ = 1$, $\tau_+ = 2$, and S_+ be a seed configuration. Let n_P and n_Y be the sizes of P and Y in bits respectively. Let $n = \max(n_P, n_Y)$. If $n_P < n$, the number needs to be padded to be n bits long with extra 0 in the P's high bit, and if $n_Y < n$, the number needs to be padded to be n bits long with extra 0 in the Y's high bit. Then, there exists some $(x_0, y_0) \in \mathbb{Z}^2$, such that $S_+(x_0+1, y_0-1) = S0$, $S_+(x_0-n, y_0-1) = E0$, $S_+(x_0+1, y_0-0) = Add$; for all $i \in \{0, 1, \cdots, n\}$, $bd_N(S_+(x_0-(i-1)), y_0-1) = x_i p_i m_i y_i$, for all other positions (x, y),$(x, y) \notin S_+$. Then the tile system $\mathbb{S}_+$ produces a unique final configuration based on S_+ and can compute the sum of two input numbers by using $\Theta(1)$ distinct tiles in linear assembly time.

Proof. Consider this tile system $\sum_+$. Let Γ_+ be composed of the tiles $\{0*, \#, \#, null\}$, $\{a<, null, null, null\}$ with the label $S0$, $\{\#, null, null, null\}$ denoted as the tile $E0$, $\{a<, a, *, null\}$ represented as the tile Add, and the basic tile set T_+, here $a \in \{0, 1\}$. Let the seed configuration $S_+ : \mathbb{Z}^2 \to \Gamma_+$ be such that

$$\begin{cases} S_+(1,-1) = S0; \\ S_+(-n,-1) = E0, \ \ and \ \ S_+(1,0) = Add; \\ \forall i \in \{0, 1, \cdots, n\}, \ \ S_+(-i,-1) = x_i p_i m_i y_i; \\ For\ all\ other\ (x,y) \in \mathbb{Z}^2, \ \ S_+(x,y) = empty. \end{cases}$$

It is clear that there is only a single position where a tile may attach to S_+. And after that tile attaches there will only be a single position where a tile may attach. By induction, because $\forall t \in T_+$, the triplet $< bd_S(t), bd_E(t) >$ is unique, and because $\tau_+ = 2$, it follows that this tile system $\mathbb{S}_+$ produces a unique final configuration on S_+.

Fig.2(a) gives a sample seed configuration for adding two n-bit input numbers. Fig.2(b) shows the final configuration of the example for adding two integers $P = (001110)_2$ and $Y = (001101)_2$ with the result $(011011)_2$. The set of tiles which is described in Fig.2(c) shows the functions of the addition system has three functions.

The addition system has three functions. First, the value of P, M and Y are arranged in the seed configuration from the lowest bit to the highest bit, and the value of X is set from the highest to the lowest bit. Here, a, b, c, d, e, f, g, k $\in \{0, 1\}$. The initial value of P is set as 0. Of course, the highest bit of X denoted as x_{n-1} is 1, so the first addition operation only needs to pass the value of Y to P for the corresponding bits. The tile types with the red color can be seen in Fig.2(c). The value "a0cd" at the bottom of the tiles are denoted as X, P, M, Y respectively. The number 1 as the input, and output of the tile is the value of x_{n-1} which is assigned to 1. The numbers of bits of P and M are more than X and Y, so "bc" in the first tile of the tile types are represented as the higher bits of P and M respectively. More importantly, I use "a/// to label the highest bit of X which should be passed to the lowest bit, then $x_i = 0$ or 1 $(0 \leq i \leq n-2)$ can be passed to the addition system to determine whether the value of Y should be added with P together. "e/// is the value passed from the highest bit, so the label of "a// will be passed to the addition system at next step.

Second, for the next addition, if $x_i = 0(0 \leq i \leq n-2)$, the lowest bit of P is 0, and the value of P should be shifted one bit from right to left, then the result of the addition operation is $2 * P$. At the same time, $x_i = 0(0 \leq i \leq n-2)$ should be passed to the higher bit of Y, and the bits of Y and M are passed to the upper layer. So here, there is no carry bit for the addition at this step. If $x_i = 1(0 \leq i \leq n-2)$, the lowest bit of P is the corresponding lowest bit of Y,

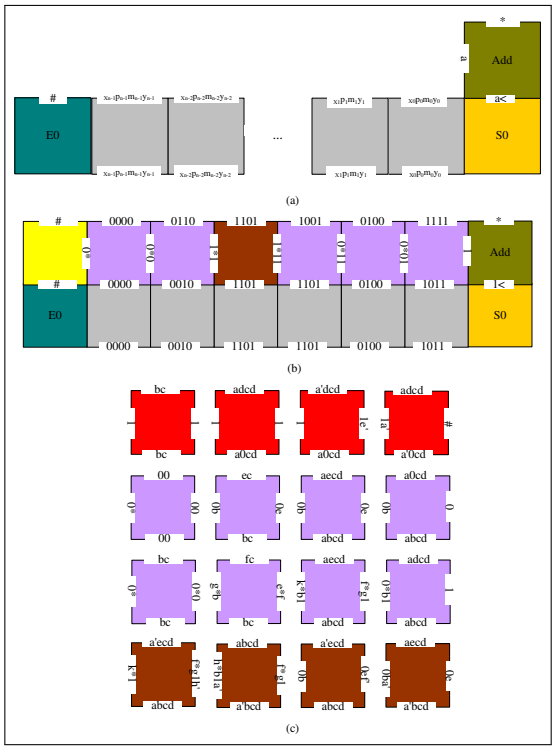


Figure 2: (a) A sample seed configuration for adding two n-bit input numbers. (b) The final configuration of the example for adding two integers $P = (001110)_2$ and $Y = (001101)_2$ with the result $(011011)_2$. (c) The basic tile types of the addition system.

then the addition is done between P and Y, here P also should be shifted one bit from right to left. So if the carry bit from the lower bit is "f*", and the shifted bit from P is "g", thus the addition result "e〃 and the carry bit "k*" can be generated at this step, synchronously the bit of P and Y labeled as "b" and 1 respectively should be passed to the higher bit. The tile types with lilac are shown this function of the addition system.

Third, this system can determine whether and what the bit of X is passed to the next addition system. If the lowest bit of X has been completed the addition operation, then it only should be passed to the final result. Otherwise, the value of x_i should be passed to the lower bit x_{i-1} which will be done the next round addition. So if the input on the right of the tiles includes the label "f〃〃, it can determine the value of x_{i-1} denoted as "a〃 will be passed to the next addition system.

3.2 Subtraction system

In this section, I will describe a system that can make subtraction between two positive integers. When the comparison result at the previous step is ">" or "=" which means the value of P is bigger than or equal to the modulus M, then this system will make the subtraction between the two integers P and M.

Theorem 3.2. Let the subtraction system $\sum_{-}$={ab, abcd, a*, *; a, b, c, d $\in \{0,1\}$ }, and T_{-} be the set of tiles as described in Fig.3(c). Let $g_{-} = 1$, $\tau_{-} = 2$, and S_{-} be a seed configuration. Let n_P and n_M be the sizes of P and M in bits respectively. Let $n = \max(n_P, n_M)$. If $n_P < n$, the number needs to be padded to be n bits long with extra 0 in the P's high bit, and if $n_M < n$, the number needs to be padded to be n bits long with extra 0 in the M's high bit. Then, there exists some $(x_0, y_0) \in \mathbb{Z}^2$, such that $S_{-}(x_0 + 1, y_0 - 1) = S0$, $S_{-}(x_0 - n, y_0 - 1) = E0$,

$S_-(x_0+1, y_0-0) = Sub$; for all $i \in \{0, 1, \cdots, n\}$, $bd_N(S_-(x_0-(i-1)), y_0-1) = x_i p_i m_i y_i$, for all other positions (x, y),$(x, y) \notin S_-$. Then the tile system $\mathbb{S}_-$ produces a unique final configuration based on S_- and can compute the difference of two input numbers with $\Theta(1)$ distinct tiles in linear time.

Proof. Consider the tile system $\sum_-$. Let Γ_- contain the tiles $\{0^*, \#, \#, \text{null}\}$, $\{>, null, null, null\}$ denoted as $S0$, $\{\#, null, null, null\}$ represented as the tile $E0$, $\{*, *, >, null\}\}$ labeled as the tile Sub, and the basic tile set T_-. Let the seed configuration $S_- : \mathbb{Z}^2 \to \Gamma_-$ be such that

$$\begin{cases} S_-(1, -1) = S0; \\ S_-(-n, -1) = E0, \ \ and \ \ S_-(1, 0) = Sub; \\ \forall i \in \{0, 1, \cdots, n\}, \ \ S_-(-i, -1) = x_i p_i m_i y_i; \\ For\ all\ other\ (x, y) \in \mathbb{Z}^2, \ \ S_-(x, y) = empty. \end{cases}$$

It is obvious that there is only a single position where a tile may attach to S_-. And after that tile attaches there will only be a single position where a tile may attach. By induction, because $\forall t \in T_-$, the triplet $< bd_S(t), bd_E(t) >$ is unique, and because $\tau_- = 2$, it follows that this tile system $\mathbb{S}_-$ produces a unique final configuration on S_-.

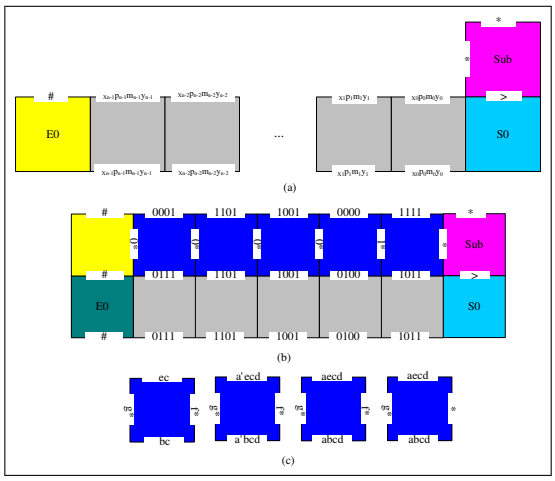


Figure 3: (a) A sample seed configuration for subtraction operation for two n-bit input numbers. (b) The final configuration for the example of subtracting two integers $P = (11010)_2$ and $M = (10001)_2$ with the result $(01001)_2$. (c) The basic tile types of this system.

Here, the value of "a, b, c, d" at the bottom of the tile is denoted as X, P, M, Y respectively, and a, b, c, d, e, f, g $\in \{0, 1\}$. The subtraction operation begins from the lowest bit to the highest bit with the label "*". "a′" is the value of X with the label which is passed to the addition system. The representation "f*" is the borrow bit from the lower bit and "g*" is the borrow bit which is generated at this step. The value of "e" is the result which can be computed by the sum of "b" and "c" minus "f". Fig.3(a) is a sample seed configuration for subtraction operation for two n-bit input numbers. Fig.3(b) shows the final configuration for the example of subtracting two integers $P = (11010)_2$ and $M = (10001)_2$ with the result $(01001)_2$, and (c) gives the basic tile types of the subtraction system.

3.3 Comparing system

This system is to check the relationship for given input string of binary bits which are represented as the addition or subtraction result P at each step and the modulus M respectively.

At the same time, the relationship can determine the next operation should be made addition or subtraction. The system compares two input numbers bit-by-bit from the highest bit to the lowest bit until the relationship which is less than (<), greater than (>) or equal (=) is determined. Here, I will describe the comparing system which uses $\Theta(1)$ distinct tiles in linear time to compare the size of two input numbers.

Theorem 3.3. Let the comparing system $\sum_R$={ab, abcd, a/bcd, >, <, =, a>, a<, a=, #; a, b, c, d$\in \{0,1\}$}, and T be the set of tiles as described in Fig.4(c). Let $g_R = 1$, $\tau_R = 2$, and S_R be a seed configuration. Let n_P and n_M be the sizes of P and M in bits respectively. Let $n = \max(n_P, n_M)$. If $n_P < n$, the number needs to be padded to be n bits long with extra 0 in the P's high bit, and if $n_M < n$, the number needs to be padded to be n bits long with extra 0 in the M's high bit. Then, there exists some $(x_0, y_0) \in \mathbb{Z}^2$, such that $S_R(x_0 - n, y_0 - 1) = S0$, $S_R(x_0 + 1, y_0 - 1) = E0$, $S_R(x_0 - n, y_0 - 0) = Com$; for all $i \in \{0, 1, \cdots, n\}$, $bd_N(S_R(x_0 - (i-1)), y_0 - 1) = x_i p_i m_i y_i$, for all other positions (x, y),$(x, y) \notin S_R$. Then the tile system $\mathbb{S}_R$ produces a unique final configuration based on S_R and can give the relationship of two input numbers.

Proof. The proof is referred to Theorem 3.1 and Theorem 3.2, so I don't restate it here.

This system mainly makes the comparison between two non-negative integers which are used in the binary form. First, for two given integers, the numbers of bits are the same. Second, the comparison begins from the most significant bit to the rightmost bit. When the $(i+1)$-th bit of the addition or subtraction result P is smaller then the modulus M, then the tile is labeled as "<", which also should be considered with the result of the i-th bit which is "<", ">" or "=", then the result which is "<" can be obtained and should be passed to the $(i-1)$-th bit. On the contrary, the final result is ">" also should be passed to the $(i-1)$-th bit no matter what the relationship of the i-th bit is. If the $(i+1)$-th bit of the addition or subtraction result P is equal to the modulus M, then the tile is labeled as "=", which also should be considered with the result of the i-th bit which are "<", ">" or "=", then the result which is "<", ">" or "=" respectively can be obtained and should be passed to the $(i-1)$-th bit. Furthermore, if the comparison result of the two integers is "<", the next operation would turn to the addition computation with $x_i = 0$ or 1. If the comparison result is ">" or "=", the next operation would turn to the subtraction computation between P and M. When the comparison result is "<" with the label of the position for the lowest bit of X, the modular multiplication is completed at this step.

Fig.4(a) is a sample seed configuration for two n-bit input numbers. Fig.4(b) shows the final configuration for the example of comparing two integers $P = (1010)_2$ and $M = (0101)_2$ with the result ">". The tiles used in this system are as follows in Fig.4(c). In each tile, the value of "a, b, c, d" is denoted as X, P, M, Y respectively.

3.4 An example

Now I take an example to verify the validity of the algorithm based on DNA tile self-assembly model introduced above. Here, for the integers X, Y, M with $0 \le X, Y < M$, I suppose $X = 11$, $Y = 13$, $M = 17$. The modular multiplication $X * Y \mod M$ is computed as the following steps:

First, X, Y, M can be represented as $X = 11 = (1011)_2$, $Y = 13 = (1101)_2$, $M = 17 = (10001)_2$ respectively. The initial value of P is set as 0 and arranged from the lowest bit to the highest bit together with Y, M. On the contrary, the bits of X are from the highest bit to the lowest bit.

Second, according to the method introduced above, I need construct the basic tile types in each of the three small systems and they are the same as the tiles described above. When all the tiles and the seed configuration are prepared, they are put together into the reaction buffer.

Figure 4: (a) A sample seed configuration for two n-bit input numbers. (b) The final configuration for the example of comparing two integers $P = (1010)_2$ and $M = (0101)_2$ with the result ">". (c) The basic tile types of the comparing system.

According to the mechanism of algorithmic DNA tile self-assembly through Watson-Crick base pairing, the self-assemble process starts at the same time with the connector tiles, so the final stage can be seen in Fig.5.

The process of the three small systems performing is shown. Here, because the highest bit of X which is denoted as $x_3 = 1$, the value of P is equal to Y, so the comparing system can check that P is smaller than M. The next addition operation begins with $x_2 = 0$, so the value of P at the previous step shifted one bit from right to left can be assigned to P at this step, which can be represented as $P = (011010)_2$, then the comparing system determines it is more than M, so the next operation is the subtraction operation between P and M, and the subtraction result should be assigned to P which is $(001001)_2$. When $x_1 = 0$, the addition system can make the sum between Y and P, here P should be shifted one bit from right to left. So the addition result also should be assigned to P which is $(011111)_2$, and it is bigger than M by the comparing system, then the subtraction result is $(001110)_2$. The computation stops until x_0 attaches to the addition system with the result of P which is smaller than M. The final result of P is $(000111)_2$ which can be obtained by repeated computations using the addition, subtraction and comparing system.

4 The nondeterministic algorithm for breaking Diffie-Hellman key exchange using self-assembly of DNA tiles

On basis of the algorithm for computing modular multiplication, I will first give the method of performing the discrete logarithm problem in the finite field $GF(p)$, then the nondeterministic algorithm for breaking Diffie-Hellman key exchange is successfully proposed.

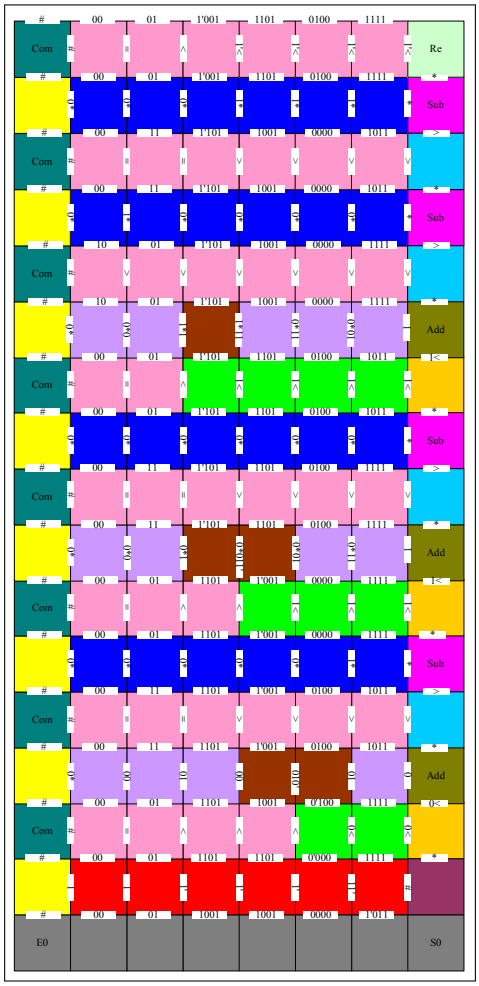


Figure 5: An example of computing modular multiplication $1011 * 1101 \bmod (10001)$.

4.1 The nondeterministic algorithm for solving the discrete logarithm problem

A group G is cyclic if there is an element $\alpha \in G$, such that for each $b \in G$, there is an integer i with $b = \alpha^i$. Such that an element α is called a generator of G. Now given a cyclic group G of order p, a primitive-root g and the element y of the group, the problem is to find an integer x such that $y \equiv g^x (\mathrm{mod}\ p)$ with $1 \leq x \leq p-1$, and this problem is called discrete logarithm problem in the finite field $GF(p)$. Considering the equation $y = g^x (\mathrm{mod}\ p)$, for given y, g and p, it is very difficult to compute the value of x. So here I solve the discrete logarithm problem by implementing the algorithm for computing the modular multiplication based on DNA tile self-assembly model.

In the process of implementing the tile self-assembly systems, many assemblies happen in parallel by creating billions of billions of copies of the participating DNA tiles, so this is simulated by an exponential number of DNA assemblies which can be converted into the space occupied by the DNA molecules, thus I expect that the procedure of computing y will run in parallel on all possible value of x under the condition $1 \leq x \leq p-1$. Then I can compare the computation result with the given value of y, finally the result of x can be read by the biological operations described in Section 3.

So first, I give the algorithm for computing modular exponentiation based on the modular multiplication introduced above. There are two differences between the two computations. Firstly, the modular exponentiation is actually a series of modular multiplications, so in the process of designing the seed configuration, it only needs to give different labels of the lowest bit of X which can be used to distinguish the value of x. Here, I consider one of the value of X as Y. Secondly, given the value of i, after the computation of $g^i (\mathrm{mod}\ p)$ is completed, there are some tiles that can convert the result of $g^i (\mathrm{mod}\ p)$ into the new value of X, and the value of Y doesn't need to change, then the assembly complexes can be used to implement the computation of $g^{i+1} (\mathrm{mod}\ p)$. Therefore I don't give too much explanation for the process of computing modular exponentiation.

On this basis, I use the nondeterministic algorithm to solve the discrete logarithm problem in the finite field $GF(p)$. It can nondeterministically guess the value of x so that a parallel implementation can be executed. I need to construct different labels to denote the value of x, which can determine the computation of modular exponentiation. Here, an example in $GF(11)$ is taken. Suppose the equation $y = 7^x (\mathrm{mod}\ 11)$, for given $y = 2$, I can get the value of x which is 3 as following steps and the final stage can be shown in Fig.6.

Step 1: The binary forms of g, p is obtained respectively, here I consider $Y = X$. The initial value of the integer P is set as 0 and arranged from the lowest bit to the highest bit together with Y, p. On the contrary, the bits of X are from the highest bit to the lowest bit. Then the basic tile types and the seed configuration are constructed and put together into the reaction buffer. According to the nondeterministic algorithm, there are many choices at the first position of the seed configuration, it can nondeterministically guess the value of x under the condition $1 \leq x \leq p-1$. In this example, the tile containing the label "11" which is denoted as the value of $x = 3$ attaches the seed configuration. So the modular exponentiation $y = 7^3 (\mathrm{mod}\ 11)$ can be parallely performed by the three small systems including addition system, subtraction system and comparing system, the assembly complexes can grow, therefore I can obtain the solutions of the problem.

Step 2: Once the self-assembly has occurred, it is necessary to extract the answer. An estimate of the length of the reporter strands can be obtained by annealing the reporter strands with the component strands, so I can extract the result strands of different lengths representing the output tiles which run through the value of y.

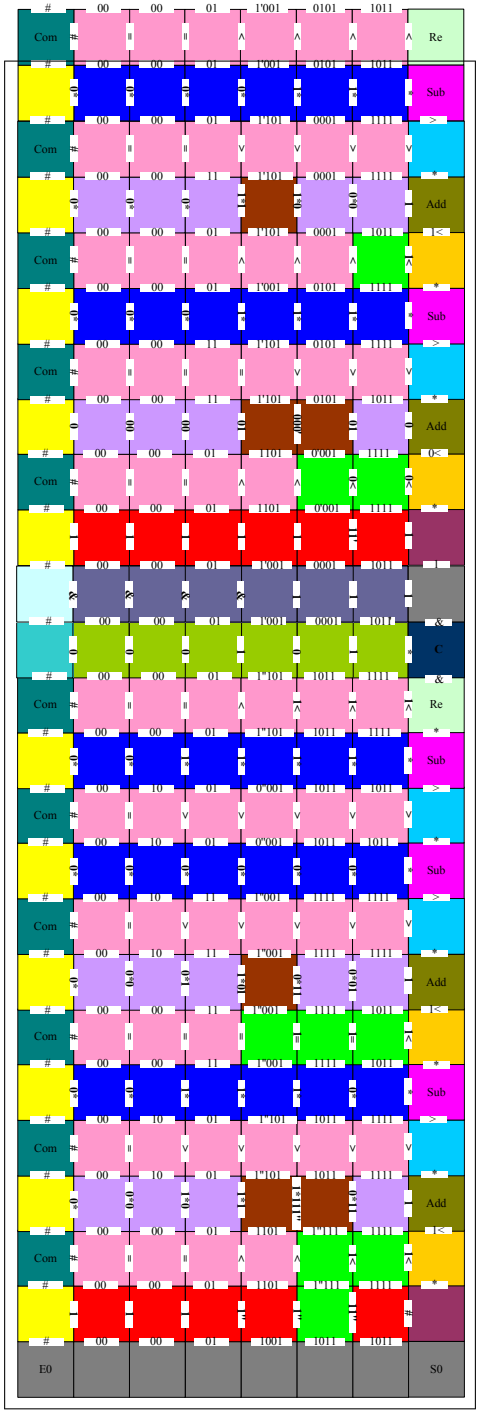


Figure 6: An example of computing modular exponentiation $y = 7^3 (\mathrm{mod}\ 11)$.

Step 3: By comparing the computation result of y for every feasible value of x with given $y = 2$, so I can obtain the solution of the discrete logarithm problem.

4.2 The nondeterministic algorithm for breaking Diffie-Hellman key exchange

Diffie-Hellman method is used for symmetric key exchange between entities. The steps of the algorithm are as follows:

(1) Key generation

(a) Generate a large random prime integer p and a generator α of the multiplicative group U_p. The set U_n is the multiplicative group of Z_n with order $\varphi(n)$, which is defined as follows:

$U_n = \{[a] \in Z_n : GCD(a, n) = 1\}$

(b) Select two random integers a and b, $1 \leq a,\ b \leq p-1$ for entity A and B, then compute $x = \alpha^a (\mathrm{mod}\ p)$ and $y = \alpha^b (\mathrm{mod}\ p)$.

(c) Broadcast the public key of A and B where A's public key is (α, x) and B's public key is (α, y).

(2) A computes the symmetric key $k_a = y^a (\mathrm{mod}\, p)$ and B computes the symmetric key $k_b = x^b (\mathrm{mod}\ p)$. A and B will use the Diffie-Hellman method to exchange symmetric keys.

In order to get the symmetric key K which is equal to k_a or k_b between the users A and B, I can give the nondeterministic algorithm for breaking the Diffie-Hellman key exchange as follows:

First, I can get value of the secret key of user A denoted as a according to the public key (α, x) and the primer p. As a matter of fact, this process can be performed by the method of solving the discrete logarithm problem in the finite field $GF(p)$ which is introduced above. I can design the all needed tiles in the computation, including the basic types of tiles, boundary tiles and the seed configuration by the binary form of the integer x, α and p. When all kinds of the tiles and the seed configuration are prepared, I puts them together into the reaction buffer, there are many choices at the first position of the seed configuration, it can nondeterministic guess the value of a under the condition $1 \leq a \leq p-1$. So the modular exponentiation $\alpha^a (\mathrm{mod}\, p)$ can be parallely performed by the three small systems, the assembly complexes can grow, therefore I can obtain the solutions of the problem. Then I can read the result of the process of the DNA tile growth using a combinational of PCR and gel electrophoresis.

Second, I can make the modular multiplication $K = y^a (\mathrm{mod}\ p)$ together with the value of a obtained at the first step and the public key (α, y) of the user B.

For example, suppose the multiplicative group U_{11} and the generator $\alpha = 7$. The user A selects one random integers a as his own secret key, at the same time, the users A and B broadcast their public key $x = 2$ and $y = 3$ respectively.

Considering the equation $x = 7^a (\mathrm{mod}\ 11)$, I can get the solution of the discrete logarithm problem in the finite field $GF(11)$ which is the secret key of the user A by the method described above, and the result of a is equal to 3. So the common key can be computed as $K = y^a (\mathrm{mod}\ p) = 3^3 (\mathrm{mod} 11) = 5$.

4.3 Complexity and probability analysis

Considering the nondeterministic algorithm for breaking Diffie-Hellman key exchange based on self-assembly of DNA tiles, the complexity of this algorithm can be computed in terms of computation time and the number of distinct tiles required. Generally, for the equation $y = g^x (\mathrm{mod}\ p)$, here, $1 \leq x \leq p-1$, suppose the number of the bits of g be k, and for the given problem, the value of k is a constant.

For a round modular multiplication, there is one addition operation and at most two subtraction operations followed by the comparison operation respectively, so the upper bound of the computation time T which is the number of assembled steps is $x(6k+2)+2 = \Theta(p)$.

Finally, these distinct tile types include the input boundary tiles and computation boundary tiles. As the basic tiles described in Section 3, the tile types are all $\Theta(1)$, so I can obtain the basic tile types are $\Theta(1)$ for a round modular multiplication, then I also consider the labels to distinguish the number of the rounds which can be determined by the size of $x(1 \leq x \leq p-1)$, therefore the tile types needed for this algorithm is $\Theta(p)$.

Now, I give the probability analysis for this algorithm.

Theorem 4.1. Let each tile that may attach to a configuration at a certain position attach there with a uniform probability distribution. For all integers $x(1 \leq x \leq p-1)$, n_p is denoted as the number of the bits of the primer p, for given integers g and y, then the probability of assembling a particular final configuration which attaches the value of x such that $y = g^x (\mod p)$ is at least $(\frac{1}{p-1})^{n_p}$.

Proof. Now I calculate the probabilities of each tile attaching in its proper position and then multiply those probabilities together to get the overall probability of a final configuration.

On one hand, there are $(p-1)$ possible tiles that may attach to the positions which represent the bits of the primer P in the seed configuration, and only one is correct, so the probability of it attaching is $P_i = \frac{1}{p-1}(1 \leq i \leq n_p)$. On the other hand, for the next two positions, there is only one tile that may attach, so the probability of the last two tiles are both $P_{n_p+1} = P_{n_p+2} = 1$.

The overall probability of a specific final configuration can be computed as the following formulas:

$\prod_{i=1}^{n_p+2} P_i = (\frac{1}{p-1})^{n_p}$

This is the probability of a nondeterministic assembly successfully identifying the solution. I use the nondeterministic computation to perform the function $y = g^x (\mod p)$, and construct three small systems to define the whole tile systems that use $\Theta(p)$ input tile types, and assemble in $\Theta(p)$ steps with probability of each assembly finding the solution at least $(\frac{1}{p-1})^{n_p}$. Therefore a parallel implementation of this tile system such as the execution of the DNA tile self-assembly model in [12], with $(p-1)^{n_p}$ seeds has at least a $(1-e^{-1})$ chance of finding the secret key of the user A or B, and one with $100(p-1)^{n_p}$ seeds has at least a $(1-e^{-100})$ chance. The probability of finding the successful solutions among the many parallel executions can be proved to be arbitrarily close to 1.

5 Summary and Conclusions

DNA tile self-assembly is looked forward to many applications in different fields. In this paper, I show how the DNA self-assembly process can be used for breaking Diffie-Hellman key exchange based on the discrete logarithm problem in the finite field $GF(p)$ by computing the modular multiplication. The advantage of our method is that once the initial strands are constructed, each operation can compute fast parallelly through the process of DNA self-assembly without any participation of manpower, thus the algorithm is proposed which can be successfully solved the modular multiplication, and then can be used to break Diffie-Hellman key exchange with the computation time complexity of $\Theta(p)$, and the probability of finding the successful solutions among the many parallel executions is proved to be arbitrarily close to 1.

Acknowledgments

This work was supported by the Research Project of Department of Education of Zhejiang Province (Y201120124) and the National Natural Science Foundation of China (Grant Nos. 60703047, 60803113, 60903105).

Bibliography

[1] L. Pan and C. Martin-Vide, Solving multidimensional 0−1 knapsack problem by P systems with input and active membranes, *Journal of Parallel and Distributed Computing*, Vol. 65, pp. 1578-1584, 2005.

[2] L. Pan and M. J. P¨Śrez-Jim¨Śnez, Computational complexity of tissue-like P systems, *Journal of Complexity*, Vol. 26, pp. 296-315, 2010.

[3] L.M. Adleman, Molecular computation of solutions to combinatorial problems, *Science*, Vol. 266ŁŹ pp. 1021-1024, 1994.

[4] L.Q. Pan, G.W. Liu , J. Xu, Solid phase based DNA solution of the coloring problem, *Progress in Natural Science*. vol. 14, pp. 104-107, 2004.

[5] A. Carbone, N.C. Seeman, Molecular tiling and DNA Self-assembly, *Springer-Verlag Berlin Heidelberg*, Vol. 2950, pp. 61-83, 2004.

[6] N.C. Seeman, DNA nanotechnology: novel DNA constructions, Annu. Rev. Biophy. *Biomol. Struct.*, Vol. 27, pp. 225-248, 1998.

[7] C. Mao, W. Sun, N.C. Seeman, Designed two dimensional DNA Holliday junction arrays visualized by atomic force microscopy, *J. Am. Chem. Soc.*, Vol. 121, pp. 5437-5443, 1999.

[8] E. Winfree, On the computational power of DNA annealing and ligation, *DNA Based Computers*, pp. 99-221, 1996.

[9] T. Eng, Linear self-assembly with hairpins generates the equivalent of linear context-free grammars, *3rd DIMACS Meeting on DNA Based Computers*, Univ. of Penn., 1997.

[10] R. Barish, P.W. Rothemund, E. Winfree, Two computational primitives for algorithmic self-assembly: copying and counting, *Nano Letters*, Vol. 12, pp. 2586-2592, 2005.

[11] P. M. Espane/s, A. Goel, Toward minimum size self-assembled counters, *Springer Science Business Media B.V.*, 2008.

[12] P.W. Rothemund, N. Papadakis, E. Winfree, Algorithmic self-assembly of DNA Sierpinski triangles, *PLoS Biology*, Vol. 12, pp. 2041-2053, 2004.

[13] M. Cook, P.W. Rothemund, E. Winfree, Self assembled circuit patterns, *DNA*, pp. 91-107, 2004.

[14] C. Mao, T.H. LaBean, J.H. Reif, Logical computation using algorithmic self-assembly of DNA triple-crossover molecules, *Nature*, Vol. 407, pp. 493-496, 2000.

[15] Y. Brun, Arithmetic computation in the tile assembly model: addition and multiplication, *Theoretical Computer Science*, Vol. 378, pp. 17-31, 2006.

[16] G.L. Michail, T.H. LaBean, 2D DNA self-assembly for satisfiability, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, Vol. 44, pp. 139-152, 1999.

[17] Y. Brun, Nondeterministic polynomial time factoring in the tile assembly model, *Theoretical Computer Science*, Vol. 395, pp. 3-23, 2008.

[18] Y. Brun, Solving NP-complete problems in the tile assembly model, *Theoretical Computer Science*, Vol. 395, pp. 31-36, 2008.

[19] A. Gehani, T.H. LaBean, J.H. Reif, In DNA Based Computers: Proceedings of a DIMACS Workshop, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 1999.

[20] R.L. Rivest, A. Shamir, L.M. Adleman, A method for obtaining digital signatures and public-key cryptosystems, *Commun. ACM*, Vol. 21, pp. 120-126, 1978.

[21] I.R. Jeong, J.O. Kwon, D.H. Lee, A Diffie-Hellman key exchange protocol without random oracles, *Springer-Verlag Berlin Heidelberg*, Vol. 4301, pp. 37-54, 2006.

[22] ANSI X9.30. Public Key Cryptography for the Financial Services Industry: Part 1: The Digital Signature Algorithm (DSA), *American National Standard Institute*, American Bankers Association, 1997.

[23] N. Koblitz. Elliptic curve cryptosystem, *Math. Comp.*, Vol. 48, pp. 203-209, 1987.

[24] G. Blakley, A computer algorithm for calculating the product $A * B \mod M$, *IEEE Transactions on Computers*, Vol. C-32, pp. 497-500, 1983.

[25] T. Acar, B.S. Kaliski, C. Koc, Analyzing and computing Montgomery multiplication algorithms, *IEEE micro*, Vol. 16, pp. 26-33, 1996.

[26] W. Diffie, M.E. Hellman, New directions in cryptography, *IEEE Transactions on Information Theory*, Vol. 22, pp. 644-654, 1976.

[27] Z.H. Chen, Breaking the Diffie-Hellman key exchange algorithm in the tile assembly model, *Chinese Journal of Computers*, Vol. 31, pp. 2116-2122, 2008.

[28] E. Winfree, Algorithmic self-assembly of DNA, *Ph.D. Thesis*, Caltech, Pasadena, CA, June 1998.

[29] P.W. Rothemund, E. Winfree, The program-size complexity of self-assembled squares, *ACM Symposium on Theory of Computing (STOC)*, pp. 459-468, 2001.

[30] H. Wang, Proving theorems by pattern recognition, *I. Bell System Technical Journal*, Vol. 40, pp. 1-42, 1961.

[31] E. Marcelo Kaihara, T. Naofumi, A hardware algorithm for modular multiplication/division, *IEEE Transactions on Computers*, Vol. 54, pp. 12-21, 2005.

[32] P.L. Montgomery, Modular multiplication without trial division, *Mathematics of Computation*, Vol. 44, pp. 519-521, 1985.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 632-644

Formation Control of Multiple Agents with Preserving Connectivity and its Application to Gradient Climbing

K.D. Do

K.D. Do
Department of Mechanical Engineering,
Curtin University of Technology,
Perth, WA 6845, Australia
E-mail: duc@curtin.edu.au

Abstract: A design of cooperative controllers that force a group of N mobile agents with limited communication ranges to perform a desired formation is presented. The proposed formation control system also preserves initial communication connectivity and guarantees no collisions between the agents. The formation control design is based on smooth step functions, potential functions, and the Lyapunov direct method. The proposed formation control system is applied to solve a gradient climbing problem where the gradient average of a distributed field is estimated over a bounded region using the field measurement by the agents.
Keywords: Formation control, collision avoidance, gradient climbing.

1 Introduction

Formation control involves controlling positions of a group of agents such that they perform desired tasks such as optimizing objective functions from measurements taken by each agent, and stabilization/tracking desired locations relative to reference point(s). Various methods have been proposed for formation control of multiple agents.

Here, three popular methods are briefly mentioned. The leader-follower method (e.g., [1], [2]) uses several agents as leaders and others as followers. This method is easy to understand and ensures formation maintenance if the leaders are disturbed. However, the desired formation cannot be maintained if followers are perturbed unless a formation feedback is implemented, [3]. The behavioral method (e.g., [4], [5]), where each agent locally reacts to actions of its neighbors, is suitable for decentralized control but is difficult in control design and stability analysis since group behavior cannot explicitly be defined. The virtual structure method (e.g., [6], [7]) treats all agents as a single entity. This method is amenable to mathematical analysis but is difficult to deal with time-varying formation structure. Research works on formation control usually utilize one or more of the above methods in a centralized or a decentralized manner. Centralized strategies (e.g., [8], [3]) use a single controller that generates collision free trajectories in the workspace. These strategies guarantee a complete solution but require high computational power and are not robust. Decentralized schemes (e.g., [9], [10], [7]) require less computational effort but have difficulties in controlling critical points, especially when collision avoidance between the agents is a must.

The control design in the above works did not put hard constraints on the controlled outputs except for those papers considered the problem of collision avoidance. Without hard constraints on controlled outputs, overshoot might result in loss of initial communication between agents due to limited communication between the agents. Hard constraints on the controlled outputs were applied to design cooperative controllers for mobile agents to preserve initial communication. These constraints on the controlled outputs were obtained through barrier Lyapunov or potential functions using non-trivial bump functions or switching control strategies in [11] for the agreement problem, [12] for the centralized approach, and [13] for the swarm aggregation.

This paper contributes two main folds. The first one is a design of smooth and bounded cooperative controllers for a group of mobile agents to perform a desired formation task. The desired formation task includes collision avoidance and communication connectivity preservation between the agents, time-varying desired formation shape, and stabilization of the desired formation shape at any reference trajectories with bounded time derivatives. The second contribution is an algorithm for estimating gradient average of a distributed field over a region in two dimensional space. This algorithm uses only the field measurement on the boundary of a region, over which the gradient average is to be estimated. The two contributions are then combined to provide an effective gradient climbing system for a group of mobile agents by allowing the reference trajectory for each agent generated based on the gradient average.

2 Preliminaries and Formation Control Objective

2.1 Smooth step function

This section presents a construction of a smooth step function. The smooth step function is to be embedded into a potential function to avoid discontinuities in the control law due to the agents' communication limitation in solving collision avoidance and connectivity preserving problems.

Definition 1. A scalar function $h(x,a,b,c)$ is said to be a smooth step function if it possesses the following properties

where $x \in \mathbb{R}$, $h\prime(x,a,b,c) = \frac{\partial h(x,a,b,c)}{\partial x}$, $h\prime\prime(x,a,b,c) = \frac{\partial^2 h(x,a,b,c)}{\partial x^2}$, a and b are constants such that $a < b$, and c is a positive constant.

Lemma 2. *Let the scalar function $h(x,a,b,c)$ be defined as*

$$h(x,a,b,c) = \frac{f(\tau)}{f(\tau) + cf(1-\tau)} \quad with \quad \tau = \frac{x-a}{b-a}, \tag{1}$$

where

$$f(\tau) = 0 \quad if \quad \tau \leq 0 \quad and \quad f(\tau) = e^{-\frac{1}{\tau}} \quad if \quad \tau > 0, \tag{2}$$

with a and b being constants such that $a < b$, and c being a positive constant. Then the function $h(x,a,b,c)$ is a smooth step function.

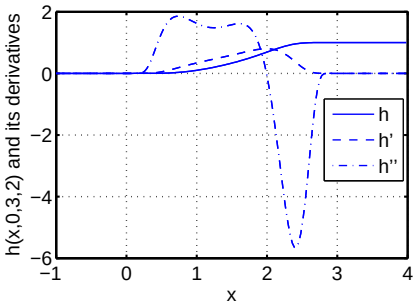


Figure 1: A smooth step function and its first and second derivatives.

Proof. Proof of this lemma follows the same lines as the proof of Lemma 1 in [7]. An illustration of a smooth step function ($a = 0, b = 3, c = 2$) is given in Figure 1.

2.2 Problem statement

Agent dynamics

We assume that the agent i has the dynamics:

$$\dot{\boldsymbol{q}}_i = \boldsymbol{u}_i,\ i \in \mathbb{N}, \tag{3}$$

where $\mathbb{N}$ is the set of all agents in the group, $\boldsymbol{u}_i \in \mathbb{R}^n$ is the control input vector and $\boldsymbol{q}_i \in \mathbb{R}^n$ the position vector of the agent i.

Formation control objective

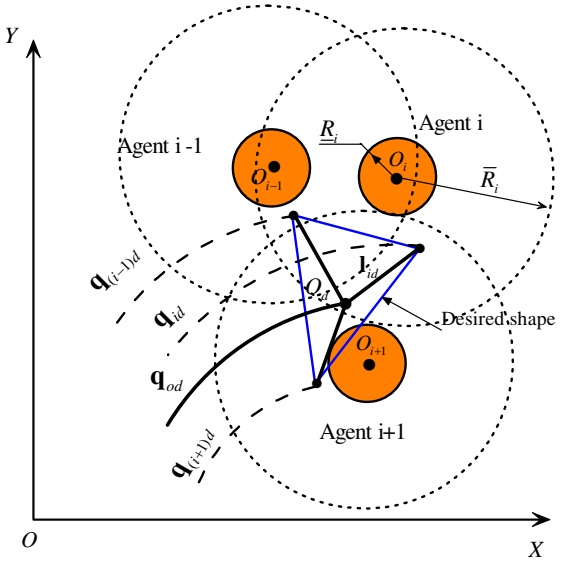


Figure 2: Formation setup.

Each agent in the group needs its reference trajectory to track. The reference trajectories can be predefined or determined from measurement data. Furthermore, each agent needs to communicate with other agents in the group to perform its cooperative mission. Therefore, before stating formation control objective we impose the following assumption on the reference trajectories, communication and initial conditions between the agents in the group:

Assumption 3.

1) The agent i has a physical safety ball, which is centered at the point O_i and has a radius $\underline{R}_i$, and has a communication ball, which is centered at the point O_i and has a radius $\overline{R}_i$, see Figure 2. The radius $\overline{R}_i$ is such that

$$\overline{R}_i \geq \underline{R}_i + \underline{R}_j + \varepsilon_{1ij}, \tag{4}$$

for all $j \in \mathbb{N}, j \neq i$, where ε_{1ij} is a strictly positive constant.

2) The reference trajectory $\boldsymbol{q}_{id}$ for the agent i is generated by

$$\boldsymbol{q}_{id} = \boldsymbol{q}_{od}(s_{od}) + \boldsymbol{l}_{id}, \tag{5}$$

where $\boldsymbol{q}_{od}(s_{od})$ is referred to as the common reference trajectory with s_{od} being the common trajectory parameter, and $\boldsymbol{l}_{id}$ is to specify a desired formation shape. The trajectory $\boldsymbol{q}_{od}$ has its bounded derivatives. The vectors $\boldsymbol{l}_{id}, i \in \mathbb{N}$ have bounded derivatives, and satisfy

$$(\underline{R}_i + \underline{R}_j + \varepsilon_{2ij}) \leq \|\boldsymbol{l}_{id} - \boldsymbol{l}_{jd}\| \leq \min(\overline{R}_i, \overline{R}_j) - \varepsilon_{2ij}, \tag{6}$$

for all $(i,j) \in \mathbb{N},\ i \neq j$, where ε_{2ij} is a strictly positive constant, and is strictly less than $\frac{\varepsilon_{1ij}}{2}$.

3)The agent i broadcasts its trajectory, $\boldsymbol{q}_i$, and its reference trajectory $\boldsymbol{q}_{id}$ in its communication ball. Moreover, the agent i can receive the trajectory, $\boldsymbol{q}_j$, broadcasted by other agents j, $j \in \mathbb{N},\ j \neq i$ in the group if the points O_j of these agents are in the communication ball of the agent i.

4) At the initial time $t_0 \geq 0$, all the agents in the group are sufficiently far but not too far away from each other in the sense that the following condition holds:

$$(\underline{R}_i + \underline{R}_j + \varepsilon_{3ij}) \leq \|\boldsymbol{q}_i(t_0) - \boldsymbol{q}_j(t_0)\| \leq (\min(\overline{R}_i, \overline{R}_j) - \varepsilon_{3ij}), \tag{7}$$

for all $(i,j) \in \mathbb{N},\ i \neq j$, where ε_{3ij} is a strictly positive constant and is strictly less than $\frac{\varepsilon_{1ij}}{2}$.

Remark 4. Item 2) in Assumption 3 defines a desired formation (by vectors $\boldsymbol{l}_{id}$) and how this desired formation moves (by the common reference trajectory $\boldsymbol{q}_{od}$). Item 3) specifies the way each agent communicates with other agents in the group within its communication range. In Figure 2, the agents i and $i-1$ are communicating with each other since the points O_{i-1} and O_i are in the communication areas of the agents i and $i-1$, respectively. Item 4) implies that at the initial time t_0 there are no collision between all the agents, and that all the agents are communicating with each other. The conditions (4), (6) and (7) are imposed to avoid conflict when solving collision avoidance and connectivity preserving problems. This is because we will design a formation control system so that $\boldsymbol{q}_i$ to track its reference trajectory $\boldsymbol{q}_{id}$.

Under Assumption 3, for each agent i design the control input vector $\boldsymbol{u}_i$ to achieve a desired formation consisting of 1) no switchings in the controllers; 2) no collisions between any agents; 3) asymptotic convergence of each agent's trajectory $\boldsymbol{q}_i$ to its reference trajectory $\boldsymbol{q}_{id}$; and 4) initial connectivity preservation. Mathematically, the objective is to design a smooth $\boldsymbol{u}_i$ to achieve:

$$\|\boldsymbol{q}_i(t)-\boldsymbol{q}_j(t)\| > (\underline{R}_i+\underline{R}_j),\ \lim_{t\to\infty}(\boldsymbol{q}_i(t)-\boldsymbol{q}_{id}(t))=0,\ \|\boldsymbol{q}_i(t)-\boldsymbol{q}_j(t)\| < \min(\overline{R}_i,\overline{R}_j), \tag{8}$$

for all $\forall t \geq t_0 \geq 0$ and $(i,j)\in\mathbb{N}$ and $j\neq i$.

3 Formation Control Design

Consider the following potential function

$$\varphi = \sum_{i=1}^{N}\left(\gamma_i + \frac{1}{2}\beta_i\right). \tag{9}$$

The aim of the goal function γ_i is to achieve asymptotic convergence of each agent's trajectory $\boldsymbol{q}_i$ to its reference trajectory $\boldsymbol{q}_{id}$. As such the function γ_i puts penalty on the tracking errors between the trajectory $\boldsymbol{q}_i$ of the agent i and its reference trajectory $\boldsymbol{q}_{id}=\boldsymbol{q}_{od}+\boldsymbol{l}_{id}$. We choose the function γ_i as:

$$\gamma_i = \frac{1}{2}\|\boldsymbol{q}_i-\boldsymbol{q}_{id}\|^2. \tag{10}$$

The purpose of the collision avoidance and connectivity preserving function β_i is to force the agent i to move away from other agents, and to maintain communication connectivity between the agent i and other agents in the group. This function is chosen as follows:

$$\beta_i = \sum_{j\in\mathbb{N}_i}\beta_{ij}, \tag{11}$$

where $\mathbb{N}_i$ is the set of all the agents in the group except for the agent i. The function $\beta_{ij}=\beta_{ji}$ is a function of $\|\boldsymbol{q}_{ij}\|^2/2$ with $\boldsymbol{q}_{ij}=\boldsymbol{q}_i-\boldsymbol{q}_j$, and possesses the following properties:

$$\begin{aligned}
&1)\quad \beta_{ij}=0,\ \beta_{ij}\prime=0,\ \beta_{ij}\prime\prime=0,\ \forall\ \|\boldsymbol{q}_{ij}\|\in\big((\underline{R}_i+\underline{R}_j+\delta_{ij}),(\min(\overline{R}_i,\overline{R}_j)-\delta_{ij})\big),\\
&2)\quad \beta_{ij}>0,\ \forall\ \|\boldsymbol{q}_{ij}\|\in\Big(\big((\underline{R}_i+\underline{R}_j),(\underline{R}_i+\underline{R}_j+\delta_{ij})\big)\cup\big((\min(\overline{R}_i,\overline{R}_j)-\delta_{ij}),\min(\overline{R}_i,\overline{R}_j)\big)\Big),\\
&3)\quad \lim_{\|\boldsymbol{q}_{ij}\|\to(\underline{R}_i+\underline{R}_j)}\beta_{ij}=\infty,\quad \lim_{\|\boldsymbol{q}_{ij}\|\to\min(\overline{R}_i,\overline{R}_j)}\beta_{ij}=\infty,\\
&4)\quad \beta_{ij}\text{ is smooth for all }\|\boldsymbol{q}_{ij}\|\in\big((\underline{R}_i+\underline{R}_j),(\min(\overline{R}_i,\overline{R}_j))\big),
\end{aligned} \tag{12}$$

where δ_{ij} is a strictly positive constant and is strictly less than ε_{2ij} specified in Assumption 3. The terms $\beta_{ij}\prime$ and $\beta_{ij}\prime\prime$ are defined as follows:

$$\begin{aligned}&\beta_{ij}\prime = \infty,\ \beta_{ij}\prime\prime = \infty,\ \text{if } \|\boldsymbol{q}_{ij}\| = \underline{R}_i + \underline{R}_j,\ \text{or } \|\boldsymbol{q}_{ij}\| = \min(\overline{R}_i, \overline{R}_j),\\ &\beta_{ij}\prime = \frac{\partial \beta_{ij}}{\partial(\|\boldsymbol{q}_{ij}\|^2/2)},\ \beta_{ij}\prime\prime = \frac{\partial^2 \beta_{ij}}{\partial(\|\boldsymbol{q}_{ij}\|^2/2)^2},\ \text{elsewhere.}\end{aligned} \tag{13}$$

Based on the smooth step function in Section 2.1, we can find many functions that satisfy all properties listed in (12). As an example, we will use the following function β_{ij}:

$$\beta_{ij} = \kappa_{ij}\left[\frac{1 - h\Big(\frac{\|\boldsymbol{q}_{ij}\|^2}{2}, \frac{(\underline{R}_i+\underline{R}_j)^2}{2}, \frac{(\underline{R}_i+\underline{R}_j+\delta_{ij})^2}{2}, c_{ij}\Big)}{\Big(\frac{\|\boldsymbol{q}_{ij}\|^2}{2} - \frac{(\underline{R}_i+\underline{R}_j)^2}{2}\Big)^2} + \frac{h\Big(\frac{\|\boldsymbol{q}_{ij}\|^2}{2}, \frac{(\min(\overline{R}_i,\overline{R}_j)-\delta_{ij})^2}{2}, \frac{\min(\overline{R}_i,\overline{R}_j)^2}{2}, c_{ij}\Big)}{\Big(\frac{\min(\overline{R}_i,\overline{R}_j)^2}{2} - \frac{\|\boldsymbol{q}_{ij}\|^2}{2}\Big)^2}\right], \tag{14}$$

where κ_{ij} and c_{ij} are positive constants, and the function $h(\bullet)$ is a smooth step function defined in Definition 1. An illustration of β_{ij} defined in (14) is given in Figure 3 with $\underline{R}_i + \underline{R}_j = 1$, $\min(\overline{R}_i, \overline{R}_j) = 11$, $\delta_{ij} = 2$, $c_{ij} = 1$, $\kappa_{ij} = 1$.

The derivative of φ along the solutions of (3) satisfies

$$\dot{\varphi} = \sum_{i=1}^{N} \boldsymbol{\Omega}_i^T(\boldsymbol{u}_i - \dot{\boldsymbol{q}}_{id}) + \sum_{i=1}^{N}\Big(\sum_{j\in\mathbb{N}_i} \beta_{ij}\prime \boldsymbol{q}_{ij}^T\Big)\dot{\boldsymbol{l}}_{id}, \tag{15}$$

where

$$\boldsymbol{\Omega}_i = \boldsymbol{q}_i - \boldsymbol{q}_{id} + \sum_{j\in\mathbb{N}_i} \beta_{ij}\prime \boldsymbol{q}_{ij}. \tag{16}$$

From (15), we design the control input $\boldsymbol{u}_i$ to make the sum $\sum_{i=1}^{N} \boldsymbol{\Omega}_i^T(\boldsymbol{u}_i - \dot{\boldsymbol{q}}_{id})$ negative definite as

$$\boldsymbol{u}_i = -k\boldsymbol{\Psi}(\boldsymbol{\Omega}_i) + \dot{\boldsymbol{q}}_{od} + \dot{\boldsymbol{l}}_{id}, \tag{17}$$

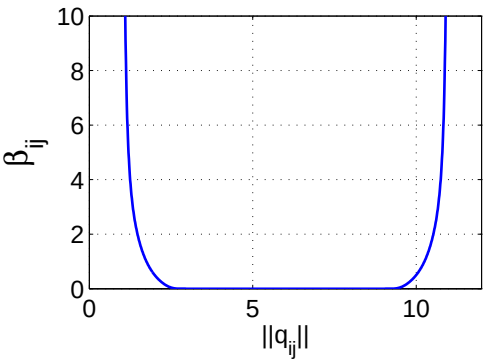


Figure 3: An illustration of β_{ij}.

where k is a positive constant, and $\boldsymbol{\Psi}(\boldsymbol{\Omega}_i)$ denotes a vector of bounded functions of elements of $\boldsymbol{\Omega}_i$ in the sense that $\boldsymbol{\Psi}(\boldsymbol{\Omega}_i) = \big[\psi(\Omega_i^1)\ ..., \psi(\Omega_i^l), ..., \psi(\Omega_i^n)\big]^T$ with Ω_i^l the l^{th} element of $\boldsymbol{\Omega}_i$, i.e., $\boldsymbol{\Omega}_i = [\Omega_i^1\ ..., \Omega_i^l ... \Omega_i^n]^T$. The function $\psi(x)$ satisfies

$$\begin{aligned}&1)\ |\psi(x)| \le M_1, && 2)\ \psi(x) = 0 \quad \text{if } x = 0, \quad x\psi(x) > 0 \text{ if } x \ne 0,\\ &3)\ \psi(-x) = -\psi(x), (x-y)[\psi(x) - \psi(y)] \ge 0, && 4)\ \Big|\frac{\psi(x)}{x}\Big| \le M_2, \Big|\frac{\partial\psi(x)}{\partial x}\Big| \le M_3, \frac{\partial\psi(x)}{\partial x}\Big|_{x=0} = 1,\end{aligned} \tag{18}$$

for all $x \in \mathbb{R}, y \in \mathbb{R}$, where M_1, M_2, M_3 are strictly positive constants. Some functions that satisfy the above properties are $\arctan(x)$ and $\tanh(x)$. The above bounds mean that the large control effort problem is avoided when the distance $\|\boldsymbol{q}_{ij}\|$ between the agent i and an agent j in the group reaches a collision limit $\underline{R}_i + \underline{R}_j$ or a connectivity preserving limit $\min(\overline{R}_i, \overline{R}_j)$.

To deal with the sum $\sum_{i=1}^{N}\Big(\sum_{j\in\mathbb{N}_i} \beta_{ij}\prime \boldsymbol{q}_{ij}^T\Big)\dot{\boldsymbol{l}}_{id}$ in (15), we observe that $\beta_{ij}\prime = 0$ for all $\|\boldsymbol{q}_{ij}\| \in \big((\underline{R}_i + \underline{R}_j + \delta_{ij}), (\min(\overline{R}_i, \overline{R}_j) - \delta_{ij})\big)$, see Property 1) of the function β_{ij} in (12). This observation motivates us to design an update law for $\boldsymbol{l}_{id}$ so that $\sum_{i=1}^{N}\Big(\sum_{j\in\mathbb{N}_i} \beta_{ij}\prime \boldsymbol{q}_{ij}^T\Big)\dot{\boldsymbol{l}}_{id} = 0$ for all time and $\dot{\boldsymbol{l}}_{id}$ tends to its desired value $\boldsymbol{v}_{id}$ asymptotically. As such, we choose:

$$\dot{\boldsymbol{l}}_{id} = H_i \boldsymbol{v}_{id}, \tag{19}$$

where

$$H_i = \prod_{j\in\mathbb{N}_i} h(\|\boldsymbol{q}_{ij}\|^2/2, (\underline{R}_i + \underline{R}_j + \delta_{ij})^2/2, (\underline{R}_i + \underline{R}_j + \delta_{ij}^v)^2/2, c_{ij})\times \\ \Big(1 - h(\|\boldsymbol{q}_{ij}\|^2/2, \min(\overline{R}_i + \overline{R}_j - \delta_{ij}^v)^2/2, (\overline{R}_i + \overline{R}_j - \delta_{ij})^2/2, c_{ij})\Big), \tag{20}$$

with δ_{ij}^v being a positive constant such that $\delta_{ij} < \delta_{ij}^v < \epsilon_{2ij}$, and $h(\bullet)$ being a smooth step function defined in Definition 1. With the choice of $\delta_{ij} < \delta_{ij}^v < \epsilon_{2ij}$, we can see that

$$\begin{aligned} &H_i = 1,\ \forall\ \|\boldsymbol{q}_{ij}\| \in \big((\underline{R}_i + \underline{R}_j + \delta_{ij}^v), (\min(\overline{R}_i, \overline{R}_j) - \delta_{ij}^v)\big), \\ &H_i = 0,\ \forall\ \|\boldsymbol{q}_{ij}\| \in \big((0, (\underline{R}_i + \underline{R}_j + \delta_{ij})) \cup (\min(\overline{R}_i, \overline{R}_j) - \delta_{ij}), \infty)\big), \\ &0 < H_i < 1,\ \text{elsewhere.} \end{aligned} \tag{21}$$

Obviously, the choice of the update law for $\boldsymbol{l}_{id}$ in (19) with H_i being satisfied (21) gives:

$$\sum_{j\in\mathbb{N}_i} \beta_{ij}\prime \boldsymbol{q}_{ij}^T \dot{\boldsymbol{l}}_{id} = 0,\ \forall\ \|\boldsymbol{q}_{ij}\| \in ((\underline{R}_i + \underline{R}_j), \min(\overline{R}_i, \overline{R}_j)). \tag{22}$$

Remark 5. 1) A careful look at the control law $\boldsymbol{u}_i$ in (17) with $\boldsymbol{\Omega}_i$ in (16) shows that the argument of the bounded $\boldsymbol{\Psi}$ (with the negative sign moved in) consists of two parts. The first part is $-(\boldsymbol{q}_i - \boldsymbol{q}_{id})$, and the second part is $-\sum_{j\in\mathbb{N}_i} \beta_{ij}\prime \boldsymbol{q}_{ij}$. The first part together with $\dot{\boldsymbol{q}}_{od} + \dot{\boldsymbol{l}}_{id}$ is referred to as the attractive force plays the role of forcing the agent i to track its reference trajectory. The second part is referred to as the repulsive force takes care of collision avoidance and connectivity preserving for the agent i with the other agents in the group. Moreover, the control $\boldsymbol{u}_i$ is a smooth function of and depend on only its own state and reference trajectory, and the states of other neighbor agents j if the agents j are sufficiently close to the agent i for collision avoidance, or are sufficiently far away from the agent i for connectivity preserving.

2) The choice of the update law in (19) ensures that when the collision avoidance or connectivity preserving is active, i.e., when the sum $\sum_{j\in\mathbb{N}_i} \beta_{ij}\prime \boldsymbol{q}_{ij}$ is non-zero, the vector $\boldsymbol{l}_{id}$ is not updated, i.e., the desired formation shape is not changed. This implies that the control law $\boldsymbol{u}_i$ gives priority to the collision avoidance and/or connectivity preserving mission or the desired formation shape updating mission whenever which mission is more important.

Substituting the control law $\boldsymbol{u}_i$ in (17) and the update law $\dot{\boldsymbol{l}}_{id}$ in (19) into (15) gives

$$\dot{\varphi} = -k \sum_{i=1}^{N} \boldsymbol{\Omega}_i^T \boldsymbol{\Psi}(\boldsymbol{\Omega}_i), \tag{23}$$

where we have used (22). On the other hand, substituting the control law the control law $\boldsymbol{u}_i$ in (17) into (3) including the update law $\dot{\boldsymbol{l}}_{id}$ in (19) results in the closed loop system:

$$\begin{aligned} \dot{\boldsymbol{q}}_i &= -k\boldsymbol{\Psi}(\boldsymbol{\Omega}_i) + \dot{\boldsymbol{q}}_{od} + \dot{\boldsymbol{l}}_{id}, \\ \dot{\boldsymbol{l}}_{id} &= H_i \boldsymbol{v}_{id}, \end{aligned} \tag{24}$$

for all $i \in \mathbb{N}$. We now present the main result of our paper in the following theorem.

Theorem 6. *Under Assumption 3, the smooth control input $\boldsymbol{u}_i =$ given in (17) and the update law $\dot{\boldsymbol{l}}_{id}$ in (19) for the agent i solve the formation control objective. In particular:*

1) There are no collisions between any agents, connectivity between the agents is maintained, and the closed loop system (24) is forward complete. The first and last inequalities in (8) hold.

2) The reference velocity $\dot{\boldsymbol{l}}_{id}$ approaches its desired reference velocity $\boldsymbol{v}_{id}$ asymptotically.

3) The trajectory $\boldsymbol{q}_i$ of each agent i tracks its reference trajectory $\boldsymbol{q}_{id}$ asymptotically, i.e., the limit in the second equation of (8) holds.

Proof. See Appendix 1.

4 Gradient climbing

4.1 Approach

In this section, we present an application of our proposed formation control to solve a gradient climbing mission in a distributed environment $\Phi(t, \boldsymbol{\eta})$. To do so, we consider each agent in the group as a mobile sensor and the network as a reconfigurable sensor array. As such, at each time t the agent i with $i \in \mathbb{N}$ in the group of N agents is equipped with a sensor that can measure $\Phi(t, \boldsymbol{q}_i)$ at the location $\boldsymbol{q}_i$. With $\Phi(t, \boldsymbol{q}_i)$, we estimate/calculate an approximation of the gradient average, $\overline{\boldsymbol{\nabla}}\Phi$, of the distributed environment over a region A bounded by a contour C, on which the agents in the group are positioned. After $\overline{\boldsymbol{\nabla}}\Phi$ is estimated/calculated, we let the gradient of the common reference trajectory $\boldsymbol{q}_{od}$ equal to $\overline{\boldsymbol{\nabla}}\Phi$. This means that the common reference trajectory $\boldsymbol{q}_{od}$ is simply generated by

$$\dot{\boldsymbol{q}}_{od} = \frac{\partial \boldsymbol{q}_{od}}{\partial s_{od}} \dot{s}_{od} = \overline{\boldsymbol{\nabla}}\Phi \dot{s}_{od}, \tag{25}$$

with some initial condition $\boldsymbol{q}_{od}(t_0)$, where $\dot{s}_{od}$ specifies how fast the desired formation moves along the common reference trajectory $\boldsymbol{q}_{od}$. For the case of gradient descent, we can use $\dot{\boldsymbol{q}}_{od} = -\overline{\boldsymbol{\nabla}}\Phi \dot{s}_{od}$ instead of (25). Moreover, we can specify the desired formation shape velocity $\boldsymbol{v}_{id}$ to change (expand/shrink/rotate) the formation shape, i.e., change the shape define vector $\boldsymbol{l}_{id}$, see (19), to improve the gradient average approximation. We propose the the desired formation shape velocity $\boldsymbol{v}_{id}$ as follows:

$$\boldsymbol{v}_{id} = -\boldsymbol{K}_{1v}(\boldsymbol{l}_{id} - \boldsymbol{l}^*_{id}) + \boldsymbol{K}_{2v}\boldsymbol{\Psi}(\overline{\boldsymbol{\nabla}}\Phi), \tag{26}$$

where $\boldsymbol{K}_{1v}$ and $\boldsymbol{K}_{2v}$ are diagonal positive definite matrices. The constant vectors $\boldsymbol{l}^*_{id}, i \in \mathbb{N}$ are chosen so that they specify the minimum desired formation shape, which is such that the condition (6) holds with $\boldsymbol{l}_{id}$ replaced by $\boldsymbol{l}^*_{id}$ for all $i \in \mathbb{N}$. The vector function $\boldsymbol{\Psi}(\overline{\boldsymbol{\nabla}}\Phi)$ is a bounded vector function of $\overline{\boldsymbol{\nabla}}\Phi$, see the paragraph just below (17). Once the common reference trajectory $\boldsymbol{q}_{od}$ and the desired formation shape $\boldsymbol{l}_{id}$ are available, the formation control design proposed in Section 3 can be used directly to drive the agents in the group. The following section gives a method to estimate an approximation of the gradient average, $\overline{\boldsymbol{\nabla}}\Phi$, of the distributed environment $\Phi(t, \boldsymbol{\eta})$ from measurements $\Phi(t, \boldsymbol{q}_i)$ on the boundary, i.e., the contour or surface C, carried out by the agents in the group. Therefore, we will present a method to calculate the gradient average of a distributed field in the following subsection.

4.2 Average gradient estimate of a distributed field

We consider a region A, see Figure 4, bounded by a contour C, such that any line through A parallel to either one of the coordinate axes intersects C in only two points. The curve C is divided by its leftmost and rightmost points ($x = a$ and $x = b$) into a lower segment C_1, described by $y = f_1(x)$, and an upper segment C_2 described by $y = f_2(x)$. With the position vector to a point P on C given by $\boldsymbol{r} = xe_x + ye_y$, where e_x and e_y are the unit vector on the OX and OY axes, respectively. The unit tangent vector at P is $\boldsymbol{t} = \frac{d\boldsymbol{r}}{ds} = \frac{dx}{ds}e_x + \frac{dy}{ds}e_y$, where ds is the differential length along C, and the unit normal vector is $\boldsymbol{n} = \boldsymbol{t} \times e_z = \frac{dy}{ds}e_x - \frac{dx}{ds}e_y = n_x e_x + n_y e_y$. For the function

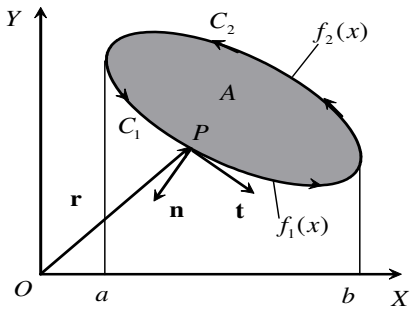


Figure 4: Coordinates for a gradient computation

$\Phi(t,x,y)$ defined in A, consider the area integral

$$\int_A \frac{\partial \Phi}{\partial y} dA = \int_a^b \Big(\Phi(t,x,f_2(x)) - \Phi(t,x,f_1(x))\Big)dx = \int_a^b \Big([\Phi]_{C_2} - [\Phi]_{C_1}\Big)dx. \tag{27}$$

As shown in Fig. 4, a positive contour integration corresponds to a counter-clockwise traversal of C. To make the first integral in (27) consistent with this connection, we write

$$\int_A \frac{\partial \Phi}{\partial y} dA = -\int_b^a [\Phi]_{C_2} dx - \int_a^b [\Phi]_{C_1} dx = -\int_C \Phi dx = -\int_C \Phi \frac{dx}{ds} ds, \tag{28}$$

which combines with $\boldsymbol{n} = n_x e_x + n_y e_y$ to yield $\int_A \frac{\partial \Phi}{\partial y} dA = \int_C \Phi n_y ds$. A similar computation gives $\int_A \frac{\partial \Phi}{\partial x} dA = \int_C \Phi n_x ds$. Therefore, we have

$$\int_A \boldsymbol{\nabla} \Phi dA = \int_C \boldsymbol{n} \Phi ds \tag{29}$$

where $\boldsymbol{\nabla}\Phi = \left[\frac{\partial \Phi}{\partial x}, \frac{\partial \Phi}{\partial y}\right]^T$. It is of interest to note that the total gradient $\int_A \boldsymbol{\nabla}\Phi dA$ of the distributed field $\Phi(t,\boldsymbol{\eta})$ over the region A is completely determined from the integral $\int_C \boldsymbol{n}\Phi ds$ carried out on the boundary C only. From (29), we can calculate the gradient average of $\Phi(t,\boldsymbol{\eta})$ over the region A as

$$\overline{\boldsymbol{\nabla}}\Phi = \frac{\int_C \boldsymbol{n}\Phi ds}{\Omega_A} \tag{30}$$

where Ω_A is the area of the region A. Usually, it is not possible to obtain an explicit result of the integral $\int_C \boldsymbol{n}\Phi ds$ because the distributed field Φ is unknown. Hence, we approximate this integral from measurement $\Phi(t,\boldsymbol{q}_i)$ at the time t and the location $(\boldsymbol{q}_i)$ by each agent i, and approximate the area Ω_A. We assume that the formation shape is a convex polygon whose vertices are at $\boldsymbol{q}_i$. The steps to calculate an approximate value of the integral $\int_C \boldsymbol{n}\Phi ds$ and the region area Ω_A are as follows:

1) Using a curve fitting method such as Spline or least square to find a best fitted and smooth contour $C(\theta)$, where θ is the curve parameter, that goes through all vertices at the time t;

2) Calculating an approximate value of $\int_C \boldsymbol{n}\Phi ds$ and Ω_A as follows:

$$\int_C \boldsymbol{n}\Phi ds \approx \sum_{i=1}^N \Phi(t,\boldsymbol{q}_i)\boldsymbol{n}(\theta_i)\Delta_{Ci},\ \Omega_A \approx \frac{1}{2}\sum_{i=1}^N \big(x_i y_{i+1} - x_{i+1} y_i\big), \tag{31}$$

where $\boldsymbol{q}_{N+1} = \boldsymbol{q}_1$, $\boldsymbol{n}(\theta_i)$ is the unit vector normal to $C(\theta)$ at θ_i corresponding to the position of the vertex $\boldsymbol{q}_i$, and Δ_{Ci} is the arc length from the middle point M_{i-1} between $\boldsymbol{q}_{i-1}$ and $\boldsymbol{q}_i$ an the middle point M_{i+1} between $\boldsymbol{q}_i$ and $\boldsymbol{q}_{i+1}$, see Fig.5.

For a special case where the formation shape is a regular simple polygon, which has the center at $\boldsymbol{q}_{od}$ and the vertices at $\boldsymbol{q}_i, i \in \mathbb{N}$, and that the contour C goes through all the vertices at the time t. Moreover, the unit vector $\boldsymbol{n}$ normal to the contour C at $\boldsymbol{q}_i$ is in the direction from $\boldsymbol{q}_{od}$ to $\boldsymbol{q}_i$ at the time t. The the integral $\int_C \boldsymbol{n}\Phi ds$ and the region area Ω_A can be approximated as

$$\int_C \boldsymbol{n}\Phi ds \approx \sum_{i=1}^N \Phi(t,\boldsymbol{q}_i)\frac{\boldsymbol{q}_i - \boldsymbol{q}_{od}}{\|\boldsymbol{q}_i - \boldsymbol{q}_{od}\|}\left\|\frac{\boldsymbol{q}_{i+1} - \boldsymbol{q}_{i-1}}{2}\right\|,$$
$$\Omega_A \approx \frac{1}{2}\sum_{i=1}^N \Big|\det([\boldsymbol{q}_i, \boldsymbol{q}_{i+1}])\Big|, \tag{32}$$

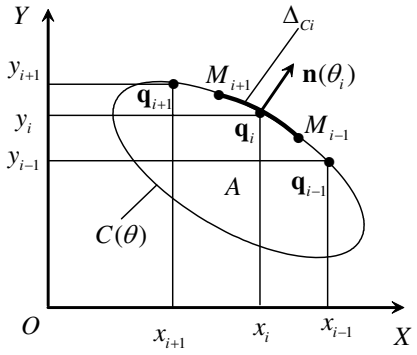


Figure 5: Coordinates for gradient average calculation.

with $\boldsymbol{q}_{N+1} = \boldsymbol{q}_1$ and $\boldsymbol{q}_{-1} = \boldsymbol{q}_N$, and $\det([\boldsymbol{q}_i, \boldsymbol{q}_{i+1}])$ is the determinant of the matrix $[\boldsymbol{q}_{i+1}, \boldsymbol{q}_i]$.

5 Simulation results

In this section, we a problem of gradient climbing by our proposed formation controller using a group of $N = 6$ identical agents. Each agent i has a physical safety radius $\underline{R}_i = 0.5$ and a communication radius $\overline{R}_i = 10$. The control design parameters are taken as $k = 4$, $\delta_{ij} = 0.5$, $\delta^v_{ij} = 0.75$, $c_{ij} = 1$, and the bounded function $\psi(.)$ taken as arctan(.).

The desired formation shape specification vectors $\boldsymbol{l}^*_{id}$ are chosen as $\boldsymbol{l}^*_{id} = R_f[\cos(\frac{2(i-1)\pi}{N}), \sin(\frac{2(i-1)\pi}{N})]^T$ with $R_f = 3$, and the gain $\boldsymbol{K}_{1v} = \text{diag}(2.5, 2.5)$. This choice of $\boldsymbol{l}^*_{id}$ means that the desired formation configuration is a polygon whose vertices uniformly distribute on a circle centered on the common reference trajectory and with a radius R_f. The initial conditions are $\boldsymbol{l}_{id}(0) = \boldsymbol{l}^*_{id}$, $\boldsymbol{q}_{od}(0) = [0\ 0]^T$, and $\boldsymbol{q}_i(0) = R_f[\cos(\frac{2(i-1)\pi}{N} + \pi), \sin(\frac{2(i-1)\pi}{N} + \pi)]^T$. These particular initial $\boldsymbol{q}_i(0)$ were chosen to illustrate the collision avoidance capability of our proposed formation control system as all the agents have to across the center of the desired formation shape to track their desired reference trajectories. The distributed environment $\Phi(t, x, y)$ is taken as $\Phi(t, x, y) = \text{e}^{-\frac{(x-15)^2+(y-15)^2}{150}}$, which has a global maximum value at $(x = 15, y = 15)$.

We set $\boldsymbol{K}_{2v} = \text{diag}(1.5, 1.5)$ to improve the gradient climbing, i.e., the desired formation shape is adapted to the distributed field. Simulation results are plotted in Figure 6. From these figures, it is seen that our proposed formation is able to achieve the objective of both formation control and gradient climbing. The control inputs $\boldsymbol{u}_i$, see sub-figure 6D, force the agents to move in such a way that collision between the agents is avoided and that communication between the agents is preserved, see sub-figure 6A where trajectories of the agents are plotted in XY-plane. These sub-figures also show that our proposed formation control performs the gradient climbing mission very well in the sense that the center of the formation shape, see the polygon of which vertices are the agents, converges to the global maximum location of the function $\Phi(t, x, y)$. Collision avoidance and communication preserving are also confirmed in sub-figure 6C, where the distances $\|\boldsymbol{q}\|_{1i}$ between the agent 1 and other agents in the group are plotted. These distances are within the range of $(1, 10)$ since $\underline{R}_i + \underline{R}_j = 1$ and $\min(\overline{R}_i, \overline{R}_j) = 10$.

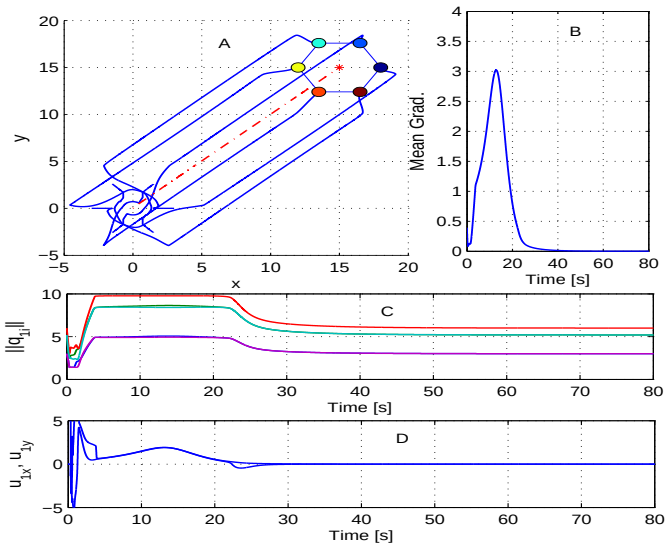


Figure 6: Simulation results with formation shape adaptation.

6 Conclusions

A constructive method has been proposed to design smooth and bounded cooperative controllers for a group of N mobile agents with limited communication to perform a desired formation. Novel potential functions encoding desired formation mission tasks with smooth step functions embedded in were constructed to design the controllers that guaranteed all equilibrium (critical) sets, except for the desired set in formation, are unstable. The proposed formation control system is applied to solve a gradient climbing problem. An extension of the proposed formation control design in this paper and those controllers designed for single underactuated ships in [17] to provide a formation control system for a group of underactuated ships is under consideration.

1 Proof of Theorem 6

Proof of no collisions, connectivity preserving, and forward completeness of the closed loop system: It is seen from (23) that $\dot{\varphi} \leq 0$. Integrating $\dot{\varphi} \leq 0$ from t_0 to t and using the definition of φ in (9) with its components defined in (10) and (11) results in

$$\varphi(t) \leq \varphi(t_0), \tag{33}$$

where $\varphi(t_0) = \sum_{i=1}^{N}(\gamma_i(t_0) + \frac{1}{2}\sum_{j\in\mathbb{N}_i}\beta_{ij}(t_0))$ and $\varphi(t) = \sum_{i=1}^{N}(\gamma_i(t) + \frac{1}{2}\sum_{j\in\mathbb{N}_i}\beta_{ij}(t))$, for all $t \geq t_0 \geq 0$. From the condition specified in item 4) of Assumption 3, and properties of β_{ij}, we have the right hand side of (33) is bounded by a positive finite constant depending on the initial conditions. Boundedness of the right hand side of (33) implies that the left hand side of (33) must be also bounded. As a result, $\beta_{ij}(\|\boldsymbol{q}_{ij}\|^2/2)$ must be smaller than some positive constant depending on the initial conditions for all $t \geq t_0 \geq 0$. From properties of β_{ij}, see (12), $\|\boldsymbol{q}_{ij}\|$, for all $(i,j) \in \mathbb{N}$ and $i \neq j$, must be in the interval $\big((\underline{R}_i + \underline{R}_j), \min(\overline{R}_i, \overline{R}_j)\big)$. Hence, there are no collisions between any agents and connectivity between agents is preserved for all $t \geq t_0 \geq 0$. Boundedness of the left hand side of (33) also implies that of $(\boldsymbol{q}_i(t) - \boldsymbol{q}_{id}(t))$ also bounded for all $t \geq t_0 \geq 0$. Moreover, from (21) we can see that $|H_i| \leq 1$. Therefore, $\|\dot{\boldsymbol{l}}_{id}(t)\| \leq \|\boldsymbol{v}_{id}(t)\|$ for all $t \geq t_0 \geq 0$. Therefore, the closed loop system (24) is forward complete.

Equilibrium set: We will use Lemma 2 in [7] to find the equilibrium set, which the trajectories of the closed loop system (24) converge to. Integrating both sides of (23) yields

$$\int_0^{\infty} \omega(t)dt = \varphi(t_0) - \varphi(\infty) \leq \varphi(t_0), \tag{34}$$

with $\omega(t) := \sum_{i=1}^{N} \boldsymbol{\Omega}_i^T(t)\boldsymbol{\Psi}(\boldsymbol{\Omega_i(t)})$, where $\boldsymbol{\Omega}_i(t)$ is given in (16), and the function $\boldsymbol{\Psi}(\boldsymbol{\Omega}_i(t))$ is the bounded vector function of $\boldsymbol{\Omega}_i(t)$ with properties listed in (18). Indeed, the function $\omega(t)$ is scalar, nonnegative and differentiable. Now differentiating $\omega(t)$ along the solutions of the closed loop system (24) and using properties of the function β_{ij} given in (12) readily show that $\big|\frac{\mathrm{d}\omega(t)}{\mathrm{d}t}\big| \leq M\omega(t)$ with M being a positive constant. Therefore Lemma 2 in [7] results in $\lim_{t\to\infty}\omega(t) = 0$, which implies from the expression of $\omega(t)$ and properties of the bounded vector function $\boldsymbol{\Psi}(\boldsymbol{\Omega}_i(t))$ in (18) that $\lim_{t\to\infty}\boldsymbol{\Omega}_i(t) = 0$. Therefore, from the expression of $\boldsymbol{\Omega}_i(t)$ the limit $\lim_{t\to\infty}\boldsymbol{\Omega}_i(t) = 0$ given in (16) implies that

$$\lim_{t\to\infty}\sum_{i=1}^{N}\left(\boldsymbol{q}_i(t) - \boldsymbol{q}_{id}(t) + \sum_{j\in\mathbb{N}_i}\beta_{ij}\prime(t)\boldsymbol{q}_{ij}(t)\right) = 0. \tag{35}$$

The limit in (35) implies that $\boldsymbol{q}(t) = [\boldsymbol{q}_1^T(t)\ \boldsymbol{q}_2^T(t), \ldots, \boldsymbol{q}_N^T(t)]^T$ can tend to $\boldsymbol{q}_d = [\boldsymbol{q}_{1d}^T\ \boldsymbol{q}_{2d}^T, \ldots, \boldsymbol{q}_{Nd}^T]^T$ denoted by the set $\boldsymbol{\Xi}_d$, since $\beta_{ij}\prime(t) = 0$ at $\boldsymbol{q}_i = \boldsymbol{q}_{id}$ and $\boldsymbol{q}_j = \boldsymbol{q}_{jd}$, for all $(i,j) \in \mathbb{N}$ and $i \neq j$ or

tend to the set $\boldsymbol{q}_c = [\boldsymbol{q}_{1c}^T\ \boldsymbol{q}_{2c}^T, \ldots, \boldsymbol{q}_{Nc}^T]^T$ denoted by the set $\boldsymbol{\Xi}_c$ as the time goes to infinity, i.e., the equilibrium sets can be $\boldsymbol{\Xi}_d$ or $\boldsymbol{\Xi}_c$. The equilibrium set $\boldsymbol{\Xi}_c$ is such that

$$\boldsymbol{\Omega}\Big|_{\boldsymbol{q}\in\boldsymbol{\Xi}_c} = \Big(\boldsymbol{q}_i - \boldsymbol{q}_{id} + \sum_{j\in\mathbb{N}_i} \beta_{ij}\prime\boldsymbol{q}_{ij}\Big)\Big|_{\boldsymbol{q}\in\boldsymbol{\Xi}_c} = 0, \tag{36}$$

for all $i \in \mathbb{N}$. Thus, we have already proved that the trajectory $\boldsymbol{q}$ can approach either the desired equilibrium set denoted by $\boldsymbol{\Xi}_d$ or the undesired equilibrium set denoted by $\boldsymbol{\Xi}_c$ 'almost globally'. The term 'almost globally' refers to the fact that the agents start from a set that includes the condition (7) and that does not coincide at any point in the undesired equilibrium set $\boldsymbol{\Xi}_c$. Therefore, we now need to prove that $\boldsymbol{\Xi}_d$ is a locally asymptotically stable set and that $\boldsymbol{\Xi}_c$ is a locally unstable set. Once this is proved, we can conclude that the trajectory $\boldsymbol{q}$ approaches $\boldsymbol{q}_d$ from almost everywhere except for from the set denoted by the condition (7) and the undesired equilibrium set $\boldsymbol{\Xi}_c$, which is unstable (to be proved below). To prepare for showing that $\boldsymbol{\Xi}_d$ is asymptotically stable and that $\boldsymbol{\Xi}_c$ is unstable. We write the first equation of the closed loop system (24) for all $i \in \mathbb{N}$ in a vector form as

$$\dot{\boldsymbol{q}} = -k\boldsymbol{\Phi}(\boldsymbol{q}, \boldsymbol{q}_d) + \dot{\boldsymbol{q}}_d \tag{37}$$

where $\boldsymbol{\Phi}(\boldsymbol{q}, \boldsymbol{q}_d) = [\boldsymbol{\Psi}^T(\boldsymbol{\Omega}_1), ..., \boldsymbol{\Psi}^T(\boldsymbol{\Omega}_N)]$. Linearizing (37) around $\boldsymbol{q}_o = [\boldsymbol{q}_{1o}^T, \ldots, \boldsymbol{q}_{No}^T]^T$, and letting the set $\boldsymbol{\Xi}_o$ contain $\boldsymbol{q}_o$ results in

$$\dot{\boldsymbol{q}} = -k\frac{\partial\boldsymbol{\Phi}(\boldsymbol{q}, \boldsymbol{q}_d)}{\partial\boldsymbol{q}}\Big|_{\boldsymbol{q}\in\boldsymbol{\Xi}_o} + \dot{\boldsymbol{q}}_d, \tag{38}$$

where $\frac{\partial\boldsymbol{\Phi}(\boldsymbol{q},\boldsymbol{q}_d)}{\partial\boldsymbol{q}} = [\Delta_{ij}]$ with $\Delta_{ij} = \frac{\partial\boldsymbol{\Psi}(\boldsymbol{\Omega}_i)}{\partial\boldsymbol{\Omega}_i}\frac{\partial\boldsymbol{\Omega}_i}{\partial\boldsymbol{q}_j}$ and

$$\frac{\partial\boldsymbol{\Omega}_i}{\partial\boldsymbol{q}_i} = \Big(1 + \sum_{i\in\mathbb{N}_i}\beta_{ij}\prime\Big)I_n + \sum_{j\in\mathbb{N}_i}\beta_{ij}\prime\prime\boldsymbol{q}_{ij}\boldsymbol{q}_{ij}^T,\ \frac{\partial\boldsymbol{\Omega}_i}{\partial\boldsymbol{q}_j} = -\beta_{ij}\prime I_{n\times n} - \beta_{ij}\prime\prime\boldsymbol{q}_{ij}\boldsymbol{q}_{ij}^T, \tag{39}$$

for all $(i, j) \in \mathbb{N}$. Let $\mathbb{N}^*$ be the set of the agents such that if the agents i and j belong to the set $\mathbb{N}^*$ then $\|\boldsymbol{q}_{ij}\| \in ((\underline{R}_i + \underline{R}_j), \min(\overline{R}_i, \overline{R}_j)$. Next we will show that the equilibrium set $\boldsymbol{\Xi}_d$ is asymptotically stable and that the equilibrium set $\boldsymbol{\Xi}_c$ is unstable.

Proof of $\boldsymbol{\Xi}_d$ being asymptotically stable: As mentioned above, to prove that the equilibrium set $\boldsymbol{\Xi}_d$ is asymptotically stable, we just need to show that $\boldsymbol{\Xi}_d$ is locally asymptotically stable. Letting $\boldsymbol{\Xi}_o$ be $\boldsymbol{\Xi}_d$ in (38), we obtain

$$\dot{\boldsymbol{q}} = -k(\boldsymbol{q} - \boldsymbol{q}_d) + \dot{\boldsymbol{q}}_d, \tag{40}$$

where we have used the fact that $\beta_{ij}\prime\big|_{\boldsymbol{q}\in\boldsymbol{\Xi}_d} = 0$ and $\beta_{ij}\prime\prime\big|_{\boldsymbol{q}\in\boldsymbol{\Xi}_d} = 0$, see Property 1) of the function β_{ij} in (12). Local asymptotic stability of the equilibrium set $\boldsymbol{\Xi_d}$ follows from (40) since the first time derivative of the function $V_d = \frac{1}{2}\|\boldsymbol{q} - \boldsymbol{q}_d\|^2$ along the solutions of (40) satisfies $\dot{V}_d = -2kV_d$.

Proof of $\boldsymbol{\Xi}_c$ being asymptotically stable: Let us define

$$\bar{\boldsymbol{q}} = [\boldsymbol{q}_{12}^T, ..., \boldsymbol{q}_{1N}^T\boldsymbol{q}_{23}^T, ..., \boldsymbol{q}_{2N}^T, ..., \boldsymbol{q}_{N-1,N}^T]^T,\ \bar{\boldsymbol{q}}_c = [\boldsymbol{q}_{12c}^T, ..., \boldsymbol{q}_{1Nc}^T\boldsymbol{q}_{23c}^T, ..., \boldsymbol{q}_{2Nc}^T, ..., \boldsymbol{q}_{N-1,Nc}^T]^T,$$
$$\beta_{ijc}\prime = \beta_{ij}\prime\big|_{\boldsymbol{q}\in\boldsymbol{\Xi}_c},\ \beta_{ijc}\prime\prime = \beta_{ij}\prime\prime\big|_{\boldsymbol{q}\in\boldsymbol{\Xi}_c},\ \boldsymbol{q}_{ijc} = \boldsymbol{q}_{ic} - \boldsymbol{q}_{jc}.$$

With the above definitions, we can see that stability of $\boldsymbol{\Xi}_c$ is equivalent to that of $\bar{\boldsymbol{\Xi}}_c = \bar{\boldsymbol{q}}_c$. Define $\boldsymbol{\Omega}_{ijc} = \boldsymbol{\Omega}_{ic} - \boldsymbol{\Omega}_{jc},\ \forall(i, j) \in \mathbb{N},\ i \neq j$ where $\boldsymbol{\Omega}_{ic} = \boldsymbol{\Omega}_i|_{\boldsymbol{q}\in\boldsymbol{\Xi}_c} = 0$, see (36). Therefore $\boldsymbol{\Omega}_{ijc} = 0$. Hence $\sum_{(i,j)\in\mathbb{N}^*}\boldsymbol{q}_{ijc}^T\boldsymbol{\Omega}_{ijc} = 0, i \neq j$, which by using (36) is expanded to

$$\sum_{(i,j)\in\mathbb{N}^*}\big(\boldsymbol{q}_{ijc}^T(\boldsymbol{q}_{ijc} - \boldsymbol{q}_{ijd}) + N\beta_{ijc}\prime\boldsymbol{q}_{ijc}^T\boldsymbol{q}_{ijc}\big) = 0 \Rightarrow \sum_{(i,j)\in\mathbb{N}^*}(1 + N\beta_{ijc}\prime)\boldsymbol{q}_{ijc}^T\boldsymbol{q}_{ijc} = \sum_{(i,j)\in\mathbb{N}^*}\boldsymbol{q}_{ijc}^T\boldsymbol{q}_{ijd} \tag{41}$$

where $i \neq j$. The sum $\sum_{(i,j)\in\mathbb{N}^*} \boldsymbol{q}_{ijc}^T \boldsymbol{q}_{ijd}$ is strictly negative since at the point F where $\boldsymbol{q}_{ij} = \boldsymbol{q}_{ijd},\ \forall (i,j) \in \mathbb{N}^*, i \neq j$ all attractive and repulsive forces are equal to zero while at the point C where $\boldsymbol{q}_{ij} = \boldsymbol{q}_{ijc}\ \forall (i,j) \in \mathbb{N}^*, i \neq j$ the sum of attractive and repulsive forces are equal to zero (but attractive and repulsive forces are nonzero). Therefore the point O where $\boldsymbol{q}_{ij} = 0,\ \forall (i,j) \in \mathbb{N}^*, i \neq j$ must locate between the points F and C for all $(i,j) \in \mathbb{N}^*, i \neq j$. That is the points F, O, and C must be co-linear. Hence, there exists a strictly positive constant b such that $\sum_{(i,j)\in\mathbb{N}^*} \boldsymbol{q}_{ijc}^T \boldsymbol{q}_{ijd} < -b$, which is substituted into (41) to yield

$$\sum_{(i,j)\in\mathbb{N}^*} (1 + N\beta_{ijc\prime}) \boldsymbol{q}_{ijc}^T \boldsymbol{q}_{ijc} < -b, i \neq j. \tag{42}$$

Since $\boldsymbol{q}_{ijc}^T \boldsymbol{q}_{ijc} > 0, \forall (i,j) \in \mathbb{N}^*, i \neq j$, there exists a nonempty set $\mathbb{N}^{**} \subset \mathbb{N}^*$ such that for all $(i,j) \in \mathbb{N}^{**}, i \neq j$, $(1 + N\beta_{ijc\prime})$ is strictly negative, i.e., there exists a strictly positive constant b^{**} such that $(1 + N\beta_{ijc\prime}) < -b^{**},\ \forall (i,j) \in \mathbb{N}^{**}, i \neq j$.

We now define a subspace $\boldsymbol{\Upsilon}$ as $\boldsymbol{\Upsilon} := \big(\boldsymbol{q}_{ij} - \boldsymbol{q}_{ijc} = 0,\ \forall\, (i,j) \in \mathbb{N} \setminus \mathbb{N}^{**}\big) \cap \big(\boldsymbol{q}_{ijc}^T (\boldsymbol{q}_{ij} - \boldsymbol{q}_{ijc}) = 0,\ \forall (i,j) \in \mathbb{N}^*, i \neq j\big)$. In the subspace $\boldsymbol{\Upsilon}$, we have

$$\bar{V}_c = \frac{1}{2} \sum_{(i,j)\in\mathbb{N}^{**}} \|\boldsymbol{q}_{ij} - \boldsymbol{q}_{ijc}\|^2, \quad \dot{\bar{V}}_c = -k \sum_{(i,j)\in\mathbb{N}^{**}} (1 + N\beta_{ijc\prime}) \|\boldsymbol{q}_{ij} - \boldsymbol{q}_{ijc}\|^2 \geq 2kb^{**} \bar{V}_c \tag{43}$$

where we have used $(1 + N\beta_{ijc\prime}) < -b^{**},\ \forall (i,j) \in \mathbb{N}^{**}, i \neq j$. Since the set $\mathbb{N}^{**}$ is nonempty, (43) implies that the equilibrium set $\bar{\boldsymbol{\Xi}}_c$ is unstable by Chetaev's Theorem (Theorem 4.3 in [15]). This implies the desired result that the equilibrium set $\boldsymbol{\Xi}_c$ is unstable. We can further explore instability of the equilibrium set $\boldsymbol{\Xi}_c$ based on (43) as follows. From (43), we have

$$\sum_{(i,j)\in\mathbb{N}^{**}} \|\boldsymbol{q}_{ij}(t) - \boldsymbol{q}_{ijc}\| \geq \sum_{(i,j)\in\mathbb{N}^{**}} \|\boldsymbol{q}_{ij}(t_0) - \boldsymbol{q}_{ijc}\| e^{kb^{**}(t-t_0)}, i \neq j,\ t \geq t_0 \geq 0. \tag{44}$$

Now assume that the equilibrium set $\boldsymbol{\Xi}_c$ is stable, i.e., $\lim_{t\to\infty} \|\boldsymbol{q}_i(t) - \boldsymbol{q}_{ic}\| = d_i, \forall i \in \mathbb{N}$ with d_i a nonnegative constant. Note that $\mathbb{N}^{**} \subset \mathbb{N}$, we have $\lim_{t\to\infty} \|\boldsymbol{q}_i(t) - \boldsymbol{q}_{ic}\| = d_i, \forall i \in \mathbb{N}^{**}$, which implies that $\lim_{t\to\infty} \sum_{(i,j)\in\mathbb{N}^{**}} \|\boldsymbol{q}_{ij}(t) - \boldsymbol{q}_{ijc}\| = d^{**}, \forall (i,j) \in \mathbb{N}^{**}, i \neq j$ with d^{**} a nonnegative constant, since $\boldsymbol{q}_{ij} = \boldsymbol{q}_i - \boldsymbol{q}_j$ and $\boldsymbol{q}_{ijc} = \boldsymbol{q}_{ic} - \boldsymbol{q}_{jc}$. This contradicts (44) for the case $\sum_{(i,j)\in\mathbb{N}^{**}} \|\boldsymbol{q}_{ij}(t_0) - \boldsymbol{q}_{ijc}\| \neq 0$, since the right hand side of (44) is divergent (so does the left hand side). For the case $\sum_{(i,j)\in\mathbb{N}^{**}} \|\boldsymbol{q}_{ij}(t_0) - \boldsymbol{q}_{ijc}\| = 0$, there would be no contradiction. However this case is never observed in practice since the ever-present physical noise would cause $\|\boldsymbol{q}_{ij}(t^*) - \boldsymbol{q}_{ijc}\|$ for some $(i,j) \in \mathbb{N}^{**}, i \neq j$ to be different from 0 at the time $t^* \geq t_0$. Proof of Theorem 6 is completed.

Bibliography

[1] A. Das, R. Fierro, V. Kumar, J. Ostrowski, J. Spletzer, and C. Taylor, "A vision based formation control framework," *IEEE Transactions on Robotics and Automation*, vol. 18, no. 5, pp. 813–825, 2002.

[2] J. Hu and G. Feng, "Distributed tracking control of leader-follower multi-agent systems under noisy measurement," *Automatica*, vol. 46, no. 8, pp. 1382–1387, 2010.

[3] M. Egerstedt and X. Hu, "Formation constrained multiagent control," *IEEE Transactions on Robotics and Austomation*, vol. 17, no. 6, pp. 947–951, 2001.

[4] T. Balch and R. C. Arkin, "Behavior-based formation control for multirobot teams," *IEEE Transactions on Robotics and Automation*, vol. 14, no. 6, pp. 926–939, 1998.

[5] R. T. Jonathan, R. W. Beard, and B. Young, "A decentralized approach to formation maneuvers," *IEEE Transactions on Robotics and Automation*, vol. 19, no. 6, pp. 933–941, 2003.

[6] H. G. Tanner and A. Kumar, "Towards decentralization of multi-robot navigation functions," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, (Barcelona, Spain), pp. 4132–4137, 2005.

[7] K. D. Do, "Bounded controllers for formation stabilization of mobile agents with limited sensing ranges," *IEEE Transactions on Automatic Control*, vol. 52, no. 3, pp. 569–576, 2007.

[8] E. Rimon and D. E. Koditschek, "Exact robot navigation using artificial potential functions," *IEEE Trans. Robot. and Automat.*, vol. 8, no. 5, pp. 501–518, 1992.

[9] D. M. Stipanovic, G. Inalhan, R. Teo, and C. J. Tomlin, "Decentralized overlapping control of a formation of unmanned aerial vehicles," *Automatica*, vol. 40, no. 8, pp. 1285–1296, 2004.

[10] R. Olfati-Saber, "Flocking for multi-agent dynamic systems: algorithms and theory," *IEEE Transactions on Automatic Control*, vol. 51, no. 3, pp. 401–420, 2006.

[11] M. Ji and M. Egerstedt, "Distributed coordination control of multi-agent systems while preserving connectedness," *IEEE Transactions on Robotics*, vol. 23, no. 4, pp. 693–703, 2007.

[12] M. M. Zavlanos and G. J. Pappas, "Potential fields for maintaining connectivity of mobile networks," *IEEE Transactions on Robotics*, vol. 23, no. 4, pp. 812–816, 2007.

[13] D. V. Dimarogonas and K. J. Kyriakopoulos, "Connectedness preserving distributed swarm aggregation for multiple kinematic robots," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1213–1222, 2008.

[14] K. D. Do, "Output-feedback formation tracking control of unicycle-type mobile robots with limited sensing ranges," *Robotics and Autonomous Systems*, vol. 57, pp. 34–47, 2009.

[15] H. Khalil, *Nonlinear Systems*. Prentice Hall, 2002.

[16] M. Krstic, I. Kanellakopoulos, and P. Kokotovic, *Nonlinear and Adaptive Control Design*. New York: Wiley, 1995.

[17] K. D. Do and J. Pan, *Control of Ships and Underwater Vehicles: Design for Underactuated and Nonlinear Marine Systems*. Springer, 2009.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 645-660

A Multimodel Approach for Complex Systems Modeling based on Classification Algorithms

N. Elfelly, J-Y Dieulot, M. Benrejeb, P. Borne

Nesrine Elfelly, Jean-Yves Dieulot,
Mohamed Benrejeb, Pierre Borne
Université des Sciences et Technologies de Lille (USTL),
Ecole Polytechnique de lille,
Ecole Nationale d'Ingénieurs de Tunis -LARA Automatique (ENIT),
Ecole Centrale de Lille (EC Lille)
Laboratoire d'Automatique, Génie Informatique et Signal
Ecole Centrale de Lille, Cité scientifique BP 48
59651 Villeneuve d'Ascq Cedex, France
E-mail:nesrine.elfelly@ed.univ-lille1.fr,
jean-yves.dieulot@polytech-lille.fr,
mohamed.nerejeb@ec-lille.fr,pierre.borne@ec-lille.fr

Abstract: In this paper, a new multimodel approach for complex systems modeling based on classification algorithms is presented. It requires firstly the determination of the model-base. For this, the number of models is selected via a neural network and a rival penalized competitive learning (RPCL), and the operating clusters are identified by using the fuzzy K-means algorithm. The obtained results are then exploited for the parametric identification of the models. The second step consists in validating the proposed model-base by using the adequate method of validity computation. Two examples are presented in this paper which show the efficiency of the proposed approach.
Keywords: complex systems , multimodel , system modeling, classification

1 Introduction

The multimodel approach has been recently developed in several science and engineering domains, with typical applications in the mechanical and chemical engineering areas, with application to modelling, control and/or fault detection e.g. [1–3]. It was introduced as an efficient and powerful method to cope with modelling and control difficulties when complex non linear and/or uncertain processes are concerned. The multimodel approach assumes that it is possible to replace a unique non linear representation by a combination of simpler models thus building a so-called model-base. Usually, each model of this base describes the considered process at a specific operating point. The interaction between the different models of the base through normalized activation functions allows the modelling of the global non-linear and complex system. Therefore, the multimodel approach aims at lowering the system complexity by studying its behavior under specific conditions. The multimodel principle is given in figure 1.

The different models of the base could be of different structures and orders but no model can represent the system in its whole operating domain. The decision unit allows the estimation of the weight of each model and thus the selection of the most relevant models at each time. As for the output unit, controlled by the decision unit, it allows the computation of the multimodel output which is obtained by the contribution of the different models' outputs. In spite of its success in many fields (academic, biomedical, etc), the multimodel approach remains confronted with several difficulties, in particular the determination of the number and parameters of the different models representative of the system and the choice of an adequate method of validities computation used for multimodel output deduction. Last years, many authors [4–7] have been

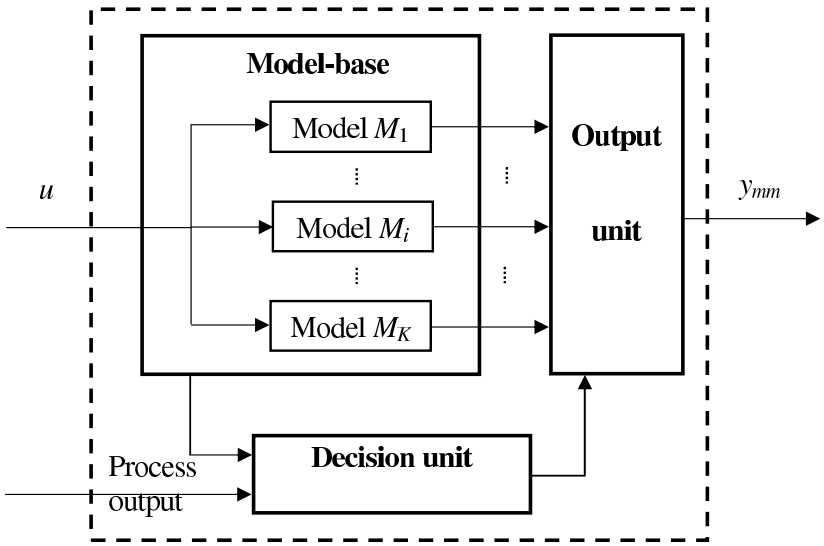


Figure 1: Multimodel approach principle.

interested by this approach. The main differences between the proposed studies are the selected method for models identification and the type of models. Linear models are mainly used, e.g. fuzzy Takagi-Sugeno models [8, 9] which are often obtained using linearization methods. However, the multimodel identification is more difficult to work out when the models of the base should be determined from only input/output data. Some results are given in [10]. In the latter case, the number of models must firstly be determined. Then, the data are classified and the models' parameters are estimated. Last, the validities of the different models are computed according to the adequate method. The issue of the selection of the appropriate method of validities' estimation was discussed in [11]. Many authors propose to apply classification algorithms in order to handle a set of dynamical models. For example, neural networks have been used to represent and control complex systems [7, 12–15]. In another hand, thanks to their ability to classify data and their simplicity, K-means algorithms have proved to be efficient for data clustering e.g. [16, 17]. In short, whereas many architectures using multiple models and neural networks have been proposed, there has not been much work on clustering techniques, based on neural networks and K-means algorithms [18], applied to traditional multimodel representation using only input/output data. The most tedious issues are related to the model-base size and the clustering procedure which aims to the determination of the operating domains. This paper thus proposes a multimodel approach for complex processes modelling based on classification algorithms. The efficiency of the proposed study is illustrated by two examples for which simulations are proposed. The issue of the determination of the number of models in the base is solved through the application of the Rival Penalized Competitive Learning (RPCL), which is an extension of the work in [19] . In the following sections, the different steps of model-base building are detailed knowing that only input/output data are available.

2 Model-base determination using classification algorithms

The proposed approach allows the determination of the number, structure and parameters of the different models of the base. Firstly, a neural network and a rival penalized competitive learning are used for the selection of the adequate number of models. The second step consists in clustering system data by using fuzzy K-means. The classification results are then used for

the parametric identification of the models.

2.1 Determination of the number of models with a Rival Penalized Competitive Learning (RPCL)

The proposed approach allows the construction of the model-base by using two clustering algorithms. The application of this approach requires firstly the determination of the number of models which will be handled by using a Rival Penalized Competitive Learning (RPCL) [20]. Secondly, the application of the fuzzy K-means algorithm for data clustering and then the characterization of the different base-models is carried out by exploiting the clustering results. Finally, the validation of the modeling strategy is considered by the use of an adequate method for validity computation allowing the generation of the multimodel output, compared to the real system output for a different set of inputs.

The models number determination requires experimental data obtained by applying an appropriate input sequence. In order to generate the different operating domains of the process, the measurements must be merged into a set of clusters by the use of a classification algorithm with unsupervised learning. Most existing clustering algorithms [21, 22] do not handle the selection of the appropriate number of clusters, which is, however, essential to the estimation and classification performance in the multimodel approach when no information is available about the operating domains and their number. However, many experimental results have shown that the RPCL algorithm automatically allocates an appropriate number of units for an input data set when they are used for clustering. The selection of the number of models in the multimodel representation requires that the excitation signal must be rich enough (a sine curve for example with the adequate frequency and amplitude added to a random signal) to take into consideration the non-linear aspect of the considered process. For tackling the issue of determination of this number, via input-output data, it is proposed to apply the learning algorithm called RPCL which allows the selection of the adequate number of operating clusters for an input data set. Thus, the extra units are gradually driven far away from the distribution of the data set when the number of units is larger than the real number of clusters in the input data set.

RPCL is an unsupervised learning strategy (proposed by Xu [23] and renewed by Tambe [24]), that automatically determines the optimal number of clusters. The principle underlying RPCL can be considered as an extension of the competitive learning based on Kohonen rule [25]. Its specificity lies in the modification, for each input vector, not only of the winner weights, but also of the weights of its rival (called second winner) so that the rival will be moved or penalized. The rate, at which the rival is penalized, is much smaller than the learning rate, e.g. [26].
Given a competitive learning neural network (Figure 2), i.e. a layer of units with the output u_i of each unit and its weight vector w_i for $i = 1 \dots N$; N is the number of output units, the RPCL algorithm can be described by the following steps.

1. Initialize weight vectors w_i randomly.

2. Take a sample x from a data set D, and for $i = 1 \dots N$, let

$$u_i = \begin{cases} 1 & \text{if} \quad i = c; \\ -1 & \text{if} \quad i = r; \\ 0 & \text{otherwise;} \end{cases} \tag{1}$$

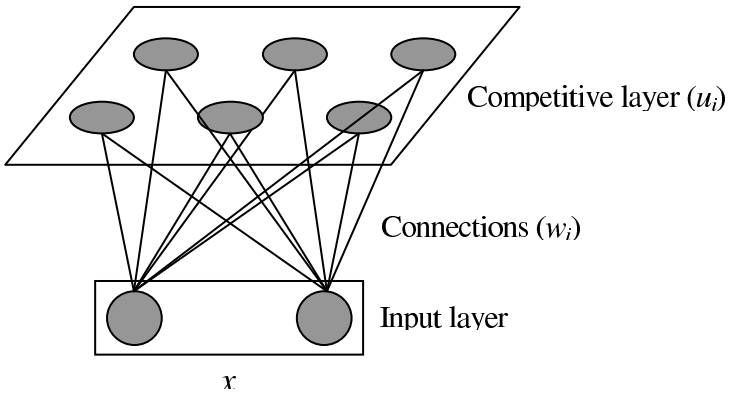


Figure 2: Competitive learning neural network.

with:

$$\gamma_c \|x - w_c\|^2 = \min_j \gamma_j \|x - w_j\|^2; \tag{2a}$$

$$\gamma_r \|x - w_r\|^2 = \min_{j \neq c} \gamma_j \|x - w_j\|^2; \tag{2b}$$

$\| * \|$: Euclidean distance;
c: index of the unit which wins the competition (winner);
w_c: weight vector of the winner;
r: second winner (rival) index;
w_r: weight vector of the rival;
γ_j: conscience factor (relative winning frequency) used to reduce the winning rate of the frequent winners. It is useful to develop a set of equiprobabilistic features or prototypes representing the input data. γ_j is calculated as follows [27]:

$$\gamma_j = \frac{n_j}{\sum_{i=1}^{N} n_i}; \tag{3}$$

where n_j refers to the cumulative number of occurrences the node j has won the competition ($u_j = 1$).

3. Update the weight vectors as follows:

$$w_j(k+1) = w_j(k) + \Delta w_j; \tag{4}$$

with:

$$\Delta w_j = \begin{cases} \alpha_c(k)(x - w_j(k)) & \text{if} \quad u_j = 1; \\ -\alpha_r(k)(x - w_j(k)) & \text{if} \quad u_j = -1; \\ 0 & \text{otherwise;} \end{cases} \tag{5}$$

$0 \leq \alpha_c(k)$ and $0 \leq \alpha_r(k) \leq 1$ are respectively the winner learning rate and the rival de-learning rate. In practice, the rates are fixed small numbers or depend on time (starting from not so small initial values and then reduced to zero in some way) with $\alpha_c(k) \gg \alpha_r(k)$ at each step. Several empirical functions have been proposed for the update of the learning and de-learning rates [27, 28].

4. Repeat steps 2 and 3 until the whole learning process has converged.

Referring to the learning results, only the units enclosed within the data set should be considered and the number of clusters could be deduced as equal to the number of the selected units.

2.2 Identification of the operating clusters

After the selection of the appropriate number of clusters, let us consider the problem of splitting up the measurements to generate the different operating domains from which the base-models will be identified.

Data classification by using fuzzy K-means algorithm

The fuzzy K-means classification algorithm has been chosen for data classification, according to its performance and easy working out.

Fuzzy K-means algorithm

Fuzzy K-means algorithm (developed by Dunn [29] and improved by Bezdek [30]) is a data clustering technique which allows each data point to belong to more than one cluster with different membership degrees (between 0 and 1) and vague or fuzzy boundaries between clusters. The aim of this method is to find an optimal fuzzy K-partition and corresponding prototypes minimizing the following objective function:

$$J_m = \sum_{i=1}^{M}\sum_{j=1}^{K} u_{ij}^m \|x_i - c_j\|^2, \;\; 1 \leq m < \infty; \tag{6}$$

with:
$\| * \|$: any norm expressing the similarity between any measured data and a cluster centre;
m: weighting exponent (real number greater than 1) which is a constant that influences the membership values;
u_{ij}: degree of membership of x_i to cluster j, such as $u_{ij} \in [0,1]$, $\sum_{j=1}^{K} u_{ij} = 1 \;\forall i$ and $0 < \sum_{i=1}^{M} u_{ij} < M \;\forall j$;
x_i: i^{th} data point;
c_j: center vector (node) of the cluster j;
M: number of observations;
K: number of clusters ($2 \leq K < M$).

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j [31].
This procedure will stop when $\max\limits_{i,j}\{|u_{ij}(k+1) - u_{ij}(k)|\} < \varepsilon$ and converges to a local minimum of J_m, where ε is a termination criterion belonging to [0,1] and k the iteration step.

The algorithm is composed of the following steps.

1. Initialize the matrix $U = [u_{ij}]$, $U(0)$.
2. At k-step, calculate the centers vectors $c_j; j = 1 \ldots K$:

$$c_j = \frac{\sum_{i=1}^{M} u_{ij}^m \, x_i}{\sum_{i=1}^{M} u_{ij}^m}. \tag{7}$$

3. Update the matrix of membership degrees $U(k)$ according to the new centers positions $U(k+1)$:

$$u_{ij} = \left[\sum_{l=1}^{K} \left(\frac{\|x_i - c_j\|}{\|x_i - c_l\|}\right)^{\frac{2}{m-1}}\right]^{-1}. \tag{8}$$

4. While $\max_{i,j}\{|u_{ij}(k+1) - u_{ij}(k)|\} \geq \varepsilon$, return to step 2.

In our study, the final clustering result is obtained by considering that a given data point belongs to the cluster to which it presents the greatest membership degree.

Parametric identification of the models in the base

The application of the clustering algorithm results in some repartition of the data set. Each cluster is represented by a set of input/output measurements which will be exploited for the identification of the different models in the base. The application of the classification algorithm results in some repartition of the data set. Each cluster is represented by a set of input/output measurements which will be exploited for the identification of the different models in the base. For this, the models orders are first estimated by using the so-called instrumental determinants' ratio-test. This method is mainly based on the conditions concerning a matrix called "information matrix" which contains the input/output measurements [32]. This matrix is described as follows:

$$Q_m = \frac{1}{N_{ob}} \sum_{k=1}^{N_{ob}} \begin{bmatrix} u(k) \\ u(k+1) \\ \vdots \\ u(k-m+1) \\ u(k+m) \end{bmatrix} \begin{bmatrix} y(k+1) & u(k+1) & \cdots & y(k+m) & u(k+m) \end{bmatrix}; \quad (9)$$

where N_{ob} is the number of observations. The Instrumental Determinants' Ratio (IDR) is given by:

$$IDR(m) = \left| \frac{det(Q_m)}{det(Q_{m+1})} \right|. \quad (10)$$

For each value of $m, (m \geq 1)$ the determination procedure of the order consists in building the matrices Q_m and Q_{m+1} and in evaluating the ratio $IDR(m)$; the retained order m is the value for which the ration $IDR(m)$ quickly increases for the first time.
Given those different orders of models, the parametric identification issue consists in calculating the values of the parameters of the corresponding model-equation, given several experimental measures which describe the dynamic behavior of the model. For this, the Recursive Least-Squares method (RLS) [32] is applied to achieve the parameters estimation.

3 Validity computation and validation of the proposed modelling scheme

The steps described in the previous paragraphs allow the design of the model-base. The purpose is to test the efficiency of the proposed modeling. For this, a validation step is worked out for some inputs different from those used for clustering. Then, the multimodel output y_{mm}, computed and compared to the real output of the studied process, is obtained through a fusion of the K models' outputs y_i weighted by their respective validity indexes v_i, as illustrated by the

system 11 and figure 3.

$$y_{mm}(k) = \sum_{i=1}^{K} y_i(k) v_i(k); \tag{11a}$$

$$\sum_{i=1}^{K} v_i(k) = 1. \tag{11b}$$

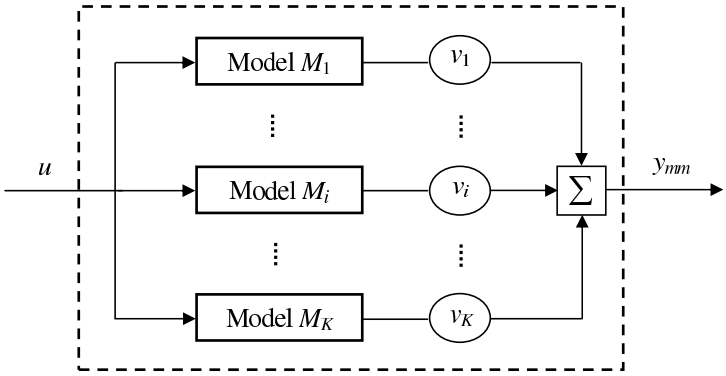


Figure 3: Fusion principle.

The validity index is a real number belonging to the interval $[0, 1]$. It represents the relevance degree of each model calculated at each instant. In the literature, several methods have been proposed to deal with the validity issue. In our study, the residues' approach is adopted for the calculation of validities. This method is based on the distance measurement between the outputs of the process and of the considered model. For example, the residue can be given by the following expression:

$$r_i = |y - y_i|; \qquad i = 1, \ldots, K; \tag{12}$$

with:
y: process output;
y_i: output of the model M_i.

If this residue value is equal to zero, the corresponding model M_i represents perfectly the process at that time; if not, the model represents the process partially.
Between the methods proposed for the calculation of validities [4, 33], only the simple and the reinforced validities approaches are here considered. In general, the expression of the validities is given by:

$$v_i = 1 - r\prime_i \,; \tag{13}$$

where $r\prime_i$ represent the normalized residues and are given by:

$$r\prime_i = \frac{r_i}{\sum_{j=1}^{K} r_j} \,. \tag{14}$$

Simple validities: the normalized simple validities v_i^{simp} are defined so that their sum must be equal to 1 at each time:

$$v_i^{simp} = \frac{v_i}{K - 1} \,. \tag{15}$$

Reinforced validities for this type of validities, the reinforcement expression $v_i^{\prime renf}$ is introduced as:

$$v_i^{\prime renf} = v_i \prod_{j=1, j\neq i}^{K} (1 - v_j)\,. \tag{16}$$

The normalized reinforced validities v_i^{renf} could be written as follows:

$$v_i^{renf} = \frac{v_i^{\prime renf}}{\sum_{j=1}^{K} v_j^{\prime renf}}\,. \tag{17}$$

The comparative study between the two considered validities [11] has shown that the selection of the suitable approach depends on clustering results i.e. the clusters structure and repartition. In fact, when there are important variations in the same cluster and when an overlapping between clusters occurs, it is worth to use the simple validities' method since it takes account of different models' outputs referring to the expression 15. In this case, no model could represent ideally the process at any time. But when the clusters present very few variations and are well separated, the reinforced validities' method is better adapted. Thanks to the reinforcement expression 16, the application of this method promotes the contribution of the most dominant model which represents at best the process behavior.

Validation of the proposed modeling scheme

Once the appropriate method of validity computation selected, the validation of the global modeling scheme is carried out through a comparison between the real and the multimodel outputs for different input sequences.

4 Simulation examples

4.1 Modelling of a mechanical manipulator

In order to underline the interest and the efficiency of the proposed approach, a first example of nonlinear system is considered, which consists of a two-link manipulator (Figure 4) [34, 35] that can be described by the following equations:

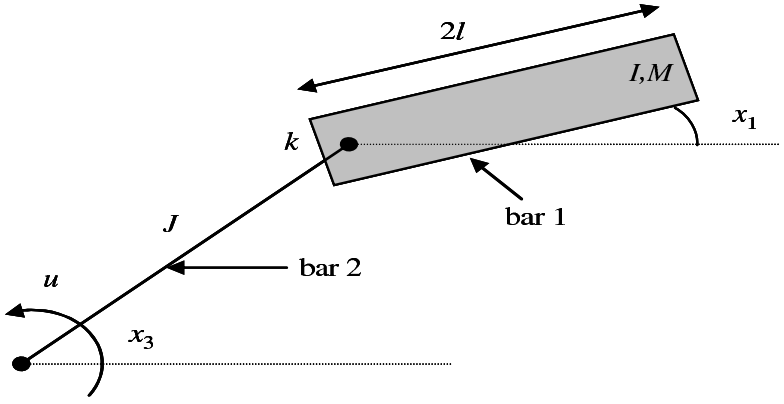


Figure 4: Mechanical system (simulation example).

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = \frac{-Mgl}{I} sin(x_1) - \frac{Mgl}{I}(x_1 - x_3), \\ \dot{x}_3 = x_4, \\ \dot{x}_4 = \frac{k}{J}(x_1 - x_3) + \frac{1}{J}u. \end{cases} \tag{18}$$

The considered system is composed of two rotating bars, where:

- x_1, x_2: angular position and angular velocity of bar 1;
- x_3, x_4: angular position and angular velocity of bar 2;
- u: torque applied to bar 2;
- $g = 9.8\ m \cdot s^{-2}$: gravity constant;
- $I = 1\ kg \cdot m^2$: moment of inertia of bar 1;
- $J = 1\ kg \cdot m^2$: moment of inertia of bar 2;
- $l = 1\ m$: half of the length of bar 1;
- $M = 1\ kg$: mass of bar 1;
- $k = 1\ Nm \cdot rad^{-1}$: elastic rigidity at the link between bars 1-2.

The normal form of the nonlinear model of the system can be written as follows:

$$\begin{cases} \dot{z}_1 = z_2, \\ \dot{z}_2 = z_3, \\ \dot{z}_3 = z_4, \\ \dot{z}_4 = a(z) + b(z)u, \end{cases} \tag{19}$$

where $a(z) = \frac{Mgl}{I} sin(z_1) z_2^2 - \frac{Mgl}{I} cos(z_1) z_3 - (\frac{k}{I} + \frac{k}{I}) z_3 - \frac{kMgl}{IJ} sin(z_1)$ and $b(z) = \frac{k}{IJ}$.

The variables z_i are related to the variables x_i through the following equations:

$$\begin{cases} x_1 = z_1, \\ x_2 = z_2, \\ x_3 = z_1 + \frac{1}{k} z_3 + \frac{Mgl}{k} sin(z_1), \\ x_4 = z_2 + \frac{1}{k} z_4 + \frac{Mgl}{k} cos(z_1) z_2. \end{cases} \tag{20}$$

In the remainder of the study, u will be considered as the input and x_1 as the output of the system. First, the system is excited by an adequate signal $u(k)$ in order to collect the measurements $x_1(k)$ and $x_1(k-1)$ at different instants. These numerical data are used for the determination of the appropriate number of operating clusters by using a neural network and the RPCL algorithm. Figure 5, which gives the results of the learning procedure, shows that by considering six units in the input layer, two centres move away from the observation space, which permits to conclude that the adequate number of clusters can be chosen equal to four.

Then, the fuzzy K-means algorithm is carried out in order to select the different operating clusters (Figure 6).

Referring to each of the four data sets relative to the clusters resulting from the implementation of the proposed classification algorithm, the orders and the parameters of the transfer

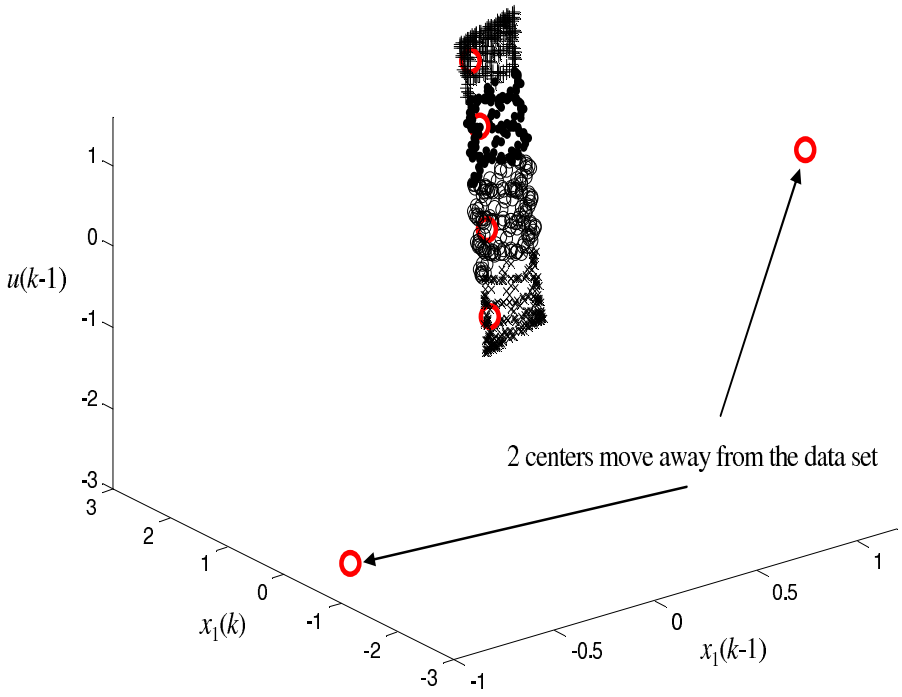


Figure 5: Determination of the number of clusters (N=6).

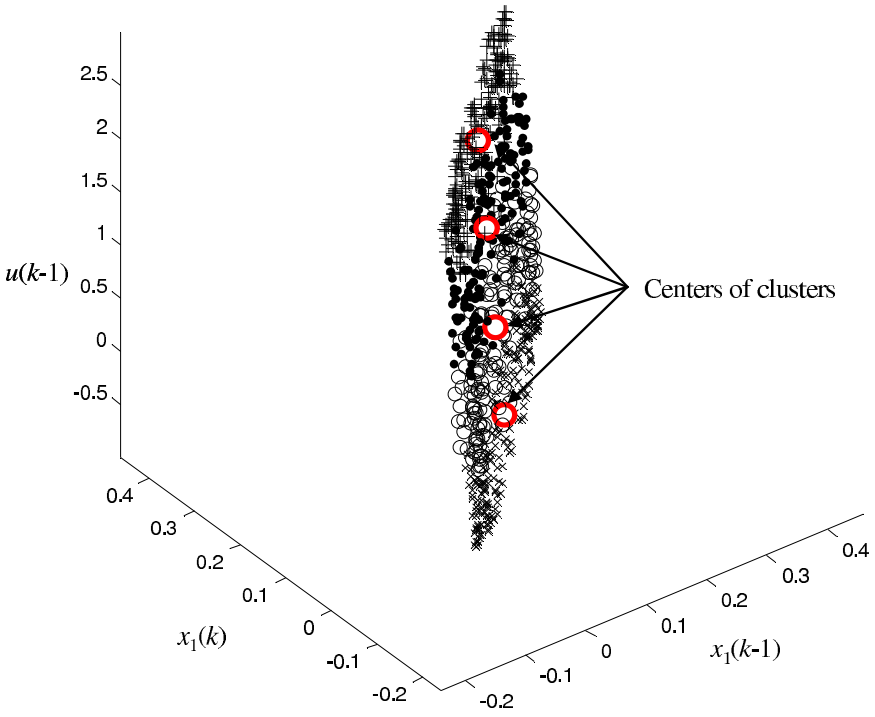


Figure 6: Classification results.

functions relative to the four models of the base are estimated by using respectively the IDR and the RLS methods.

The study presented in [11,19] leads to the selection of the reinforced validities method since the classification results (Figure 6) present well-separated clusters with few variations. In order

to give prominence to the capacity of the identified models to reproduce the operating system in different domains, many input sequences have been considered, two of which are presented in this paper and are described by the following equations:

$$u_1(k) = sin(0.5k). \tag{21}$$

$$u_2(k) = 1 + 0.5sin(0.6k). \tag{22}$$

Referring to the designed model-base, the multimodel output is computed for each input sequence (by fusion of the models' outputs weighted by the reinforced validities) and is then compared to the real system output. The results are given in Figure 7 and Figure 8 where x_1 is the real output, and x_{1mm} the multimodel output obtained by using the fuzzy K-means algorithm.

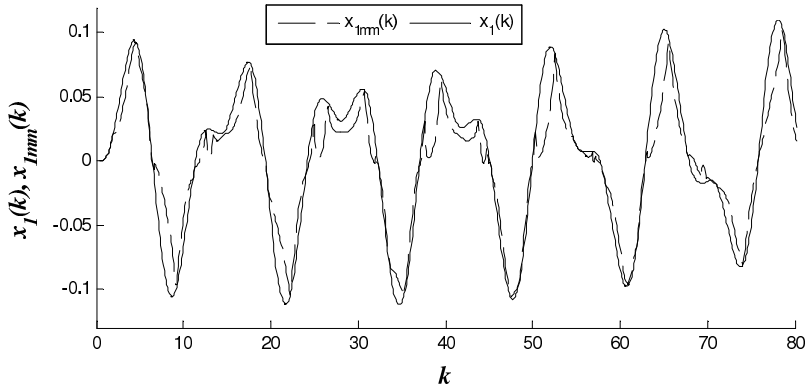


Figure 7: Real and multimodel outputs for the mechanical manipulator (input sequence u_1).

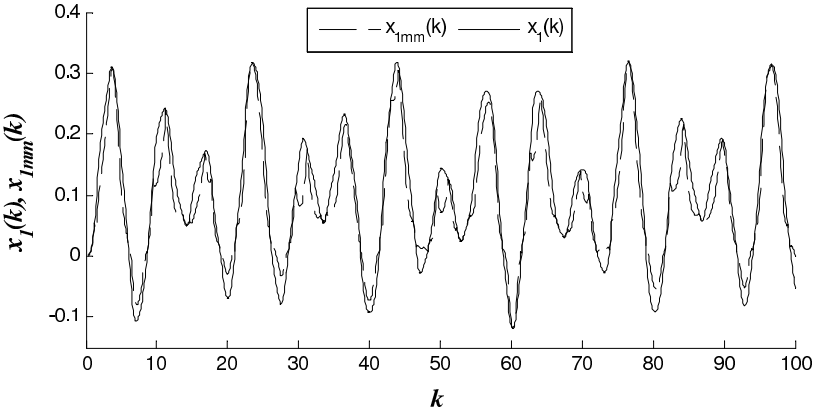


Figure 8: Real and multimodel outputs for the mechanical manipulator (input sequence u_2).

Introducing the NRMSE (Normalized Root Mean Square Error):

$$NRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{N}(y_{mm}-y)^2}{N}}}{y_{max} - y_{min}}, \tag{23}$$

where y_{mm}, y are the multimodel and process output and N is the number of samples; one obtains an accuracy of $NRMSE = 0.001$. A fair comparison was made in a previous paper [39] with an

academic system which was controlled in [6] with a multimodel method where 64 models were proposed instead of only 4 for the method proposed in this paper. The results where only 3 times better for the method in [39] which needed 16 times more models and complicated workout and tuning, but both methods outperformed the results.

It can be noticed that the implementation of the proposed multimodel approach allows an acceptable modelling of the studied system respective to its complexity. In fact, by considering the first input sequence u_1, the multimodel output follows the real output and the error between the two signals shaves over the time. Besides, even with a more complicated system output (obtained by considering the input sequence u_2) with important oscillations and for a longer time interval, the results given by the proposed modelling approach (Figure 8) are satisfactory.

4.2 Modeling of a bioreactor

This second example shows the relevance of our method compared to a "black-box" multimodel technique with a large model base as provided in the work by [6], illustrated on the same benchmark, a bioreactor model. This model has been used in several nonlinear modeling or control issues [6, 36, 37]. Substrate is fed continuously with a constant feedrate F to the well-mixed bioreactor which has a constant volume V. Microbial growth follows a Contois model [38], the microorganisms and substrate concentration, respectively x_1 and x_2, are supposed to be small. The discrete-time mass-balance equations are derived as follows:

$$\begin{cases} x_1(k+1) = x_1(k) + 0.5\frac{x_1x_2}{x_1+x_2} - 0.5u(k)x_1(k), \\ x_2(k+1) = x_2(k) - 0.5\frac{x_1x_2}{x_1+x_2} - 0.5u(k)x_2(k) + 0.05u(k), \\ y(k) = x_1(k) \end{cases} \tag{24}$$

where $y(k)$ is the model output, $u(k) = F/V$ is the dilution rate, normalized to 1. As in the previous example, it is assumed that no a priori information is given on the possible (real) model structure and the system is adequately excited to generate a set of input/output measures which will allow to describe the whole process dynamics. As was explained before, the rival penalized competitive learning algorithm is applied, and the number of models was found to be equal to 10. Experimental measurements are then classified by using the fuzzy k-means algorithm. Based on the learning results, where small overlapping between clusters was found, the reinforced validities were embedded in the multimodel structure. In terms of modelling accuracy, the obtained results were found to be quite similar to those given in [6] where 196 models were needed to represent the system dynamics . This enlightens the interest of the proposed approach as a high number of models yields a considerable computational burden. Figure 9 provides the real output y and the multimodel output y_{mm} by considering the following input sequence: $u(k) = 0.5 + 0.1sin(2k)$. The results are confirmed through another input sequence u_2 shown on figure 10 and figure 11 shows a good agreement between multimodel and real outputs.

5 Conclusion

In this paper, a new approach for complex systems modelling is proposed and an experimental validation is presented. This approach is applicable when dealing with complex, strongly nonlinear and uncertain systems. It allows the determination of the base of models, which are representative of the system in different operating domains, by using two classification algorithms and two methods of structural and parametric identification. The issue of selection of the models' number is solved by using a neural network and a Rival Penalized Competitive Learning (RPCL). Once this number determined, the data collected on the system are then clustered by using

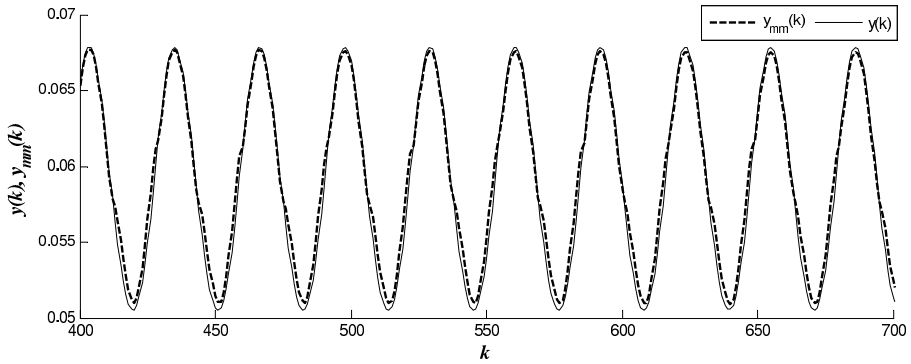


Figure 9: Real and multimodel outputs for the bioreactor model (input sequence u_1).

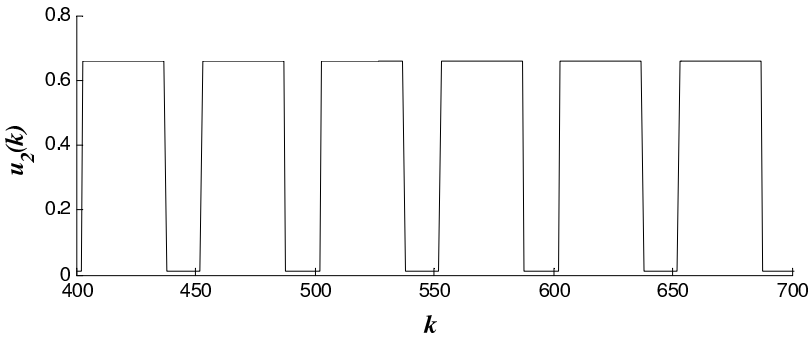


Figure 10: Bioreactor input sequence u_2.

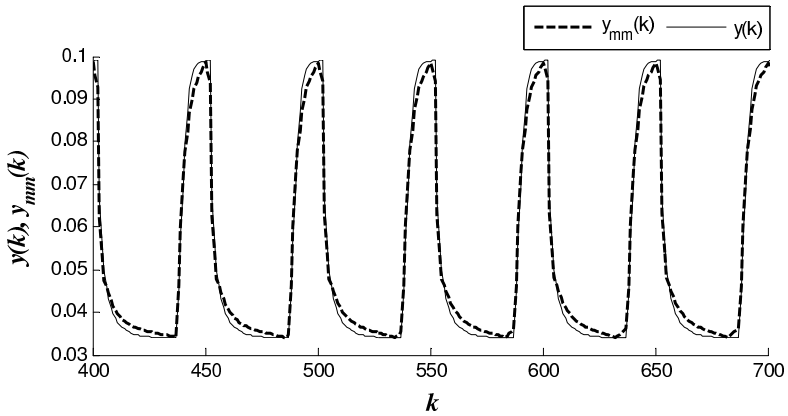


Figure 11: Real and multimodel outputs for the bioreactor model (input sequence u_2).

the fuzzy K-means algorithm. The classification results are exploited for the identification of models. Finally, a validation procedure is worked out in order to demonstrate the ability of the proposed modelling structure to reproduce the system response in different operating domains. The proposed approach has been implemented and applied for a complex mechanical system and for a bioreactor. The obtained results seem to be interesting and prove the efficiency of the proposed modelling strategy.

Bibliography

[1] Alavandar S., Nigam M.J., Neuro-Fuzzy based Approach for Inverse Kinematics Solution of Industrial Robot Manipulators, *INT J COMPUT COMMUN*, 3(3):224-234, 2008.

[2] J. M. Böling and D. E. Seborg and J. P. Hespanha, Multi-model adaptive control of a simulated pH neutralization process, *Control Engineering Practice*, Vol. 15, pp. 663-672, 2007.

[3] M. Rodrigues and D. Theilliol and M. Adam-Medina and D. Sauter A fault detection and isolation scheme for industrial systems based on multiple operating models, *Control Engineering Practice*,16,225-239,2008.

[4] F. Delmotte, L. Dubois, P. Borne, A General Scheme for Multi-Model Controller using Trust, *Mathematics and Computers in Simulation*, Vol. 41, pp. 173-186, 1996.

[5] T. A. Johansen, B. A. Foss, Editorial: Multiple model approaches to modelling and control, *International Journal of Control*, Vol. 72, pp. 575, 1999.

[6] J. Cho and J. C. Principe and D. Erdogmus and M. A. Motter, Quasi-sliding mode control strategy based on multiple-linear models, *Neurocomputing*, 70, 960-974,2007.

[7] I. S. Baruch and R. B. Lopez and J-L. Olivares and J-M. Flores, A fuzzy-neural multi-model for nonlinear systems identification and control, *Fuzzy sets and systems*, 159, 2650-2667, 2008.

[8] T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modelling and control, *IEEE Transactions on Systems Man and Cybernetics*, Vol. 15, pp. 116-132, 1985.

[9] M. Benrejeb, D. Soudani, A. Sakly, P. Borne, New discrete TSK fuzzy systems characterization and stability domain, *INT J COMPUT COMMUN*, 1(4):9-19, 2006.

[10] P. Borne and M. Benrejeb, On the Representation and the Stability Study of Large Scale Systems, *INT J COMPUT COMMUN*, 3(S):55-66, 2008.

[11] N. Elfelly and J-Y. Dieulot and P. Borne, A neural approach of multimodel representation of complex processes, *INT J COMPUT COMMUN*,3(2):149-160, 2008.

[12] M. Ronen and Y. Shabtai and H. Guterman, Hybrid model building methodology using unsupervised fuzzy clustering and supervised neural networks, *Biotechnology and Bioengineering*, 77, 420-429, 2002.

[13] W. Yu, Multiple recurrent neural networks for stable adaptive control, *Neurocomputing*, 70, 430-444, 2006.

[14] Y. Fu and T. Chai, Nonlinear multivariable adaptive control using multiple models and neural networks, *Automatica*, 43, 1101-1110, 2007.

[15] G.D. Manioudakis and E.N. Demiris and S.D. Likothanassis, A self-organized neural network based on the multi-model partitioning theory, *Neurocomputing*, 37, 1-29, 2001.

[16] D. Dembélé and P. Kastner, Fuzzy C-means method for clustering microarray data, *Bioinformatics*,19, 973-980, 2003.

[17] P. Hore and L.O. Hall and D.B. Goldgof and W. Cheng, Online Fuzzy C Means, *Annual Meeting of the North American Fuzzy Information Processing Society, NAFIPS 2008*,1-5, 2008.

[18] Z.K. Xue and S.Y. Li, Multi-model modelling and predictive control based on local model networks, *Control and Intelligent Systems*, 34, 105-112, 2006.

[19] N. Elfelly and J.-Y. Dieulot and M. Benrejeb and P. Borne, A new approach for multimodel identification of complex systems based on both neural and fuzzy clustering algorithms, *Engineering Applications of Artificial Intelligence*, 23, 1064-1071, 2010.

[20] L. Xu and A. Krzyzak and E. Oja, Unsupervised and supervised classification by rival penalized competitive learning, *Pattern Recognition*, 11, 496-499, 1992.

[21] A.K. Jain and R.C. Dubes, Algorithms for Clustering Data, *Prentice-Hall Inc., Upper Saddle River, NJ*, 1988.

[22] B. Mirkin, Mathematical Classification and Clustering, *Kluwer Academic Press, Boston-Dordrecht*, 1996.

[23] L. Xu and A. Krzyzak and E. Oja, Rival penalized competitive learning for cluster analysis RBF and curve detection, *IEEE Transactions on Neural Networks*,4, 636-649,1993.

[24] S.S. Tambe and B.D. Kulkarni and P.B. Deshpande, Elements of artificial neural networks with selected application on chemical engineering, and chemical and biological sciences, *Simulation & Advanced Controls Inc., Louisville-KY, USA*, 1996.

[25] T. Kohonen, The self-organizing map, *IEEE Proceedings*, 78, 1464-1480, 1990.

[26] P. Borne and M. Benrejeb and J. Haggège, Les réseaux de neurones, *Editions Technip, Paris, France*, 2007.

[27] T. Murlidharan Nair and C. L. Zheng and J. Lynn Fink and R. O. Stuart and M. Gribskov, Rival Penalized Competitive Learning (RPCL): a topology-determining algorithm for analyzing gene expression data, *Computational Biology and Chemistry*, 27, 565-574, 2003.

[28] I. King and L. Xu and L. Chan, Using rival penalized competitive clustering for feature indexing in Hong Kong's textile and fashion image database, *Proceedings of the IEEE International Joint Conference on Neural Networks*, 1, 237-240, 1998.

[29] J. C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *Journal of Cybernetics*, 3, 32-57, 1973.

[30] J.C. Bezdek, Pattern recognition with fuzzy objective function algorithms, *Plenum Press, New York*, 1981.

[31] S. Nascimento and B. Mirkin and F. Moura-Pires, A fuzzy clustering model of data and fuzzy C-means, *The Ninth IEEE International Conference on Fuzzy Systems*, 1, 302-307, 2000.

[32] R. Ben Abdennour and P. Borne and M. Ksouri and F. M'sahli, Identification et commande numérique des procédés industriels, *Editions Technip, Paris, France*, 2001.

[33] D. J. Leith, W. E. Leithead, Analytic framework for blended multiple model systems using linear local models, *International Journal of Control*, Vol. 72, pp. 605-619, 1999.

[34] J.V. Salcedo and M. Martínez, Design of PDC fuzzy controllers under persistent disturbances and application in mechanical systems, *Advances in Engineering Software*, 39, 937-946, 2008.

[35] C.-W. Chen, Stability conditions of fuzzy systems and its application to structural and mechanical systems, *Advances in Engineering Software*, 37, 624-629, 2006.

[36] J. P. Gauthier and H. Hammouri and S. Othman, A simple observer for nonlinear systems applications to bioreactors, *IEEE Transactions on Automatic Control*, 37, 875-880, 1992.

[37] G. Bastin and D. Dochain, On-line estimation and adaptive control of bioreactors, *Elsevier, Amsterdam*, 1980.

[38] D. Contois, Kinetics of bacterial growth relationship between population density and specific growth rate of continuous cultures, *Journal of Genetic Macrobiol*, 21, 40-50, 1959.

[39] N. Elfelly and J.-Y. Dieulot and M. Benrejeb and P. Borne, Multimodel control design using unsupervised classifiers, *Studies in Informatics and Control*, ISSN 1220-1766, vol. 21 (1), pp. 101-108, 2012.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 661-673

The Research of Differentiated Service and Load Balancing in Web Cluster

A. Gao, Q. Pan, Y. Hu

Ang Gao, Quan Pan,Yansu Hu
School of Automation
Northwest Polytechnical University
Xi'an 710072, China
E-mail: gaoang@nwpu.edu.cn,
quanpan@nwpu.edu.cn, huyansu@gmail.com

Abstract: Differentiated service, as a key solution to meet the heterogenicity of Web clients' QoS requirements, has been widely used to optimize the server utilization without over-providing resources. Based on the relative differentiated service, this paper treats the application of proportional delay as a optimal control problem, and focuses on the cluster-side architecture improvement as well as QoS controller design. A load balancing Web cluster architecture supported differentiated service is proposed and implemented. By system identification and resource optimal control, the front-end dispatcher could adjust the resource quotas assigned to different classes in every single back-end server, and Multi-class based Maximum Idle First load balancing strategy is designed to ensure a fair resource consumption among back-end nodes. As a result, the end-to-end delay is controlled and proportional delay is guaranteed. The experiments demonstrate that no matter using Round-Robin, Least Connection Scheduling or Maximum Idle First load balancing strategy, the proposed resource optimal controller could hold the relationship among different classes. Compared to Round-Robin and Least Connection First Scheduling, Maximum Idle First strategy increases the cluster throughput by 33% and reduces the average delay by 21%.
Keywords: Differentiated Service, Maximum Idle First, Load Balancing, Proportional Delay Guarantee

1 Introduction

With the dramatic explosion of online information, the Internet is undergoing a transition from a data communication infrastructure to a service intergraded utility. The increase of Web applications, Web clients and HTTP requests on the Internet makes the Web server systems often suffer from huge pressure of heavy workload. The Deployment of Web cluster system keeps increasing to meet the demand for availability, scalability and stability of the diversified performance demands of clients.

Web cluster organizes a number of Web servers as a logical entirety to enhance the storage and processing capacity. The cluster also shows a good expansibility, whose capability can be easily tuned by changing the number of back-end server, which are connected by high speed local area network. Clients' HTTP requests are well-proportioned and transparently dispatched to back-end. Server nodes work concurrently and the responsiveness as well as the reliability of Web sites are improved (see [17]).

To take full use of every server's processing resources, a major issue is how to arrange each server node appropriate requests according to its capability. However, the cluster system scale is limited by the financial cost of Web sits and IDC (Internet Digital center), and the load characteristics of Web sites is often affected by the browsing habits, geographic distribution and breaking news. It is impossible to accurately predict the peak load and prepare enough

computing resources. So it is not cost-effective to allocate excessive computing resources for a Web site to accommodate the potential peak. Even the large-scale clusters, there will still be the case of overload. Now, the Internet has become a commercial product, the growth of e-commerce is creating demand for services with financial incentives for the service providers, i.e., the economic transaction is more important than a simple browsing, and the premium users also expect better quality services. So, the other major problem Web cluster system facing is how to meet the Service Level Agreements(SLAs) with their clients without excessively over-provisioning resources (see [21]).

As an initial effort, a feedback control mechanism is designed to achieve Proportional Delay Differentiation Service and Load Balancing (PDDS-LB) in a Web cluster system. First, according to the general Web cluster system framework and the HTTP processing procedure, a load balancing Web cluster architecture is proposed. Second, considering the residue delay[1] is the main factor affecting the users' experience in a Web application, with the aid of system identification and optimal control, we design a feedback controller, which periodically re-allocates the processing resources to keep the residue delay ratio around the set point. Finally, we present Multi-class based Maximum Idle First load balancing strategy(MIF) to achieve efficient and fair resource consumption. Experiment results show that our mechanism is effective, the total throughput increases by 33%, and the average delay reduces by 23%.

2 The Overview of Cluster

2.1 The Architecture and Forwarding Technology

Figure 1 gives the framework of Web cluster.The front-end is called *Dispatcher*, which is the entrance of the cluster system. The clients' HTTP requests first reach dispatcher, then are distributed to back-end servers according to the load balancing strategy. The the dispatcher can select a back-end according to the Request-URL in HTTP Message and/or other information in entity-header fields, such as *User-Agent*, *From*, *Host*, etc. According to the layer of dispatcher, clusters can be divided into three types (see [17]): *L4/2 Cluster-L4 Switcher,L2 Forwarding*; *L4/3 Cluster-L4 Swithcer,L3 Forwarding*; *L7 Cluster- Application Layer Forwarding.*

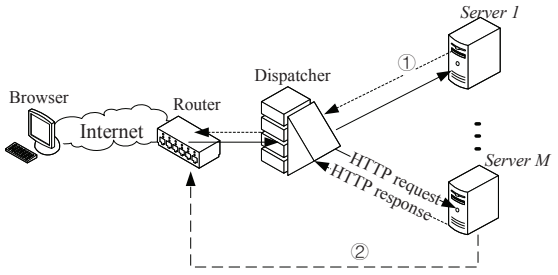


Figure 1: The Processing of HTTP Request in Web Cluster

The dispatcher should establish TCP connection with clients and back-ends concurrently. Although compared with the hardware-accelerated forwarding method used by L4/2 and L4/3 cluster, L7 cluster has the limitation of larger processing overhead. But it is still a promising implementation of Web server cluster (see [11]). It could not only combine L4/2, L4/3 forwarding technology (see [3]), but also take usage of application information to enforce the content-aware dispatcher and combine the priority scheduling with the processing/threads-based Web QoS control scheme (see [18]).

[1]residue delay is consist of connecting time and processing time.

From the view of forwarding technology used by dispatcher, there are also three types of clusters: *Reverse Proxy*, *TCP Splicing* and *TCP Handoff*. The most widely-used is Reverse Proxy which is shown in Figure 1 Arrow ①. The dispatcher distributes HTTP requests to back-ends and transmits the responses back to clients. In order to avoid the redundant data copies and enhance the forwarding efficiency of reverse proxy, TCP Splicing is enforced in operation system kernel. As Arrow ② shown in Figure 1, by means of TCP Handoff protocol, the TCP connection established between dispatcher and clients is transfer to back-ends, and the response is directly sent back to the clients without relaying of dispatcher, which increases the potential throughput of the system.

TCP splicing and TCP Handoff are both proposed to enhance the forwarding efficiency. However, none of them is software-compatible because the implements require the amendments of TCP/IP protocol stack and system kernel in dispatcher and/or back-end servers. For the reasons above, our research focus on L7 cluster whose dispatcher running as Reverse Proxy.

2.2 The Related Work

The current related work mainly focus on the load balancing strategy and algorithm. While the researches of L4 cluster are just on how to uniformly spread the requests. Besides this, L7 cluster considers the contents of the request as well.

Typical L4 load balancing strategies are as follows: Round Robin (RR), i.e. executes $(i+1)$ mod n for every request and selects back-end according to the result; Weighted Round Robin (WRR), i.e. every back-end servers' processing capacity is in accordance with its weight. The larger weight the more times requests be sent; Least Connection Scheduling (LCS) and Weighted Least Connection Scheduling (WLCS), which forward the requests to the server with least active TCP connections; Source Hashing Scheduling (SHS) and Destination Hashing Scheduling (DHS) are statics mapping algorithms, the IP source address or destination address of HTTP request is hashed and mapped to a certain back-end server.

There are two common scheduling strategy used in L7, Client-aware Policy (CAP) (see [4]) and Locality Aware Request Distribution (LARD)(see [13]). CAP is actually a kind of RR scheduling supported client-side classification. Same kind of requests are spreading uniformly without considering their content. LARD is server-status based load balancing strategy, in the threshold of stability, the same URL are forwarding to the same server to achieve a higher hit rate.

However, no matter L4 or L7 cluster research does not concern the issue of differentiated service and the related study is few. [21] gives the Demand-driven Service Differentiation(DDSD) approach, which improves Web Cluster model into a multi-queueing system, such as $M/M/1/\infty$ and $M/G/1/\infty$ models proposed by [19,20]. By selecting Stretch Factor, it treats the re-allocation of cluster resources as a SLAs constrained optimization problem. However, because queueing model is based on the premise of Poisson arrival rate and exponential distribution of processing time, queueing model can not accurately describe the of HTTP traffic feature when the arrival rate and leave rate does not match(see [2, 9]). Dynamic Partitioning (DynamicPart) algorithm is proposed by Casalicchio (see [1]). By feedback control, the back-ends is dynamically servicing different kinds of requests, which transforms a best-effort Web cluster into a QoS-enhanced system. But the performance isolation is enforced at the level of server host, it can only provide coarse-grain control at host-level, which disturbs the on-line extension and the resources utilization.

Our research purpose is to provide a differentiated services to increase the resources utilization of Web cluster at the level of processing/thread. First, the resource consumption of every back-end should be coordinated and load-balanced. Then the resource management and scheduling

should meet the QoS requirement (see [5]).

3 QoS-Enhanced Web Cluster

3.1 PDDS-LB Architechure

The processing/thread-based Web QoS control is used in the researches of single Web server. The processing or thread pool is partitioned into several parts to support the performance isolation. The processing or thread number in each part is called *quota*. The HTTP requests are classified into different business flows (1...N). By adjusting the quotas of every class, the performance isolation and differentiation service are achieved (see [8,14]). The initial ideal of PDDS-LB is a unified scheduling of all back-ends' processings/threades and a fair consumption of different servers' resource to the same flows. As shown in Figure 2, PDDS-LB supports differentiated service in two layers:

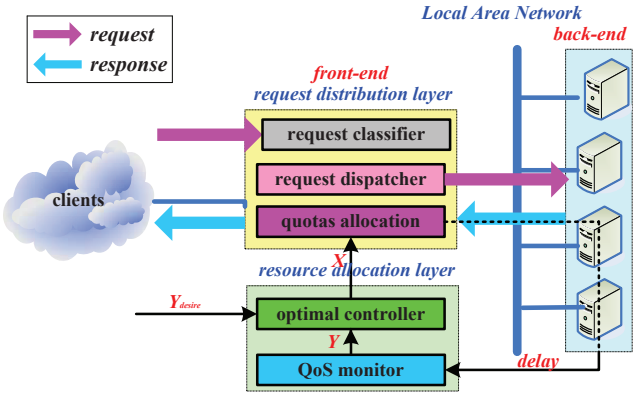


Figure 2: PDDS-LB Architecutre

In the resource allocation layer, the QoS monitor perceives the average residue delay of each class [2] $\ell_i, (1 \leq i \leq N)$ while the back-ends report their status periodically (detailed in Section 5). The optimal controller adjusts every back-ends's quota $c_{i,j}, (1 \leq j \leq M)$ assigned to different business flows.

In the request distribution layer, the requests are separated into different priorities according to the specified strategy and SLAs. The request dispatcher selects a back-end for each request by the load balancing algorithm.

The front-end dispatcher is the key element of the cluster, which establishes TCP connection with back-end and clients simultaneously. Back-end servers deploy the same contents and can service all the requests. Classification strategy back-ends used is consistent with the front-end. Different requests flows passed into the corresponding connecting queueing wait to be served in the manner of First-In-First-Out(FIFO).

Heartbeat is original used to make a high amiability of cluster infrastructure (front-end and back-ends)(see [15,16]). In PDDS-LB, it is referenced as a daemon to provide the communication between client infrastructure. At every heartbeat time, the status reporter sends the status messages to front dispatcher and the quotas of business flows are adjusted according to the optimal controller's output.

[2] residue delay is the sum of connecting time w and processing time χ.

3.2 PDDS-LB Software Component

Considered our previous work (see [8,9]), we implement and deploy PDDS-LB in the Apache plantform. the back-end server is modified into the multi-processing queue architecture. On the dispatcher, *mod_ proxy* (see in [6]) is activated as reverse proxy, and *mod_ proxy_ balancer* is used to translate the request URL into absolute URL in back-end server.

1. *mod_ cluster* is modified to achieve communication between dispatcher and back-end servers. At every heartbeat time, back-end servers initiatively report their status. Meanwhile, *mod_ clsuter* in dispatcher implements the function of resource allocation layer. By optimal control, the refreshed quotas of classes are sent back to back-end servers.

2. *mod_ proxy_ balancer* is modified to enforce the functions of request distribution layer. *mod_ cluster* calculates the busyness of every back-end server according to the their status and URLs are redirected to a appropriate server. Then Apache httpd obtains responses from it.

4 Cluster Resource Management

4.1 System Model Identification

Supposed the cluster consists of M back-end servers, serving N classes of business flows. Every back-end runs $C_j (j = 1, \ldots, M)$ service threads:

$$C_j = \sum_{i=1}^{N} c_{i,j}, \varsigma_i = \sum_{j=1}^{M} c_{i,j}, 1 \le i \le N, 1 \le j \le M, \tag{1}$$

where $c_{i,j}$ is the number of threads assigned to class i in j^{th} back-end, i.e. the quota that server j assigns to class i. ς_i is the number of threads assigned to class i in the whole cluster. At every sampling time, the optimal controller adjusts $c_{i,j}$ to hold Equation (2):

$$\frac{\ell_i}{\ell_l} = \frac{\delta_i}{\delta_l}, 1 \le i \le N, 1 \le l \le N, \tag{2}$$

where δ_i is the class i's priority assigned by SLA. The smaller δ_i, the higher its priority. As shown in Figure 2,the controlled object is threads of the whole cluster. For the constrain of Equation (1), the input has $I = N \times M$ independent variables. At the k^{th} sampling time, the input is:

$$\mathbf{X}(k) = [c_{1,1}(k), c_{1,2}(k), \ldots, c_{1,M}(k), \ldots, c_{N-1,1}(k), c_{N-1,2}(k), ..., c_{N-1,M}(k)]^{\mathrm{T}}.$$

Define $y_i(k)$, $y_{i_{desire}}$ are the *normalized residue delay* and its expected value:

$$y_i(k) = \frac{\ell_i(k)}{\sum_{l=1}^{N} \ell_l(k)}, y_{i_{desire}} = \frac{\delta_i}{\sum_{l=1}^{N} \delta_l}, 1 \le i \le N. \tag{3}$$

Because $\sum_{i=1}^{N} y_i(k) = 1$ and $\sum_{i=1}^{N} y_{idesire} = 1$, the output has $O = N - 1$ independent variables. So let

$$\begin{aligned} \mathbf{Y}(k) &= [y_1(k), y_2(k), \ldots, y_{N-1}(k)]^{\mathrm{T}}, \\ \mathbf{Y}_{desire} &= [y_{1_{desire}}, y_{2_{desire}}, \ldots, y_{N-1_{desire}}]^{\mathrm{T}}, \end{aligned}$$

where $\mathbf{Y}(k)$ is the measurement output. According to the derivation $\mathbf{E}(k)$ between $\mathbf{Y}(k)$ and $\mathbf{Y}_{desire}$, the optimal controller adjusts the threads quotas of every classes to guarantee the relative relationship constant.

Strictly speaking, we require a discrete and nonlinear model for PDDS-LB. However, such a nonlinear model is not amenable to the straight forward theoretical design and analysis (see [12]). SO the following linear model is used to approximate the system. Supposing r-order could be precise enough, the corresponding difference equation is:

$$\mathbf{Y}(k) = \sum_{j=1}^{r} [a_j \mathbf{Y}(k-j) + b_j \mathbf{X}(k-j)], \tag{4}$$

and the Z-domain transformation is:

$$\mathbf{A}(z^{-1})\mathbf{Y}(k) = \mathbf{B}(z^{-1})\mathbf{X}(k) + \varepsilon(k), \tag{5}$$

where $\varepsilon(k) = [\varepsilon_1(k), \varepsilon_2(k), ..., \varepsilon_{N-1}(k)]^T$ is O-order not related white noise sequence with mean zero.

$$\begin{aligned} \mathbf{A}(z^{-1}) &= \mathbf{I} - \mathbf{A}_1 z^{-1} - ... - \mathbf{A}_r z^{-r}, \mathbf{A}_i \in \mathbf{R}^{O\times O}, 0 < i \le r. \\ \mathbf{B}(z^{-1}) &= \mathbf{B}_1 z^{-1} + ... + \mathbf{B}_r z^{-r}, \mathbf{B}_j \in \mathbf{R}^{O\times I}, 0 < j \le r. \end{aligned}$$

Because of observation error and system noise, define $\boldsymbol{\varepsilon}(k) = [\varepsilon_1(k), \varepsilon_2(k), ..., \varepsilon_{n-1}(k)]^{\mathrm{T}}$ be O-order white noise sequence. Equation (5) can be rewritten as:

$$\mathbf{Y}(k+1) = \boldsymbol{\theta}\boldsymbol{\Phi}(k) + \boldsymbol{\varepsilon}(k+1), \tag{6}$$

where $\boldsymbol{\theta} = [\mathbf{B}_1 ..., \mathbf{B}_r, \mathbf{A}_1, ...\mathbf{A}_r], k \ge r-1$, $\boldsymbol{\Phi}(k) = [\mathbf{X}^{\mathrm{T}}(k), ..., \mathbf{X}^{\mathrm{T}}(k-r+1), \mathbf{Y}^{\mathrm{T}}(k), ..., \mathbf{Y}^{\mathrm{T}}(k-r+1)]^{\mathrm{r}}$, $\boldsymbol{\theta} \in R^{O\times[O\times 2\times r]}$.

Recursive least square (RLS) estimate algorithm is used to calculate parameter matrix $\boldsymbol{\theta}$, and the white noise-similarity of pseudo-random sequence is used as impulse to fully stimulate the system . In the system identification experiment, the cluster is composed of 4 back-end server, each of which runs 100 threads serving two classes of business flows. i.e. $N = 2$, $M = 4$,$C_j = 100, j = 1, \ldots, 4$, $\boldsymbol{Y}_{desire} = [y_{1_{desire}}] = [1/3]$. In order to fully stimulate the system, at every sampling time, the quotas assigned to every classes $\boldsymbol{X}(k+1) = [c_{1,1}(k+1), c_{1,2}(k+1), ..., c_{1,4}(k+1)]^{\mathrm{T}}$ are adjusted according to the current value of pseudo-random sequence. The pseudo-random sequence is generated as Equation (7). Set $p = 7$, $q = 12$, the relationship between $\epsilon(k)$ and $\mathbf{X}$ is shown in Table 1.

$$\epsilon(k) = \epsilon(k-p) + \epsilon(k-q) (\mathrm{mod} 4) \tag{7}$$

Table 1: Relation of $\epsilon(k)$ and $\mathbf{X}$ when $p = 7$, $q = 12$

$\epsilon(k)$	$c_{1,j}(k+1)$	$c_{2,j}(k+1)$
0	25	75
1	40	60
2	60	40
3	75	25

The system identification experiment lasts 5000 seconds with the sampling time $T = 30s$ and we got 150 sets of effective data. Supposing $\hat{\boldsymbol{\theta}}_q$ is the estimation of $\boldsymbol{\theta}$ from the former q ($q \ge r-1$)

sampled data. After the $q+1^{th}$ sampling time, $\hat{\boldsymbol{\theta}}_q$ can be revised as:

$$\hat{\boldsymbol{\theta}}_{q+1} = \hat{\boldsymbol{\theta}}_q + \frac{[\mathbf{Y}(k+1) - \hat{\boldsymbol{\theta}}_q \boldsymbol{\Phi}(k)]\boldsymbol{\Phi}^{\mathrm{T}}(k)\mathbf{P}_q}{\lambda + \boldsymbol{\Phi}^{\mathrm{T}}(k)\mathbf{P}_q\boldsymbol{\Phi}(k)}, \tag{8}$$

where

$$\mathbf{P}_{q+1}^{-1} = \mathbf{P}_q^{-1} + [1 + (\lambda - 1)\frac{\boldsymbol{\Phi}^{\mathrm{T}}(k)\mathbf{P}_q\boldsymbol{\Phi}(k)}{(\boldsymbol{\Phi}^{\mathrm{T}}(k)\boldsymbol{\Phi}(k))^2}]\boldsymbol{\Phi}(k)\boldsymbol{\Phi}^{\mathrm{T}}(k),$$

$\mathbf{P}_q$ is covariance matrix and λ is forgetting factor. $\boldsymbol{\Phi}$, $\mathbf{Y}$ are measured by QoS monitor. By selecting an appropriate $\hat{\boldsymbol{\theta}}_0$ and $\mathbf{P}_0$, we could get the estimation of parameter matrix. The criteria for $\hat{\boldsymbol{\theta}}_0$ and $\mathbf{P}_0$ is

$$\begin{cases} \hat{\boldsymbol{\theta}}_0 = \boldsymbol{\iota}, \boldsymbol{\iota} \text{ is sufficiently small real vector} \\ \mathbf{P}_0 = \alpha^2\mathbf{I}, \alpha \text{ is sufficiently large real number} \end{cases} \tag{9}$$

The loss function is defined as Equation (10) to describe the variance between identified residue delay proportion and its measured value.

$$j(m) = \sum_{k=m}^{M+m-1} ||\mathbf{Y}(k+1) - \hat{\boldsymbol{\theta}}_q\boldsymbol{\Phi}(k)||^2, \tag{10}$$

where $||\bullet||$ is vector norm, and R is sample size.

The system order r is desided by F-test. Supposed m_1 and m_2 are adjacent orders of system, statistics variable H is constructed as follows:

$$H(m_1, m_2) = \frac{j(m_1) - j(m_2)}{j(m_2)}\frac{M - 2m_2}{2(m_2 - m_1)}. \tag{11}$$

If M is large enough and $m_2 > m_1$, $H(m_1, m_2)$ obeys F-distribution.

$$H \sim F(2(m_2 - m_1), M - 2m_2).$$

The following pseudo-code used for $j(m)$ will be generated:

1. **Begin**
2. Set m be the maximum possibel order, i.e. $m = m_{MAX}$.
3. $q = 0$, $k = m - 1$, chose an appropriate $\hat{\boldsymbol{\theta}}_0$ and $\mathbf{P}_0$.
4. $\boldsymbol{\Phi}(k)$ is constructed according to the former $[k - m + 1, k]$ sample data.
5. $\hat{\boldsymbol{\theta}}_{q+1}$ and $\mathbf{P}_{q+1}$ is calculated as Equation (8).
6. $k = k + 1$, $q = q + 1$.
7. IF $k \leq M$, goto *4*; Else $\hat{\boldsymbol{\theta}}_{M-m+1}$ is the RLS estimation of this m-order system.
8. $j(m)$ is calculated as Equation (10).
9. $m = m - 1$.
10. IF $m \geq 1$, goto *3*.
11. **End**.

Then $\mathbf{J} = [j(1), j(2), ..., j(m_{MAX})]^{\mathrm{T}}$ is obtained. Given the degree of confidence $\alpha = 5\%$, where is $F_{0.05}(2, 144) \approx 3.05$. Because $H(2, 3) = 2.29 < F_{0.05}(2, 144)$, which means there is no

significant reduction of the loss function when system order changes from 2 to 3. So the PDDS-LB model can be modeled as a second-order linear time-invariant system. The corresponding RLS estimation of parameter matrix $\boldsymbol{\theta}$ is:

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= [\hat{\boldsymbol{B}}_0,\hat{\boldsymbol{B}}_1,\hat{\boldsymbol{A}}_1,\hat{\boldsymbol{A}}_2], \\
\hat{\boldsymbol{B}}_1 &= [b_{1,1}, b_{1,2}, b_{1,3}, b_{1,4}] = [-0.0004, -0.0005, -0.0004, -0.0004] \\
\hat{\boldsymbol{B}}_2 &= [b_{2,1}, b_{2,2}, b_{2,3}, b_{2,4}] = [0.0049, 0.0060, 0.0056, 0.0050] \\
\hat{\boldsymbol{A}}_1 &= a_1 = 0.4528 \\
\hat{\boldsymbol{A}}_2 &= a_2 = 0.0922
\end{aligned}$$

Figure 3 is the compare of identified value vs. actual measurement of $\boldsymbol{Y}(k)$. The identified value of $\boldsymbol{Y}(k)$ is closed to the measured value, so 2-order linear MIMO model is appropriate to describe PDD-LB.

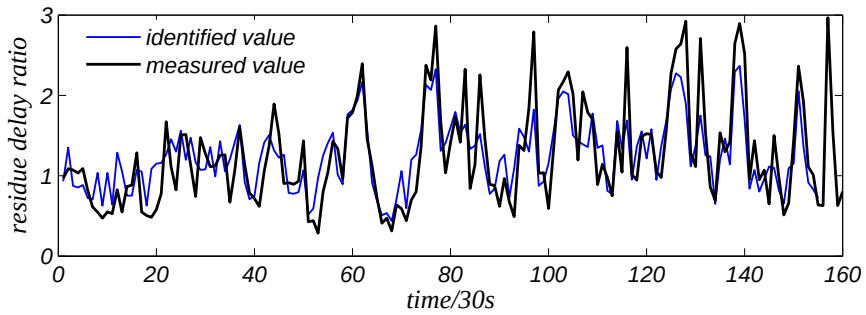


Figure 3: The identified value of $\boldsymbol{Y}(k)$ is closed to actual measurement

4.2 Optimal Controller Design

The resource reallocation of process/thread can be treated as an optimization problem which minimizes the deviation between $y_i(k)$ and $y_{i_{desire}}$, while penalizes large changes in the control variables to save the overhead (see in [12]). So we construct the following quadratic cost function:

$$\mathfrak{L} = \mathfrak{E}\{||\boldsymbol{W}[\boldsymbol{Y}(k+1) - \boldsymbol{Y}_{desire}]||^2 + ||\boldsymbol{Q}[\boldsymbol{X}(k) - \boldsymbol{X}(k-1)]||^2\}, \tag{12}$$

where $\boldsymbol{W}$ is $O \times O$ positive-definite weighting matrix and $\boldsymbol{Q}$ is $I \times I$ positive-definite penalty matrix. Diagonal elements of $\boldsymbol{W}$ represent the priorities of the corresponding classes, the smaller $w_{i,i}$, the more discriminated against the class i . In our application, $\boldsymbol{W}$, $\boldsymbol{Q}$ are diagonal matrix and unit matrix. The derivative of $\mathfrak{L}$ about $\boldsymbol{X}$ is zero when at its minimum, so there is(the derivation of equation see in [10, 12]):

$$\begin{aligned}
\frac{\partial \mathfrak{L}}{\partial \boldsymbol{X}(k)} &= 2(\boldsymbol{W}\hat{\boldsymbol{B}}_0)^{\mathrm{T}}\boldsymbol{W}[\hat{\boldsymbol{\theta}}\tilde{\boldsymbol{\Phi}}(k) - \boldsymbol{Y}_{desire}] + 2(\boldsymbol{W}\hat{\boldsymbol{B}}_0)^{\mathrm{T}}\boldsymbol{W}\hat{\boldsymbol{B}}_0\boldsymbol{X}(k) \\
&+ 2\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X}(k) - 2\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X}(k-1) = 0,
\end{aligned} \tag{13}$$

By solving the Equation (13), we can obtain the following optimal control law at the k^{th} sampling time

$$\begin{aligned}
\boldsymbol{X}^*(k) &= [(\boldsymbol{W}\hat{\boldsymbol{B}}_0)^{\mathrm{T}}\boldsymbol{W}\hat{\boldsymbol{B}}_0 + \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}]^{-1}\{(\boldsymbol{W}\hat{\boldsymbol{B}}_0)^{\mathrm{T}}\boldsymbol{W}[\boldsymbol{Y}_{desire} - \hat{\boldsymbol{\theta}}\tilde{\boldsymbol{\Phi}}(k)] \\
&+ \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X}(k-1)\}.
\end{aligned} \tag{14}$$

5 Maximum Idle First

The resource optimal control is used to calculate the quotas ς_i assigned to class i, while the load balancing strategy is to assure a fair assumption of threads, avoiding some back-end starvation or overload. Most researches of load balancing are on the promise of Possion arrival distribution and the exponential service time distribution, such as RR and LCS. However, the tendency that Web pages is on a going to contains more and more dynamic objects and database operation makes processing time hard to meet the exponential distribution, and all of these strategies can not support differentiated service. In order to precisely estimate the load status of each classes in back-end servers and dispatch new coming requests equally, we propose a load balancing strategy called Maximum Idle First (MIF).

The status of j^{th} back-end server are N sequence of two-tuples $< n_{i,j}, \tau_{i,j} >$ďżˇďżˇ$i = 1, \ldots, N$, which includes the connection queue length n_{ij} and the number of idle threads in the server pool τ_{ij}. The residue delay ℓ_i is proportional to the queue length $n_{i,j}$ and inverse proportional to the number of service threads, i.e. $\tau_{i,j}$ (see in [8, 14]). For the clients supported HTTP 1.1 , a Web session is always a sequence of HTTP requests on a consistent TCP connection as a manner of Pipeline, so there is

$$\ell_{i,j} = \alpha n_{i,j} \mathfrak{E}[\rho_{i,j}]/\tau_{i,j}, \tag{15}$$

$$\boldsymbol{idle}_{i,j} \propto 1/\ell_{i,j}. \tag{16}$$

Let $\boldsymbol{idle}_{i,j}$ be the idle degree of service for class i in server j, which is inverse proportional to the residue dalay. α is a planform dependence translation parameter. $\mathfrak{E}[\rho_{i,j}]$ is the mathematic expecting of Web session size. (see in [7]),there is :

$$\mathfrak{E}[\rho_{i,j}] = \mathfrak{E}[v_{i,j}]\mathfrak{E}[u_{i,j}], 1 \le i \le N, 1 \le j \le M \tag{17}$$

$v_{i,j}$,$u_{i,j}$ are the size and the number of the embedded requests from a Web session. Because every back-end servers deploys the same content and the self-similarity of the requests, $\mathfrak{E}[\rho_{i,j}]$ is a constant value. When a request belonged to class i arrivals at the front-end dispatcher, *mod_ proxy_ banlancer* calculates each back-end's $\boldsymbol{idle}_{ij}$, then redirects this request to the server, which has the maximum idle degree of service.

6 Experiment Evaluation

6.1 Configuration of Experiment

The test-bed is developed to evaluate the PDDS-LB mechanism, which consists of 9 computers connected together via 1Gbps Ethernet. There are 4 back-ends running Apache-2.0.53, which each has 100 concurrent threads, and 4 Linux machines simulate 120×4 different clients using SURGE-1.00a. Considering the front-end may be a new bottleneck, we configure the Apache-2.2.63 on the front-end with 1000 concurrent threads, which is large enough. In our experiment, HTTP requests are classified into 2 business flows based on the clients' IP address,i.e. class 1 and class 2. Set the expected delay ratio is $\ell_1/\ell 2 = 1/2$, $\boldsymbol{Y}_{desire} = [1/3]$, i.e. the class 1's residue delay should be the half of the class 2's.

Define the relative variance $\Psi(Y)$ be a control performance metric. A smaller $\Psi(Y)$ indicates a better stability that controller can keep $\boldsymbol{Y}(k)$ at $\boldsymbol{Y}_{desire}$.

$$\Psi(Y) = \frac{\sqrt{\sum_{k=1}^{I} ||\boldsymbol{Y}(k) - \boldsymbol{Y}_{desire}||^2 / I}}{\boldsymbol{Y}_{desire}}, \tag{18}$$

6.2 Result and Analysis

We design two sets of contrast experiments to evaluate the PDDS-LB cluster's ability of differentiation,throughput and dalely under different load balancing strategy.

Firstly, We evaluate the optimal controller's ability of differentiation under different load balancing strategy and sampling time. As shown in Figure 4(a), 4(b) and 4(c), each of them are the results of $T = 20s$ and $T = 30$ when using RR, LCS and MIF algorithm. The optimal controller launches at 800 seconds, and the measured residue delay ratio of class 1 and class2 is gradually settled around the expected value. According to the Equation (18), $\Psi(Y)$ are calculated and shown in Figure 4(d). Compared Figure 4(a), 4(b) and 4(c), the following conclusion can be summarized:

1. No matter using which load balancing algorithms or which sampling time, our optimal controller can provide proportional delay services. This proves the correctness and feasibility of PDDS-LB cluster model.

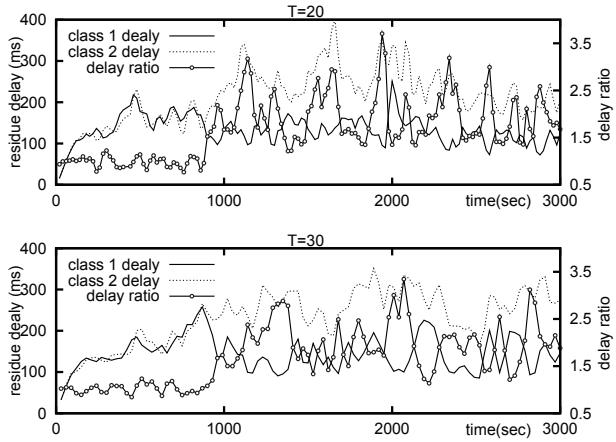
2. ℓ_i^{-800} and ℓ_i^{+800} are used respectively to compare average residue delay before and after 800 sec, there is:
$$\ell_2^{+800} - \ell_2^{-800} \geq \ell_1^{-800} - \ell_1^{+800}, \tag{19}$$
It means that the increase of class 2's delay is more than the decrease of class 1's. Nevertheless, it is worth making for the reasons in Section 1.

3. In equilibrium state, i.e. $\boldsymbol{Y}(k) = \boldsymbol{Y}_{desire}$, there are
$$\ell_{1,MIF}^{+800} < \ell_{1,RR}^{+800} \approx \ell_{1,LCS}^{+800}, \tag{20}$$
$$\ell_{2,MIF}^{+800} < \ell_{2,RR}^{+800} \approx \ell_{2,LCS}^{+800}, \tag{21}$$
where $\ell_{i,RR}^{+800}$, $\ell_{i,LCS}^{+800}$, $\ell_{i,MIF}^{+800}$ are the average residue delay of class i in stable sate under RR, LCS and MIF strategies, Equation (20)and (21) proves that MIF has less delay in the stable state.

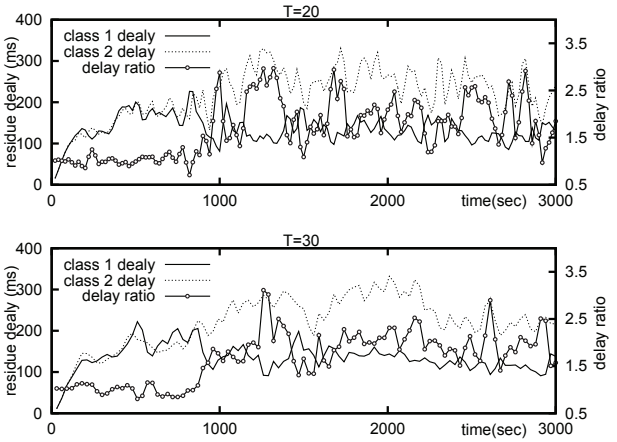
4. As shown in figure 4(d), when sampling time $T = 20s$, the delay ratio jitters severely. The possible reason is the delay jitter caused by large files, and which are more apparent in small sampling time.

Then, a further comparation is enforced to evaluate the different balancing strategy on the system throughput and residue delay. When $T = 20$, as the increase of the client's TCP connection requests arrival rate, the throughput and residue time also enhance, while the two classes of business still keep differentiate relationship. The Figure 5(a) and 5(b) are the histogram of throughput and the residue delay with the width of 20 TCP connections/sec. As can be seen, when the TCP connection arrive rate is 25/sec the system is saturated. Both in LCS and RR, the throughput of class 1 and 2 are 80 requests/sec and 40 requests/sec respectively, while in MIF, the throughput of class 1 can increase to 120 requests/sec. At the same time, in LCS and RR, the average delay of class 1 and 2 are 100ms and 280ms, while the average delay of class 2 reduce to 200ms in MIF.

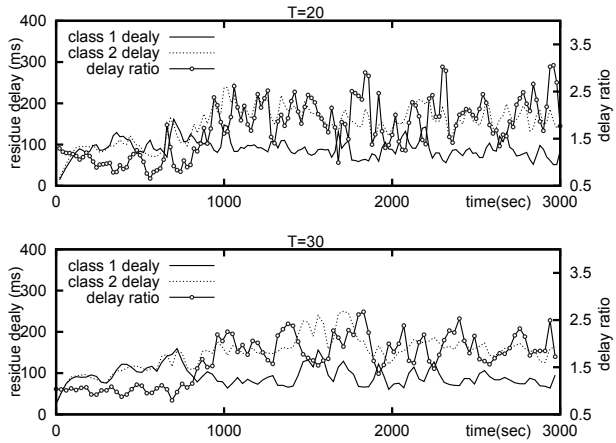
This is not only because MIF consider the impact of the connection queue length and idle threads on the residue delay, but also due to the fine-grained state feedback. The status of back-ends contain details of every class , we can design a better dispatch strategy. Although LCS and RR can dispatch requests to all back-ends evenly, since lack of the fine-grained state feedback, they cannot maximize the resource utilization. For example, if the thread quotas of class 1 will be exhausted, or its connection queue will overflow on back-end j, while the resources



(a) RR



(b) LCS



(c) MIF

$\Psi(Y)$	RR	LCS	MIF
20s	0.3843	0.3810	0.31925
30s	0.3831	0.3759	0.28000

(d) $\Psi(Y)$

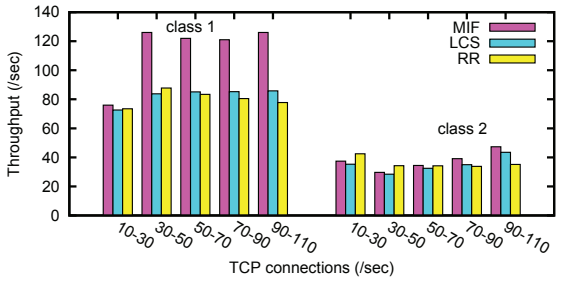
Figure 4: Differentiation effects under Different load balancing strategy and sampling time

of other classes are idle, LCS and RR tend to redirect requests to this back-end. Now, there is a new request, which is exactly belongs to class 1, resource shortage on back-end j will further deteriorate, however, other back-ends may be in idle status.

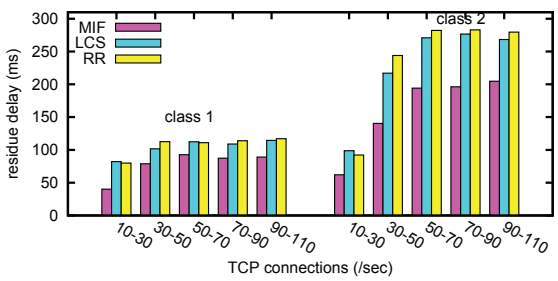
7 Conclusion

In this paper, we design a feedback control mechanism to achieve Proportional Delay Differentiation Service and Load Balancing in a Web Cluster System. By recursive least square estimation and F-test, PDDS-LB is model as a second-order liner time-invariant system. We construct the cost function and obtain the optimal law by derivation of cost function. Experimental evaluations have shown that our mechanism achieves PDDS, while the low priority class is not over-sacrificed. With the aid of MIF, we maximize the resource utilization of the cluster system.

However, our experiments are just in the condition of ideal transport layer, in the real network, link status is complex and changeable ,the factors that influence QoS interweave each other, such as bandwidth, cache, I/O etc. As part of out ongoing work, we are exploring a integrated resource management to adapt to more complex environment.



(a) Throughput



(b) Delay

Figure 5: Compared with LCS and RR, the average delay reduces 21% and the total throughput increases 33% in MIF.

Bibliography

[1] M. Andreolini, E. Casalicchio, M. Colajanni, and M. Mambelli. A cluster-based web system providing differentiated and guaranteed services. *Cluster Computing*, 7(1):7–19, 2004.

[2] G. Ang, M. Dejun, H. Yansu, and P. Wenping. Proportional Delay Guarantee in Web QoS Based on Predictive Control. In *Information Science and Engineering (ICISE), 2009 1st International Conference on*, pages 1789–1792. IEEE, 2009.

[3] G. Apostolopoulos, D. Aubespin, V. Peris, P. Pradhan, and D. Saha. Design, Implementation, and Performance of a Content-Based Switch. In *IEEE INFOCOM*, volume 3, pages 1117–1126. Citeseer, 2000.

[4] E. Casalicchio and M. Colajanni. A client-aware dispatching algorithm for web clusters providing multiple services. In *Proceedings of the 10th international conference on World Wide Web*, page 544. ACM, 2001.

[5] L. Cherkasova and M. Karlsson. Scalable web server cluster design with workload-aware request distribution strategy WARD. In *wecwis*, page 0212. Published by the IEEE Computer Society, 2001.

[6] JBoss Community. http://www.jboss.org/mod_cluster.

[7] A. Gao, D. Mu, and Y. Hu. A QoS Control Approach in Differentiated Web Caching Service. *Journal of Networks*, 6(1):62–70, 2011.

[8] A. Gao, D. Mu, Y. Hu, and W. Pan. Proportional Delay Guarantee in Web QoS Based on Predictive Control. *1st International Conference on Information Science and Engineering,ICISE2009*, 1:1789–1792, 2009.

[9] A. Gao, H. Zhou, Y. Hu, D. Mu, and W. Hu. Proportional Delay Differentiation Service and Load Balancing in Web Cluster Systems. In *INFOCOM IEEE Conference on Computer Communications Workshops, 2010*, pages 1–2. IEEE, 2010.

[10] M. Karlsson, X. Zhu, and C. Karamanolis. An adaptive optimal controller for non-intrusive performance differentiation in computing services. In *Control and Automation, 2005. ICCA'05. International Conference on*, volume 2, pages 709–714. IEEE.

[11] A.F. Liu. *Research on the Web Server Cluster Technologies*. PhD thesis, Central South Univercity, 2005.

[12] X. Liu, X. Zhu, P. Padala, Z. Wang, and S. Singhal. Optimal multivariate control for differentiated services on a shared hosting platform. In *Proc. of the 46th IEEE Conf. on Decision and Control*. Citeseer, 2007.

[13] V.S. Pai, M. Aron, G. Banga, M. Svendsen, P. Druschel, W. Zwaenepoel, and E. Nahum. Locality-aware request distribution in cluster-based network servers. In *Proceedings of the eighth international conference on Architectural support for programming languages and operating systems*, pages 205–216. ACM, 1998.

[14] W. Pan, D. Mu, H. Wu, and Q. Sun. Proportional Delay Differentiation Service in Web Application Servers: A Feedback Control Approach. *International Journal of Intelligent Information Technology Application*, 1(1):37–42, 2008.

[15] A. Robertson. Linux-HA heartbeat system design. In *Proceedings of the 4th annual Linux Showcase & Conference-Volume 4*, pages 20–20. USENIX Association, 2000.

[16] A. Robertson. The evolution of the Linux-HA project. In *UKUUG LISA/Winter Conference High-Availability and Reliability*, 2004.

[17] T. Schroeder, S. Goddard, and B. Ramamurthy. Scalable Web server clustering technologies. *IEEE network*, 14(3):38–45, 2000.

[18] Z.G. Shan and C. Ling. Performance Evaluation of QoS-Aware Load Balancing of Web-server Clusters. *Journal of System Simulation in chinese*, 17:184–189, 2005.

[19] X. Wu, M. Li, and J. Wu. Enhanced Demand-driven Service Differentiation Algorithm in Web Clusters. In *IEEE International Conference on e-Business Engineering, 2006. ICEBE'06*, pages 386–391, 2006.

[20] Z. Xing, P.L. Yan, and C.C. Guo. Proportional Stretch Factor Differentiated Service in Heterogeneous Web Server Cluster. *Computer Science in chinese*, 33(010):61–65, 2006.

[21] H. Zhu, H. Tang, and T. Yang. Demand-driven service differentiation in cluster-based network servers. In *IEEE INFOCOM*, volume 2, pages 679–688. Citeseer, 2001.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 674-687

Impact of Network Infrastructure Parameters to the Effectiveness of Cyber Attacks Against Industrial Control Systems

B. Genge, C. Siaterlis, M. Hohenadel

Béla Genge, Christos Siaterlis and Marc Hohenadel
Institute for the Protection and Security of the Citizen
European Commission, Joint Research Centre
Via E. Fermi, 21027, Ispra (VA), Italy
E-mail: {bela.genge, christos.siaterlis, marc.hohenadel}@jrc.ec.europa.eu

Abstract:
The fact that modern Networked Industrial Control Systems (NICS) depend on Information and Communication Technologies (ICT), is well known. Although many studies have focused on the security of SCADA systems, today we still lack the proper understanding of the effects that cyber attacks have on NICS. In this paper we identify the communication and control logic implementation parameters that influence the outcome of attacks against NICS and that could be used as effective measures for increasing the resilience of industrial installations. The implemented scenario involves a powerful attacker that is able to send legitimate Modbus packets/commands to control hardware in order to bring the physical process into a critical state, i.e. dangerous, or more generally unwanted state of the system. The analysis uses a Boiling Water Power Plant to show that the outcome of cyber attacks is influenced by network delays, packet losses, background traffic and control logic scheduling time. The main goal of this paper is to start an exploration of cyber-physical effects in particular scenarios. This study is the first of its kind to analyze cyber-physical systems and provides insight to the way that the cyber realm affects the physical realm.
Keywords: cyber attacks, Industrial Control Systems, SCADA, security.

1 Introduction

Modern Critical Infrastructures (CI), e.g. power plants, water plants and smart grids, rely on Information and Communication Technologies (ICT) for their operation since ICT can lead to cost optimization as well as greater efficiency, flexibility and interoperability between components. In the past CIs were isolated environments and used proprietary hardware and protocols, limiting thus the threats that could affect them. Nowadays CIs, or more specifically Networked Industrial Control Systems (NICS), are exposed to significant cyber-threats; a fact that has been highlighted by many studies on the security of Supervisory Control And Data Acquisition (SCADA) systems [1, 2].

The recently reported Stuxnet worm [3] is the first malware specifically designed to attack NICS. Its ability to reprogram the logic of control hardware in order to alter physical processes demonstrated how powerful such threats can be. Stuxnet was a concrete proof of a successful cyber-physical attack but by no means a trivial attack. It required a thorough knowledge of the physical system, software and OS vulnerabilities. In this paper we consider an adversary with a lower level of sophistication that instead of reprogramming the highly specialized hardware (PLCs) as in the Stuxnet case, he exploits the ability of control hardware to communicate with remote stations using well-established protocols such as TCP, that are normally used by the operator to read sensor values, e.g. pressure, temperature, and control actuators, e.g. valves. Furthermore, we identify the communication and control logic implementation parameters that influence the outcome of cyber attacks against NICS and that could be used as effective measures for increasing the resilience of industrial installations. In our study the attacker is located

outside the plant, somewhere in the Internet, from where he/she exploits the capability of control hardware to communicate with remote stations. Although direct connections between control hardware and the Internet are usually avoided in NICS implementations [5], the attacker might install a malware on a station with an Internet connection located within the corporate network that could be used to forward messages to the control hardware. We consider that the control hardware is running legitimate code designed to keep the physical process in a normal operating point, while the attacker tries to bring it to a *critical state*, i.e. dangerous, or more generally unwanted state of the system [6]. The method employed by the attacker is a repeated transmission of commands to the control hardware that open/close specific valves. The main contribution of this paper is that it is the first of its kind to analyze cyber-physical systems and provides insight to the way that the cyber realm affects the physical realm.

The real applicability of the implemented scenario is confirmed not only by Stuxnet, but by other past events as well. One example in this sense is a 2002 penetration test done by a security firm for a power company located in California. The testers parked their van outside a remote substation, where they noticed a wireless antenna. Without leaving their vehicle they managed to connect to the system and within 20 minutes they have not only mapped the entire network, but also "they were talking to the business network and had pulled off several business reports" [4]. By taking these events and adapting them to our scenario we can further imagine that the testers are in fact attackers that install a remotely accessible gateway and launch their attack from the anonymity provided by the Internet.

The attack scenario has been implemented with the help of our previously developed framework [7] that uses simulation for the physical components and an emulation testbed based on Emulab [8, 9] to recreate the cyber part of NICS, e.g., SCADA servers, corporate network, etc. In the implemented scenario we have used the model of a Boiling Water Power Plant developed by Bell and Åström [10].

The paper is structured as follows. After an analysis of related work in Section 2 we present a brief overview of the experimentation framework in Section 3. The methodology and implemented experiments are presented in Section 4 and 5, respectively. The paper concludes in Section 6.

2 Related Work

An approach where real sensors and actuators, combined with simulated PLCs and communication protocols were used to study cyber-physical systems has been proposed by Queiroz, *et al.* [15]. Their study showed that while PLCs are under a DoS attack, operators might take delayed or wrong decisions that could disrupt the operation of the plant. A similar experiment has also been documented by Davis, *et al.* [16] that used the PowerWorld server to study the effects of communication delays between the physical process and human operators. In the same direction, the work of Chabukswar, *et al.* [17] proved that a DDoS attack against communication nodes between controllers and sensors causes the PLCs to take wrong decisions based on old sensor values. Cárdenas, *et al.* [18], went further by not only documenting the effects of DoS attacks on sensors, but also proposing a new detection mechanism and possible countermeasures.

The previously mentioned approaches have demonstrated the effectiveness of DoS attacks, but without reaching a sophistication level that would have allowed the attacker to reprogram the low level control logic of the PLCs. This fact sets an important barrier in terms of knowledge, skills and efforts required by the attacker, as was the case of Stuxnet, where developers had also knowledge of the PLC code, OS and hardware details. In this category we find the work of Nai Fovino, *et al.* [5] that have proposed an experimental platform for studying the effects of cyber attacks against NICS. In their paper the authors describe several attack scenarios, including DoS attacks and worm infections that send Modbus packets to control hardware. Although the

authors provide a wide set of countermeasures, they do not identify communication parameters that affect the outcome of the attacks. Moreover, in our analysis we have also identified installation-specific parameters that can directly affect the resilience of physical processes.

3 Framework Overview

After providing a brief description of a typical NICS architecture, this section presents a short overview of our previously developed experimentation framework used in our experiments.

3.1 Process Control Architecture Overview

In modern NICS architectures, one can identify two different control layers: (i) the physical layer composed of all the actuators, sensors, and generally speaking hardware devices that physically perform the actions on the system, e.g. open a valve, measure the voltage in a cable; (ii) the cyber layer composed of all the ICT devices and software which acquire the data, elaborate low level process strategies and deliver the commands to the physical layer. The cyber layer typically uses SCADA protocols to control and manage the physical devices within the cyber layer. The "distributed control system" of the cyber layer is typically split among two networks: the *control network* and the *process network*. The process network usually hosts all the SCADA (also known as SCADA Masters) and HMI (Human Machine Interface) servers. The control network hosts all the devices which, on the one side control the actuators and sensors of the physical layer and on the other side provide the "control interface" to the process network. A typical control network is composed of a mesh of PLCs (Programmable Logic Controller). From an operational point of view, PLCs receive data from the physical layer, elaborate a "local actuation strategy", and send back commands to the actuators. PLCs execute also the commands that they receive from the SCADA servers (Masters) and additionally provide, whenever requested, detailed physical layer data.

3.2 Framework Architecture

The previously developed framework [7] follows a hybrid approach, where the Emulab-based testbed recreates the control and process network of NICS, including PLCs and SCADA servers, and a software simulation reproduces the physical processes. The architecture, as shown in Figure 1, clearly distinguishes 3 layers: the cyber layer, the physical layer and a link layer in between. The cyber layer includes regular ICT components used in SCADA systems, while the physical layer provides the simulation of physical devices. The link layer, i.e. cyber-physical layer, provides the "glue" between the two layers through the use of a shared memory region.

The physical layer is recreated through a soft real-time simulator that runs within the SC (Simulation Core) unit and executes a model of the physical system. The simulator's execution time is strongly coupled to the timing service of the underlying operating system (OS). Throughout the paper the term *time step* is used to denote the time between two successive executions of the physical model in the simulator. The cyber layer is recreated by an emulation testbed that uses the Emulab architecture and software [8, 9] to automatically and dynamically map physical components, e.g. servers, switches to a virtual topology. Besides the process network, the cyber layer also includes the control logic code that in the real world is run by PLCs. The control code can be run sequentially or in parallel to the physical model. In the sequential case, a *tightly coupled* code (TCC) is used, i.e. code that is running in the same memory space with the model, within the SC unit. In the parallel case a *loosely coupled* code (LCC) is used, i.e. code that is running in another address space, possibly on another host, within the R-PLC unit (Remote

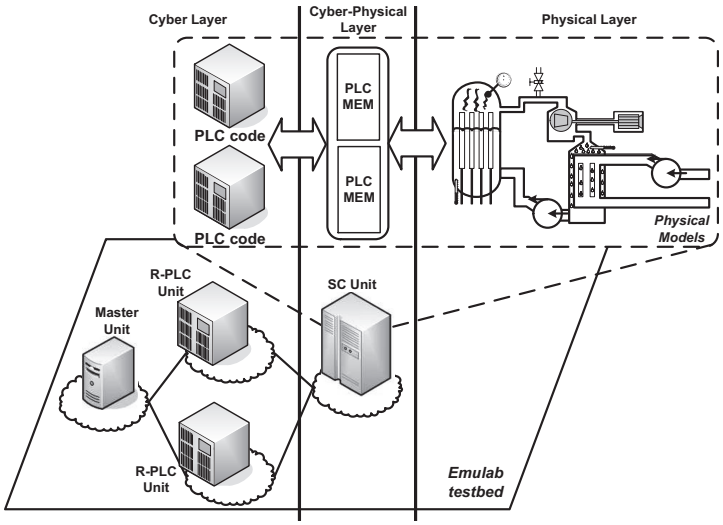


Figure 1: Experimentation framework architectural overview

PLC). The unit that implements global decision algorithms based on the sensor values received from the R-PLC units is also present in the proposed framework as the *Master* unit. The cyber-physical layer incorporates the PLC memory, seen as a set of registers typical of PLCs, and the communication interfaces that "glue" together the other two layers. Prototypes of SC, R-PLC and Master Units have been developed in C# (Windows) and have been ported and tested on Unix-based systems (FreeBSD, Fedora and Ubuntu) with the use of the *Mono* platform. Matlab Simulink was used as the physical process simulator (physical layer). From Simulink models the corresponding 'C' code is generated using Matlab RTW. The communication between SC and R-PLC units is handled by .NET's binary implementation of RPC (called *remoting*) over TCP. For the communication between the R-PLC and Master units, we used the Modbus over TCP protocol.

4 Methodology

In this section we provide a description of the methodology we used, including a description of the scenario and experimental setup.

4.1 Scenario

Previous security events [3, 4] involving NICS showed that attackers can (easily) compromise stations located within an installation's internal network. These stations can then be used as gateways for downloading malicious code and for remotely controlling other stations, including control hardware such as PLCs. The implemented scenario assumes that there is already a compromised station providing access to PLCs. This is used by the attacker to send remote Modbus commands in order to bring the physical process into a critical state.

The physical process used in our experiments is the Boiling Water Power Plant (BWPP) model developed by Bell and Åström [10]. Within this context the critical state is given by an increased steam pressure that is more than twice the value of a typical operating point. Although to the best of our knowledge the literature does not mention anything about the consequences of running the process with these parameters, we assumed that this might cause physical damages

and therefore could become a desired target for the attacker. Furthermore, we assumed that the attack is conducted for a limited time period of 10 minutes, as immediately after it is started the process experiences significant deviations from its normal operating point. Consequently, alarms might be turned ON and human operators might intervene by switching OFF devices or disconnecting equipment. As shown later in this paper, the 10 minute time period is more than enough for the attacker to bring the physical process into a critical state.

Our experimental results presented in the next section show that the outcome of the attack is affected not only by network delays, packet losses and background traffic, but also by the execution of PLC control code and the speed of control valves. These can be used as effective measures to increase the resilience of physical processes confronted with cyber attacks.

4.2 Experimental Setup

The previously described scenario has been implemented in the Joint Research Centre's (JRC) Experimental Platform for Internet Contingencies (EPIC) laboratory. The Emulab testbed included nodes with the following configuration: FreeBSD OS 8, AMD Athlon Dual Core CPU at 2.3GHz and 4GB of RAM. We have emulated network delays, packet losses and background traffic with the *Dummynet* [11, 12] and *iperf* [13] software in order to recreate a dynamic and unpredictable environment such as the Internet.

As shown in Figure 2, the experimental setup consisted of 6 hosts: 1 host for running the SC unit, 3 hosts for running the R-PLC units, 1 host for running the Compromised Station and 1 host for the attacker. The attacker uses the Compromised Station to forward Modbus packets to R-PLC units and finally to write the PLC memory within the SC unit. The control code that is in charge of maintaining the BWPP at a constant operating point has been implemented as TCC code, where *TCC1* controls the fuel valve, *TCC2* controls the steam valve and *TCC3* controls the feed-water valve. The role of the R-PLC units is simply to enable access to the physical model using the Modbus/TCP protocol. Although R-PLC units can also run control code, the decision to implement control code as TCCs has been made based on the granularity of the process model execution step that needs to be less than 1ms. This is needed in order to emulate PLC *tasks* that can be scheduled to run within milliseconds. Consequently, the chosen time step for the physical model was of 0.5ms.

Network delays, background traffic and packet losses, specific to a dynamic environment such as the Internet, have been emulated between the attacker's station and the Compromised Station. The limited bandwidth and communication capabilities of PLCs have been emulated with 10Mbit/s Lans (*Lan2* and *Lan1*, respectively), while using a 100Mbit/s Lan (*Lan0*) between R-PLCs and the SC unit in order to maximize the performance of the interaction between R-PLC units and the BWPP model.

4.3 Boiling Water Power Plant

As already mentioned, in this paper we use the Boiling Water Power Plant (BWPP) model developed by Bell and Åström in [10]. It models a 160MW oil-fired electric power plant based on the Sydsvenska Kraft AB plant in Malmö, Sweden. The operation of the process is controlled by three valves, i.e. fuel valve, steam valve and feed-water valve, while the operator is able to monitor the process by reading three sensors: steam pressure, water level and generated electricity. The following equations describe the dynamics of the physical process [10]:

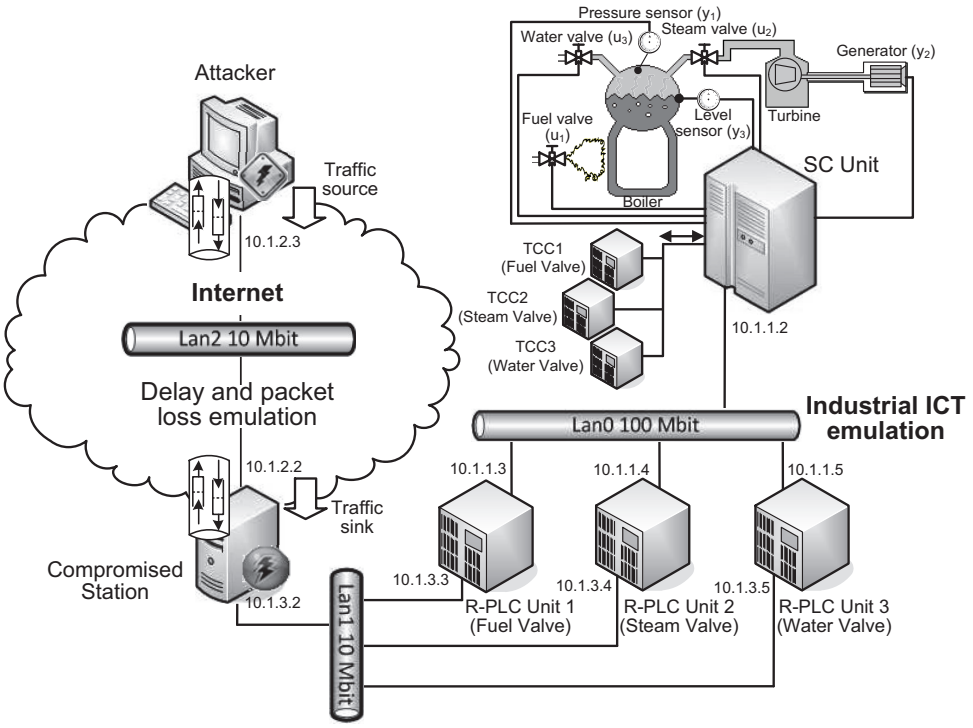


Figure 2: Experimental setup

$$
\begin{aligned}
&\dot{x}_1 = -0.0018u_2x_1^{9/8} + 0.9u_1 - 0.15u_3,\\
&\dot{x}_2 = (0.073u_2 - 0.016)x_1^{9/8} - 0.1x_2,\\
&\dot{x}_3 = (141u_3 - (1.1u_2 - 0.19)x_1)/85,\\
&y_1 = x_1,\\
&y_2 = x_2,\\
&y_3 = 0.05(0.1307x_3 + 100s_q + e_r/9 - 67.975),\\
&s_q = \tfrac{(1-0.001538x_3)(0.8x_1-25.6)}{x_3(1.0394-0.0012304x_1)},\\
&e_r = (0.85u_2 - 0.147)x_1 + 45.59u_1 - 2.514u_3 - 2.096,\\
&0 \le u_i \le 1 (i = 1, 2, 3),\\
&\dot{u}_1 \le 0.007/sec, -1.0 \le \dot{u}_2 \le 0.1/sec, \dot{u}_3 \le 0.05/sec,
\end{aligned}
\tag{1}
$$

where x_1, x_2 and x_3 denote the steam pressure (kg/cm^2), electric output (MW), and fluid density (kg/m^3), respectively; y_1, y_2 and y_3 denote the outputs of the model, where y_3 is the water level deviation (m); u_1, u_2 and u_3 are the valve positions for fuel flow, steam flow and feed-water flow, respectively; s_q and e_r denote the steam quality and evaporation rate (kg/s), respectively.

In our experiments we used a normal operating point for the BWPP in which the pressure equals 108 kg/cm^2, that was achieved by keeping the three valves in the following *normal valve positions* (NVP): $u_1 = 0.34$, $u_2 = 0.69$ and $u_3 = 0.433$ [14]. As a consequence, TCCs, that implement the control logic code, must maintain constant the position of the three valves in order to provide a constant steam pressure of 108 kg/cm^2. For the critical state we consider a pressure of 250 kg/cm^2 that is more than twice the value of the steam pressure of the process running in the previously mentioned normal operating point. The attacker is able to bring the

BWPP into the critical state by continuously sending Modbus packets (around 100/sec for each valve) that keep the steam and feed-water valves in the CLOSED position ($u_2 = 0$ and $u_3 = 0$, respectively) and the fuel valve in the OPENED position ($u_1 = 1$). Throughout this paper we use the term *attacker's valve positions* (AVP) to denote the position of the three valves in the attacker's setting.

5 Attacks and Analysis

In this section we analyze the influence of several parameters on the cyber attacks launched remotely from the Internet, as described in the scenario from the previous section. The analysis is conducted in two phases: in the first phase we analyze the ability of the attacker to maintain the control valves in the AVP; in the second phase we analyze the ability of the attacker to increase the steam pressure to 250 kg/cm^2, thus actually bringing the plant into the critical state. While the goal of the second phase is clear, the goal of the first one needs further explanation. The actual goal of the first phase is to analyze the reaction of PLCs in terms of commands sent to the valves, reaction that might provide assistance in the rationale of the results in the second phase. In both phases we have measured the influence of: PLC task scheduling (TS) every 100ms and 1ms; network delays of 0s, 0.5s and 3s; packet losses of 1%, 5% and 10%; and background traffic of 2.5Mbit/s, 5Mbit/s and 10Mbit/s. Such extreme values for network delays, i.e. 3s, and packet losses, i.e. 10%, can rarely be measured over the Internet (possibly over satellite links or multiple intermediate proxies). Nevertheless, we have included them in our analysis in order to justify our statements related to the required magnitude of these parameters for influencing the outcome of the attack. For each configuration setting, representing a combination of PLC TS, network delays, packet losses and background traffic we have run one experiment for 10 minutes. In total, we have run 540 experiments in 9 hours.

5.1 The Effect of the Cyber Attack on the Position of Control Valves

The implementation of Modbus over TCP allows attackers to remotely control the three valves, i.e. fuel valve, steam valve, feed-water valve, thus providing a certain anonymity to the attacker. Nevertheless, this status is compensated by fluctuating parameters such as network delays, packet losses and background traffic that can have a major effect on the outcome of the attack. In this sub-section we analyze the effects of these parameters and two different TS, i.e. 100ms and 1ms, on the position of the three control valves.

100ms Task Scheduling

The effect of network delays on the position of control valves for a 100ms TS time is given in Figure 3. Each sub-figure shows the position of a specific control valve during the 10 minute attack, starting with the NVP shown at $t = 0$ minutes and following with the changes induced by the attacker that, as shown by most of the figures, bring the valves into the AVP. The differences we observe in the behavior of the three valves are caused by the motion speed of each valve that is different in each case (Equation 1). Thus, the fastest to open and close is the steam valve (0.1/sec for opening and 1/sec for closing) followed by the feed-water valve (0.05/sec) and by the fuel valve (0.007/sec). Consequently, the greater the speed, the higher the fluctuations that we see in Figure 3 and the higher the ability of PLCs to maintain the NVP.

One of the main conclusions from our results (Figure 3) is that network delays are beneficial to the physical process when confronted with cyber attacks. Nevertheless, the attacker is able to bring the valves into the AVP even for network delays of 3s. The attacker is most successful with

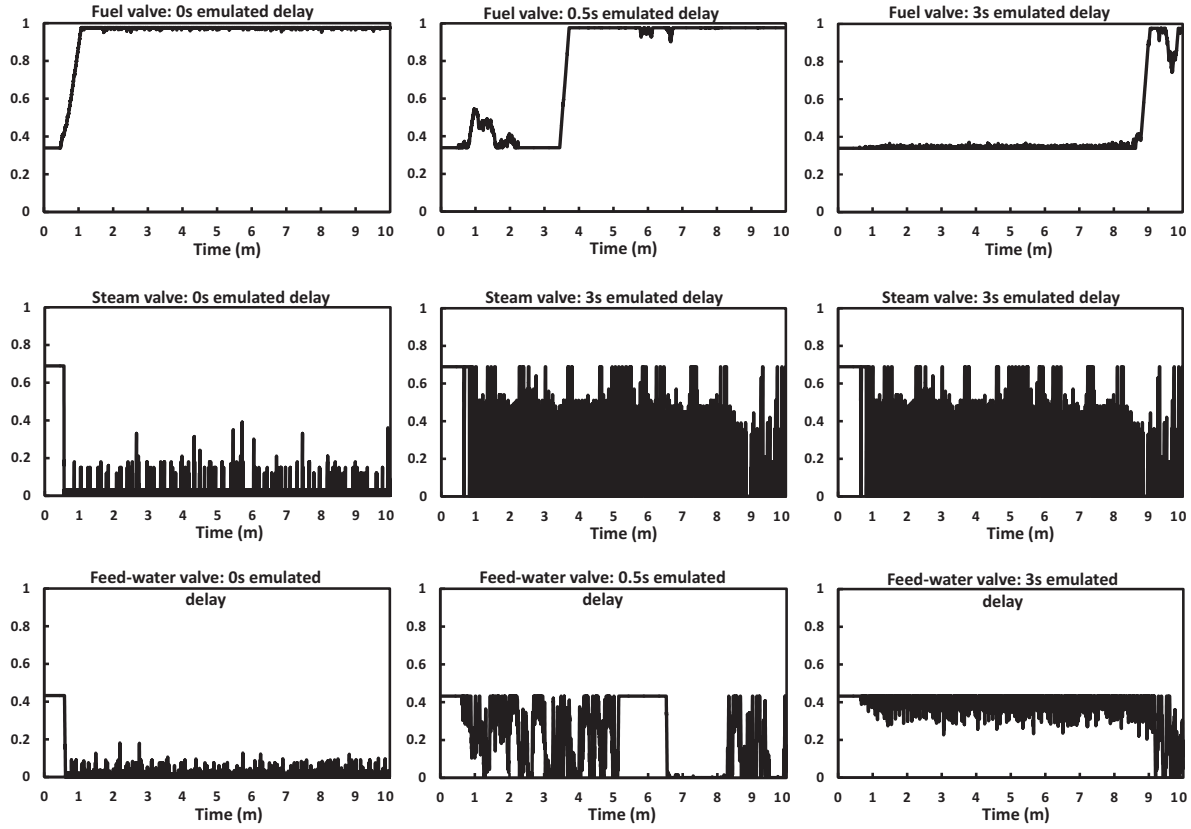


Figure 3: Effect of network delays on control valve positions for 100ms TS, 1% loss rate and 2.5Mbit/s background traffic

the fuel valve as this has the lowest speed and the control code is running only once every 100ms, while the attacker sends one Modbus command every 10ms, i.e. 100/sec. For the other two valves the PLCs are able to produce a slight deviation from the AVP; however, these are unable to bring the valves back to the NVP. As we increase the network delays we notice that PLCs are causing larger deviations from the AVP and in the extreme case of 3s the attacker is able to produce only small deviations from the NVP for the fuel and feed-water valves. Nevertheless, the attacker is still able to close the steam valve, as its closing speed (1/sec) exceeds 10 times its opening speed (0.1/sec). Based on these results we can conclude that only extreme network delays, e.g. 3s, have a major influence on the outcome of the attack. In addition, an attacker could successfully exploit a different opening and closing speed of valves, that could be interpreted as a slower reaction of the PLCs.

Going further, in Figure 4 we show the effect of packet losses on the position of the valves for the same TS of 100ms. By increasing the packet loss from 1% (Figure 3) to 5% and 10% (Figure 4) we also increase the deviations from AVP. Nevertheless, even with higher packet losses PLCs are unable to maintain the NVP, not even in the case of slower valves such as the fuel and feed-water valves. As shown in the same figure, we have also experimented with extreme packet losses of 10%. However, as the attacker uses a 10 times higher packet rate than the control code scheduling rate, even in this case PLCs are unable to keep the valves in the NVP.

We have also investigated the effect of background traffic on the three valves, shown in Figure 5. By increasing the background traffic from 2.5Mbit/s (Figure 3) to 5Mbit/s (Figure 5) we do not notice major effects, as the maximum network capacity is 10Mbit/s and the 100 Modbus

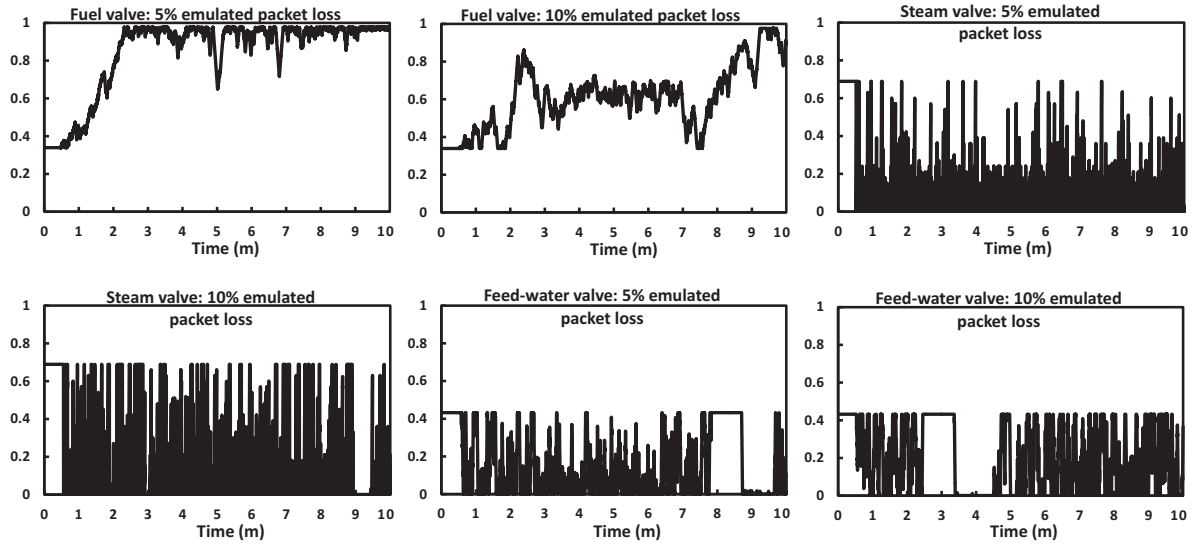


Figure 4: Effect of packet loss on plant valve positions for 100ms TS, 0s emulated delay and 2.5Mbit/s background traffic

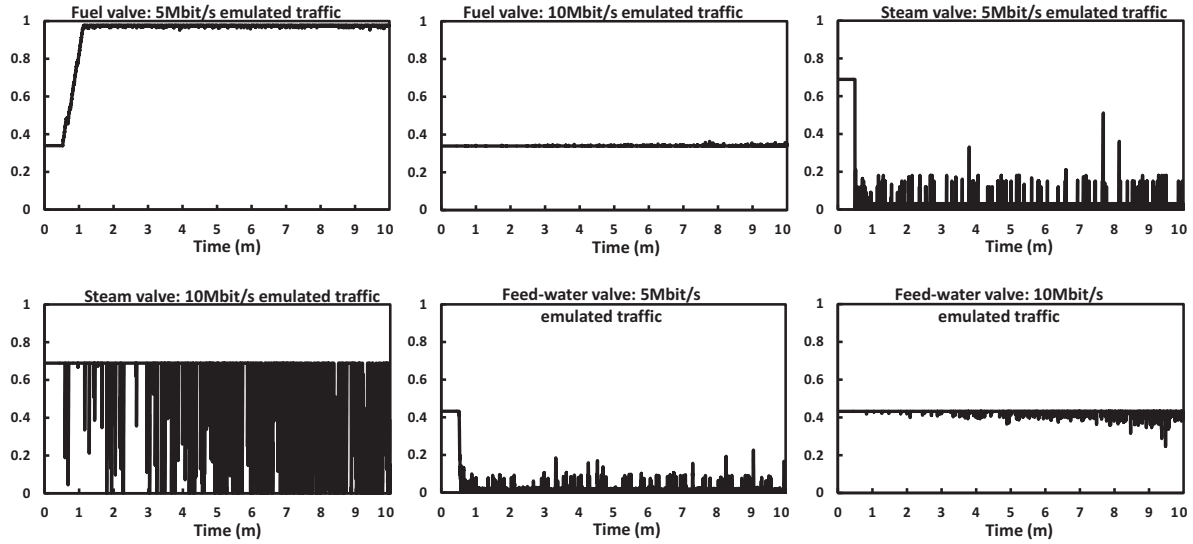


Figure 5: Effect of background traffic on plant valve positions for 100ms TS, 0s emulated delay and 1% packet loss

messages sent every second for each valve generate a traffic of around 140Kbit/s, with a total of 420Kbit/s for all three valves. Nevertheless, when the background traffic reaches the network capacity fewer messages reach the PLCs which are able to maintain the NVP with only small deviations. However, even in this later extreme case the steam valve is affected by the attacker, as its closing speed is 10 times higher than its opening speed, and counteracting one single packet received from the attacker requires 10 executions of the control code.

1ms Task Scheduling

The previous results have shown that if PLCs have a control code TS of 100ms they are not effective in maintaining the NVP. Moreover, the outcome of the attack is affected only by extreme

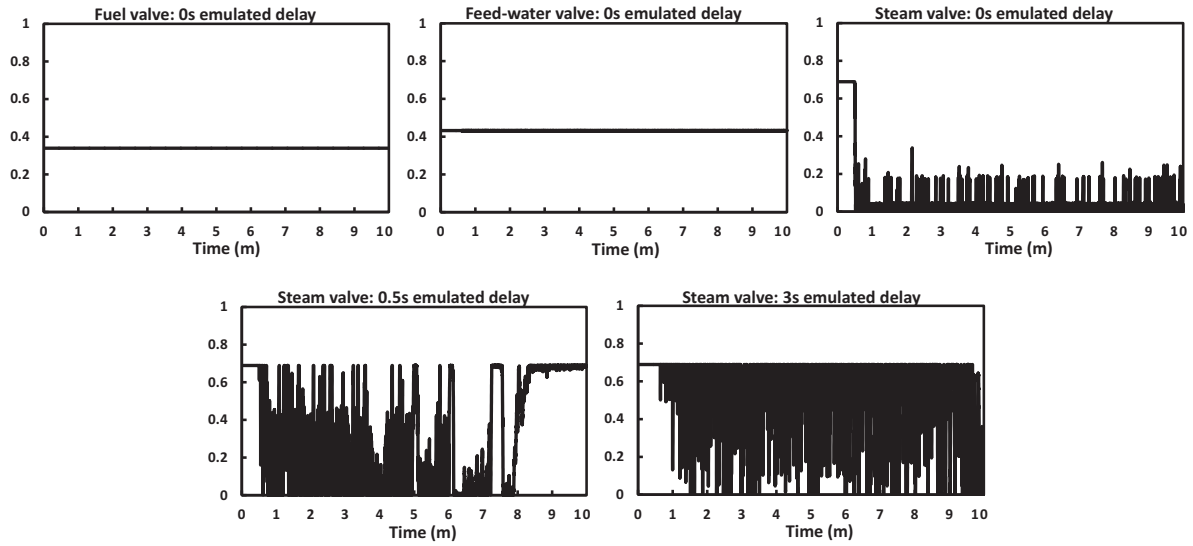


Figure 6: Effect of network delays on plant valve positions for 1ms TS, 1% loss rate and 2.5Mbit/s background traffic

cases of network delays and background traffic. For this reason we have also experimented with a TS of 1ms that could significantly improve the reaction of the PLCs.

For a TS of 1ms, even in the case of 0s emulated delay, the attacker is only able to produce insignificant deviations from NVP for fuel and feed-water valves, as shown in Figure 6. Nevertheless, the steam valve is still affected by the attack and the attacker is able to bring it to the AVP. However, in this case larger network delays show a significant effect on the steam valve, as the attacker is unable to maintain the steam valve in the AVP for delays larger than 3s.

As the fuel and feed-water valves present only insignificant changes already shown in Figure 6, we further focus our attention on the steam valve. We have repeated the experiments for different packet losses and background traffic, with the results shown in Figure 7. As in the case of the previous results, larger packet losses do not improve the response of the PLCs. Nevertheless, the extreme case of 10Mbit/s for background traffic ensures small deviations from the NVP for the first 5 minutes of the experiment. In the second half more packets reach the PLC that produce larger and larger deviations and finally after 8 minutes are able to bring the valve into the AVP.

The results from this sub-section have shown that the attacker is able to affect the NVP for all three valves if the PLCs use a TS of 100ms. Within this setting network delays and background traffic are major factors that influence the attack, as also illustrated in Table 1. Nevertheless, the attacker is still able to produce major deviations from NVP for valves with an opening and closing speed difference. For this reason, a better solution would be to provide a smaller TS combined with an equal opening and closing speed of valves. As shown by the results, with a smaller TS, the fuel and feed-water valves experience insignificant deviations from NVP with minimum emulated delays and background traffic. In contrast, the speed difference of the steam valve still causes major deviations from NVP, that are reduced by only extreme network delays of 3s and a background traffic of 10Mbit/s. Lowering the value of the TS below 1ms could be considered a measure for a more resilient control code. However, this is only possible if the control code's execution time is smaller than this value.

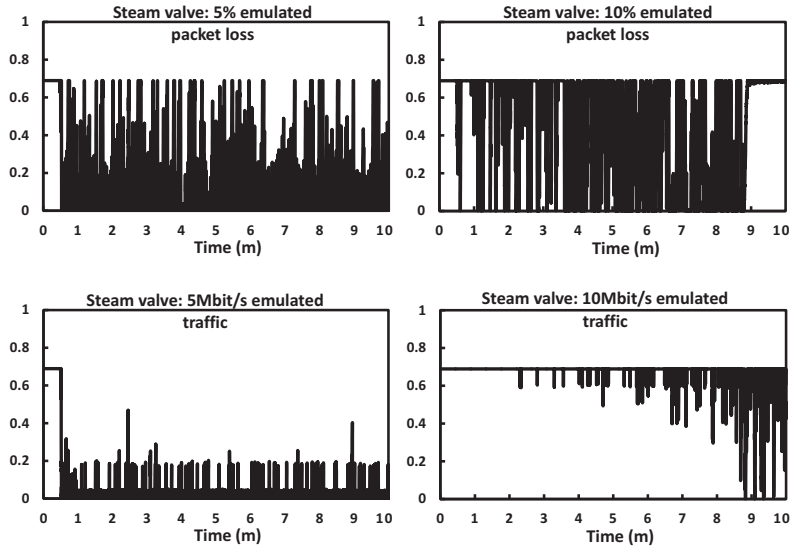


Figure 7: Effect of packet loss and background traffic on the steam valve position for 1ms TS and 0s emulated delay

5.2 The Effect of the Cyber Attack on the Steam Pressure

In the previous sub-section we have shown that the attacker can cause major deviations from the NVP for all three valves. In this sub-section we analyze the effects that the previously presented deviations have on the steam pressure. Following the same experimental strategy, we have first recorded the steam pressure for a TS of 100ms followed by a TS of 1ms.

100ms Task Scheduling

The results from Figure 8 show that for an emulated network delay of 0s, the attacker is able to increase the steam pressure above 250 kg/cm^2 after only 2.5 minutes the attack is started. Moreover, the attacker is able to bring the process into a critical state even for a network delay of 0.5s. Only extreme network delays of 3s show a major impact and prevent the successful outcome of the attack, although the attacker is still able to produce major deviations from NVP, as previously shown in Figure 3. With larger packet losses the process reaches the critical state after 4 minutes for a 5% loss and 6 minutes for 10% loss, as shown in the same figure. The reason for this behavior is that all three valves still experience major deviations from their NVP, as shown in Figure 4, which causes the pressure to increase immediately after the attack is started. The background traffic affects the outcome of the attack only in extreme cases when it reaches the maximum network capacity. Otherwise, the attacker is able to open and close the valves, as shown in Figure 5, and to bring the process into the critical state.

1ms Task Scheduling

By decreasing the TS time to 1ms the attacker is not able to reach his goal in neither of the settings included in this study (Figure 9). Nevertheless, it is still able to increase the steam pressure to a maximum value of 234 kg/cm^2 for minimal network delay, packet loss and background traffic. The reason behind this is that the attacker is still able to completely close the steam valve, although it is able to cause only negligible deviations from the NVP for the other two valves, as shown in Figure 6.

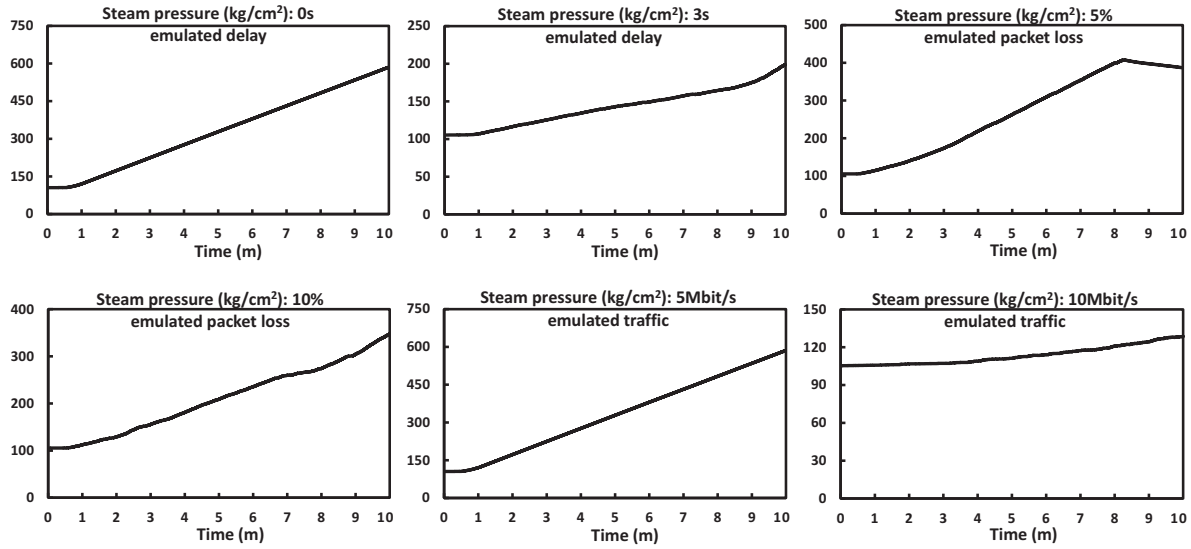


Figure 8: Effect of network delays, packet loss and background traffic on the steam pressure for 100ms TS

Table 1: Average valve positions and maximum pressure (P) during cyber attacks

Parameters		Fuel valve (Target: 0.34) 100ms	1ms	Steam valve (Target: 0.68) 100ms	1ms	Feed-water valve (Target: 0.43) 100ms	1ms	Max P (kg/cm^2) (Target: 108) 100ms	1ms
Delay (s)	0	0.93	0.34	0.04	0.05	0.03	0.433	585	234
	0.5	0.75	0.34	0.11	0.31	0.21	0.433	459	173
	3	0.4	0.34	0.25	0.51	0.39	0.433	198	131
Loss (%)	5	0.84	0.34	0.1	0.3	0.14	0.433	407	204
	10	0.61	0.34	0.17	0.46	0.2	0.433	347	145
Traffic (Mbit/s)	5	0.92	0.34	0.04	0.05	0.03	0.433	586	234
	10	0.34	0.34	0.55	0.67	0.43	0.433	128	110

By increasing the emulated network delays the maximum pressure induced by the attacker reduces gradually from 173 kg/cm^2 for 0.5s to 131 kg/cm^2 for 3s (Table 1). However, larger packet losses and background traffic do not produce major differences in the outcome of the attack, unless extreme values are used. Nevertheless, a 5% packet loss is still able to reduce the maximum pressure to 204 kg/cm^2, while a 10% packet loss reduces the maximum pressure to 145 kg/cm^2. In the extreme case of 10Mbit/s background traffic the maximum pressure is reduced to 110 kg/cm^2.

The results from this section have shown that by decreasing the TS to 1ms the physical process is able to react more efficiently to cyber attacks. Network delays, packet losses and background traffic have also shown to have an influence on the attack. Nevertheless, by using a TS of 100ms these are able to affect the outcome of the attack only in extreme cases. Consequently, designers should consider using a lower TS whenever possible in order to prevent the successful outcome of similar attacks.

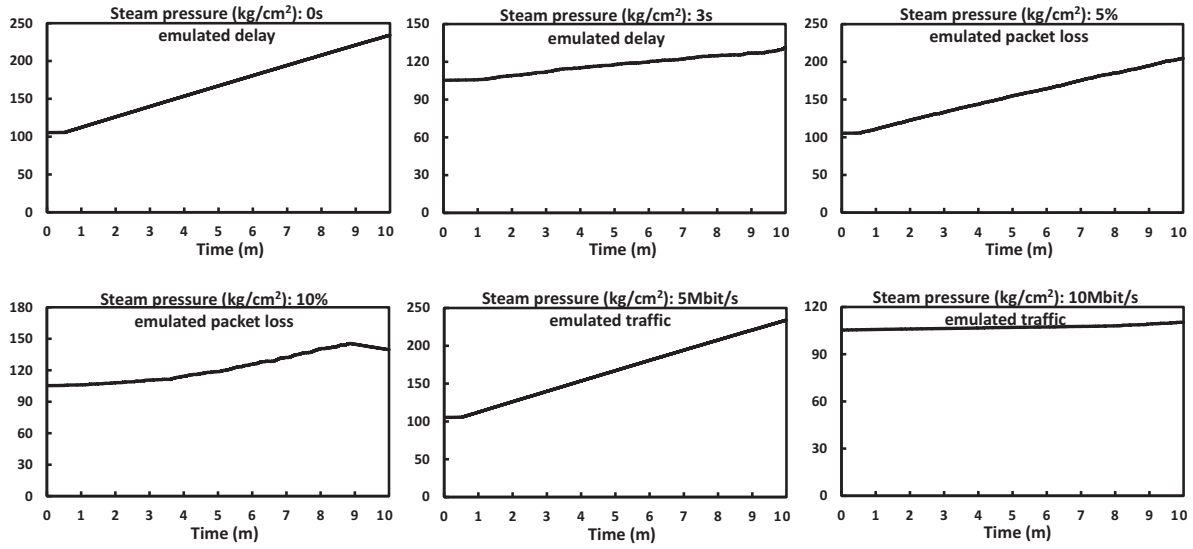


Figure 9: Effect of network delays, packet loss and background traffic on the steam pressure for 1ms TS

6 Conclusion

The study presented in this paper showed that cyber attacks exploiting knowledge on NICS can use regular Modbus packets to bring the physical process into a critical state. Within this scenario we evaluated the impact of network and installation-specific parameters on cyber attacks targeting a power plant. The experimental results showed that while communications parameters such as network delays, packet losses and background traffic have a limited effect on the attack, task scheduling and properties of physical processes, i.e. the speed of control valves, can become effective measures for increasing the resilience of physical processes. The main contribution of this paper is that it identifies two key parameters that could be adopted at design-time to increase the resilience of physical processes confronted with cyber attacks. The first one, i.e. control code task scheduling, provides engineers an efficient mechanism to counterbalance disturbances caused by malicious command packets, while the second one, i.e. the speed of control valves, provides insight into the way that an attacker might manipulate knowledge on physical properties to bring the process into a critical state. Such properties should be taken into account at process design time, which will lead to a more resilient physical process.

Bibliography

[1] S. East, J. Butts, M. Papa, S. Shenoi, A Taxonomy of Attacks on the DNP3 Protocol, in *Proceedings of IFIP Advances in Information and Communication Technology*, 311:67–81, 2009.

[2] T.C. Aseri, N. Singla, Enhanced Security Protocol in Wireless Sensor Networks, *International Journal of Computers Communications & Control*, 6(2):214–221, 2011.

[3] The Symantec Stuxnet Dossier, 2010, http://www.wired.com /images_blogs/threatlevel/2010/11/w32_stuxnet_dossier.pdf

[4] A.S. Brown, SCADA vs. the Hackers - Can Freebie Software and a Can of Pringles Bring Down the U.S. Power Grid?, *Mechanical Engineering*, 124(12), 2002.

[5] I. Nai Fovino, M. Masera, L. Guidi, G. Carpi, An Experimental Platform for Assessing SCADA Vulnerabilities and Countermeasures in Power Plants, in *Proceedings of Human System Interactions*, pp. 679–686, 2010.

[6] I. Nai Fovino, A. Carcano, T. De Lacheze Murel, M. Masera, A. Trombetta, Distributed Critical State Detection System for Industrial Protocols, in *Proceedings of IFIP International Conference on Critical Infrastructure Protection*, pp. 95–110, 2010.

[7] B. Genge, C. Siaterlis, I. Nai Fovino, M. Masera, A Cyber-Physical Experimentation Environment for the Security Analysis of Networked Industrial Control Systems, *Computers & Electrical Engineering*, In Press, 2012.

[8] B. White, J. Lepreau, L. Stoller, R. Ricci, S. Guruprasad, M. Newbold, M. Hibler, C. Barb, A. Joglekar, An Integrated Experimental Environment for Distributed Systems and Networks, in *Proceedings of the 5th symposium on Operating systems design and implementation*, pp. 255–270, 2002.

[9] C. Siaterlis, A. Garcia, B. Genge, On the Use of Emulab Testbeds for Scientifically Rigorous Experiments, *IEEE Communications Surveys & Tutorials*, PP(99):1–14, 2012.

[10] R.D. Bell, K.J. Åström, Dynamic Models for Boiler-Turbine Alternator Units: Data Logs and Parameter Estimation for a 160MW Unit, *Lundt Institute of Technology, Report TFRT–3192*, Sweden, 1987.

[11] L. Rizzo, Dummynet: A Simple Approach to the Evaluation of Network Protocols, *ACM Computer Communication Review*, 27(1):31–41, 1997.

[12] M Carbone, L. Rizzo, Dummynet Revisited, *ACM SIGCOMM Computer Communication Review*, 40(2):12–20, 2010.

[13] NLANR/DAST, Iperf: The TCP/UDP Bandwidth Measurement Tool, http://sourceforge.net/projects/iperf/

[14] W. Tan, H.J. Marquez, T. Chen, J. Liu, Analysis and Control of a Nonlinear Boiler-Turbine Unit, *Journal of Process Control*, Elsevier, 15(8):883–891, 2005.

[15] C. Queiroz, A. Mahmood, J. Hu, Z. Tari, X. Yu, Building a SCADA Security Testbed, in *Proceedings of the International Conference on Network and System Security*, pp. 357–364, 2009.

[16] C.M. Davis, J.E. Tate, H. Okhravi, C. Grier, T.J. Overbye, D. Nicol, SCADA Cyber Security Testbed Development, in *Proceedings of the North American Power Symposium*, pp. 483–488, 2006.

[17] R. Chabukswar, B. Sinopoli, G. Karsai, A. Giani, H. Neema, A. Davis, Simulation of Network Attacks on SCADA Systems, *First Workshop on Secure Control Systems*, April, 2010.

[18] A. Cárdenas, S. Amin, Z.S. Lin, Y.L. Huang, Chi-Y. Huang, S. Sastry, Attacks Against Process Control Systems: Risk Assessment, Detection, and Response, in *Proceedings of the ACM Symposium on Information, Computer and Communications Security*, pp. 355–366, 2011.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 688-695

An Electromagnetism-Like Approach for Solving the Low Autocorrelation Binary Sequence Problem

J. Kratica

Jozef Kratica
1. Mathematical Institute, Serbian Academy of Sciences and Arts
Kneza Mihaila 36/III, 11 000 Belgrade, Serbia
E-mail: jkratica@mi.sanu.ac.rs

Abstract: In this paper an electromagnetism-like approach (EM) for solving the low autocorrelation binary sequence problem (LABSP) is applied. This problem is a notoriously difficult computational problem and represents a major challenge to all search algorithms. Although EM has been applied to the topic of optimization in continuous space and a small number of studies on discrete problems, it has potential for solving this type of problems, since movement based on the attraction-repulsion mechanisms combined with the proposed scaling technique directs EM to promising search regions. Fast implementation of the local search procedure additionally improves the efficiency of the overall EM system.
Keywords: low autocorrelation binary sequence problem, electromagnetism-like metaheuristic, combinatorial optimization.

1 Introduction

Low autocorrelation binary sequence problem (LABSP) is a very hard combinatorial optimization problem with quite a simple formulation. The mathematical formulation of LABSP is based on a binary sequence s of length n. Let $s \in \{-1, 1\}^n$, i.e. s be represented by (s_1, s_2, ... , s_n), where $s_i \in \{-1, 1\}$ for $1 \leq i \leq n$. Each sequence s is associated with the value of its energy function which is defined as follows:

$$\begin{aligned} E(s) &= \sum_{j=1}^{n-1} C_j^2(s), \text{ where} \\ C_j(s) &= \sum_{i=1}^{n-j} s_i s_{i+j} \end{aligned} \tag{1}$$

Now, the low autocorrelation problem for binary sequences with length n, can be formulated as finding a sequence s of length n whose energy function is as minimal as possible. The second measure of quality of the sequence s is a merit factor

$$F(s) = \frac{n^2}{2E(s)}, \tag{2}$$

defined by Bernasconi in [2]. Mathematically, LABSP can be formulated as $\max_{s \in \{-1,1\}^n} F(s)$. Both formulations are equivalent, and either of them can be used when it is convenient.

2 Previous work

LABSP has been deeply studied since the 1960s by both the communities of physics and artificial intelligence. There are two reasons behind this interest:

- It arises in many diverse areas including statistical mechanics and configuration state analysis [2], calibration of surface profile metrology tools [1], satellite and space applications [8], digital signal processing [16], etc.;

- LABSP is also a significant challenge of exact and/or heuristic applications, since it is known that the problem has "bit-flip" neighborhood structure of combinatorial landscapes [5, 6]. With this type of neighborhood, it is extremely steep around the optimum, which is sometimes referred to as "golf hole" landscapes and it poses a very difficult optimization problem. In that case, small changes in argument values usually cause drastic difference in objective value. For example, alteration of only one bit in binary sequence s can affect the objective value change by several tens of percents. From these reasons the LABSP is also listed as a problem 005 in the CSPLIB library.

Although Golay in [9] estimated that $\lim_{n\to\infty} F(s) = 12.32$, it is not well enough, because for dimensions between 21 and 60 the merit factor varies from $F(s) = 5.627$ for $n = 23$ up to $F(s) = 9.85$ for $n = 27$, which is obviously far from the estimated limit 12.32.

The state-of-the-art exact method given in [12,13] is based on exhaustive search and it solves problem optimally up to $n = 60$. The experimental research was carried out for several days on a multiprocessor cluster of 160 CPUs. Up to now it is the largest dimension with known optimal solution.

A hybrid evolutionary approach described in [4] combines the evolutionary search described in [14] and Kerninghan-Lin heuristic defined in [11]. That evolutionary approach uses a specially defined termination criterion based on statistical analysis of known optimal solutions and their asymptotic behavior.

A detailed analysis of different stand-alone local search strategies is given in [7]. That analysis is later used in embedding the best local search strategy within other metaheuristic approaches. The results indicate that pure evolutionary algorithm cannot cope with the complexity of the problem and the assistance of local-search operators it is required to provide optimal or suboptimal results consistently. As a best choice for solving LABSP a memetic algorithm endowed with a tabu search local searcher is proposed, and that approach consistently finds optimal sequences in considerably less time than approaches previously reported in the literature.

Another metaheuristic method for solving LABSP, based on the stochastic local search (SLS), is presented in [10]. In-depth analysis of LABSP fitness landscape and the white-box visualization get insights on how SLS can be effective and lead to a slightly better strategy.

Local search algorithm described in [15], on the other hand, uses a quite different strategy compared to previous local search approaches, which is based on the randomized form of backtracking. In that way, the optimization problem is reduced to a series of constraint satisfaction problems which are to be solved iteratively, with decreasing upper bounds on the given objective function. Experimental results indicate that the algorithm is time consuming. For example, the average running time for $n = 40$ is over 1000 seconds.

3 Proposed EM method

An electromagnetism-like (EM) metaheuristic is a powerful algorithm for global optimization that converges rapidly to optimum [3]. The method is also used for combinatorial optimization as a stand-alone approach or as an accompanying algorithm for other methods.

EM is a population based algorithm that can solve nonlinear optimization problems. In the following text each member p_k, $k = 1, 2, ..., m$ of the population maintained by the algorithm will be referred to as EM point (or solution), and the population itself will be referred as a solution set.

The proposed EM algorithm for solving LABSP is given by the following pseudo code:

EM points in the first iteration are randomly initialized from $[-1, 1]^n$ (function Random_Init()).

```
Algorithm EM
Random_Init()
while iteration < max_iteration do
    foreach EM point p_k in the solution_set do
        Calculate_objective_value(p_k)
        Local_search(p_k)
        Scale_solution(p_k)
    Calculate_charge_and_forces()
    Move()
```

Algorithm 1: EM pseudo code

For a given EM point p_k, sequence s is obtained by rounding, i.e. $s_i = \begin{cases} 1, & p_{ki} \geq 0 \\ -1, & p_{ki} \leq 0 \end{cases}$, for each coordinate $i = 1, ..., n$. Energy $E(s)$ and merit factor $F(s)$ are computed by using (1) and (2).

3.1 Local search and scaling

This step is used to move the sample points towards the local optima that are near them. Points are pushed towards the local valleys using a neighborhood search procedure. The local search method used in this algorithm is simple but effective. Regarding the importance of the local search step, it is described in Algorithm 2.

The proposed local search procedure uses the first improvement strategy, which means that when an improvement is detected, the improvement is immediately applied and local search continues. If for each member of sequence s swap produces energy value greater or equal than the original one, the local search ends with no further improvement.

In this implementation, scaling procedure is also applied, which additionally moves points towards solutions obtained by local search. It is considered only with some factor $\lambda \in (0, 1)$ to prevent falling into a local optimum and become trapped there. An EM point p_k is moved by the following formula

$$p_k \leftarrow \lambda \cdot p_k\prime + (1 - \lambda) \cdot p_k \tag{3}$$

where $p_k\prime$ denotes sequence s of the k-th EM point in the current iteration when the local search procedure finished its work.

Choosing appropriate value of the scale factor λ is significant for governing the search process. In the extremal case, when λ is close to 1, the search process will likely fall into a local optimum and become trapped. Another extremal case, when λ is equal to 0, obviously represents no-scaling situation. Experiments showed that $\lambda = 0.1$ is a good compromise which yields satisfactory results.

3.2 Attraction-repulsion mechanism

As can be seen from the literature, the strength of the EM algorithm lies in the idea of directing the sample points towards local optima utilizing an attraction-repulsion mechanism. Therefore, after applying the local search procedure to each solution in the current population, the solutions must be moved towards promising regions in order to get closer to the optimal solution.

Function Local_search(p_k)

repeat
  $impr \leftarrow$ **false**
  $i \leftarrow 0$
  while $i < n$ **and not** $impr$ **do**
    $i \leftarrow i + 1$
    for $j = 1$ *to* n **do**

$$C_j^{tmp} = \begin{cases} C_j - 2 \cdot s_j \cdot s_{i+j}, & j < n - i \\ C_j^{tmp} = C_j, & j \geq n - i \end{cases}$$

$$C_j^{new} = \begin{cases} C_j^{tmp} - 2 \cdot s_j \cdot s_{j-i}, & j \geq i \\ C_j^{tmp}, & j < i \end{cases}$$

    $E^{new}(s) = \sum_{j=1}^{n-1} (C_j^{new}(s))^2$
    if $E^{new}(s) < E(s)$ **then**
      $impr \leftarrow$ **true**
      $s_i \leftarrow -s_i$
      for $j = 1$ *to* n **do**
        $C_j \leftarrow C_j^{new}$
      $E(s) \leftarrow E^{new}(s)$
until not $impr$
$F(s) \leftarrow \frac{n^2}{2E(s)}$

Algorithm 2: Local search pseudo code

In this process, each sample point is considered as a charged particle. The charge of each sample point is calculated by the following formula:

$$q_i = \exp\left(-n \frac{f(p_i) - f(p^{best})}{\sum_{k=1}^{m} f(p_k) - f(p^{best})}\right). \tag{4}$$

The amount of charge relates to the value of the objective function ($f(p_k) = E(s)$) at the point, which also determines the magnitude of attraction or repulsion of the point over the sample population.

According to the superposition principle of electromagnetism theory, the force exerted on a point via another point is inversely proportional to the distance between the points and directly proportional to the product of their charges. Mathematically, the power of attraction or repulsion of charges is calculated as follows:

$$F_i = \sum_{j=1, j \neq i}^{m} F_i^j, \text{ where}$$
$$F_i^j = \begin{cases} \left(\frac{q_i q_j}{||p_j - p_i||^2}\right) \cdot (p_j - p_i), & f(p_j) < f(p_i) \\ \left(\frac{q_i q_j}{||p_j - p_i||^2}\right) \cdot (p_i - p_j), & f(p_j) \geq f(p_i) \end{cases} \tag{5}$$

where $\|p_i - p_j\|$ is the Euclidean distance between EM points p_i and p_j.

As mentioned before, by using the Move() procedure of the electromagnetism approach, current solutions are shifted towards the best ones. All the EM points are moved with the exception of the current best solution. Detailed explanations about movement are given in Algorithm 3.

```
Function Move()
foreach EM point p_k in the solution_set do
    if p_k ≠ p_best then
1       β ← Random[0, 1]
2       F_i ← F_i/‖F_i‖
3       for i = 1 to n do
4           if F_ki > 0 then
                p_ki ← p_ki + β · F_ki · (1 − p_ki)
5           else
                p_ki ← p_ki + β · F_ki · (p_ki + 1)
```

Algorithm 3: Move pseudo code

As can be seen from Algorithm 3, the movement of each EM point is in the direction of total force exerted on it by a random step length β. This length is generated from uniform distribution between [0,1]. As can be seen in [3], the candidate solutions have a nonzero probability to move to the unvisited solution along this direction when random step length is selected. Moreover, normalizing the total force exerted on each candidate solution implies that infeasible solutions cannot be produced.

4 Experimental results

In this section, the proposed EM solution procedure on LABSP is tested for n up to 40 nodes, for which the optimal solutions are known in the literature.

Each numerical experiment was repeated 20 times and the results are summarized in Table 1, which is organized as follows:

- The first three columns contain n, optimal solution value (merit factor $F(s)$) and the EM best solution obtained in 20 runs;
- The average running time (t) and number of iterations *iter* used to reach the final EM solution for the first time are given in the fourth and fifth columns, while the total running time t_{tot} necessary to finish EM is given in the sixth column.
- The last two columns (*agap* and σ) contain information on the average solution quality: *agap* is a percentage gap defined as $agap = \frac{1}{20}\sum_{i=1}^{20} gap_i$, where $gap_i = 100 * \frac{EM_{best} - EM_i}{EM_{best}}$ and EM_i represents the EM solution (merit factor $F(s)$) obtained in the i-th run, while σ is the standard deviation of gap_i, $i = 1, 2, ..., 20$, obtained by formula $\sigma = \sqrt{\frac{1}{20}\sum_{i=1}^{20}(gap_i - agap)^2}$.

The computational results were performed on an Intel 2.5 GHz single processor with 1GB memory, under Windows operating system. All EM runs were made with the following empirically determined parameters: $m = 10$, $iter_{max} = 100000$ and $\lambda = 0.1$. These values cause most charges to exhibit convergent behavior with a few individuals diverging, thereby providing a good balance between local and global search. In this case all these values were chosen experimentally as a matter of convenience because they provide good results.

Table 1: Computational results

n	Opt_{sol}	EM_{best}	t (sec)	t_{tot} (sec)	$agap$ (%)	σ (%)
3	4.500000	opt.	0.0010	1.5503	0.000	0.000
4	4.000000	opt.	0.0010	4.3613	0.000	0.000
5	6.250000	opt.	0.0010	4.8269	0.000	0.000
6	2.571429	opt.	0.0010	11.6996	0.000	0.000
7	8.166667	opt.	0.0010	9.8925	0.000	0.000
8	4.000000	opt.	0.0010	16.4588	0.000	0.000
9	3.375000	opt.	0.0010	17.3339	0.000	0.000
10	3.846154	opt.	0.0010	20.3198	0.000	0.000
11	12.100000	opt.	0.0017	16.8205	18.462	34.585
12	7.200000	opt.	0.0010	17.4573	0.000	0.000
13	14.083333	opt.	0.0031	17.3925	11.429	25.806
14	5.157895	opt.	0.0010	26.4173	0.000	0.000
15	7.500000	opt.	0.0900	22.1761	1.739	7.715
16	5.333333	opt.	0.0010	23.1151	0.000	0.000
17	4.515625	opt.	0.0031	25.3597	0.000	0.000
18	6.480000	opt.	0.0052	28.1081	2.424	7.453
19	6.224138	opt.	0.0249	23.8753	1.212	3.681
20	7.692308	opt.	0.3983	26.3590	8.235	12.230
21	8.480769	opt.	0.0865	26.3949	19.226	12.095
22	6.205128	opt.	0.0266	33.8676	3.404	7.048
23	5.627660	opt.	0.1937	30.9995	2.745	3.847
24	8.000000	opt.	0.2528	32.9262	19.003	13.390
25	8.680556	opt.	0.8880	32.2269	13.759	12.523
26	7.511111	opt.	1.6412	36.4104	4.887	9.350
27	9.851351	opt.	0.9698	35.7136	29.084	18.230
28	7.840000	opt.	1.1606	37.8760	17.338	11.389
29	6.782258	opt.	3.7152	36.8581	6.531	6.318
30	7.627119	opt.	1.7120	42.7611	13.471	10.877
31	7.171642	opt.	2.6340	43.7440	9.258	6.876
32	8.000000	opt.	2.7331	46.6714	15.540	11.667
33	8.507813	opt.	4.4168	49.8097	15.667	9.911
34	8.892308	opt.	6.9574	51.4971	25.079	9.309
35	8.390411	opt.	1.8435	52.5394	19.264	8.260
36	7.902439	opt.	2.4458	55.0151	17.239	8.781
37	7.959302	opt.	6.4090	55.7682	15.507	7.856
38	8.298851	opt.	2.6012	61.6151	20.118	9.361
39	7.681818	opt.	7.8379	62.7660	13.623	8.275
40	7.407407	opt.	9.4135	73.1222	14.369	6.279

Observing the data shown in Table 1, it is remarkable that EM identifies optimal solutions in all cases. Moreover, the EM performs very efficiently, since the total running time was less than 74 seconds with one million objective function evaluations. Note that, most of this time is spent after the EM reach optimal solution merely to satisfy the finishing criterion. Also mind that in the case when $n = 40$, search space is 2^{40} and EM searched only $1.17 \cdot 10^{-7}$ part of it to reach the optimal solution.

5 Conclusions and Future Works

In this article, a hybrid approach combining an electromagnetism-like method (EM) with a scaling technique for solving the LABSP is proposed. The fast local search procedure and the applied scaling scheme were adapted to facilitate the use of EM to boost the performance of the proposed algorithm. To show the efficiency of the proposed hybrid EM, a number of experiments was carried out and the results were compared with the optimal solutions taken from the literature. The obtained results clearly indicate that EM is a useful tool for solving this problem.

As a direction for future studies, it can be interesting to parallelize the EM and run it on a powerful multiprocessor computer. Another orientation of future research can be incorporation of this method in some exact solution framework.

Acknowledgments

This research was partially supported by Serbian Ministry of Education and Science under grants 174010 and 174033.

Bibliography

[1] S.K. Barber, P. Soldate, E.H. Anderson, R. Cambie, W.R. McKinney, P.Z. Takacs, D.L. Voronov, V.V. Yashchuk, Development of Pseudorandom Binary Arrays for Calibration of Surface Profile Metrology Tools, *Journal of Vacuum Science and Technology B: Microelectronics and Nanometer Structures*, Vol.27, No.6, pp.3213–3219, 2009.

[2] J. Bernasconi, Low Autocorrelation Binary Sequences: Statistical Mechanics and Conguration State Analysis, *Journal Physique*, Vol.48, pp.559–567, 1987.

[3] S.I. Birbil, S.C. Fang, An Electromagnetism-like Mechanism for Global Optimization, *Journal of Global Optimization*, Vol.25, pp.263–282, 2003.

[4] F. Brglez, X.Y. Li, M.F. Stallman, B. Militzer, Reliable Cost Prediction for Finding Optimal Solutions to LABS Problem: Evolutionary and Alternative Algorithms, *Fifth International Workshop on Frontiers in Evolutionary Algorithms*, Cary, NC, USA 2003.

[5] A.V. Eremeev, C.R. Reeves, Non-parametric Estimation of Properties of Combinatorial Landscapes, *Lecture notes on Computer Science*, Vol.2279, pp.31–40, 2002.

[6] F. Ferreira, J. Fontanari, P. Stadler, Landscape Sstatistics of the Low Autocorrelated Binary String Problem, *Journal of Physics A: Mathematical and General*, Vol.33, pp.8635–8647, 2000.

[7] J.E. Gallardo, C. Cotta, A.J. Fernandez, Finding Low Autocorrelation Binary Sequences with Memetic Algorithms, *Applied Soft Computing*, Vol.9, No.4, pp.1252–1262, 2009.

[8] R. Garello, N. Boujnah, Y. Jia, Design of Binary Sequences and Matrices for Space Applications, *Proceedings of the 2009 International Workshop on Satellite and Space Communications - IWSSC'09*, art. no. 5286416, pp.88–91, 2009.

[9] M.J.E. Golay, The Merit Factor of Long Low Autocorrelation Binary Sequences, *IEEE Transactions on Information Theory*, Vol.28, pp.543–549, 1982.

[10] S. Halim, R.H.C. Yap, F. Halim, Engineering Stochastic Local Search for the Low Autocorrelation Binary Sequence Problem, *Lecture Notes in Computer Science*, Vol.5202, pp.640–645, 2008.

[11] B.W. Kernighan, S. Lin, An Efficient Heuristic Procedure for Partitioning Graphs, *Bell System Technical Journal*, pp.291–307, 1970.

[12] S. Mertens, Exhaustive Search for Low-autocorrelation Binary Sequences, *Journal of Physics A: Mathematical and General*, Vol.29, pp.473–481, 1996.

[13] S. Mertens, H. Bauke, Ground States of the Bernasconi Model with Open Boundary Conditions, Available online at `http://odysseus.nat.uni-magdeburg.de/~mertens/bernasconi/open.dat`

[14] B. Militzer, M. Zamparelli, D. Beule, Evolutionary Search for Low Autocorrelated Binary Sequences, *IEEE Transactions on Evolutionary Computation*, Vol.2, pp.34–39, 1998.

[15] S. Prestwich, Exploiting Relaxation in Local Search for LABS, *Annals of Operations Research*, Vol.156, pp.129-141, 2007.

[16] A. Ukil, Low Autocorrelation Binary Sequences: Number Theory-Based Analysis for Minimum Energy Level, Barker codes, *Digital Signal Processing: A Review Journal*, Vol.20, No.2, pp.483–495, 2010.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 696-708

P2P Resource Sharing in Wired/Wireless Mixed Networks

J. Liao

Jianwei Liao
College of Computer and Information Science
Southwest University of China
400715, Beibei, Chongqing, China
E-mail: liaojianwei@il.is.s.u-tokyo.ac.jp

Abstract: This paper presents a new routing protocol called Manager-based Routing Protocol (MBRP) for sharing resources in wired/wireless mixed networks. MBRP specifies a manager node for a designated sub-network (called as a group), in which all nodes have the similar connection properties; then all manager nodes are employed to construct the backbone overlay network with ring topology. The manager nodes act as the proxies between the internal nodes in the group and the external world, that is not only for centralized management of all nodes to a certain extent, but also for avoiding the messages flooding in the whole network. The experimental results show that compared with Gnutella2, which uses super-peers to perform similar management work, the proposed MBRP has less lookup overhead including lookup latency and lookup hop count in the most of cases. Besides, the experiments also indicate that MBRP has well configurability and good scaling properties. In a word, MBRP has less transmission cost of the shared file data, and the latency for locating the sharing resources can be reduced to a great extent in the wired/wireless mixed networks.
Keywords: wired/wireless mixed network, resource sharing, manager-based routing protocol, backbone overlay network, peer-to-peer.

1 Introduction

Peer-to-Peer technology (P2P) is a widely used network technology, the typical P2P network relies on the computing power and bandwidth of all participant nodes, rather than a few gathered and dedicated servers for central coordination [1, 2]. According to the research and analysis on Internet traffic management conducted by ipoque Germany, P2P applications dominate Internet traffic from 50% to 90%, and the statistics from Chinese sources reveal that P2P traffic currently accounts for 70% of China's total network traffic [3]. This indicates that resource sharing via various P2P techniques contributes to the major part of resource sharing on the Internet [4].

In general, P2P systems implement an abstract overlay network, built at application layer on top of the native or physical network topology [2]. From the architecture view [5], P2P systems are generally divided into structured systems, unstructured systems [6] and hybrid systems [7]. A structured P2P system employs a globally consistent protocol to ensure that any node can efficiently find some of the peers that have the desired resources, even though the file is extremely rare. However, since DHT-like (Distributed Hash Table, DHT) data structure is employed for maintaining the whole structured P2P system, the scalability is a critical problem. Chord [8] is a typical structured P2P system. The unstructured P2P system is formed when the overlay links are established arbitrarily. In an unstructured P2P system, if a peer wants to lookup a piece of desired resources in the network, the query has to be flooded through the network to find the peers who have the desired sharing resources as many as possible. However, the unstructured P2P system uses flooding queries to discover the target objects, which may introduce lots of network traffic. Bittorrent [10] is a well-known unstructured P2P system. In addition, many structured P2P systems use stronger peers (super-peers or super-nodes [11]) as servers, and the client-peers are connected in a star-like fashion to a single super-peer. This architecture can simplify the

network architecture, but super-peers hold all routing information, even though a local search is also conducted by a relevant super-peer, thus the super-peers are apt to be overloaded. As examples for hybrid networks can be named modern implementations of Gnutella2 [12] and the eDonkey network [13].

Nowadays, various modern devices can access Internet by different resorts, and these devices also want to share resources with others, it is quite clear they can employ the P2P techniques. But different kinds of devices have the different properties and purposes of use, thus the sharing targets are inequality. For instance, widely used handheld devices may share a several megabytes mp3 audio file, and rarely share a size up to several gigabytes video file. But for normal desktops and laptops, the latter sharing is quite common; therefore, treating different kinds of peers as same is not an ideal strategy in the heterogeneous networks.

In this paper, we propose, implement and evaluate a new routing protocol called MBRP (Manager-based Routing Protocol) for constructing P2P resource sharing networks, in which there are many disparate devices. As a matter of fact, MBRP has been inspired by the hybrid P2P architecture, but in MBRP, the target sharing resources might be replicated and stored on the manager nodes, and the message forwarding may not conducted by manager nodes. In addition, although the P2P protocols mentioned above are scalable and efficient, they were designed originally for wired networks and are generally not suitable for wireless networks, in which nodes join and depart much more frequently. The main idea of MBRP is to organize the diverse devices into different groups according to their properties such as location, wired or wireless etc., and then appoint a manager node for each group to communicate with other groups. Since the devices in the same group have the similar properties, the expected resources are stored and shared in the same group with quite high probability, only the desired resource is not in the group, the inter-group communication is launched.

This paper is organized as follows: we present the design and implementation of MBRP in Section 2; the evaluation experiments and results are shown in Section 3; finally, we make concluding remarks in Section 4.

2 The Design and Implementation of MBRP

As various Internet-connected devices have the different properties including bandwidth, connection resorts and storage capacity, the proposed MBRP first divides all participant nodes into several groups according to their properties such as wired or wireless connection, then elects a manager node for each group for communicating with the external nodes belonging to other groups. Different from traditional hybrid P2P systems, in MBRP, the nodes may communicate with other internal nodes belonging to the same group directly without any intervention from the manager node. Besides, the manager nodes might cache the replicas of hot sharing resources to reduce the lookup and transmission overhead for the sharing objects.

2.1 The Architecture of MBRP Network

Figure 1 shows the topology of a heterogeneous network built by resorting to MBRP. All nodes are divided into several groups according to their connection property, i.e. wired or wireless access. In each group, there is only one proxy node called manager node, which is responsible for the management of other internal nodes in the same group. For instance, the internal nodes who want to communicate with other nodes belonging to the different groups, are supposed to resort to their manager node. In addition, all manager nodes are connected into a ring, a similar topology to Chord [8], but all internal nodes in the same group can connect to each other via different topologies.

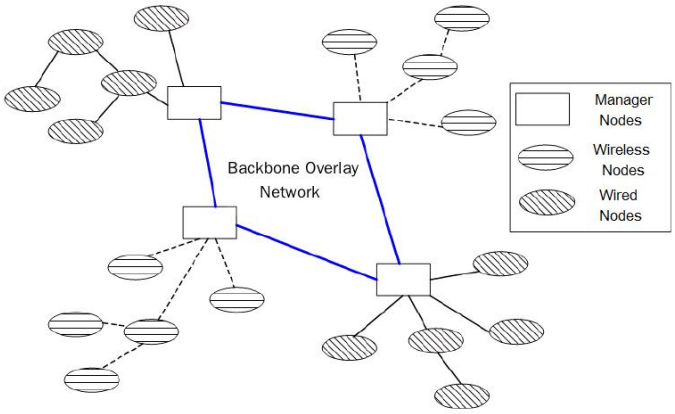


Figure 1: The Topology of Wireless/Wired Mixed Network via MBRP

2.2 Application Routing Protocol

We assume that MBRP is built on IP-based wide area network, we use a m-bit hash function to calculate the unique m-bit group ID. All internal nodes in the same group have their unique group ID, which is also used by the manager node as its own private ID to communicate with other manager nodes. For the purpose of routing in the whole network, like routing in Chord, each manager node holds its predecessor and finger table [14].

$$T_k = BNID + 2^{k-1} \ mod \ 2^m \quad (\forall k, where \ 1 \leq \ k \leq \ m) \tag{1}$$

Equation 1 is employed to calculate the manager node's successors, where *BNID* is the manager node ID, and m is the number of bits of manager node ID, T_k is the ID of the first successor; then, it calls the **find_successor**(T_k) function, which is shown in Figure 2, to calculate the next successor $T_k\prime$; finally, each manager node has m successors in its finger table when the **find_successor** function has been called *m-1* times. Figure 3A shows the finger table of node *N5512*, in which some example successors of node *N5512* are demonstrated.

```
n.find_successor (id)
    if (id ∈ (n, suceessor))
        return successor;
    else
        return successor.find_successor (id);
```

Figure 2: Find Manager Node's Successor

For the nodes who want to publish some sharing files after joining into the group, they are supposed to add the group ID as a part of the identification of the sharing files. Figure 3B shows that an internal node belonging to group *N2124* has published a file named as *K4814*, but the published name sent to the other manager nodes is *N2124#K4814*. For other manager nodes, that means the sharing file *K4814* is located in the *N2124* group.

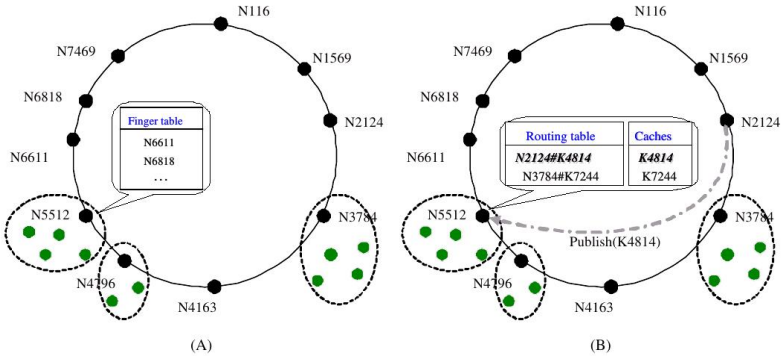


Figure 3: (A) Finger Table of *N5512*; (B) Routing Table and Cache Table of *N5512*

Creating Routing Table

As shown in Figure 3B, for publishing file *K4814*, the manager node *N2124* computes the first successor according to Equation 1, therefore, *N2124* publishes *K4814* to its successor manager node, i.e. *N5512*; then *N5512* adds the new entry *N2124#K4814* to its own routing table; finally, it makes a replica of *K4814* and an associated cache entry in the cache table if the cache strategy is enabled.

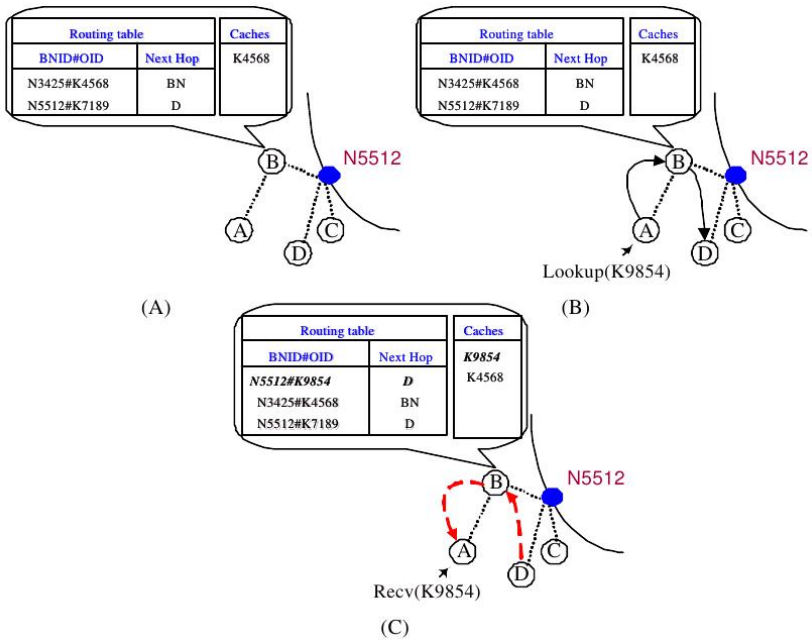


Figure 4: (A) Routing Table of Internal Node B; (B) Lookup Originated from Internal Node A to B and D; (C) Lookup Succeed Message Passed Back From D to B and A

While an internal node lookups a piece of sharing resource, which is in the different groups, then the lookup process can be taken over by its manager node. Otherwise, while the sharing file is in the same group, the desired file data can be transferred within the group directly. Thus, each internal node also holds a routing table to show the routines within the group. However, quite different from the manager node, the IDs of the internal nodes are their IP addresses; thus,

the routing table is different as well. The routing table of the internal node B (B represents node's IP address) is shown in the Figure 4. *BNID#OID* stands for the manager node ID and the sharing resource ID, and the Next Hop means the next node for getting the sharing resource. For instance, in order to obtain the sharing resource *K7189* located in group *N5512*, internal node B should send the lookup request to the next hop, i.e. the internal node D. The symbol *BN* in the routing table represents the ID of the manager node that in the same group.

Lookup Algorithm

The lookup algorithm will be described from 4 parts: sending a request, receiving a request, manager node routing on the backbone overlay network and receiving a lookup hit message. Figure 5A shows the algorithm of sending a lookup request. If a node wants to locate a piece of sharing resource labeled as *obj_ id*, then it calls **Send_LookupReq(obj_id)** to send the lookup request. First, the **find_cache(obj_id)** function is called to make sure whether the sharing object is in its own cache or not; if not,the **find_routetable(obj_id)** function will be called to obtain the related routing information; while there is no related routing information, it calls **get_approximated(obj_id)** to obtain the feasible routing information to find the next hop n'. Finally, **Send_Req(obj_id, n′)** is called to send the lookup request to the next hop n'. Figure 4A shows an example of sending a lookup request for the sharing object *K9854*, since there is no corresponding entry in the route table, the **get_approximated(K9854)** function is called to obtain the feasible routing entry and the lookup request is forwarded to the selected node D.

```
// send a lookup request to find the obj_id
Send_LookupReq (obj_id)
    r = nil;
    n' = nil;
    if (find_cache (obj_id) is not nil)
        r = get_routetable (obj_id);
    else
        r = get_approximated (obj_id);
    n' = r.nexthop;
    Send_Req (obj_id, n');
```

```
// receive a lookup request from the predecessor
Recv_LookupReq (obj_id)
    r = nil;
    n"= nil;
    if (find_cache (obj_id) is not nil)
        Send_Lookup_Hit (src, obj_id, data);
        return;
    if (find_routetable (obj_id) is not nil)
        r = get_routetable (obj_id);
    else
        r = get_approximated (obj_id);
    n" = r.nexthop;
    Send_Forward (obj_id, n");
```

Figure 5: (A) Sending a Lookup Request; (B) Receiving a Lookup Request

The algorithm of receiving a lookup request from the predecessor is almost same to that of sending a lookup request. It receives a lookup request, and then processes like sending a lookup request, the pseudo-code of the algorithm is shown in Figure 5B. While the target sharing object is found, then the **Send_LookupHit(obj_id, src, data)** function will be called to transfer the resource to the request node. Figure 4B shows an example of the procedure while the internal node B handles the received lookup request from node A. Since there is no corresponding routing entry in node B's routing table, it calls **get_approximated(obj_id)** to find the feasible next hop, i.e. node D, and forwards the request to it. The process of handling a received lookup request does not stop until the resource is found or timeout (i.e. maximum hop count exceeded).

While the lookup request is not fulfilled within the group, it will be forwarded to the manager node, the **Core_Ring_Route(obj_id)** function shown in Figure 6A, will be called by the group's manager node to handle the request from other manager nodes. After receiving the

```
// routing in backbone overlay network
Core_Ring_Route (obj_id)
    r = nil;
    n'= nil;
    b' = nil;
    if (find_cache (obj_id) is not nil)
        Send_Lookup_Hit (src, obj_id, data);
        return;
    if (find_local (obj_id) is not nil)
        n' = get_local (obj_id);
        Send_Forward (obj_id, n');
    else
        if (find_routetable (obj_id) is not nil)
            r = get_routetable (obj_id);
            b' = r.bnid;   // the manager ID
        else
            b' = find_successor (obj_id);
    Send_Core_Ring_Forward (obj_id, b');
```

```
// receive a object hit message
Recv_LookupHit (src, obj_id, data)
    update_routetable (src, obj_id);
    if (src is not equal to me)
        Send_Lookup_Hit _BackRoute
            (src, obj_id, data);
    if (find_caches (obj_id) is nil)
        append_caches (data);
    return;
```

Figure 6: (A) Routing in Backbone Overlay Network; (B) Algorithm of Receiving a Lookup Hit

lookup request, the manager node checks the target object is in its own group or not. If the object is in its group, then the request is forwarded to the corresponding internal node; otherwise, it forwards the request to other corresponding manager node whose group has the target object or the successor manager node in the finger table.

The algorithm of receiving a lookup hit message is shown in Figure 6B, while the manager node receives the lookup hit message, it first updates its routing table to label the new routine to the target object; then if the cache mechanism is enabled, it also makes a replica of the target object on its local disk.

From the above description, we can see that while the target sharing object is found, then the expected file will be transferred to the request node by the reverse routing path. At the same time, if cache strategy is enabled, one replica of this target object is made and stored in each manager node on the routing path for quick responses to the future lookup requests. We should mention that the number of cached replicas in the manager node is configured and limited, LRU is used to evict an existing replica and append the new replica into the local cache.

2.3 Dynamic Registration

Since the handheld devices have the roaming property, that means they might change their groups and manager nodes frequently, we have adopted a mechanism like IP mobility support [15] called dynamic handover, to allow the handheld device to register to a new manager node after the roaming. Figure 7 illustrates the dynamic registration in following steps:

1. The handheld device issues a registration request with its current IP address, the former IP address and the former manager node ID (hashed value from the manager node's IP address) to the new manager node for registration after it roams to another group.

2. The new manager node creates a new internal node record, and replies the handheld device with message about the registration has been handled. The handheld device updates its manager node and tries to re-build its routing table again.

3. The new manager node notifies the device's former manager node that the device has joined into its group and the old record should be removed; then the previous manager

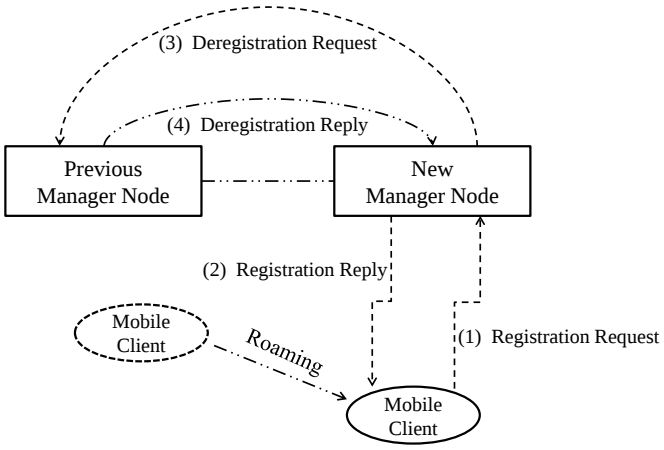


Figure 7: Dynamic Handover Algorithm

node deletes the corresponding record, and broadcasts that de-registration message to its all internal nodes to require them to update their routing tables.

4. The previous manager node replies to the new manager node with the message of the de-registration has been handled.

2.4 Manager Node Election

The manager node plays a critical role in the MBRP network, while the manager node fails or departs from the group, a new manager node is supposed to be elected. As a matter of fact, the main principle for electing a new manager node is the well relaying property, which means lots of internal nodes choose it as the next hop in their routing table entries; if more than one internal nodes have the same relaying property, then one of them will be selected randomly.

On the premise that network is still working when the manager node has exited or failed, the internal node, who first detects the failure or exit of the manager node, broadcasts the election request to other internal nodes. We assume the messages related to election are never lost, then all internal nodes belonging to the same group should take part in the process of election.

1. After receiving the election request, all internal nodes check the possible manager node candidate(s) according to the relaying property, and then reply to the issuer of the election request with the candidate(s). It is possible that the internal nodes may send several candidates who have the same relaying property;

2. The issuer of election request collects all replies from the internal nodes, then determines which node is the unique manager node;

3. The result of election will be broadcast to the corresponding internal nodes;

4. The new created manager node should be insert into the backbone overlay network simply like inserting a node into a ring topology network. In general, the new manager node replaces the failure one in the ring of the backbone overlay network;

5. All nodes in the group (including the manager node), which have a new manager node, remove the records with previous manager node and update the manager node information;

6. The new manager node broadcasts that the former manager node is not working, to other manager nodes on the backbone network and requires them to update the routing table. At the same time, the new manager node re-builds its routing table and finger table.

We should mention that because all nodes in the group should re-build their routing tables, and routines on the backbone overlay network are supposed to be updated as well, the overhead brought by electing a manager node is not trivial.

3 Experiments and Evaluation

The NS2 [16] was employed as our experimental platform while analyzing the performance and overhead on both the MBRP system and its comparison counterpart. Much more exactly, the module Gnutellasim in NS2 is used for our experiments. The hybrid P2P system Gnutella2, which uses super-peers (called hubs) to manage the internal nodes in the same group, has been selected as our comparison counterpart while evaluating the overhead, such as network traffic, in the manager nodes in the MBRP network system.

Moreover, the manager nodes play significant roles in our proposed mechanism MBRP, so that they are supposed to have enough bandwidth and processing power. Because wireless network is connected to the wired network via a gateway (also called access point), in our experiments, we selected such nodes as manager nodes for wireless groups. Regarding to wired groups, we simply appointed the fixed servers as manager nodes. Thus, all nodes can join the whole system by registering to their own manager nodes.

3.1 Overhead on Manager Node

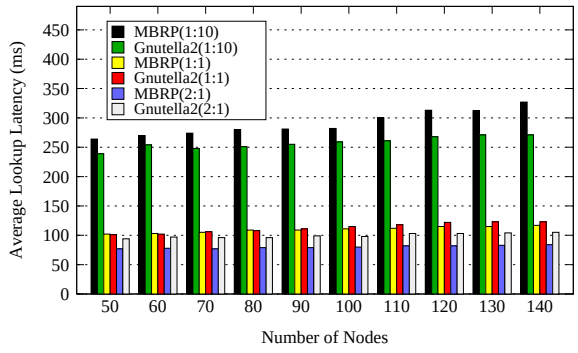


Figure 8: The Average Lookup Latency

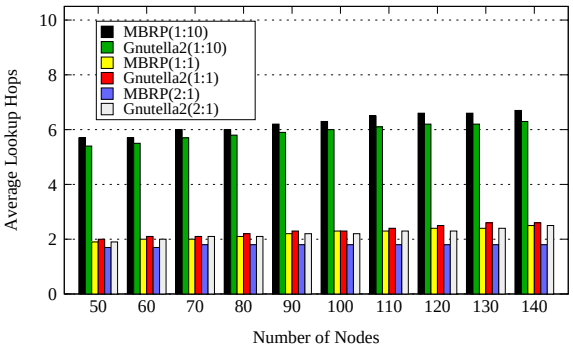


Figure 9: The Average Lookup length

We have conducted the comparison experiments between Gnutella2 and the proposed MBRP, to show the overhead of introducing the manager nodes in MBRP. For both target systems, the overhead of locating the local objects (belonging to the same group) and locating the external objects (belonging to the different groups) are different, we adopted different Internal/External Access Ratios, i.e. 1:10, 1:1, 2:1, to show the different overhead; each group has 50 internal nodes and each experiments stands 200 seconds.

We measured both the average lookup latency and the average lookup hops for Gnutella2 and MBRP with different access ratios. Figure 8 reports the results of average lookup latency (lower is better), except for the case of Internal/External Access Ratio is 1:10, while no less than half of lookups for sharing resources are hit within the group (i.e. access ratios are 1:1 and 2:1), the lookup latency introduced by MBRP is less than Gnutella2, especially while the access

ratio is 2:1, MBRP reduces around 20% lookup latency. That is because all communications within the group should be handled by the super-peers in Gnutella2, but the internal nodes may communicate with each other in MBRP.

Figure 9 presents the results of average lookup length(i.e. hot count, lower is better), the similar tendency to the average lookup latency between Gnutella2 and MBRP. In addition, from Figures 8 and 9, we can conclude that MBRP needs just a little more time while increasing the total number of nodes, this shows that MBRP has well scalability.

3.2 Overhead on Backbone Overlay Network

NS2-Gnutellasim was also employed to show the traffic on the backbone overlay network with the different network properties. In order to accelerate the access speed to the sharing resources, MBRP applies cache mechanism to store the hot resources in the manager nodes when these resources are transferred via/to them. In this section, we will inspect that different cache strategies bring about what kind of benefits to the lookup latency and negative effect on the traffic overhead in backbone overlay network respectively. The following cache strategies have been taken into consideration:

1. **No Cache**, which represents that the cache mechanism is disabled.

2. **Unlimited Cache**, which means the manager nodes can cache unbounded replicas.

3. **Weighted Cache**, which indicates that the manager nodes can cache limited resources, while the ceiling is reached, some existing cached resources should be swapped out to make space for the new replicas with LRU cache algorithm.

Because in both of Gnutella2 and the MBRP mechanisms, all inter-group messages are handled by the manager nodes or super-peers, the traffic overhead on backbone overlay network of Gnutella2 is same to that of MBRP without cache. We do not report the experimental results regarding to Gnutella2 in the following experiments.

In these experiments, GT-ITM [17] is used to construct the network topology, which has 500 manager nodes, the size of the sharing resource is 1024 byte, Wired/Wireless access Ratio is 1:1, and the duration of each experiment is 500 seconds. Since P2P is an application level protocol, we only care about application level packets rather than other level packets.

Average Traffic Overhead

We defined the average traffic overhead as the traffic on backbone overlay network divided by the size of sharing resource. For instance, in order to transfer a sharing file (default size as 1024 byte), which introduces 4096 byte total traffic on backbone overlay network, then the average traffic overhead is 4.

Figure 10 shows the average traffic overhead by using different cache strategies, the X axis stands for the number of successful lookups, in other words, during 500 seconds for conducting an experiment, how many successful lookups have been completed; the Y axis represents the average traffic overhead. Without doubt, No Cache strategy works worst, meanwhile Unlimited Cache mechanism works best. Figure 11 shows the total traffic in the all manager nodes, which has the similar trend to that of average traffic, In addition, Figures 10 and 11 also show that with the increasing the maximum number of caches, the total traffic goes down slowly.

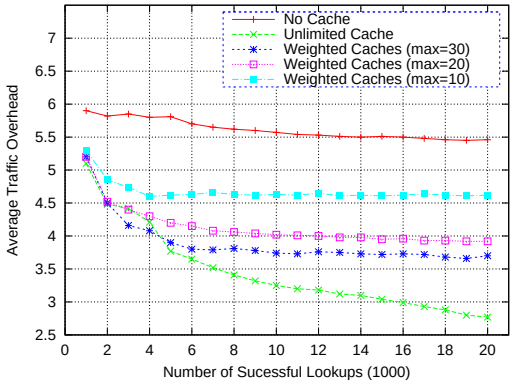


Figure 10: Average Traffic Overhead

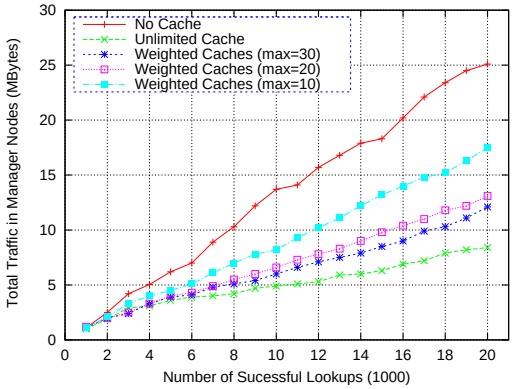


Figure 11: Total Traffic in Manager Nodes

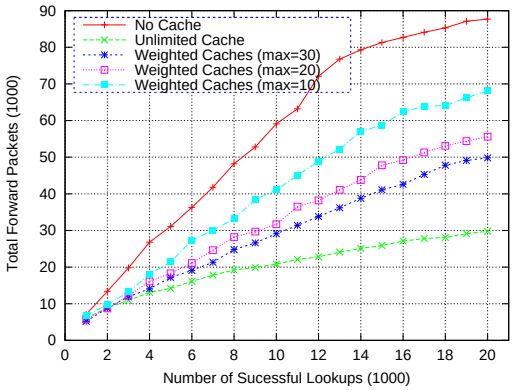


Figure 12: Total Forward Packets on Backbone Overlay Network

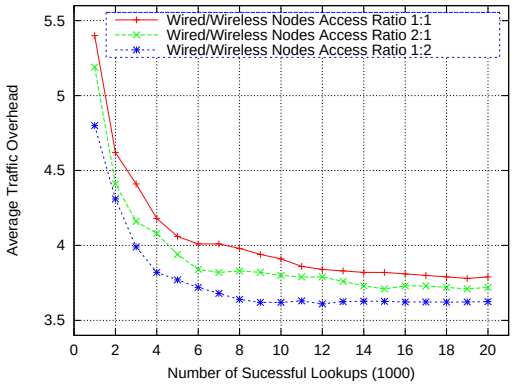


Figure 13: Average Traffic Overhead with Different Wired/Wireless Access Ratio

Total Traffic Overhead

Figure 12 shows the total forward packets on the backbone overlay network. While the number of cached resources is increasing, then less packets is forwarded on the backbone overlay network. That is because the more cached replicas, the more lookup hits can be obtained within the group.

Overhead with Different Wired/Wireless Access Ratios

In the previous experiments, we fixed the wired/wireless access ratio as 1:1. In this section, we will discuss the ratios are 1:1, 2:1 and 1:2 respectively. First, we configured the cache strategy as Weighted Caches (max=30). Then we repeated to measure the average traffic overhead, total traffic in manager nodes and total forward packets on the backbone overlay network.

Figures 13, 14 and 15 show the relevant results of overhead on backbone overlay network with different wired/wireless access ratios separately. From these figures, we can see that while the ratio is 1:2, the MBRP with cache mechanism can achieve considerable performance improvement because of a major part of lookup hits occurred in the groups.

3.3 Discussion

From the above experiments, we can see that MBRP has low latency, and the cache mechanism is also suitable for a large amount of accessing to the sharing resources in wireless/wired mixed networks. In addition, MBRP keeps a good scaling property, it employs manager nodes

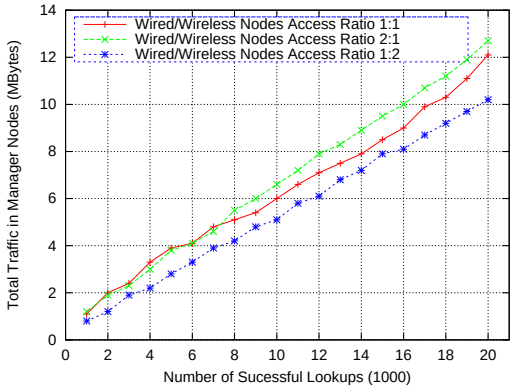


Figure 14: Total Traffic in Manager Nodes with Different Wired/Wireless Access Ratio

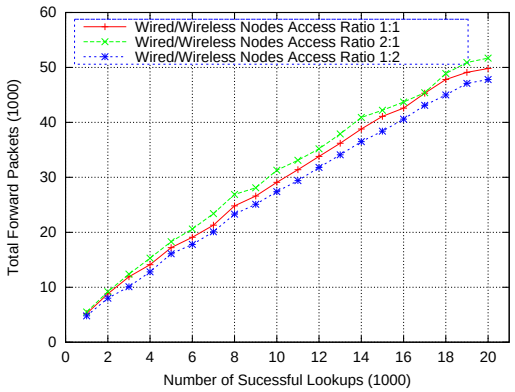


Figure 15: Total Forward Packets with Different Wired/Wireless Access Ratio

to construct the backbone overlay network, and then other nodes can register to the manager node to join into the whole network. The manager nodes are responsible for the communication between the different groups, therefore, both traffic overhead and lookup hops do not increase drastically even though lots of the new nodes join into the whole network suddenly.

4 Concluding Remarks and Future Work

A new routing protocol named as Manager-based Routing Protocol (MBRP) has been proposed, implemented and evaluated in this paper. All nodes in the network have been partitioned into several groups according to their properties, wired or wireless access for instance; then a manager node is elected for each group and in charge of communication between the internal nodes in the group and external nodes belonging to other groups. From our experimental results, compared with Gnutella2, which has super-peers for different groups, except the Internal/External Locating Ratio is 1:10, in which MBRP performs a little worse than Gnutella2; in other cases, MBRP outperforms than Gnutella2. In addition, since the nodes in the same group, which have the similar properties, are mostly like to share the same kind of resources, the most of resource sharing cases may occur within the group. Namely, MBRP can work well in the heterogeneous networks, in which internal accesses might more than external ones.

Furthermore, for responding quickly to the lookup requests for the hot resources, MBRP adopts caching the hot resources in the group while transferring the target objects from external groups. Therefore, the future lookups for these cached resources can be fulfilled in the local group. The cache strategy reduces not only the lookup latency and lookup node hops, but also

the network traffics on the backbone overlay network. Consequently, the system performance can be upgraded greatly.

However, the current design of MBRP still has it limitations, although we assume that the sharing resources are not modified frequently, modification of the resources really happen, thus we need to consider how to maintain the consistency between the cached copies and the original ones in the near future. In addition, to determine which manager nodes for storing the copies, and make the cache mechanism much more effective is another aspect of our future work.

Bibliography

[1] G. Fox, Peer-to-peer networks, Computing in Science and Engineering, ISSN 1521-9615, 3(3):75-77, 2001.

[2] R. Schollmeier , A definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications, Proceedings of the First International Conference on Peer-to-Peer Computing, pp.101-102, 2001.

[3] Eric Bangeman, P2P responsible for as much as 90 percent of all Net traffic (http://arstechnica.com/old/content/2007/09/p2p-responsible-for-as-much-as-90-percent-of-all-net-tra_c.ars, 2007

[4] M. Parameswaran, A. Susarla, A.B. Whinston, P2P networking: an information sharing alternative, Computer, ISSN 0018-9162, 34(7):31-38, 2001.

[5] Analoui M., Sharifi M., Rezvani M.H., Probabilistic Proximity-aware Resource Location in Peer-to-Peer Networks Using Resource Replication, International Journal of Computers Communications & Control, ISSN 1841-9836, 5(4):447-457, 2010.

[6] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker , Search and replication in unstructured peer-to-peer networks, Proceedings of the 16th international conference on Supercomputing,ICS'02, pp.84-95, 2002.

[7] Beneventano, Domenico and Bergamaschi, Sonia and Guerra, Francesco and Vincini, Maurizio , Querying a super-peer in a schema-based super-peer network, Proceedings of the 2005/2006 international conference on Databases, information systems, and peer-to-peer computing, pp.12-25, 2007.

[8] M. Kelaskar, V. Matossian, P. Mehra, D. Paul, M. Parashar, A Study of Discovery Mechanisms for Peer-to-Peer Application, Proceedings of the 2nd IEEE/ACM Intenational Symposium on Cluster Computing and the Grid,IEEE Computer Society Washington, DC, USA, pp.444, 2002.

[9] Sameh Elfiansary , Luc Onana Alima , Per Brand , Seif Haridi , Efficient Broadcast in Structured P2P Networks, Proceedings of 2nd International Workshop On Peer-To-Peer Systems, 2003.

[10] BitTorrent, http://www.bittorrent.com

[11] J. Lin, M. Yang, Robust Super-Peer-Based P2P File-Sharing Systems, the Computer Journal, ISSN 0010-4620, 53(7):951-968, 2010.

[12] M. Ripeanu , Peer-to-peer architecture case study: Gnutella network, Proceedings of the First International Conference on Peer-to-Peer Computing, pp.99-100, 2002.

[13] S.B. Handurukande et al , Peer sharing behaviour in the eDonkey network, and implications for the design of server-less file sharing systems, Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems 2006, pp. 359-371, 2006.

[14] Stoica, Ion and Morris, Robert and Karger, David and Kaashoek, M. Frans and Balakrishnan, Hari, Chord: A scalable peer-to-peer lookup service for internet applications, SIGCOMM Comput. Commun. Rev. ISSN 0146-4833, 31(4):149-160, 2001.

[15] IP Mobility Support for IPv4, http://tools.ietf.org/html/rfc3344

[16] Rupali Bhardwaj and V.S. Dixit and Anil Kr. Upadhyay, An Overview on Tools for Peer to Peer Network Simulation, International Journal of Computer Applications, ISSN 0975-8887, 1(1):70-76, 2010.

[17] E. W. Zegura, GT-ITM: Georgia Tech internetwork topology models (software), http://www.cc.gatech.edu/fac/Ellen.Zegura/gt-itm/gt-itm.tar.gz, 1996.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 709-720

Distributed Collaborative Processing under Communication Delay over Wireless Sensor and Actuator Networks

L. Mo, B. Xu

Lei Mo, Bugong Xu
Key Laboratory of Autonomous Systems and Network Control,
Ministry of Education
College of Automation Science and Engineering
South China University of Technology
Guangzhou 510640, China
E-mail: mo.lei@mail.scut.edu.cn, aubgxu@scut.edu.cn

Abstract:
In wireless sensor and actuator networks (WSANs), the sensor nodes are involved in gathering information about the physical phenomenon, while the actuator nodes take decisions and then perform appropriate actions upon the environment. The collaborative operation of sensor and actuator nodes brings significant advantages over WSNs, including improved accuracy and timely actions upon the sensed phenomena. However, unreliable wireless communication and finding a proper control strategy cause challenges in designing such network control system. In order to accomplish effective sensing and acting tasks, efficient coordination mechanisms among different nodes are required. In this paper, the coordination and communication problems in WSANs are studied. First, we formulate the mathematical models for the WSANs system. Then, a predictor-controller algorithm based on distributed estimation is adopted to mitigate the effects of network-induced delay. Finally, we apply a collaborative processing mechanism to meet the desired system requirements and improve the overall control performance. This approach will group the sensor and actuator nodes to work in parallel so as to reduce the computation complexity and enhance the system reacting time. Simulation results demonstrate the effectiveness of our proposed method.
Keywords: Wireless sensor and actuator networks, Distributed estimation, Collaborative processing.

1 Introduction

Wireless sensor and actuator networks(WSANs) comprise groups of sensor and actuator nodes that are connected with wireless medium. It is an important extension of wireless sensor networks (WSNs), allowing actuator nodes within the network to make autonomous decisions and then perform appropriate actions in response to the sensor nodes measurements [1]. Thus, the novel network architecture performs not only 'read' operations, but also 'write' operations, which brings about unique and new challenges that need to be addressed [2]. In order to satisfy the requirements introduced by the coexistence of sensor and actuator nodes, multiple coordination levels among nodes are required to implement, which can be defined as: Sensor-Sensor(S-S), Sensor-Actuator(S-A) and Actuator-Actuator(A-A). The S-S coordination is similar to the scheme already used in wireless sensor networks applications. Thus, in this paper we mainly focus on the S-A and A-A coordination.

Due to the unreliable wireless communications, system noise and time-delay are the common phenomenons which will influence the overall system performance. To this end, it is quite necessary for the nodes to perform the estimation and compensation algorithms of the required information [3]. A queuing strategy is introduced both in controllers and actuator nodes in [4], and the time delay between controllers and actuator nodes is compensated by multi-step control increment given by the algorithm of general predictive control. The work given in [5]

presents a real-time architecture for automated WSANs, the delay bound of S-A communication is maintained by the distributed mechanisms for S-A event reporting and self-aware coordination. In [6] a general reliability-centric framework for event reporting in WSANs is proposed, which contains an fault-tolerant event data aggregation algorithm and a delay-aware data transmission protocol. In [7], the authors study the model-based predictive networked control systems that compensate random delays and data loss of the communication, and use a predictive control scheme to avoid performance loss. Our work is motivated by the above studies. The key difference is that we focus on the S-A delay, not the uniform network nodes delay. Moreover, we apply a predictor-controller algorithm based on the state estimation to mitigate the detrimental effects of the communication delay. In this context, the model of the WSANs system needs to be analyzed in detail.

Finding a proper control strategy is still the core in designing the A-A coordination [8], This process involves which actuator node should be scheduled to execute a specific task and how to adjust its actuation to meet the desired system requirements. According to the way data is routed among different actuator nodes, control strategies can be categorized into the distributed control (DC) and centralized control (CC) scheme [9]. In the DC scheme, the control decision of a signal actuator node relies on the local information received from its neighbor nodes rather than global information [10]. Then it can achieve a superior performance in modularity, integrated diagnostics, quick and easy maintenance and low cost. In [11] a framework of optimizing a collaborative sensing and actuation system is built for environment control, the sensor is set in the actuator node and the control objection is to balance the energy saving against the spatial smoothness of the control signals. In [12], the authors propose two control schemes in WSANs for building-environment control systems, a CC scheme in which control decisions are made based on global information, and a DC scheme that enables distributed actuator nodes to make decisions locally. In [8] a new distributed estimation and collaborative control scheme is proposed for industrial control systems with WSANs, which can achieve robust control against inaccurate system parameters. In this paper, we focus on the problem of utilize distributed sensor measurements to design control strategies in order to elicit a desired response from the monitored environment. Our methodology incorporates a dynamic clustering schedule into the collaborative estimation and control framework, which can minimize the control error and improve the control quality.

The remainder of this paper is organized as follows: Section 2 models the WSANs system. While Section 3 provides a delay compensation algorithm. Then a distributed collaborative proceeding method is designed in Section 4. At last, the results of simulations conducted to explore the performance of proposed algorithms are demonstrated in Section 5.

2 System Models

We consider the WSANs system that are employed to the industrial instrumentation and control applications. The control objective is to adjust the system variables to meet our requirements. A set of static sensor and actuator nodes that are spread throughout the region of interest (ROI) to detect and track events and take necessary actions. Let x denote the system variable of our concern, such as temperature, brightness, humidity, sound, pressure, vibrations, etc. in different parts of the field. Let S^A represent the set of actuator nodes, with $n_a = |S^A|$. Let S^S represent the set of sensor nodes, with $n_s = |S^S|$. We make the following assumption of our network: (1) Sensor node is the time-driven device, input reception or output transmission is controlled by a sample time, while the actuator node is the event-driven device depends on the control techniques used; (2) Sensor and actuator nodes are aware of their geographical position; (3) The network is synchronized by means of one of the existing synchronization protocols [13];

(4) The randomly varying delays between S-A are bounded; (5) The system is observable and controllable.

Let S_i denote the ith sensor node, each node has the following model:

$$z_i(k) = h_i x_i(k) + \nu_i(k), i = 1, ..., n_s \tag{1}$$

where h_i and $\nu_i(k)$ are the observation item and measurement noise, respectively. Assume that $\nu_i(k)$ is a zero-mean Gaussian white noise with $E\{\nu_i(k)\} = 0$, $E\{\nu_i(k)\nu_j^T(l)\} = r_i(k)\delta_{kl}\delta_{ij}$, where $\delta_{kl} = 1$ if $k = l$, and $\delta_{kl} = 0$, otherwise. Then the matrix form of Eq.(1) is:

$$Z(k) = HX(k) + \nu(k) \tag{2}$$

where $Z(k) = [z_1(k), ..., z_{n_s}(k)]^T$, $H = diag[h_1, ..., h_{n_s}]$, $X(k) = [x_1(k), ..., x_{n_s}(k)]^T$ and $\nu(k) = [\nu_1(k), ..., \nu_{n_s}(k)]^T$.

Let A_j denote the jth actuator node, f_j denote its output which influences its ambient plant state, u_j denote the control signal that is used to adjust A_j's actuation. The change of each actuator node's actuation is assumed linearly proportional to the control signal received by this node, which is modeled as:

$$f_j(k) = gu_j(k), j = 1, ..., n_a \tag{3}$$

where g is the transfer function of A_j. Here, we consider a scenario with homogenous actuator nodes. Then the matrix form of Eq.(3) is:

$$F(k) = GU(k) \tag{4}$$

where $F(k) = [f_1(k), ..., f_{n_a}(k)]^T$, $G = diag[g_1, ..., g_{n_a}]$ and $U(k) = [u_1(k), ..., u_{n_a}(k)]^T$.

Here, we used two sets to indicate the interaction between the sensor and actuator nodes: the associated sensor nodes of $A_i, \forall i \in \{1, ..., n_a\}$:

$$S_{A_i} = \{S_j | d_{ij} \neq 0, j = 1, ..., m_s\} \tag{5}$$

and the influenced actuator nodes of $S_j, \forall j \in \{1, ..., n_s\}$:

$$S_{S_j} = \{A_i | d_{ij} \neq 0, i = 1, ..., m_a\} \tag{6}$$

where the parameter d_{ij} represents the relation between the ith and the jth nodes:

$$d_{ij} = \begin{cases} 1, & \text{influenced} \\ 0, & \text{isolated} \end{cases} \tag{7}$$

Eq.(5) and Eq.(6) show that the sensor nodes in set S_{A_i} will transmit the sensing data to A_i, while the actuator nodes in set S_{S_j} will influence the plant state monitored by S_j.

In WSANs, the sensor and actuator nodes are usually linked with wireless medium, since the actuator nodes are connected with each other directly and much more powerful than the ordinary sensor nodes, communication delay between S-A become a general problem of such network control system [14]. The S-A delays do not only degrade the system performance, but can also destabilize the system [15]. The delay system at sample step k has the following dynamics:

$$Z_a(k) = Z(k - \Delta_k) \tag{8}$$

The finite non-negative integers Δ_k represent the S-A delays at the kth step, $Z_a(k)$ is the sensing data received by the actuator nodes with communication delays.

In the course of the practice, the variation of plant state x_i at time t is caused by the output transferred form the actuator nodes and its ambient environment [12], under that assumption we have:

$$\frac{dx_i}{dt} = \sum_{1\le k\le n_s, k\neq i} \alpha_{ki}(x_k - x_i) + \sum_{1\le l\le n_a} \beta_{li}(f_l - x_i) \tag{9}$$

where α_{ki} and β_{li} are coefficients relating to the state-transfer efficiency. So, the plant state equation can be written in the matrix form as follows:

$$\frac{dX}{dt} = \Phi X(t) + \Psi F(t) \tag{10}$$

where $\Phi \in R^{n_s \times n_s}$, $\Psi \in R^{n_a \times n_a}$, we assume that F is constant within each step, i.e., $F(t) = F(k), t \in [kT, (k+1)T)$. Since Φ and Ψ are coefficients or constants, then the dynamic system can be modeled as:

$$X(k+1) = AX(k) + BF(k) \tag{11}$$

where $A = e^{\Phi}T$ and $B = \Phi^{-1}(e^{\Phi T} - 1)\Psi$.

3 Delay Compensation Algorithm

The main idea of the delay compensation algorithm is to utilize an observer to estimate the plant states and a multi-step predictor to compute predictive control inputs based on the past sensor measurements. The block diagram of the delay compensation algorithm is shown in Figure 1.

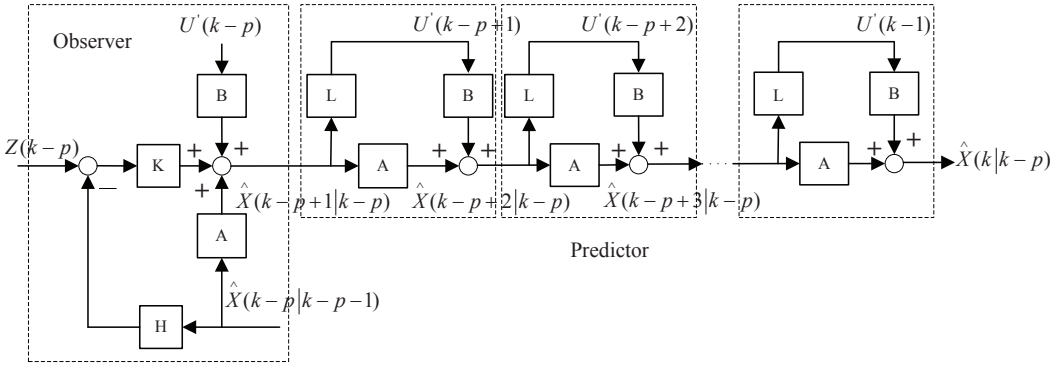


Figure 1: Block diagram of the delay compensation algorithm.

In order to keep the track of past measurements, received sensing data have to be stored in a p length FIFO (First-In-First-Out), denoted as Q, and p is the upper bound of Δ_k. Thus, the S-A delay is transformed to a constant delay, which is much easier to control than the random delay systems [15]. The delay compensation algorithm is delineated below:

Observer model:

$$\hat{X}(k-p+1 \mid k-p) = A\hat{X}(k-p \mid k-p-1) + BU'(k-p) + K(k-p)(Z(k-p) - H\hat{X}(k-p \mid k-p-1)) \tag{12}$$

Predictor model:

$$\hat{X}(k \mid k-p) = A\hat{X}(k-1 \mid k-p) + BU'(k-1) \tag{13}$$

Control law:

$$U'(k) = L(k)\hat{X}(k \mid k-p) \tag{14}$$

In [16], the authors have proved that the resulting closed-loop equations can be expressed as:

$$\begin{bmatrix} X(k+1) \\ \hat{E}(k-p+1) \end{bmatrix} = \begin{bmatrix} A+BL(k) & * \\ 0 & A-K(k-p)H \end{bmatrix} \begin{bmatrix} X(k) \\ \hat{E}(k-p) \end{bmatrix} \tag{15}$$

where:

$$\hat{E}(k) = X(k) - \hat{X}(k \mid k-1) \tag{16}$$

If K and L are constant, then Eq.(15) determines the stability of the delay compensator due to the separation of the controller and observer. Since the performance of the observer and predictor are highly dependent on the model certainty, then the dynamic model of the plant has to be very precise.

4 Distributed Collaborative Processing

4.1 Dynamic clustering schedule

In order to maximize the network lifetime and data throughput, and provide load balancing and fault tolerance [17], the clustering schedule should be established. In this paper, based on the characteristics of the current events, an event-triggered dynamic clustering schedule is designed for WSANs. If there is no event occurs, the nodes follow a static sleep schedule. When an event occurs, the sensor nodes whose sensing range cover it will be activated, and transmit the sensing data to each coordinator, then the coordinator organizes its neighbor nodes into a working cluster to take a proper action till the error signal becomes zero. During this process, the coordinator will act as the cluster head and the neighbor nodes will be selected as the cluster members. Then the control decision is made by the cluster head according to the fusion data aggregate from the associated sensor node and cluster members. For sensor node's coordinator is the nearest actuator node, since the closer the actuator node to the sensor node is, the earlier the actuator node is informed, thus the quicker the actuator node reacts and the earlier action to be initiated. Here, neighbor nodes can be defined as the actuator nodes which are within the communication range of coordinator and the associated sensor nodes are activated. So, the energy constrained sensor node does not need to transmit its readings to multiple actuator nodes. Instead, the coordinator will receive this message and relay it to its neighbor nodes to come up with an appropriate actuation. The process of dynamic clustering schedule is shown in Figure 2. Here, we assume that the data route from source sensor node to terminal actuator node is in one hop, $|S_{S_j}| = 1, j = 1, .., n_s$ and $|S_{A_i}| = 1, i = 1, .., n_a$.

4.2 Collaborative processing algorithm

Consider the control objective which is to meet the set points $X^* = [x_1^*, ..., x_{n_s}^*]^T$. In order to balance the control requirements against the spatial smoothness of the control signals, we define the control objective of A_i as:

$$J_i(k+1) = \frac{\alpha}{2} \sum_{j \in N_{ai}, j \neq i} (e_i(k+1) - e_j(k+1))^2 + \frac{1-\alpha}{2} e_i^2(k+1) \tag{17}$$

where N_{ai} is the neighbor nodes set of A_i, $e_i(k+1) = x_i(k+1) - x_i^* = a_{ii}\hat{x}_i(k) + b_{ii}g_iu_i(k) - x_i^*$ and $e_j(k+1) = x_j(k+1) - x_j^* = a_{jj}\hat{x}_j(k) + b_{jj}g_ju_j(k) - x_j^*, (j \in N_{ai}, j \neq i)$.

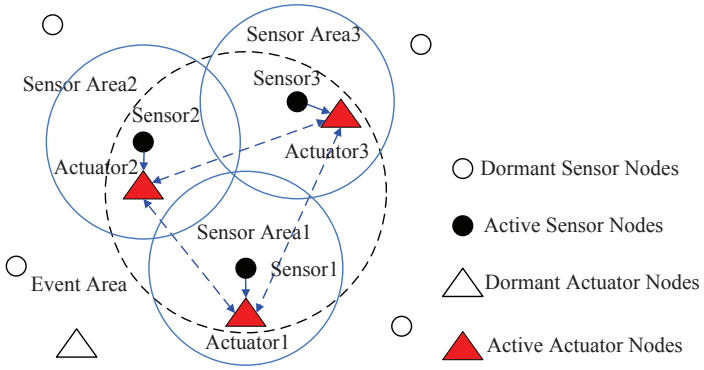


Figure 2: Dynamic clustering schedule.

In order to minimize $J_i(k+1)$, the gradient descending method can be used. The partial derivative of $J_i(k+1)$ with respect to $u_i(k)$ is calculated as:

$$\frac{\partial J_i(k+1)}{\partial u_i(k)} = [\alpha \sum_{j \in N_{ai}, j \neq i} (e_i(k) - e_j(k)) + (1-\alpha)e_i(k)]b_{ii}g_i \tag{18}$$

For each $A_i, i \in (1, ..., n_a)$, its control law $u_i(k)$ is updated by:

$$u_i(k+1) = u_i(k) + \Delta u_i = u_i(k) - \varepsilon \frac{\partial J_i(k+1)}{\partial u_i(k)} \tag{19}$$

where ε is a positive step size called the learning step length. If ε is small, the convergence speed of the objective function J_i will be slow. If ε is too large, it often leads to unstable. So it is important how to choose the proper ε.

In order to investigate the stability of Eq.(19), we rewrite Δu_i as:

$$\Delta u_i = -\varepsilon \frac{\partial J_i(k+1)}{\partial u_i(k)} = -\varepsilon \frac{\partial J_i(k+1)}{\partial e_i(k+1)} \frac{\partial e_i(k+1)}{\partial u_i(k)} = -\varepsilon b_{ii} g_i \frac{\partial J_i(k+1)}{\partial e_i(k+1)} = -\lambda_i \frac{\partial J_i(k+1)}{\partial e_i(k+1)} \tag{20}$$

We define the learning error as:

$$\Delta e_i = -\lambda_i \frac{\partial J_i(k+1)}{\partial e_i(k+1)} \tag{21}$$

where

$$\frac{\partial J_i(k+1)}{\partial e_i(k+1)} = \alpha \sum_{j \in N_{ai}, j \neq i} (e_i(k+1) - e_j(k+1)) + (1-\alpha)e_i(k+1) \tag{22}$$

Let the array $[\partial J_1(k+1)/\partial e_1(k+1), ..., \partial J_{n_a}(k+1)/\partial e_{n_a}(k+1)]^T$ to be zero, then it can be represented as:

$$DE(k+1) = 0 \tag{23}$$

Here, D is a $n_a \times n_a$ positive definite matrix, $E(k+1) = [e_1(k+1), ..., e_{n_a}(k+1)]^T$ and the elements of D satisfy the following equation:

$$|d_{ii}| - \sum_{j=1, j \neq i}^{n_a} |d_{ij}| = 1 - \alpha \tag{24}$$

We define the residual error as: $R(k) = E(k) - E^*$, where E^* is the solution of $DE^* = 0$. From Eq.(21), we have:

$$\begin{aligned} R(k) &= R(k-1) - \lambda(DE(k-1)) = R(k-1) - \lambda(DE(k-1) - DE^*) = (I - \lambda D)R(k-1) \\ &= \prod_{i=1}^{k}(I - \lambda D)R(0) = Y\prod_{i=1}^{k}(I - \lambda\Lambda)Y^T R(0) = Y\prod_{i=1}^{k}(I - \varepsilon\Lambda')Y^T R(0) \end{aligned} \quad (25)$$

where $D = Y\Lambda Y^T$, $\Lambda = diag(\eta_1, ..., \eta_{n_a})$, $\lambda = diag(\lambda_1, ..., \lambda_{n_a})$, and $\eta_1, ..., \eta_{n_a}$ are the eigenvalues of D, so we can get:

$$\Lambda' = BG\Lambda = diag(b_{11}g_1\eta_1, ..., b_{n_a n_a}g_{n_a}\eta_{n_a}) = diag(\sigma_1, ..., \sigma_{n_a}) \quad (26)$$

If we select $0 < \varepsilon < 2/max(\sigma_i), 1 \le i \le n_a$, then $R(k) \longrightarrow 0$ as $k \longrightarrow \infty$.

Eq.(19) is a completely distributed collaborative processing method, there has no need a sink to help in the coordination of the sensor and actuator nodes. Instead, each actuator node combines itself and neighbor nodes' messages to access the control law and pursuit the optimal solutions step by step.

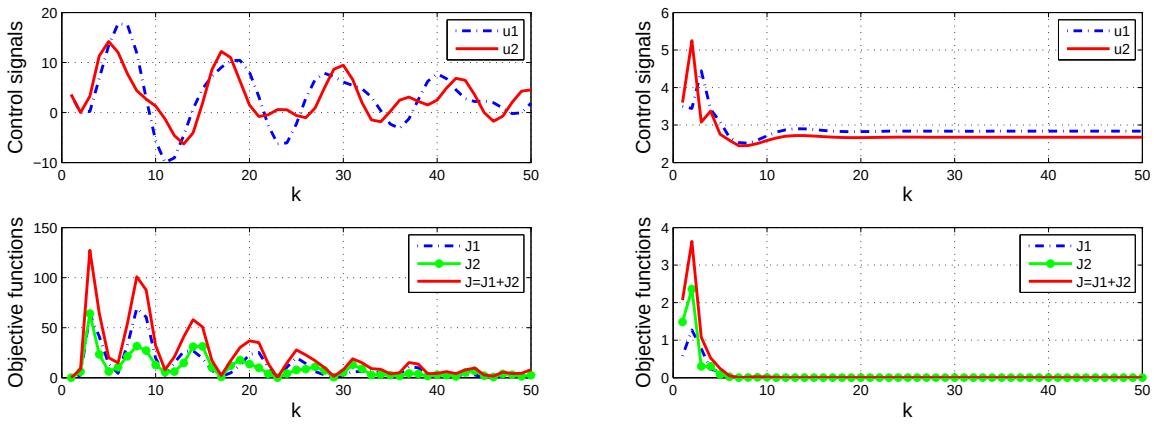
5 Numerical Examples

Let's consider a simple Humility, Ventilation, Air Conditioning (HVAC) control system for temperature control with two sensor nodes ($n_s = 2$) and two actuator nodes ($n_a = 2$). The control arm is to meet the set points $X^* = [16(^\circ C), 18(^\circ C)]$. The system parameters are:

$$A = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.9 \end{bmatrix}, B = \begin{bmatrix} 0.57 & 0 \\ 0 & 0.68 \end{bmatrix}, H = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, K = \begin{bmatrix} -0.68 & 0 \\ 0 & -0.57 \end{bmatrix}, L = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$$

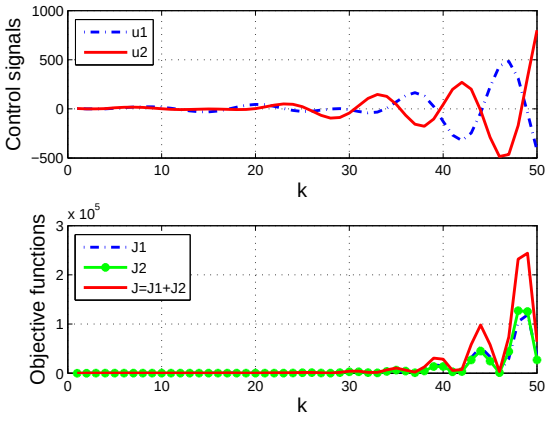
where K and L satisfy the stability condition according to Eq.(15).

The effectively actuation of actuator nodes are highly depend on the precision of the sensing data. The longer time delay between sensing and acting is, the bigger estimation error introduced. The increasing control decision error does not only degrade the system performance, but also can destabilize the system, just as shown in Figure 3(a) and 3(c). Figure 3(b) and 3(d) clearly show that the compensated system are less oscillatory than those of the uncompensated system. The predictor-controller compensation algorithm provides a valid way to estimate the sensing data with latency, reduce the estimation bias and enhance the precision of feedback control.

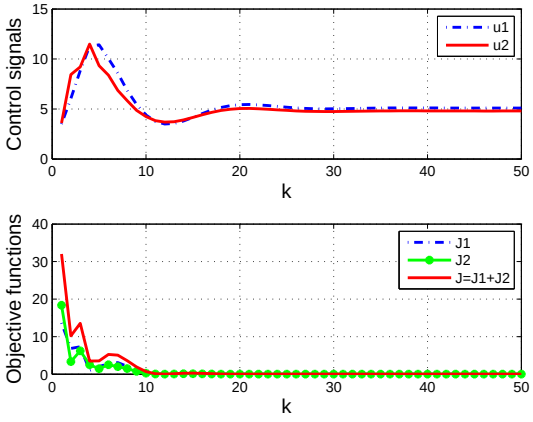


(a) $\varepsilon = 1.63$, $\alpha = 0.5$, $\Delta_k = 1$, $p = 0$

(b) $\varepsilon = 1.63$, $\alpha = 0.5$, $\Delta_k = 1$, $p = 1$



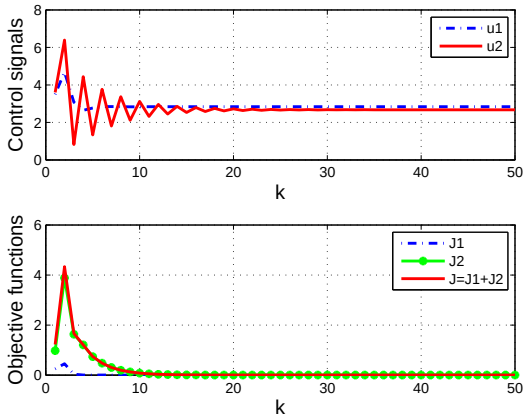
(c) $\varepsilon = 1.63$, $\alpha = 0.5$, $\Delta_k = 2$, $p = 0$



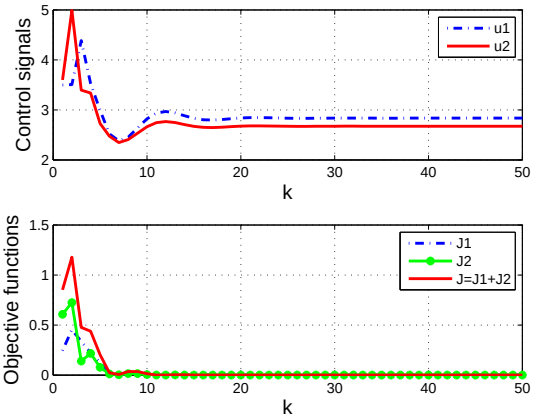
(d) $\varepsilon = 1.63$, $\alpha = 0.5$, $\Delta_k = 2$, $p = 2$

Figure 3: Dynamic responses of uncompensated ($\Delta_k \neq 0, p = 0$) and compensated system ($p = \Delta_k \neq 0$).

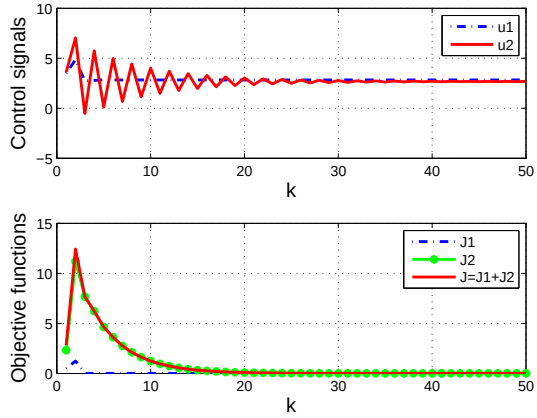
In Eq.(17), α is a collaborative factor between 0 and 1. When $\alpha = 0$, the neighbor nodes' messages are not taken into consideration, but if we select $\alpha \neq 0$, the collaborative processing among different nodes are introduced. Moreover, α can also performs as a smooth factor, it will reduce the control overshooting and stabilize the system from oscillating. The performance of compensated system with and without collaborative processing method are shown in Figure 4. It is obviously seen that the proposed method can greatly improve the system performance, which can smooth the actuator control signal and accelerate the system convergence speed.



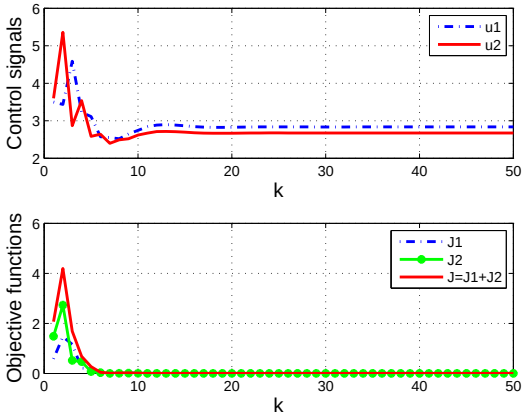
(a) $\varepsilon = 2.17$, $\alpha = 0$, $\Delta_k = 0$, $p = 0$



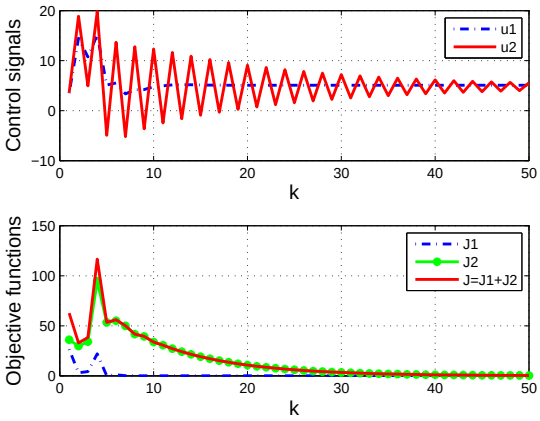
(b) $\varepsilon = 2.17$, $\alpha = 0.5$, $\Delta_k = 0$, $p = 0$



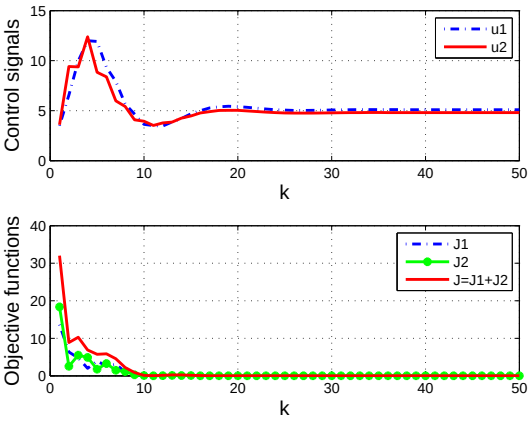
(c) $\varepsilon = 1.74$, $\alpha = 0$, $\Delta_k = 1$, $p = 1$



(d) $\varepsilon = 1.74$, $\alpha = 0.5$, $\Delta_k = 1$, $p = 1$

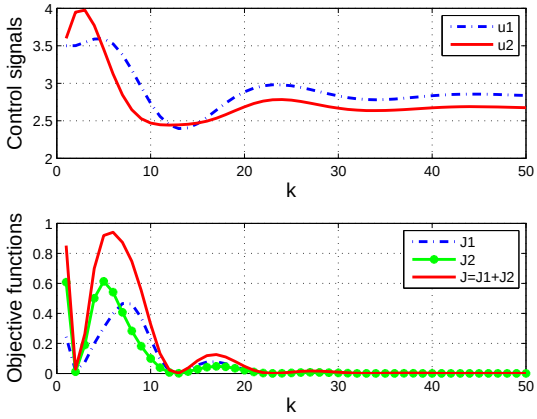


(e) $\varepsilon = 1.96$, $\alpha = 0$, $\Delta_k = 2$, $p = 2$

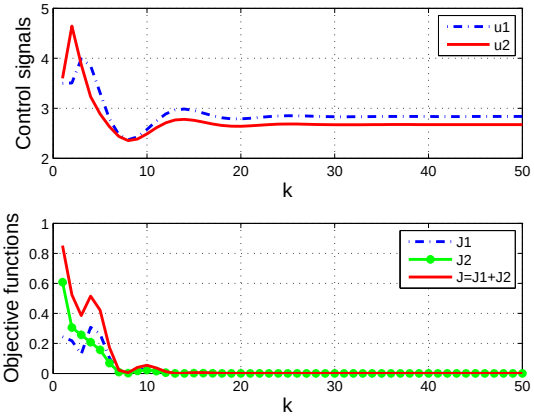


(f) $\varepsilon = 1.96$, $\alpha = 0.5$, $\Delta_k = 2$, $p = 2$

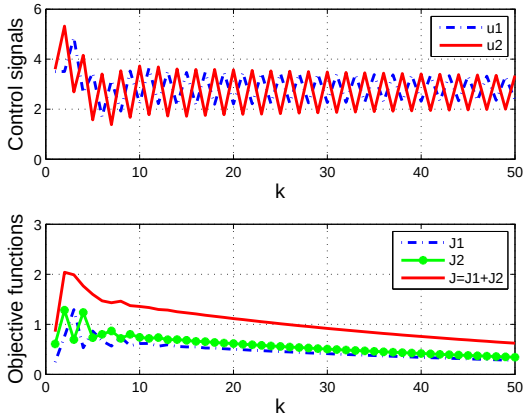
Figure 4: Dynamic responses of real-time ($\Delta_k = 0$) and compensated system ($p = \Delta_k \neq 0$) under different α.



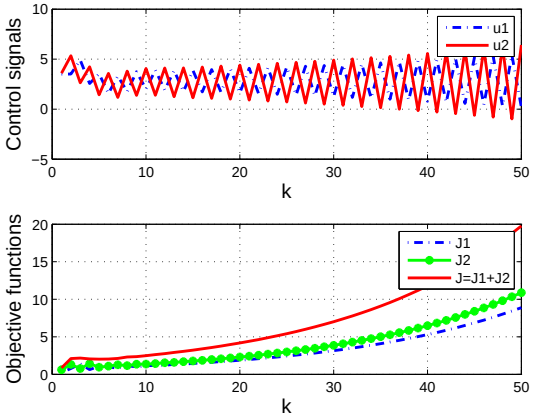
(a) $\varepsilon = 0.54$, $\alpha = 0.5$, $\Delta_k = 0$, $p = 0$



(b) $\varepsilon = 1.63$, $\alpha = 0.5$, $\Delta_k = 0$, $p = 0$

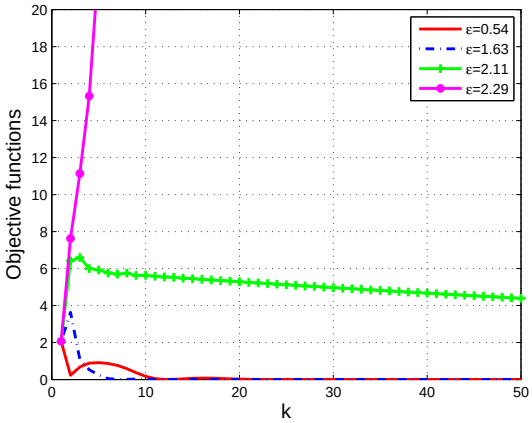


(c) $\varepsilon = 2.24$, $\alpha = 0.5$, $\Delta_k = 0$, $p = 0$

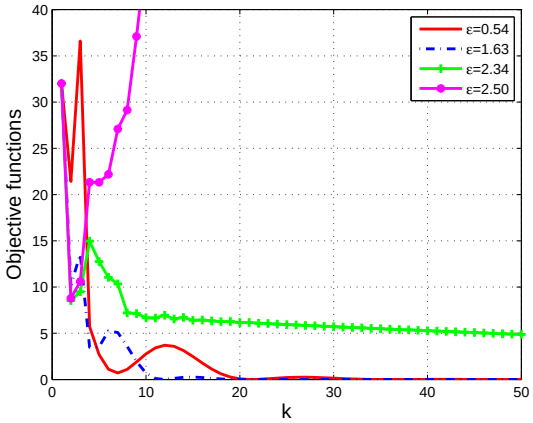


(d) $\varepsilon = 2.72$, $\alpha = 0.5$, $\Delta_k = 0$, $p = 0$

Figure 5: Dynamic responses of real-time system ($\Delta_k = 0$) under different ε.

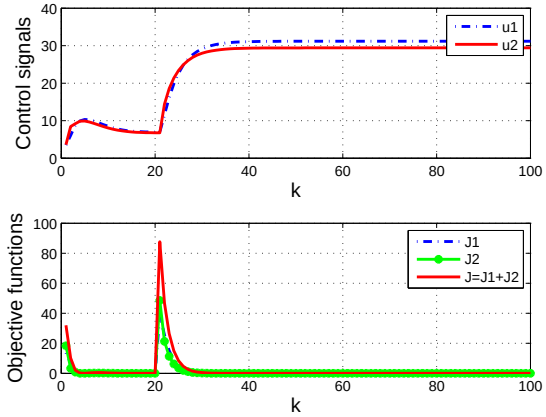


(a) $\alpha = 0.5$, $\Delta_k = 1$, $p = 1$



(b) $\alpha = 0.5$, $\Delta_k = 2$, $p = 2$

Figure 6: Dynamic responses of compensated system ($p = \Delta_k \neq 0$) under different ε.



$\varepsilon = 1.63$, $\alpha = 0.5$, $\Delta_k = 10$, $p = 10$

Figure 7: Dynamic responses of system with multi-step delay compensation.

The learning step length ε plays an important role in the gradient descending method. Figure 5 compares the responses of real-time system under different ε ($\varepsilon = 0.54, 1.63, 2.24, 2.72$). We could observe that the bigger ε is, the faster convergence speed of J can be achieved. But if ε exceeds the critical value $\varepsilon_c = 2.24$, i.e., $\varepsilon = 2.72$, the system become unstable. Figure 6 shows the dynamic responses of compensated system with collaborative processing method under different ε. Δ_k between S-A are set at one-step and two-step, respectively. We can get that $\varepsilon_c = 2.11$ and $\varepsilon_c = 2.34$ are the critical step lengths, and the stable control can be achieved within those values.

Comparing Figure 6(a) and 6(b), we could see that, the bound of J is influenced by the delay step Δ_k. The variance of J increased as the longer S-A latency. Figure 7 shows that when the system suffers a multi-step delay, such as $\Delta_k = 10$, the proposed compensation and control scheme is also useful. J will tend to be zero eventually and control signals both converge to their stable states.

6 Conclusions

In this paper, we focus on the communication and control problems in WSANs. We argue that the system performance is closely related to the communication delay and control strategy. In order to mitigate the detrimental effects of the S-A latency, a delay compensation algorithm based on the state estimation is applied to this system. Then, a distributed collaborative processing method is proposed to control actuator option in a coordinate way to accomplish the desired tasks. We formulate it as an optimization problem and utilize gradient descending algorithm to calculate the optimal control law for actuator nodes. On this basis, we discuss the control strategy parameters that relate to the system performance and provide a guide line how to choose properly. In our framework, the proposed collaborative processing method does no need a central sever and make an optimum usage of the available resources, which can be easily applied in the industrial automation systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant 61174070, the Specialized Research Fund for the Doctoral Program under grant 20110172110033.

Bibliography

[1] I. F. Akyildiz, I. H. Kasimoglu, Wireless Sensor and Actor Networks: Research Challenges. *Ad Hoc Networks*, Vol.2, No.4, pp.351-367, 2004.

[2] R. Vedantham, Z. Zhuang, R. Sivakumar, Hazard Avoidance in Wireless Sensor and Actor Networks. *Computer Communications*, Vol.29, No.13, pp.2578-2598, 2006.

[3] A. Deshpande, C. Guestrin, S. R. Madden, Resource-Aware Wireless Sensor-Actuator Networks. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, Vol.28, No.1, pp.40-47, 2005.

[4] T. Wang, L. Zhou, P. Han, Q. Zhang, Complete Compensation for Time Delay in Networked Control System based on GPC and BP Neural Network. *Proceedings of 2007 International Conference on Machine Learning and Cybernetics*, Hongkong, China, 637-641, 2007.

[5] Y. Zeng, C. J. Sreenan, G. Zheng, A Real-time Architecture for Automated Wireless Sensor and Actuator Networks. *The Fifth International Conference on Wireless and Mobile Communications*, Cannes/La Bocca, French Riviera, France, 1-6, 2009.

[6] E. Ngai, Y. Zhou, M. R. Lyu, J. Liu, A Delay-aware Reliable Event Reporting Framework for Wireless Sensor-Actuator Networks. *Ad Hoc Networks*, Vol.8, No.7, pp.694-707, 2010.

[7] A. Onat, T. Naskali, E. Parlakay, O. Mutluer, Control Over Imperfect Networks: Model-Based Predictive Networked Control Systems. *IEEE Transactions on Industrial Electronics*, Vol.58, No.3, pp.905-913, 2011.

[8] J. Chen, X. Cao, P. Cheng, Y. Xiao, Y. Sun, Distributed Collaborative Control for Industrial Automation With Wireless Sensor and Actuator Networks. *IEEE Transactions on Industrial Electronics*, Vol.57, No.12, pp.4219-4229, 2010.

[9] R. I. Erica, V. G. Luis, Cooperation Mechanism Taxonomy for Wireless Sensor and Actor Networks. *IEEE Transactions on Industrial Electronics*, Vol.264, pp.62-73, 2008.

[10] T. Melodia, D. Pompili, V. C. Gungor, I. F. Akyildiz, Communication and Coordination in Wireless Sensor and Actor Networks. *IEEE Transactions on Mobile Computing*, Vol.6, No.10, pp.1116-1128, 2007.

[11] M. Nakamura, A. Sakurai, S. Furubo, H. Ban, Collaborative Processing in Mote-Based Sensor/actuator Networks for Environment Control Application. *Signal Processing*, Vol.88, No.7, pp.1827-1838, 2008.

[12] X. Cao, J. Chen, Y. Xiao, Y. Sun, Building-Environment Control With Wireless Sensor and Actuator Networks: Centralized Versus Distributed. *IEEE Transactions on Industrial Electronics*, Vol.57, No.11, pp.3596-3606, 2010.

[13] B. Sundararaman, U. Buy, A. Kshemkalyani, Clock Synchronization for Wireless Sensor Networks: A Survey. *Ad Hoc Networks*, Vol.3, No.3, pp.281-323, 2005.

[14] V. C. Gungor, Ö. B. Akan, I. F. Akyildiz, A Real-time and Reliable Transport Protocol for Wireless Sensor and Actor Networks. *IEEE Transactions on Networking*, Vol.16, No.2, pp.359-370, 2008.

[15] M. Chow, Y. Tipsuwan, Network-Based Control Systems: A Tutorial. *The 27th Annual Conference of the IEEE Industrial Electronics Society*, Denver, USA, pp.1593-1602, 2001.

[16] R. Luck, A. Ray, Experimental Verification of a Delay Compensation Alrorithm for Integrated Communication and Control Systems. *International Journal of Control*, Vol.59, No.6, pp.1357-1372, 1994.

[17] B. McLaughlan, K. Akkaya, Coverage-based Clustering of Wireless Sensor and Actor Networks. *IEEE International Conference on Pervasive Services*, Istanbul, Turkey, 45-54, 2007.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 721-732

Specification and Validation of a Formative Index to Evaluate the Ergonomic Quality of an AR-based Educational Platform

C. Pribeanu

Costin Pribeanu
National Institute for Research and Development in Informatics - ICI Bucharest
Romania, 011455 Bucuresti, Bd. Maresal Averescu, 8-10
E-mail: pribeanu@ici.ro

Abstract:
The ergonomic quality of educational systems is a key feature influencing both the usefulness and motivation for the learner. Desktop Augmented Reality (AR) systems are featuring specific interaction techniques that may create additional usability issues affecting the perceived ease of use. Measuring key usability aspects and understanding the causal relationships between them is a challenge that requires formative measurement models specification and validation. In this paper we present an evaluation instrument based on two main formative indexes that are capturing specific usability measures for two AR-based applications. The formative indexes are forming a second order formative construct that acts as predictor for both the general ease of use and ease of learning how to operate with the application.
Keywords: formative measurement model, formative index, augmented reality, usability, ergonomic quality.

1 Introduction

Educational systems based on desktop AR technologies are creating an appealing user experience for the learner by integrating real life objects into computer environments. Touching and holding real objects is increasing the students' motivation to learn and could better support active and collaborative learning [18], [22]. As AR technologies become more wide-spread, there is an increasing interest in their ergonomic quality. Designing for usability is not easy in emerging technologies, like AR systems, which are featuring novel interaction techniques [5], [15].

The ISO standard 9126-1 defined usability as the capability of a software system to be easy to understand, easy to learn how to operate with, easy to operate with, and attractive, when used under specified conditions [19]. By ergonomic quality we refer to the first three usability aspects: ease of understanding, ease of learning how to operate, and ease of operating with a software system. How to measure and improve the usability of interactive systems is a key research topic in HCI. A research challenge is to better understand the relationships between different usability measures as well as between usability and other factors of interest [17].

In a previous work we developed a measurement model that was grounded in the technology acceptance models (TAM) theory [9] in order to explain the causal relations between various factors influencing the intention to use of an AR-based educational platform [3], [4]. Although the structural model was useful to test some typical TAM hypotheses the variance explained was small and several items targeting specific usability aspects were eliminated in order to achieve the unidimensionality required by a reflective measurement model. Moreover, reflective measurement assumes the same antecedents for reflective indicators (as manifest variables) so causal relations are estimated at construct level [8], [24]. These shortcomings suggest looking for an alternative modeling approach.

In this paper we present a measurement model for the evaluation of the ergonomic quality of applications developed onto an AR-based educational platform. The Augmented Reality Teaching Platform (ARTP) was developed in the framework of the ARiSE (Augmented Reality for

School Environments) European project. Two AR applications implementing learning scenarios for Biology and Chemistry were developed and tested onto ARTP.

The measurement model consists in two sets of formative indicators that are measuring two dimensions of the ergonomic quality of desktop AR applications: the quality of visual and auditory perception and the ease to operate and collaborate in a constrained space. The two indexes are forming a second order formative construct. In order to achieve identification requirements we used as outcome variables a reflective construct measuring the perceived ease of learning how to use ARTP and a general reflective item measuring the overall ease of use. The formative measurement model was estimated on the Biology scenario data. Then we cross validated the models on the Chemistry scenario data.

The rest of this paper is organized as follows. In the following section we describe the formative measurement models and discuss some methodological aspects related to the specification, identification and validity. In section 3 we present and discuss the estimation results with the Biology scenario data. In section 4 we present the results of a confirmatory assessment of the formative measurement model using the Chemistry scenario data and we comparatively discuss the results for each scenario. The paper ends with conclusion and future research directions.

2 The formative measurement model

2.1 Reflective vs. formative measurement models

A measurement model describes the relationships between a construct (latent variable) and its measures (indicators, items) while a structural model describes the relationships between different constructs [12], [13]. The causal relation between a construct and its measures could be from construct to measures (reflective model) or from measures to construct (formative model). There are distinct characteristics of each measurement model that were systematically presented and discussed in detail in [6], [10], [12], [20].

In the reflective measurement model the indicators are manifest variables of the latent variable. A change in the constructs is reflected in simultaneous changes in all indicators. As such, the items are interchangeable and elimination of one of them doesn't change the construct domain. Measures should be positively correlated and the measurement model should have convergent and discriminate validity.

In the formative measurement model the measures are defining the conceptual meaning of the construct. Indicators are not interchangeable since each is capturing a distinct cause. Since the measures are defining the construct, a census of indicators is recommended [6]. There are no assumptions on unidimensionality and correlations between indicators. However, collinearity should be avoided. Indicators don't have an error term and items are intercorrelated. Although there is an error terms at construct level this is not a measuring error but a disturbance accounting for other causes not specified by the model [11]. The nomological net of formative indicators could differ as this is a distinct feature of the formative measurement [20].

A formative measurement model taken in isolation is under identified and cannot be estimated. Jarvis et al. and Diamantopoulos et al. recommend achieving identification based on a 2+ rule: specifying effects (outcomes) of the formative constructs on at least two other variables that are reflectively measured [12], [20]. The outcome variables could be: two reflective indicators (MIMIC model), two reflective constructs, or a reflective construct and a reflective variable. The selection of the outcome variables is just as important as is the selection of indicators [11], [14]. According to Wilcox et al., the selected effect variables are determining the empirical meaning of the formative construct and the set of indicators [26].

The proper specification of the measurement model is a precondition before analyzing and

assigning a meaning to the structural model [1]. According to Jarvis et al., there are many studies in literature that are based on inappropriate specification of the measurement models [20]. In recent years, there is an ongoing debate regarding the formative versus reflective specification of various constructs and the appropriateness of measurement scales that are frequently used in different domains [12].

Taking the appropriate measurement perspective is not a simple issue. As pointed out by Jarvis and colleagues, based on the analysis of 178 papers published in four top journals in marketing research, there are about 29% cases (reported at 1192 constructs) of misspecification [20]. Moreover, the authors themselves experienced difficulties in classifying 14% of constructs featuring both reflective and formative characteristics. Wilcox and colleagues argued that a construct is not inherently formative or reflective so the researcher has a choice to take a perspective or another [26]. In this respect, the specification of alternative models is useful since is providing with more insights into the field of study.

2.2 Experiment, samples and data analysis

ARTP is a "seated" AR environment: users are looking to a see-through screen where virtual images are superimposed over the perceived image of a real object placed on the table [27]. Two AR-based applications were developed onto this platform (see Figure 1).

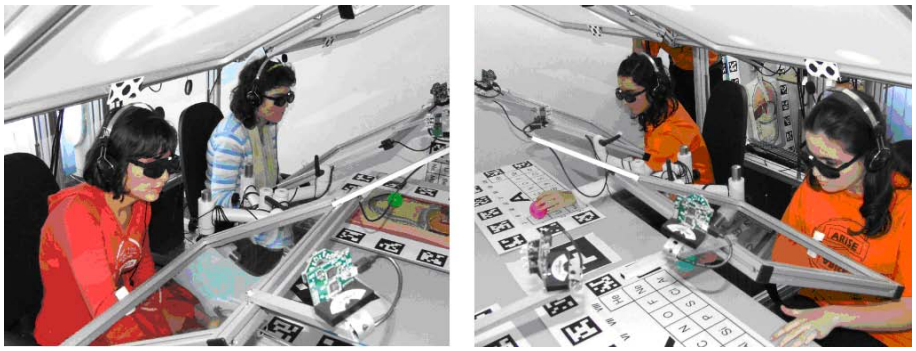


Figure 1: Students testing the ARTP learning scenarios: Biology (left) and Chemistry (right)

The first application implemented a Biology learning scenario for secondary schools. The implemented paradigm was "3D process visualization of hidden processes" and was targeted at enhancing the students' understanding and motivation to learn the human digestive system. The real object is a flat torso of the human body. A pointing device having a colored ball on the end of a stick and a remote controller Wii Nintendo as handler has been used as interaction tool that serves for three types of interaction: pointing on a real object, selection of a virtual object and selection of a menu item.

The second application implemented a Chemistry scenario. The implemented paradigm was "building with guidance" and was targeted at enhancing the students' understanding and motivation to learn the periodic table of Chemical elements, the structure of atoms / molecules, and the chemical reactions. The real objects were the periodic table of chemical elements and four sets of colored balls symbolizing atoms. The remote controller Wii Nintendo has only been used as interaction tool for confirming a selection.

The test was conducted in 2008, on the ICI's platform which is equipped with 4 ARTP modules. A total number of 139 students (13-14 years old), from which 65 boys and 74 girls tested the platform. All were 8th grade students enrolled in 3 general schools in Bucharest. None of them was familiar with the AR technology. The students came in groups of 7-8, accompanied

by a teacher. Each student tested the platform twice: once for the Biology scenario and second time for the Chemistry scenario. Each scenario consists of a demo lesson and a number of exercises.

After testing, the students were asked to answer a usability questionnaire by rating the items on a 5-point Likert scale (1-strongly disagree, 2-disagree, 3-neutral, 4-agree, and 5-strongly agree). The questionnaire has 28 closed items and 2 open questions, asking users to describe the most 3 positive and most 3 negative aspects. The first 24 closed items are targeting various dimensions of the ARTP such as ergonomics and usability (items 1-14), perceived utility (items 15-17), perceived enjoyment (items 18-21) and intention to use (items 22-24). The last four items were to assess how the students overall perceived the platform as being easy to use, useful for learning, enjoyable to learn with, and exciting.

In order to estimate the new measurement model we used the Biology scenario data. We analyzed the initial sample of 139 observations for normality (skewness and kurtosis), univariate and multivariate outliers. We transformed the data (square root extraction) and we repeated the analysis and successively removed 9 observations. The final sample has 130 observations that present moderate deviations from normality. In order to cross validate the model on another sample, we used the Chemistry scenario data. We performed the same data analysis procedure on the initial sample and successively removed 11 observations. The final sample has 128 observations with moderate deviations from normality.

2.3 Model specification and identification

According to our knowledge, there are few approaches to formative index construction for the usability and / or ease of use [21]. Although the perceived ease of use and perceived usability are frequently used in information systems research, in almost all studies they are specified as reflectively measured constructs. As such, their indicators have a limited contribution (as manifest variables) to explain the effect of usability problems.

Since the objective of this study is to analyze the relationships between different aspects related to the ergonomic quality of the ARTP, 15 items in the usability questionnaire are of interest, from which 11 are formative measures and 4 are reflective measures. The 15 items (presented in Annex 1) are grouped into four constructs and a single item measure:

- The quality of visual and auditory perception (ERG-P): clear observation and superposition, easy to read the information on the screen, and easy to understand the vocal explanations.
- The ease of interaction and collaboration and collaboration (ERG-O): comfortable work place, easy to select a menu item with the remote control, easy to correct errors, and easy to collaborate with colleagues.
- The ease of adjusting the devices and accessories (ERG-A), i.e. the see-through screen, stereo glasses and head phones.
- The ease of learning (PEOL): easy to understand, easy to learn and easy to remember how to use ARTP.
- The general item measuring the overall ease of use (PEOU1).

The first three constructs are composite indexes measuring distinct usability aspects that are specific to an AR-based learning application. As such, the indicators are not interchangeable and elimination of any of them will alter the conceptual domain of the construct. For example, if we analyze the three items measuring the quality of the visual perception, each is targeting a

different usability aspect. The clarity of observation through the screen is a hardware issue while the clarity of superposition between the augmentation and the real object is a software issue. Reading the information on the screen relates to augmentation, messages to the user and menu items.

Note that apart from the specific AR devices and accessories there are also several usability aspects which are specific to a given application. For example, in the Biology scenario the user selects an organ by pointing on flat torso of the digestive system which is a real object shared by to students staying face-to-face. In the Chemistry scenario, the students create a molecule by bringing together several colored balls symbolizing atoms. In this respect, the interaction with the remote control, the correction of mistakes (selection errors) and the collaboration between students depend on the real objects registered with the application. Therefore a formative model is an appropriate measurement perspective.

The ergonomic quality of ARTP is a multidimensional construct conceptualized as a composite of formative indexes. Each dimension is a formative index measuring a set of specific usability aspects. Each index is assumed to have a significant positive influence on two general usability aspects: the perceived ease of learning how to use ARTP (the construct PEOL) and on the overall ease of use (the general item PEOU1).

2.4 Validity of the formative indexes

According to recent studies, there are several criteria to assess the validity of formative indexes [6], [10], [12], [14]: adequate coverage of the construct's domain, absence of multicollinearity, indicator validity, significant γ-coefficients, complete mediation of effects, significant influence (β-coefficients) on outcome variables, and acceptable fit with the data.

Although a census of indicators is ideal to cover the scope of a formative index, this is rarely possible. In our model, each index is addressing a distinct aspect of the ergonomic quality of ARTP. Since formative indicators are also capturing critical usability aspects as indicated in previous studies (e.g. [23]) the coverage of the domain is acceptable.

The collinearity of formative indicators was analyzed with the VIF (Variation Inflation Factor) statistic for each index. VIF values were in the range 1.183-1.946 for the Biology scenario, respectively 1.085-1.715 for the Chemistry scenario bellow the 3.3 cut-off value [12].

The general item PEOU1 is an overall measure of the ergonomic quality of ARTP which qualify it for using as criterion validity. An analysis using Pearson's rho indicated that there are significant positive linear relationships between PEOU1 and the formative indicators of ERG-P and ERG-O but no significant correlations with the formative indicators of ERG-A. Nevertheless, in both samples ERG-A indicators are positively correlated with the formative item ERGO1. This suggests that ERG-A is not a distinct dimension of the ergonomic quality of ARTP but only an antecedent of a formative indicator measuring the comfort with the workplace. A regression analysis on the Biology data sample showed that ERGA1 and ERGA2 are two antecedents of ERGO1 (standardized coefficients β_{ERGA1} =0.191, sig=0.046 and β_{ERGA2}=0.185, sig=0.039). The regression analysis on the Chemistry data sample confirmed this finding (β_{ERGA1}=0.156, sig=0.083 and β_{ERGA2}=0.255, sig=0.005).

In order to estimate the formative indexes we used a MIMIC model and a structural model presented in Figure 2. The models were estimated using AMOS 17.0 [2]. Each index has n formative indicators, more specifically n=4 for ERG-P and ERG-O, and n=3 for ERG.

There are four outcome variables in the MIMIC model. Three of these reflective indicators are further grouped in the structural model that features 2 outcome variables: the general item PEOU1 (overall ease of use) and the reflective construct PEOL (ease of learning how to operate). All outcome variables are closely related to the focal construct as they measure general aspects

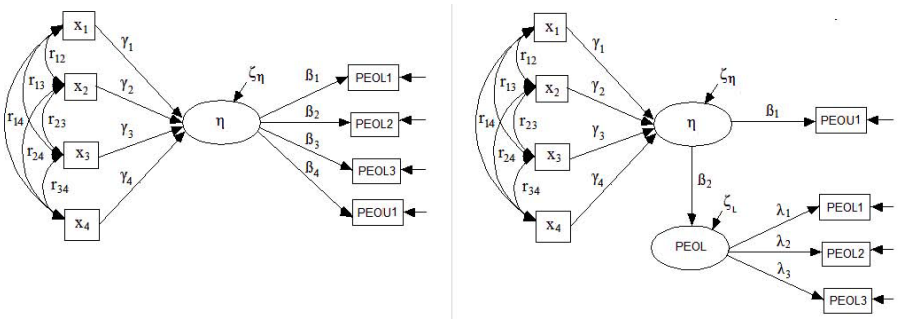


Figure 2: Estimation of formative indexes with MIMIC (left) and structural models (right)

of the perceived ergonomic quality.

There are three general hypotheses assessed with these models:

1. There is a significant contribution of the formative indicators to the composite index ($x_i \rightarrow \eta$, i=1...n).
2. There is a significant positive influence of the composite index on the perceived ease of learning how to use ARTP ($\eta \rightarrow$PEOL1, $\eta \rightarrow$PEOL2, $\eta \rightarrow$PEOL3 in the MIMIC model, respectively $\eta \rightarrow$PEOL in the structural model).
3. There is a significant positive influence of the composite index on the overall ease of use ($\eta \rightarrow$PEOU1).

Since the structural model includes a reflectively measured construct, the internal consistency and convergent validity should be assessed. The scale reliability and unidimensionality were analyzed with SPSS 16.0 and Amos 17.0. The consistency of scale (Cronbach's alpha) was 0.701 for the Biology scenario and 0.704 for the Chemistry scenario which is acceptable. Convergent validity was assessed by examining the standardized factor loadings, composite reliability, and average variance extracted for PEOL in each scenario [16]. Almost all factor loadings are over the minimum recommended level of 0.60. The composite reliability was 0.711 for the Biology scenario and 0.704 for the Chemistry scenario, above the minimum recommended value of 0.70 in each scenario. The average variance extracted was 0.456 for the Biology scenario and 0.438 for the Chemistry scenario. Overall, PEOL construct has an acceptable convergent validity.

3 Estimation results on the Biology scenario data

3.1 First order formative indexes

The results of MIMIC and structural model estimations for ERG-P and ERG-O are presented in Table 1. All γ-coefficients are significant at $p<0.05$ level thus supporting the first hypothesis. There are small differences between the magnitudes of γ-coefficients in the two models. The variance of the error term associated with the formative index is small in each model, so the formative index is sound and each formative item has a distinct contribution to the explained variance [11].

Fit indices are acceptable, over the recommended values [16]: χ^2=1.115, DF=13, χ^2/DF=1.624, GFI=0.962, CFI=0.974, SRMR=0.036 (ERG-P, structural model), and χ^2=22.963, DF=13, χ^2/DF=1.766, GFI=0.960, CFI=958, SRMR=0.042 (ERG-O, structural model).

In both models all β-coefficients are significant ($p<0.001$), which supports the last two hypotheses. The influence of formative indexes is stronger on the perceived ease of learning how to

Table 1: Estimation results for ERG-P and ERGO - Biology scenario

ERG-P	MIMIC model		Structural model	
	γ/β	sig.(p)	γ/β	sig.(p)
Contribution				
ERGP1	.33	<.001	.36	<.001
ERGP2	.31	0.001	.30	0.002
ERGP3	.20	0.010	.21	0.010
ERGP4	.27	0.002	.29	0.002
Effect variables				
PEOU1	.63	<.001	.63	<.001
PEOL			.91	<.001
PEOL1	.64	<.001		
PEOL2	.73	<.001		
PEOL3	.61	<.001		
Variance expl.				
ERG-O	71%		78%	
PEOL			83%	

ERG-O	MIMIC model		Structural model	
	γ/β	sig.(p)	γ/β	sig.(p)
Contribution				
ERGO1	.22	0.018	.27	0.006
ERGO2	.22	0.017	.21	0.030
ERGO3	.29	0.003	.30	0.003
ERGO4	.33	<.001	.33	0.001
Effect variables				
PEOU1	.62	<.001	.66	<.001
PEOL			.87	<.001
PEOL1	.63	<.001		
PEOL2	.71	<.001		
PEOL3	.63	<.001		
Variance expl.				
ERG-O	54%		62%	
PEOL			75%	

use ARTP than on the general ease of use. This means that once the user understands and learns how to use the system he finds it easy to use. The highest contributions to ERG-P have the first two items (clarity of observation through the see-through screen and accuracy of superposition). The most important contribution to ERG-O has the last item related to the ease of collaboration with colleagues. The ease of correcting the mistakes proved also to be an important measure for the Biology scenario.

3.2 Second order formative index

ERG-P and ERG-O are two distinct dimensions of the ergonomic quality of ARTP that are forming a second order formative construct (ERG). We used the scores of the first order constructs (the predicted values of the multiple regression) as formative indicators in the second order construct. Similar approaches are described in [7], [8]. The estimation results are presented in Table 2.

Table 2: Estimation results for second order construct - Biology scenario

ERG	MIMIC model		Structural model	
	γ/β	sig.(p)	γ/β	sig.(p)
Contribution				
ERG-P	.65	<.001	.68	<.001
ERG-O	.27	0.006	.30	0.004
Effect variables				
PEOU1	.63	<.001	.63	<.001
PEOL			.90	<.001
PEOL1	.64	<.001		
PEOL2	.71	<.001		
PEOL3	.61	<.001		
Variance expl.				
ERG	75%		84%	
PEOL			81%	

The γ-coefficients are significant in each model. The contribution of the first dimension is much higher showing that the quality of visual perception is a critical requirement for the desktop AR systems. The analysis of modification indices showed that the formative index is completely mediating the effects of its items.

Both β-coefficients are significant ($p < 0.001$), which supports the last two hypotheses. The variance of the error term associated with the formative index is 0.009 (medium effect). The magnitude of the error term is suggesting some other aspects not covered by the indicators.

Fit indices are acceptable, over the recommended values: χ^2=11.759, DF=7, χ^2/DF=1.679, GFI=0.972, CFI=0.984, SRMR=0.032 (structural model).

4 Cross validation of the formative indexes on the Chemistry scenario

4.1 First order formative indexes

The results of estimation are presented in Table 3. Almost all γ-coefficients are significant at p<0.05 level thus supporting the first hypothesis. There is only one exception: ERGP4 in the MIMIC model, where the γ-coefficient is signiffficant at p<0.10 level. There are relatively small differences between the contributions of each item in each model.

Table 3: Estimation results for ERG-P and ERG-O - Chemistry scenario

ERG-P	MIMIC model		Structural model	
	γ/β	sig.(p)	γ/β	sig.(p)
Contribution				
ERGP1	.23	0.016	.29	0.010
ERGP2	.28	0.009	.31	0.012
ERGP3	.24	0.010	.32	0.004
ERGP4	.20	0.053	.24	0.047
Effect variables				
PEOU1	.55	<.001	.61	<.001
PEOL			.75	<.001
PEOL1	.70	<.001		
PEOL2	.67	<.001		
PEOL3	.52	<.001		
Variance expl.				
ERG-O	47%		67%	
PEOL			56%	

ERG-O	MIMIC model		Structural model	
	γ/β	sig.(p)	γ/β	sig.(p)
Contribution				
ERGO1	.25	0.009	.24	0.018
ERGO2	.27	0.003	.30	0.002
ERGO3	.22	0.021	.24	0.016
ERGO4	.36	<.001	.38	<.001
Effect variables				
PEOU1	.48	<.001	.49	<.001
PEOL			.93	<.001
PEOL1	.71	<.001		
PEOL2	.70	<.001		
PEOL3	.55	<.001		
Variance expl.				
ERG-O	49%		55%	
PEOL			86%	

In both models β-coefficients are significant (p<0.001), which supports the last two hypotheses. The variance of the error term associated with the formative index is 0.022 (0.008) for ERG-P and 0.021 (0.016) for ERG-O. Since the magnitude of the error term is small and all indicator coefficients are significant, the formative index is sound and each formative item has a distinct contribution to the explained variance.

Fit indices are acceptable, over the recommended values [16]: χ^2=15.154, DF=13, χ^2/DF=1.624, GFI=0.973, CFI=0.990, SRMR=0.038 (ERG-P, structural model), and χ^2=22.392, DF=13, χ^2/DF=1.722, GFI=0.958, CFI=0.947, SRMR=0.048 (ERG-O, structural model).

The influence of formative indexes is stronger on the perceived ease of learning how to use ARTP than on the general ease of use. The highest contributions to ERG-P have the items ERGP2 (accuracy of superposition) and ERGP3 (understanding the vocal explanation). The contribution of ERGP3 shows the importance of vocal explanations for students. The most important contribution to ERG-O has the last item related to the ease of collaboration with colleagues. The ease of selecting a menu item proved also to be an important measure for the Chemistry scenario.

4.2 Second order formative index

The results of structural model estimation are presented in Table 4. Both γ-coefficients are significant. The contribution of each dimension is similar for the Chemistry scenario. The analysis of modification indices showed that the index is completely mediating the effects of its items.

Table 4: Estimation results for second order construct - Chemistry scenario

ERG	MIMIC model		Structural model	
	γ/β	sig.(p)	γ/β	sig.(p)
Contribution				
ERG-P	.45	<.001	.55	<.001
ERG-O	.47	<.001	.49	0.004
Effect variables				
PEOU1	.53	<.001	.55	<.001
PEOL			.83	<.001
PEOL1	.70	<.001		
PEOL2	.68	<.001		
PEOL3	.53	<.001		
Variance expl.				
ERG	63%		80%	
PEOL			69%	

Both β-coefficients are significant ($p<0.001$), which supports the hypotheses. The variance of the error term associated with the formative index is 0.016 (0.222) which means a medium to large effect. The magnitude of the error term is suggesting some other aspects not covered by the indicators.

Fit indices are acceptable, over the recommended values: χ^2=14.932, DF=7, χ^2/DF=2.133, GFI=0.964, CFI=0.960, SRMR=0.049 (structural model).

4.3 Comparison of results and discussion

The estimation of formative indexes on the Chemistry scenario data cross validated the measurement model and enables a comparison between the two implemented scenarios. The variances explained by the structural models for the formative indexes are higher for the Biology scenario than for the Chemistry scenario.

The variance explained by the model for the second order construct is slightly higher for the Biology scenario. The contribution of ERG-P to the super ordinate index is higher than the contribution of ERG-O in both scenarios but the relative importance is much higher for the Biology scenario. The variance explained by the model for the outcome variable PEOL is also higher for the Biology scenario (81% vs. 69%).

As regarding the ERG-P index, the comparison reveals that understanding of vocal explanations (ERGP3) is the most important item for the Chemistry scenario and the less important for the Biology scenario. This is explained by the fact that the Chemistry demo lesson and exercises were more difficult for students so a clear understanding of the lesson and how to perform the exercises was critical. The accuracy of superposition between the projection and the real object (ERGP2) has a higher importance for Biology.

As regarding ERG-O, the comparison reveals that the ease of collaboration with colleagues (ERGO4) is the most important item for both scenarios. Selecting a menu item (ERGO2) was easy for the Biology scenario (lowest γ-coefficient) and difficult the Chemistry scenario. This is explained by the fact that the students had to use both hands to manipulate the colored balls (symbolizing atoms) so handling also the remote control became more difficult. Correcting the mistakes (ERGO3) was more difficult for the Biology scenario because of frequent selection errors when students tried to select a small organ.

In both scenarios, ERG-A had a significant positive influence on the formative indicator ERGO1, showing that the ease to adjust the see-through screen and stereo glasses is influencing the comfort on the work place.

5 Conclusion and future work

The main contribution of this study is a measurement model for the perceived ease of use of the ARTP featuring a second order formative index with two dimensions: the quality of visual and auditory perception and the ease of interaction and collaboration. These indexes are antecedents of a reflective construct measuring the perceived ease of learning how to use ARTP. The latter could be then integrated in structural models that are based solely on reflective scales.

There are several strengths and limitations of this study. An outcome of this research is the integration of almost all items related to the perceived ease of use that were eliminated in a previous work [4] for unidimensionality and convergent validity reasons. The new measurement model includes 12 of 15 items related to the ergonomic quality. As such, it provides a wider perspective on the ergonomic quality and enables the analysis of specific usability aspects. Second, the estimation of a formative measurement model provides a more detailed information (at indicator level) shedding light on usability aspects that are critical for ARTP and a given learning scenario. Third, the formative indexes were specified and validated with a structural model that addressed all general aspects related to the perceived ergonomic quality: ease of understanding, ease of use and ease of operating with a software system. Since all variables are strongly related to the focal construct the structural model is well supporting an external validity. Up to now, there is no similar model developed for the ergonomic quality of a software system. Fourth, the model was estimated and cross validated on two different samples which enables a comparison between scenarios and makes it possible to further integrate and discuss in more detail the answers at open questions (qualitative data).

As regarding the limitations, the sample used in this study was collected from only 6 classes (3 Romanian schools), having a limited representativeness. Second, both samples are small, at limit for SEM (Structural Modeling Equation) requirements. Third, the convergent validity of the relatively measured construct is at limit (acceptable for an exploratory study). Fourth, the breadth of formative indicators is inherently limited since the evaluation questionnaire was indented to capture the main usability aspects. Fifth, there are inherent limitations since the methodology regarding formative indexes estimation and validation is not mature yet. The usability questionnaire used to collect the data was conceptualized in 2007 while the main recommendations for formative indexes development have been published only in 2008.

Based on this work we intend to develop a new evaluation questionnaire having both formative and reflective items. The questionnaire will be used for the evaluation of a new version of the Chemistry application which is currently under development.

Acknowledgements

This work was supported by the research projects TEHSIN 503/2009 and ARiSE FP6-027039.

Bibliography

[1] Anderson, J.C., Gerbing, D.W. Structural Equation Modelling in Practice: A Review and Recommended Two-Step Approach. *Psychological Bulletin 103(3)*, 411-423, 1988.

[2] Arbuckle, J.L. *AMOS 16.0 User's Guide*. Amos Development Corporation, 2007.

[3] Balog, A., Pribeanu, C. Developing a measurement model for the evaluation of AR-based educational systems. *Studies in Informatics and Control 18(2)*, 137-148, 2009.

[4] Balog, A., Pribeanu, C. The Role of Perceived Enjoyment in the Students' Acceptance of an Augmented Reality Teaching Platform: a Structural Equation Modelling Approach . *Studies in Informatics and Control 19(3)*, 319-330, 2010.

[5] Bach, C., Scapin, D., Obstacles and perspectives for Evaluating mixed Reality Systems Usability. *Proceedings of IUI-CADUI Conference 2004*, 72-79, 2004.

[6] Bollen, K., Lennox, R. Conventional wisdom on measurement: a structural perspective. *Psychological Bulletin 110(2)*, 305-314, 1991.

[7] Bruhn, M., Georgi, D., Hadwich, K. Customer equity management as formative second order construct. *Journal of Business Research 61*, 1292-1301, 2008.

[8] Cadogan, J., Souchon, A., Procter, D. The quality of market-oriented behaviors: Formative index construction. *Journal of Business Research 61*, 1263-1277, 2008.

[9] Davis, F.D. Perceived usefulness, perceived easy of use, and user acceptance of information technology. *MIS Quaterly 13*, 319-340, 1989.

[10] Diamantopoulos, A., Winklhofer, H. Index construction with formative indicators : an alternative to scale development. *Journal of Marketing Research 28*, 269-277, 2001.

[11] Diamantopoulos, A. The error term in formative measurement models : interpretation and modeling implications. *Journal of Modeling in Management 1(1)*, 7-17, 2006.

[12] Diamantopoulos, A., Riefler, P., Roth, K. Advancing formative measurement models. *Journal of Business Research 61*, 1203-1218, 2008

[13] Edwards, J., Bagozzi, R. On the nature and direction of relationship between constructs and measures. *Psychological Methods 5(2)*, 155-174, 2000.

[14] Franke, G., Preacher, K., Rigdon, E. Proportional structural effects of formative indicators. *Journal of Business Research 61*, 1229-1237, 2008.

[15] Gabbard, J., Swann, E. Usability engineering for augmented reality: Employing user-based studies to inform design. *IEEE Transactions on Visualization and Computer Graphics 14(3)*, 513-525, 2008.

[16] Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., Tatham, R.L. *Multivariate Data Analysis*, Prentice Hall, 2006.

[17] Hornbaek, K. Current practice in measuring usability: Challenges to usability studies and research. *Int. J. Human Computer Studies. 64*, 79-102, 2006.

[18] Huang, H.M., Rauch, U., Liaw, S.S. Investigating learners' attitude towards virtual reality learning environments: based on a constructivist approach. *Computers & Education 55*, 1171-1182, 2010.

[19] ISO 9126-1:2001 *Software Engineering - Software product quality.* Part 1: Quality Model

[20] Jarvis, C.B., Mackezie, S., Podsakoff, M. A critical review of construct indicators and measurement models misspecification in marketing and consumer research. *Journal of Consumer Research 30*, 199-218, 2003.

[21] Konradt, U., Christophersen, T., Schaefer-Kuelz, U. Predicting user satisfaction, strain and system usage of employee self-services. *Int. J. of Human-Computer Studies 64*, 1141-1153, 2006.

[22] Krauss, M., Riege, K., Winter, M., Pemberton, L. Remote Hands-On Experience: Distributed Collaboration with Augmented Reality. *Proceedings EC-TEL 2009*, LNCS 5794, Springer, 226-239, 2009

[23] Pribeanu, C., Balog, A., Iordache, D.D. Measuring the usability of augmented reality e-learning systems: a user-centered evaluation approach.Chapter 14: *Software and Data Technologies, CCIS 47*, Corderiro, H., Shiskov B, Ranchordas A, Helfert M (Eds.), Springer, 175-186, 2009.

[24] Ruiz, D.M., Gremler, D., Washburn, J., Carrion, G.C. Service value revisited: specifying a high order formative measure. *Journal of Business Research 61*, 1278-1291, 2008.

[25] Tabachnick, B. G., Fidell, L. S. . *Using Multivariate Statistics*, 5th ed. Boston: Allyn and Bacon, 2007.

[26] Wilcox, J., Howell, R., Breivik, E. Questions about formative measurement. *Journal of Business Research 61*, 1219-1228, 2008.

[27] Wind, J., Riege, K., Bogen M. SpinnstubeŽ: A Seated Augmented Reality Display System, *Virtual Environments: Proc. IPT-EGVE - EG/ACM Symposium*, 17-23, 2007.

Annex 1 Constructs and items

ERG	MM	Items	Variables
Quality of of visual and auditory perception (ERG-P)	F	ERGP1	Observing through the screen is clear
	F	ERGP2	The superposition between projection and the real object is clear
	F	ERGP3	Understanding the vocal explanations is easy
	F	ERGP4	Reading the information on the screen is easy
Ease of interaction and collaboration (ERG-O)	F	ERGO1	The work place is comfortable
	F	ERGO2	Selecting a menu item is easy
	F	ERGO3	Correcting the mistakes is easy
	F	ERGO4	Collaborating with colleagues is easy
Ease of adjusting devices (ERG-A)	F	ERGA1	Adjusting the "see-through" screen is easy
	F	ERGA2	Adjusting the stereo glasses is easy
	F	ERGA3	Adjusting the head phones is easy
Perceived ease of learning to operate (PEOL)	R	PEOL1	Understanding how to operate with ARTP is easy
	R	PEOL2	Learning how to operate with ARTP is easy
	R	PEOL3	Remembering how to operate with ARTP is easy
*** General item	R	PEOU1	Overall, I find the system easy to use

Note: MM(Measurement Model): F (Formative) / R (Reflective)

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 733-743

Structural Regular Multiple Criteria Linear Programming for Classification Problem

Z. Qi, Y. Shi

Zhiquan Qi
Research Center on Fictitious Economy & Data Science,
Chinese Academy of Sciences, Beijing 100190, China
E-mail: qizhiquan@gucas.ac.cn

Yong Shi
1. Research Center on Fictitious Economy & Data Science,
Chinese Academy of Sciences, Beijing 100190, China and
2.College of Information Science & Technology,
University of Nebraska at Omaha Omaha, NE 68182, USA
E-mail: yshi@gucas.ac.cn

Abstract:
Classification problem has attracted an increasing amount of interest. Various classifiers have been proposed in the last decade, such as ANNs, LDA, and SVM. Regular Multiple Criteria Linear Programming (RMCLP) is an effective classification method, which was proposed by Shi and his colleagues and have been applied to handle different real-life data mining problems. In this paper, inspired by the application potential of RMCLP, we propose a novel Structural RMCLP (called SRMCLP) method for classification problem. Unlike RMCLP, SRMCLP is sensitive to the structure of the data distribution and can construct more reasonable classifiers by exploiting these prior data distribution information within classes. The corresponding optimization problem of SRMCLP can be solved by a standard quadratic programming. The effectiveness of the proposed method is demonstrated via experiments on synthetic and available benchmark datasets.
Keywords: classification, RMCLP, structural information of data, SVM

1 Introduction

For the last decade, the researchers have extensively developed various optimization techniques to deal with the classification problem in data mining or machine learning. Support Vector Machine (SVM) ([1,2]) is one of the most popular methods. However, Applying optimization techniques to solve classification has seventy years history. Linear Discriminant Analysis(LDA) ([3]) was first proposed in 1936. Mangasarian ([4]) has proposed a large margin classifier based on linear programming in 1960's. From 1980's to 1990's, Glover proposed a number of linear programming models to solve discriminant problems with a small sample size of data ([5,6]). Recently, Shi and his colleagues([7]) extend Glover's method into classification via multiple criteria linear programming (MCLP), and then various improved algorithms were proposed one after the other ([8, 9]). These mathematical programming approaches to classification have been applied to handle many real world data mining problems, such as credit card portfolio management ([11, 12]), bioinformatics ([13]), firm bankruptcy ([14]), and etc.

Recently, how to apply the structural information of data to build a good classifier is a new research focus. Many new large margin classifiers based on structural information have been proposed. Exploiting clustering algorithms to extract the structural information embedded with classes is one popular strategy [15–17]. The structured large margin machine (SLMM) [15] is a representative work based on the strategy. Firstly, SLMM explores the structural information within classes by Ward's agglomerative hierarchical clustering method on input

data [18], and then introduces the related structure information into the constraints. Finally, SLMM can be solved by a sequential second order cone programming (SOCP). Experimentally, SLMM is superior to support vector machine minimax probability machine (MPM) [19] and maximin margin machine(M4) [20]. However, as we all know, solving the involved SOCP problem is more difficult than the Quadratic Programming Problem (QPP) as in SVM, so SLMM has more higher computational complexity than traditional SVM. Consequently, a novel structural support vector machine (SRSVM) was proposed by Xue et. al [17]. Unlike SLMM, SRSVM exploits the classical framework of SVM rather than as constraints in SLMM and the corresponding optimization problem can still be solved by the QPP. SRSVM has been shown to be theoretically and empirically better in generalization than SVM and SLMM.

In this paper, inspired by the success of SRSVM and the application potential of RMCLP, we propose a novel Structural RMCLP (called SRMCLP) method for classification problem. Unlike RMCLP, SRMCLP is sensitive to the structure of the data distribution and can construct more reasonable classifiers by exploiting these prior data distribution information within classes.

The remaining parts of the paper are organized as follows. Section 2 introduces the basic notions and formulation of MCLP; Section 3 describes in detail our proposed Algorithms; All experimental results are shown in section 4; Conclusions are given in the last section.

2 Background

We give a brief introduction of MCLP in the following. For classification about the training data

$$T = \{(x_1, y_1), \cdots, (x_l, y_l)\} \in (\Re^n \times \mathcal{Y})^l, \tag{1}$$

where $x_i \in \Re^n, y_i \in \mathcal{Y} = \{1, -1\}, i = 1, \cdots, l$, data separation can be achieved by two opposite objectives. The first objective separates the observations by minimizing the sum of the deviations (MSD) among the observations. The second maximizes the minimum distances (MMD) of observations from the critical value [6]. The overlapping of data u should be minimized while the distance v has to be maximized. However, it is difficult for traditional linear programming to optimize MMD and MSD simultaneously. According to the concept of Pareto optimality, we can seek the best trade-off of the two measurements [11, 12]. So MCLP model can be described as follows:

$$\begin{aligned}
\min_{u} \quad & e^T u \quad \& \quad \max_{v} \; e^T v, && && (2)\\
\text{s.t.} \quad & (w \cdot x_i) + (u_i - v_i) = b, && \text{for } \{i | y_i = 1\}, && (3)\\
& (w \cdot x_i) - (u_i - v_i) = b, && \text{for } \{i | y_i = -1\}, && (4)\\
& u, v \geq 0, && && (5)
\end{aligned}$$

where $e \in R^l$ is a vector whose all elements are 1, w and b are unrestricted, u_i is the overlapping and v_i the distance from the training sample x_i to the discriminator $(w \cdot x_i) = b$ (classification separating hyperplane). By introducing penalty parameter $c, d > 0$, MCLP has the following version

$$\begin{aligned}
\min_{u,v} \quad & ce^T u - de^T v, && && (6)\\
\text{s.t.} \quad & (w \cdot x_i) + (u_i - v_i) = b, && \text{for } \{i | y_i = 1\}, && (7)\\
& (w \cdot x_i) - (u_i - v_i) = b, && \text{for } \{i | y_i = -1\}, && (8)\\
& u, v \geq 0, && && (9)
\end{aligned}$$

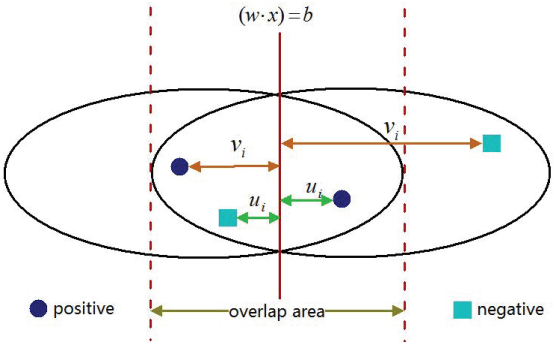


Figure 1: Geometric meaning of MCLP.

The geometric meaning of the model is shown in Figure 1.

A lot of empirical studies have shown that MCLP is a powerful tool for classification. However, we cannot ensure this model always has a solution under different kinds of training samples. To ensure the existence of solution, recently, Shi et al proposed a RMCLP model by adding two regularized items $\frac{1}{2}w^T Hw$ and $\frac{1}{2}u^T Qu$ on MCLP as follows (more theoretical explanation of this model can be found in [8]):

$$\min_{z} \quad \frac{1}{2}w^T Hw + \frac{1}{2}u^T Qu + de^T u - ce^T v, \tag{10}$$

$$\text{s.t.} \quad (w \cdot x_i) + (u_i - v_i) = b, \qquad \text{for } \{i|y_i = 1\}, \tag{11}$$

$$(w \cdot x_i) - (u_i - v_i) = b, \qquad \text{for } \{i|y_i = -1\}, \tag{12}$$

$$u, v \geq 0, \tag{13}$$

where $z = (w^T, u^T, v^T, b)^T \in R^{n+l+l+1}$, $H \in R^{n\times n}$, $Q \in R^{l\times l}$ are symmetric positive definite matrices. Obviously, the regularized MCLP is a convex quadratic programming.

Compared with traditional SVM , we can find that the RMCLP model is similar to the Support Vector Machine model in terms of the formation by considering the minimization of overlapping of the data. However, RMCLP tries to measure all possible distances v from the training samples x_i to separating hyperplane, while SVM fixes the distance as 1 (through bounding planes $(w \cdot x) = b \pm 1$) from the support vectors. Although the interpretation can vary, RMCLP addresses more control parameters than the SVM, which may provide more flexibility for better separation of data under the framework of the mathematical programming. In addition, different with SVM, RMCLP considers all the samples to solve classification problem. These make RMCLP have stronger insensitivity to outliers.

3 Structural Regular Multiple Criteria Linear Programming for Classification Problem

3.1 Extracting Structural Information within Classes

Following the strategy of the SLMM and SRSVM, SRMCLP also has two steps. The first step is to extract structural information within classes by some clustering method; the second step is the model learning. In order to compare the main different of the second step between SRMCLP and the other two methods, here we adopt the same clustering method: Ward's linkage clustering(WIL) [15–18], which is one of the hierarchical clustering analysis. A main advantage of WIL is that clusters derived from this method are compact and spherical, which provides a

meaningful basis for the computation of covariance matrices [15]. Concretely, if S and T are two clusters with means μ_S and μ_T, the Ward's linkage $W(S,T)$ between clusters S and T can be computed as [15]

$$W(S,T) = \frac{|S| \cdot |T| \cdot \|\mu_S - \mu_T\|}{|S| + |T|}. \tag{14}$$

Initially, each sample is considered as a cluster. The Ward's linkage of two samples x_i and x_j is $W(x_i, x_j) = \|x_i - x_j\|^2/2$. When two clusters are being merged to a new cluster A', the linkage $W(A', C)$ can be conveniently derived from $W(A,C), W(B,C)$ and $W(A,B)$ by [15]

$$W(A', C) = \frac{(|A| + |C|)W(A,C) + (|B| + |C|)W(B,C) - |C|W(A,B)}{|A| + |B| + |C|}. \tag{15}$$

During the hierarchical clustering, the Ward's linkage between clusters to be merged increases as the number of clusters decreases [15]. A relation curve between the merge distance and the number of clusters can be drawn to represent this process. The optimal number of clusters can be determined by finding the knee point. Furthermore, the WIL can also be extended to the kernel space. More details of WIL can be found in [15].

3.2 Model Learning

We obtained two groups of P and N clusters in class C_P and C_N by the first step, i.e., $P = P_1 \bigcup \cdots P_i \bigcup \cdots P_{C_P}, N = N_1 \bigcup \cdots N_j \bigcup \cdots N_{C_N}$. Consider the optimization(10), choosing H, Q to be identity matrix and introducing $\frac{1}{2}b^2$ and $\frac{1}{2}w^\top \Sigma w$ into the object function, SRMCLP can be be formulated as:

$$\begin{aligned} \min_z \ & \frac{1}{2}\|w\|^2 + \frac{1}{2}c_1 w^\top \Sigma w + \frac{1}{2}\|u\|^2 + \frac{1}{2}b^2 + c_2 e^T u - c_3 e^T v, \\ \text{s.t.} \ & (w \cdot x_i) + (u_i - v_i) = b, \qquad \text{for } \{i | y_i = 1\}, \\ & (w \cdot x_i) - (u_i - v_i) = b, \qquad \text{for } \{i | y_i = -1\}, \\ & u, v \geq 0, \end{aligned} \tag{16}$$

where $z = (w^T, u^T, v^T, b)^T \in R^{n+l+l+1}$, $c_1, c_2, c_3 \geq 0$ are the pre-specified penalty factors, $\Sigma = \Sigma_+ + \Sigma_-$, where $\Sigma_+ = \Sigma_{P_1} + \cdots, +\Sigma_{P_{C_P}}$, $\Sigma_- = \Sigma_{N_1} \cdots, +\Sigma_{N_{C_N}}$, Σ_{P_i} and Σ_{N_j} are respectively the covariance matrices corresponding to the i th and j th clusters in the two classes, $i = 1, \cdots, C_P$, $j = 1, \cdots, C_N$. Obviously, the regularized SRMCLP is a convex quadratic programming. By introducing its Lagrange function

$$L(w,u,v,b,\alpha,\beta,\eta) = \frac{1}{2}\|w\|^2 + \frac{1}{2}c_1 w^\top \Sigma w + \frac{1}{2}\|u\|^2 + \frac{1}{2}b^2 + c_2 e^T u - c_3 e^T v + \tag{17}$$

$$\sum_{i=1}^{l} \alpha_i (y_i((w \cdot x_i) - b) + u_i - v_i) - \sum_{i=1}^{l} \beta_i u_i - \sum_{i=1}^{l} \eta_i v_i, \tag{18}$$

where $\alpha_i, \beta_i, \eta_i \in R$ are the Lagrange multipliers, Therefore the dual problem of (39) can be formulated as

$$\begin{aligned} \max_{w,u,v,b,\alpha,\beta,\eta} \ & L(u,v,w,b,\alpha,\beta,\eta), \\ \text{s.t.} \ & \nabla_{w,u,v,b} L(u,v,w,b,\alpha,\beta,\eta) = 0, \\ & \beta_i, \eta_i \geq 0, i = 1, \ \cdots, l. \end{aligned} \tag{19}$$

From (19) we get

$$\nabla_w L = (I + c_1\Sigma)w + \sum_{i=1}^{l} y_i\alpha_i x_i = 0, \tag{20}$$

$$\nabla_{u_i} L = u_i + c_2 + \alpha_i - \beta_i = 0, \quad i = 1, \quad \cdots, l, \tag{21}$$

$$\nabla_{v_i} L = -c_3 - \alpha_i - \eta_i = 0, \quad i = 1, \quad \cdots, l, \tag{22}$$

$$\nabla_b L = b - \sum_{i=1}^{l} y_i\alpha_i = 0, \tag{23}$$

where I is an identity matrix. Substituting the above equations into problem (19), the dual problem can be expressed as

$$\max_{\alpha,u,b} \quad -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j [x_i^\top (2I + c_1\Sigma) x_j] - \frac{1}{2}\sum_{i=1}^{l} u_i^2, \tag{24}$$

$$\text{s.t.} \quad \sum_{i=1}^{l} y_i\alpha_i = 0, \quad i = 1, \cdots, l, \tag{25}$$

$$-c_2 - u_i \le \alpha_i \le -c_3, \quad i = 1, \cdots, l. \tag{26}$$

Solving the convex quadratic programming problem , we can obtain its solution.

$$w = -(I + c_1\Sigma)^{-1} \sum_{i=1}^{l} y_i\alpha_i x_i, \tag{27}$$

$$b = \sum_{i=1}^{l} y_i\alpha_i. \tag{28}$$

$$\tag{29}$$

So the decision function can be formulated as follows

$$f(x) = -\sum_{i=1}^{l} y_i\alpha_i x_i^\top (I + c_1\Sigma)^{-1} x + b. \tag{30}$$

Applying the kernel trick, we also extend the linear SRMCLP to the nonlinear case. Introduce the kernel function $K(x, x') = (\Phi(x) \cdot \Phi(x'))$, where $\Phi(\cdot)$ is a mapping from the input space R^n to Hilbert space $\mathcal{H}$

$$\Phi : \begin{array}{c} R^n \to \mathcal{H}, \\ x \to \Phi(x). \end{array} \tag{31}$$

Then the optimization problem of SRMCLP in the kernel space can be described as:

$$\max_{\alpha,u,b} \quad -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j [\Phi(x_i)^\top (2I + c_1\Sigma^\Phi)\Phi(x_j)] - \frac{1}{2}\sum_{i=1}^{l} u_i^2, \tag{32}$$

$$\text{s.t.} \quad \sum_{i=1}^{l} y_i\alpha_i = 0, \quad i = 1, \cdots, l, \tag{33}$$

$$-c_2 - u_i \le \alpha_i \le -c_3, \quad i = 1, \cdots, l. \tag{34}$$

$2\Phi(x_i)^\top I\Phi(x_j) = K(x_i, x_j)$, so we only need to consider how to compute the kernel matrix $c_1\Phi(x_i)^\top\Sigma^\Phi\Phi(x_j)$. Suppose T_{P_i} is a matrix corresponding to the cluster P_i, $T_{P_i} \in \Re^{P_i\times n}$, in which the k-th row is $x_k^\top$. O_{P_i} is a mean matrix of cluster P_i, $O_{P_i} \in \Re^{P_i\times n}$. Each row of O_{P_i} is the same, i.e.

$$\mu_{P_i} = \frac{1}{P_i}\sum_{x_k\in P_i} x_k. \tag{35}$$

The related covariance matrix for cluster P_i can be expressed as

$$\Sigma^\Phi_{P_i} = \frac{1}{P_i}(\Phi(T_{P_i}) - \Phi(O_{P_i}))^\top(\Phi(T_{P_i}) - \Phi(O_{P_i})). \tag{36}$$

So we obtain

$$\begin{aligned}\Phi(x_i)^\top\Sigma^\Phi_+\Phi(x_j) =&(\frac{1}{\sqrt{P_i}}(\Phi(T_{P_i}) - \Phi(O_{P_i}))\Phi(x_i))^\top\\ &(\frac{1}{\sqrt{P_i}}(\Phi(T_{P_i}) - \Phi(O_{P_i}))\Phi(x_j))\\ =&(\frac{1}{\sqrt{P_i}}(K(T_{P_i}, x_i) - K(O_{P_i}, x_i))^\top\\ &(\frac{1}{\sqrt{P_i}}(K(T_{P_i}, x_j) - K(O_{P_i}, x_j)).\end{aligned} \tag{37}$$

Similarly, $\Phi(M)^\top\Sigma^\Phi_-\Phi(M)$ of F_Φ can computed

$$\begin{aligned}\Phi(x_i)^\top\Sigma^\Phi_-\Phi(x_j) =&(\frac{1}{\sqrt{P_i}}(K(T_{N_i}, x_i) - K(O_{N_i}, x_i))^\top\\ &(\frac{1}{\sqrt{P_i}}(K(T_{N_i}, x_j) - K(O_{N_i}, x_j)),\end{aligned} \tag{38}$$

where T_{N_i} is a matrix of cluster N_i, O_{N_i} is a mean matrix of cluster N_i.

SRMCLP has a similar structure of RMCLP, We can easily proof that RMCLP are the special case of SRMCLP. Suppose the variance-covariance matrix of each cluster is $\Sigma_{P_i} = \sigma_{N_j} = I$, $i = 1, \cdots C_P$, $j = 1, \cdots, C_N$. For an example of linear SRMCLP, the primal optimization problem (39) of SRMCLP becomes

$$\begin{aligned}\min_z\ &\frac{1}{2}\|w\|^2 + c_1\frac{C_P + C_N}{2}\|w\|^2 + \frac{1}{2}\|u\|^2 + \frac{1}{2}b^2 + c_2e^Tu - c_3e^Tv,\\ \text{s.t.}\ &(w\cdot x_i) + (u_i - v_i) = b, \qquad \text{for } \{i|y_i = 1\},\\ &(w\cdot x_i) - (u_i - v_i) = b, \qquad \text{for } \{i|y_i = -1\},\\ &u, v \geq 0,\end{aligned} \tag{39}$$

It is not difficult to see that the optimization problem (10) is equivalent to one of the primal problem of RMCLP.

4 Experiments

We compare the SRMCLP against RMCLP and SRSVM [16,17] on various data sets in this section.

The testing accuracies of all experiments are computed using standard 10-fold cross validation. c_1, c_2, c_3 and RBF kernel parameter σ are all selected from the set $\{2^i|i = -7, \cdots, 7\}$

by 10-fold cross validation on the tuning set comprising of random 10% of the training data. Once the parameters are selected, the tuning set is returned to the training set to learn the final decision function. The "quadprog" function is used to solve the QP problems in SRMCLP, RMCLP and SRSVM. The "1 vs r" method [2] is used to solve the multi-class classification. All algorithms are implemented by using MATLAB 2010. The experiment environment: Intel Core i7-2600 CPU, 4 GB memory.

4.1 Toy data

In the subsection, we use a 2-D toy data to show the intuitive performance of SRMCLP. The 2-D toy data is the synthetic XOR dataset [17], which is a typical linearly nonseparable problem in classification and randomly generated under two Gaussian distributions in each class. In practise, samples in each class are designed to two clusters P_1, P_2 and N_1, N_2(the number of samples in each cluster is equal), and each gaussian distribution contains 100 samples. We respectively use $10\%, 20\%, 30\%, 50\%$ of data in each cluster as the training set, and others for testing. The comparative results of SRMCLP and SRSVM are shown in Figure 2.

In the XOR dataset, the positive class and negative class have both the horizontal distribution and the vertical distribution. How to fully exploit these prior knowledge will be a very difficult task. From Figure 2, we can find that SRMCLP's discriminant boundaries basically enclose those of RMCLP, which means that SRMCLP has better generalization performance than RMCLP. Figure 3, we can also find that the accuracy's different of these methods decreases with the increase of training samples. Those show that SRMCLP can fully exploit these prior structural information to design a more reasonable classifier.

4.2 UCI datasets

In this subsection, we perform these methods on the UCI datasets [21]. For each dataset, we randomly select the same number of data from different classes to compose a dataset. 50% percent of each extracted dataset are for training, 50% for testing. The results are shown in the Table 1. From the Table 1, we can draw the conclusion as follows: 1) SRMCLP and SRSVM have the better predictive ability than RMCLP in all cases. This shows that these priori structural information embedded in classes has a great help to improve the classification performance of the classifier. 2) SRMCLP is superior to SRSVM in most cases. This shows SRMCLP is a strong competitive method.

5 Conclusion

In this paper, we proposed a novel Structural RMCLP (called SRMCLP) method for classification problem. Unlike RMCLP, SRMCLP is sensitive to the structure of the data distribution and can construct more reasonable classifiers by exploiting these prior data distribution information within classes. The corresponding optimization problem of SRMCLP can be solved by a standard quadratic programming. The effectiveness of the proposed method is demonstrated via experiments on synthetic and available benchmark datasets and applications on the decision supporting system. In the future work, we will apply SRMCLP into other actual classification problems such as stock forecast, credit card analysis to further test its effectiveness.

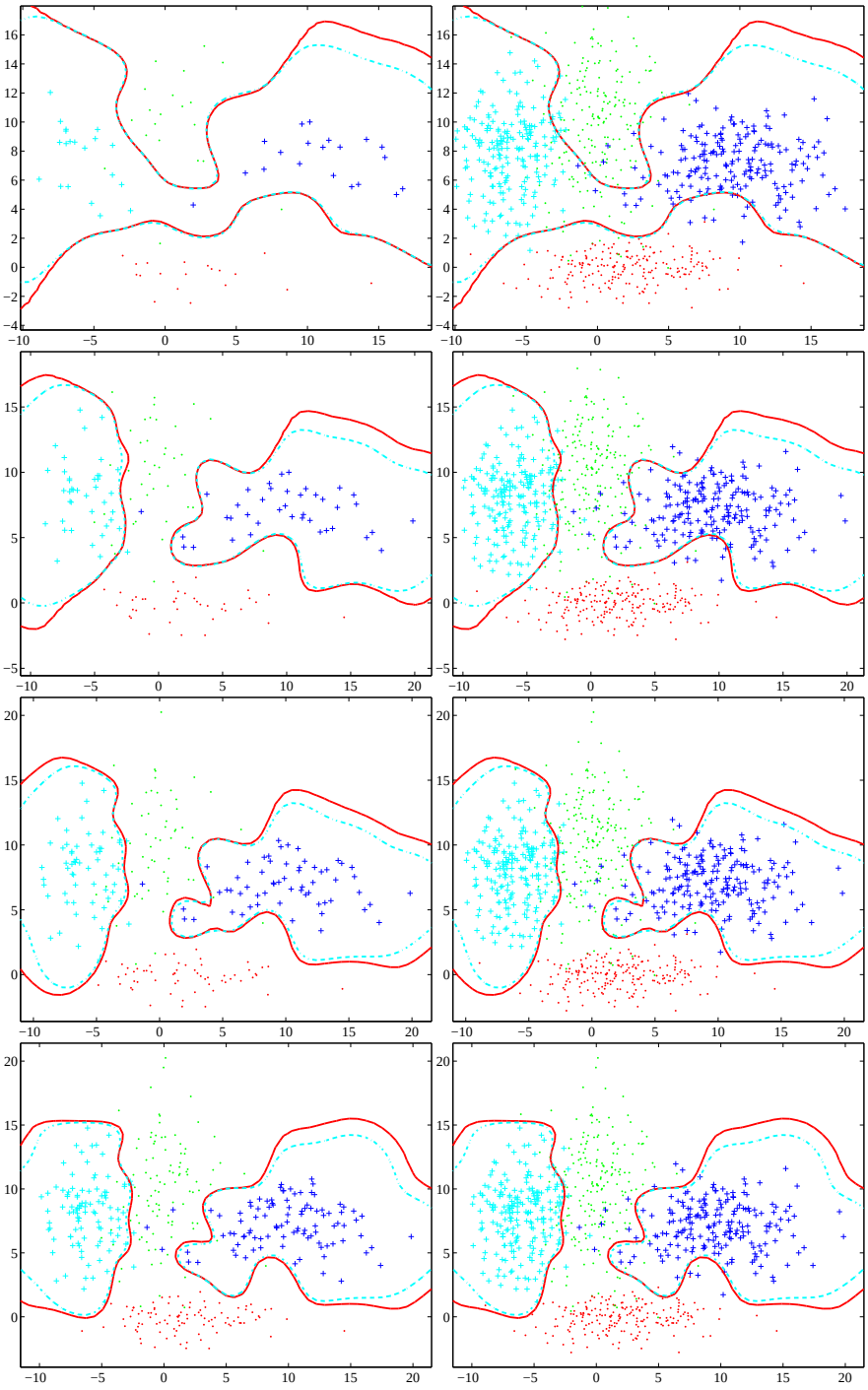


Figure 2: The performance of SRMCLP and RMCLP in the case of RBF case. The first column and second column are the results on the training set and testing set. Each row is the result on $10\%, 20\%, 30\%$ and 50% training sets, respectively. The magenta dotted curve and red solid curve denote the hyperplanes of SRMCLP and RMCLP, respectively.

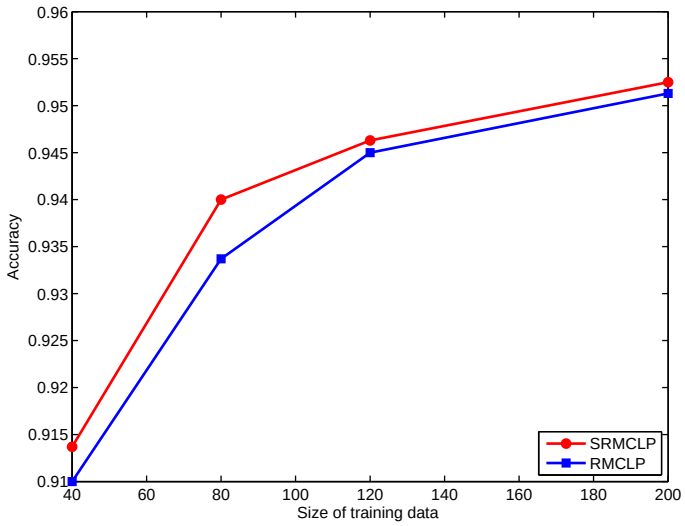


Figure 3: Accuracy of SRMCLP and RMCLP on the XOR dataset

Table 1: The testing accuracy and training times on UCI datasets

Datasets	SRMCLP Accuracy	SRSVM Accuracy	RMCLP Accuracy
Hepatitis (155×19)	**79.91±2.55**	79.83±1.27	77.82±4.22
Australian (690×14)	68.18±2.12	**69.32±2.31**	67.73±1.56
BUPA liver (345×6)	68.12±1.34	**68.96±1.09**	67.61±2.71
CMC (844×9)	**65.71±2.54**	65.27±2.35	64.53±3.62
Credit (690×19)	**76.36±2.18**	76.11±2.61	75.82±2.87
Diabetis (768×8)	63.98±1.90	**64.19±2.43**	62.44±3.47
Flare-Solar (1066×9)	**59.11 ±2.98**	58.43±2.77	57.96±2.51
German (1000×20)	**63.91±1.93**	63.84±1.88	62.55±2.86
Heart-Statlog (270×14)	**76.13 ±2.50**	76.04±2.47	76.21±2.83
Image (2310×18)	83.18 ±2.57	**83.44±1.40**	82.64±2.88
Ionosphere (351×34)	**76.93 ±2.61**	76.55±2.63	76.43± 3.26

6 Acknowledgment

This work has been partially supported by grants from National Natural Science Foundation of China(NO.70921061, NO.10601064), the CAS/SAFEA International Partnership Program for Creative Research Teams, Major International(Ragional) Joint Research Project(NO.71110107026), the President Fund of GUCAS.

Bibliography

[1] Vapnik V.N. The Nature of Statistical Learning Theory. 2nd ed. New York: Springer, 2000.

[2] Deng N.Y., Tian Y.J. Support vector machines: Theory, Algorithms and Extensions. Science Press, Beijing, 2009.

[3] Fisher R.A. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics 7(2): 179-188, 1936.

[4] Mangasarian O.L. Generalized support vector machines. Advances in Large Margin Classifiers. Cambridge, MA: MIT Press, 2000.

[5] Freed N., Glover F. Simple but powerful goal programming models for discriminant problems. European Journal of Operational Research 7: 44-60, 1981.

[6] Freed N., Glover F. Evaluating alternative linear programming models to solve the two-group discriminant problem. Decision Science 17: 151-162, 1986.

[7] Olson D., Shi Y. Introduction to Business Data Mining. McGraw-Hill/Irwin, 2007.

[8] Shi Y., Tian Y.J., Chen X.J., Zhang P. A Regularized Multiple Criteria Linear Program for Classification. In: ICDM Workshops, pp 253-258, 2007.

[9] Kou G., Shi Y., Wang S.Y. Multiple criteria decision making and decision support systems - Guest editor's introduction. Decision Support Systems 51(2): 247-249, 2011.

[10] Zhang D., Tian Y.J., Shi Y. A regression method by multiple criteria linear programming. In: 19th International Conference on Multiple Criteria Decision Making (MCDM), pp 7-12, 2008.

[11] Shi Y., Wise W., Lou M. Multiple Criteria Decision Making in Credit Card Portfolio Management. Multiple Criteria Decision Making in New Millennium, pp 427-436, 2001.

[12] Shi Y., Peng Y., Xu W. Data mining via multiple criteria linear programming: applications in credit card portfolio management. International Journal of Information Technology and Decision Making 1: 131-151, 2002.

[13] Zhang J., Zhuang W., Yan N. Classification of HIV-1 Mediated Neuronal Dendritic and Synaptic Damage Using Multiple Criteria Linear Programming. Neuroinformatics 2: 303-326, 2004

[14] Kwak W., Shi Y., Eldridge S. Bankruptcy prediction for Japanese firms: using multiple criteria linear programming data mining approach. International Journal of Data Mining and Business Intelligence, 2006.

[15] Yeung D., Wang D., Ng W., Tsang E., Wang X., Structured large margin machines: sensitive to data distributions, Machine Learning 68 (2): 171-200, 2007.

[16] Xue H., Chen S., Yang Q., Structural support vector machine, in: The 15th International Symposium on Neural Networks, pp. 501–511, 2008.

[17] Xue H., Chen S., Yang Q., Structural Regularized Support Vector Machine: A Framework for Structural Large Margin Classifier, Neural Networks, IEEE Transactions on 22 (4): 573–587, 2011.

[18] Ward, J.R. Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association, 58 (301): 236–244, 1963.

[19] G. R. G. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, M. I. Jordan, A robust minimax approach to classification, Journal of Machine Learning Research, 3: 555–582, 2002.

[20] Kzhuang K. H., Yang H. , King I., Learning large margin classifiers locally and globally, in: In The Twenty-First International Conference on Machine Learning, pp. 401-408, 2004.

[21] Murphy P.M. and Aha D.W. UCI machine learning repository, 1992.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 744-758

Minimum Cycle Time Analysis of Ethernet-Based Real-Time Protocols

J. Robert, J.-P. Georges, E. Rondeau, T. Divoux

Jérémy Robert, Jean-Philippe Georges,
Eric Rondeau, Thierry Divoux
Centre de Recherche en Automatique de Nancy,
Université de Lorraine, CNRS UMR 7039
Campus Sciences, BP 70 239, F-54506 Vandœuvre-lès-Nancy, France
E-mail: jeremy.robert@univ-lorraine.fr, eric.rondeau@univ-lorraine.fr,
jean-philippe.georges@univ-lorraine.fr, thierry.divoux@univ-lorraine.fr

Abstract:
The Ethernet standard is a standard solution for interconnecting industrial devices despite its intrinsic drawbacks, particularly its nondeterministic medium access method. Many Ethernet-based commercial solutions available (COTS - Components Off the Shelves) on the market guarantee time performance. This means that user selection of one particular solution is a critical decision, but the choice often depends more on political strategizing with an industrial device manufacturer than on the intrinsic performance of Ethernet-based interfaces. The objective of this paper is to provide a formal behavioural analysis of each Ethernet-based solution, in order to facilitate comparison.
Keywords: Real-time systems, performance analysis, embedded systems.

1 Introduction

Fieldbuses interconnecting industrial equipment are typically designed by Programmable Logic Controller (PLC) manufacturers. This is mainly attributable to specific constraints on industrial communication, which require a high degree of expertise. Fieldbuses must be robust in the noisy environments produced by the plant (physical layer), deterministic in guaranteeing data refresh during controller cycle periods (link layer), and capable of exchanging information between all types of industrial devices (application layer). Manufacturers provide different solutions to satisfy the first two constraints, without necessarily considering the final constraint to be relevant in business terms. This has resulted in the specification of a large number of fieldbus standards as described in [1].

A new trend endorsed particularly by the IAONA (Industrial Automation Open Networking Alliance) consortium was to promote the Ethernet network as a standard for industrial communications. The expected benefits are less costly network installations, because equipment is available off-the-shelf, and the avoidance of interoperability problems, because Ethernet technology is broadly used. Other advantages are that Ethernet is a well-known protocol, which is widely implemented, and its performance improves continuously with technological evolution (especially bandwidth).

However, access to the Ethernet medium relies on the nondeterministic CSMA/CD algorithm, which applies a stochastic method to resolving collisions and cannot guarantee that message transmissions will be received in bounded time. Consequently, the native Ethernet protocol cannot be directly implemented in a plant with severe time constraints, and many Ethernet-based solutions have been proposed to overcome this issue, as mentioned in [2–4]. However, if these Ethernet-based solutions are adapted to industry requirements, they lead to two types of problem. The first is that different solutions are not interoperable because different fieldbuses are

developed by each manufacturer. The second problem is that time performances are insufficiently evaluated or compared.

According to [5], different Ethernet products can be summarily classified into three main categories: the native Ethernet standard (Ethernet/IP, Modbus/TCP), Ethernet solutions using the priorities defined in IEEE802.1D/Q, and Ethernet-based solutions that incorporate new scheduling features in ASIC/FPGA (EtherCAT, Profinet IRT).

The last approach enables the elimination of all collisions and simplifies transmission time estimation, as described in [5, 6]. The authors compared EtherCAT with Profinet IRT in a simplified context using analytic models. This analysis was improved and refined in section 2 and extended to two other well-known industrial Ethernet products: the Modbus/TCP solution and Ethernet/IP. In section 3, we present several scenarios that use the analytic models to facilitate assessment of different Ethernet products.

2 Estimation of minimum cycle time

2.1 Introduction

The general objective of this study was to compare the time performance of the major industrial Ethernet products available on the market. This comparison could be achieved only in a common application context. Thus, the specification of the communication scenario was based on one controller (for example, a PLC), interconnecting sensors, and actuators in an Ethernet network. The controller was treated as the communication master that initiated all dialogues with slave nodes (sensors and actuators). The controller was characterized by its controller cycle time period, which was divided into three steps, as shown in Figure 1: - the sensor data refresh time, in the controller memory, - the processing time, and - the actuator update time.

Steps (1) and (3) represent communication periods, which should be less than the difference between the controller cycle time period and the processing time period. Thus, the time performance of each Ethernet product was compared according to a constraint named *minimum cycle time*, which was defined as:

The minimum cycle time was the communication time required by the controller to both collect and update the data memories of all sensors and actuators.

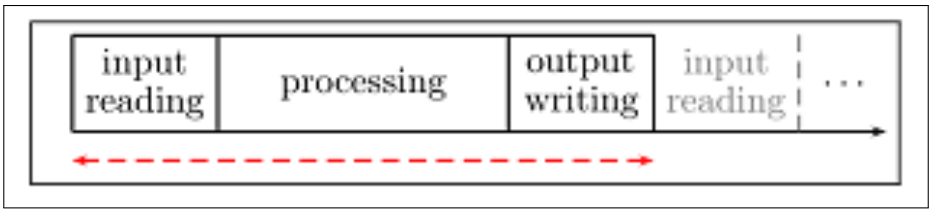


Figure 1: PLC cycle time

In the following section, the analytic models of minimum cycle time are elaborated for EtherCAT, Profinet IRT, Modbus/TCP, and Ethernet/IP. These models all used the following parameters: the transmission delay, the network device latency, the propagation delay, the link capacity, the payload, and the number of slaves. The notations of these are given in table 1. It was assumed that there were no transmission errors and that the network was dedicated to the PLC application and was not shared with other applications.

Table 1: Notations

Terms	Notation	Units
Minimum cycle time	Γ	s
Delivery time (from frame i)	d_i	s
Transmission delay	τ	s
Network device latency	ℓ	s
Propagation delay	δ	s
Link capacity	C	$bits/s$
Payload	x	$bytes$
Number of network devices (slaves)	n	–

2.2 EtherCAT

The EtherCAT network was developed by the Beckhoff company (type 12 in standard IEC 61158, [7, 8]). In theory, EtherCAT cards are standard Ethernet interfaces. In practice, specific hardware (FPGA, Field-Programmable Gate Array, or ASIC, Application-Specific Integrated Circuit) is used to mitigate the frame forwarding delay. The EtherCAT network adds a master/slave protocol over the Ethernet. A frame is sent by the master and slaves can read and write data *on the fly*. The duration of reading or writing operations corresponds only to the network device latency ℓ, which is independent of frame size and the same for all slaves. A logical ring is defined between the slaves such that when a frame reaches the last slave in the ring it is returned to the master via all the slaves. The space–time diagram shown in Figure 2 illustrates the behaviour of EtherCAT communications.

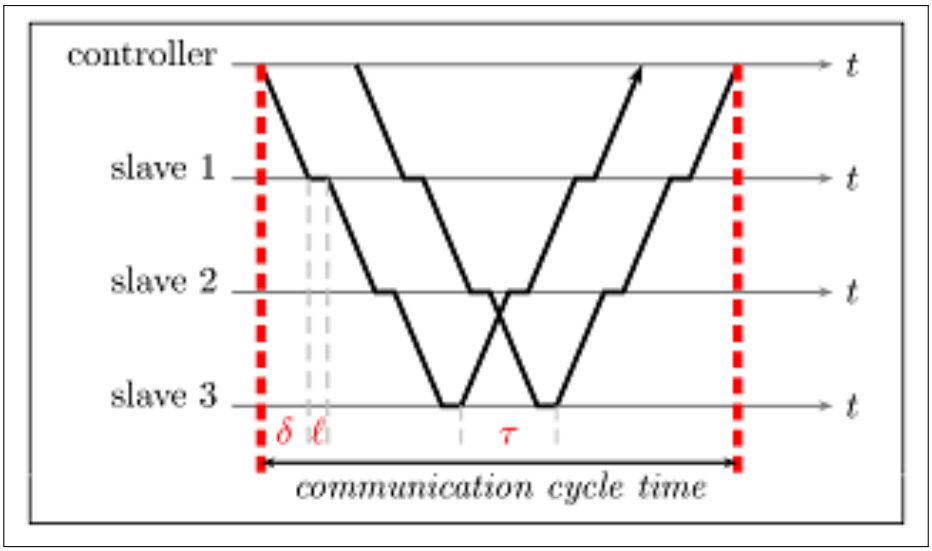


Figure 2: EtherCAT space-time diagram

The EtherCAT protocol can support both line and ring topologies. Because the line topology is used mainly in the industrial framework, this topology was investigated in our study (Figure 3).

The EtherCAT datagram is directly encapsulated inside the basic Ethernet frame as shown in Figure 4. An EtherCAT frame is composed of a header specifying the length of the frame and a list of datagrams. The number of datagrams depends on the number of slaves. A datagram is

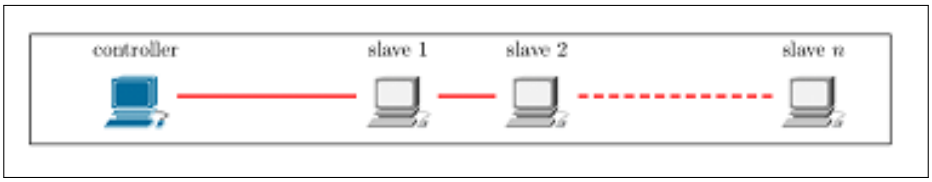


Figure 3: EtherCAT line topology

defined for each slave, and it contains the command type and associated data.

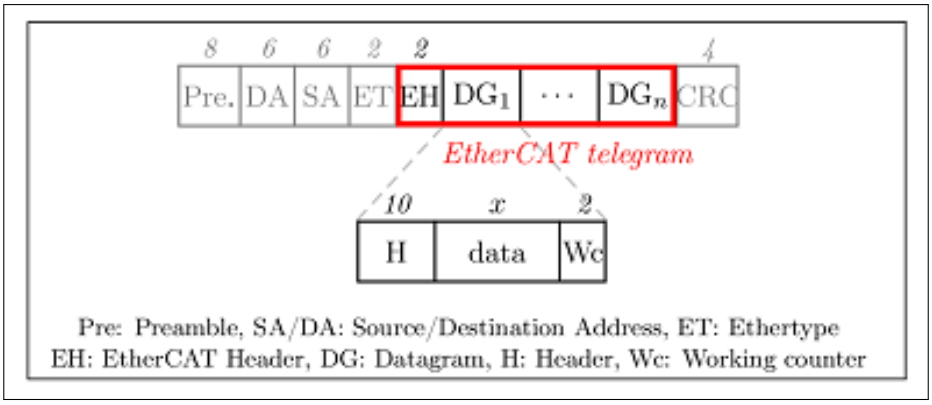


Figure 4: EtherCAT frame *(field lengths are given in bytes)*

In this study, our analysis of EtherCAT performance considered the following hypothetical test scenario: - the topology was a line, - the initialization step was ignored and only cyclic communication was studied, - the master sent only one frame per cycle, and - the payload x was the same for each slave.

The minimum cycle time shown in Figure 2 was determined for this scenario.

The link transmission delay is the ratio between frame size and link capacity C. The total frame size can be divided into two parts: - a constant value that equals the sum of the Ethernet protocol (26 *bytes*), the interframe gap (corresponding to the time of 12 *bytes*), and the EtherCAT header (2 *bytes*) and - a variable value that depends on the slave number n, the amount of data to transport x, and the header (12 *bytes*).

The link transmission delay is:

$$\tau = \frac{8\,(40 + \max\,(44, n\,(12 + x)))}{C}. \tag{1}$$

The term 44 in the equation (1) was added to ensure the minimum data size defined by the Ethernet protocol. If the EtherCAT telegram length was less than 46 *bytes*, an equivalent amount of padding was inserted in the Ethernet frame. The EtherCAT telegram already included a 2 *bytes* header, which meant that there was no padding requirement when the length of the datagram sequence was larger than 44 *bytes*.

As shown in Figure 2, the cycle time was estimated using the expression:

$$\begin{aligned} \Gamma &= (2n-1)\,\ell + 2n\delta + \tau \\ &= (2n-1)\,\ell + 2n\delta + \frac{8\,(40 + \max\,(44, n\,(12 + x)))}{C}. \end{aligned} \tag{2}$$

It should be noted that (2) considers only one frame. Because the Ethernet payload size depends directly on the number of slaves and the Ethernet frame size cannot exceed 1 526 *bytes*

(and therefore the data field 1 500 *bytes*), (2) is only valid if the number of devices interconnected to the network is less than:

$$n \leq n_{max} = \left\lfloor \frac{1500 - EH}{12 + x} \right\rfloor,$$

where EH is the EtherCAT header size (2 *bytes*) and n_{max} is the maximum number of datagrams, of length x, which can be included in a single frame (in the following, we have assumed that $x \leq 1486$ *bytes*).

In general, the number of slave devices on the network can be greater than the frame size capacity. This means that the controller has to send more than one frame in a cycle time. In fact, the number of Ethernet frames required to support n devices with a constant payload x is given by:

$$k = \left\lceil \frac{n}{n_{max}} \right\rceil.$$

Consequently, (2) now integrates a different transmission time for each frame and finally gives:

$$\Gamma = (2n - 1)\,\ell + 2n\delta + \frac{8}{C}\Big(40k + (k-1)\,n_{max}\,(12 + x)\Big) + \frac{8}{C}\max\Big(44, (n - (k-1)\,n_{max})\,(12 + x)\Big) \tag{3}$$

The final term of expression (3) was used to differentiate cases where the last frame generated padding.

Similar results are given in [5] but, in contrast to this earlier work, expression (3):

- takes into account the *on the fly* minimum cycle time mechanism proposed by EtherCAT; the main advantage of this is that a device can begin frame forwarding before complete reception of a frame (in contrast to, e.g., store-and-forward mode), which significantly reduces the forwarding time as shown in Figure 2,
- considers the use of padding, as defined by Ethernet,
- integrates the time required to forward the information sent from devices to the controller, as shown in Figure 2, and
- considers cases where the number of slaves and their payload requires the utilization of several frames.

The accuracy of the EtherCAT synchronization mechanism was reported in [9], which shows that this issue need not be considered because it was estimated as equal to a few nanoseconds.

2.3 Profinet IRT

Introduction

The Profinet protocol was developed by the Siemens company (type 10 in standard IEC 61158, [7, 8]). Profinet IRT manages real-time communications. However, standard Ethernet cards cannot be used because Profinet IRT requires the operation of specific hardware on slaves (ASIC type – 2 or 4 ports inbuilt switch). Profinet IRT is based on the time-slice mechanism, which specifies two modes, the asynchronous mode and the isochronous mode.

The asynchronous and isochronous modes are used for unconstrained traffic and real-time traffic, respectively. Our study dealt only with real-time traffic. Thus, only the isochronous mode was analysed. For more information on these two modes, the reader can refer to [10]. In

the isochronous mode, the master sends one data frame to each device and each device replies to the master. Profinet can support line, star, and ring topologies. Only the line topology (see Figure 5), with full-duplex links, was analysed in this study (as for EtherCAT).

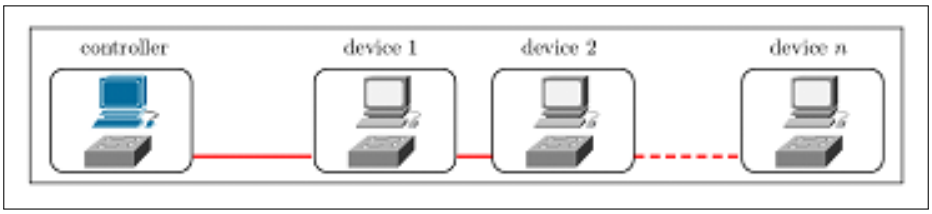


Figure 5: Profinet line topology

In this study, the Profinet IRT used the *slipstreaming effect*, where the controller began by sending frames to the most remote slave in the line topology, and then to the second remote slave slave and so on, until it reached the nearest slave. This mechanism enabled a reduction in the cycle time by minimizing the transmission time.

The *slipstreaming effect* was also applied to exchanges from slaves to the controller. The links were set up in full-duplex mode. The global communication scheme is given in Figure 6.

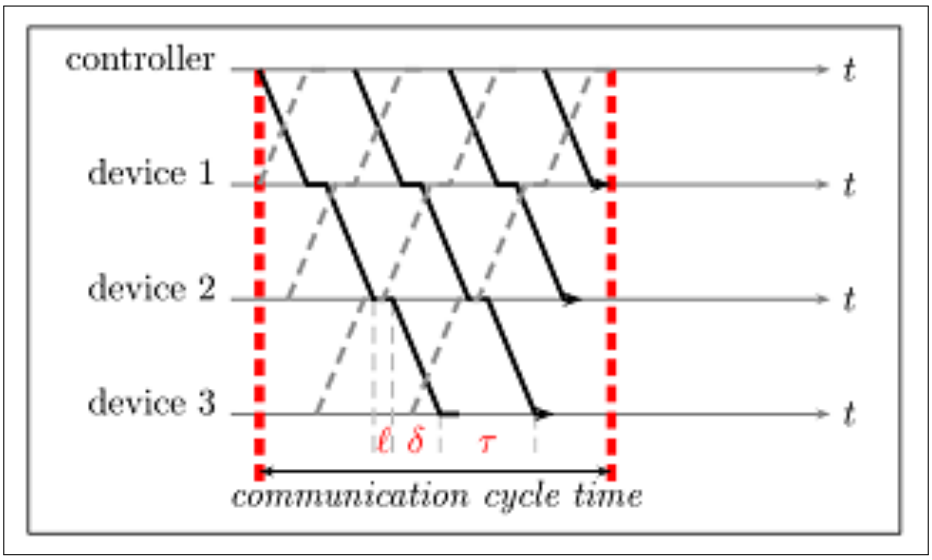


Figure 6: Profinet IRT space-time diagram with *slipstreaming effect*

The space–time diagram shown in Figure 6 illustrates the optimum use of the Profinet IRT protocol. Devices were assumed to be synchronized using a clock synchronization protocol, such as the IEEE 1588 standard. The IEEE 1588 standard generates synchronization frames (PTP) between devices, but these frames were not relevant to this study.

Minimum cycle time estimation

The slave devices periodically sent their messages to the controller at the same times as messages were sent by the controller to slaves.

Only the positive characteristics of the *slipstream effect* were considered. This required that $\tau \geq \delta + \ell$, as noted by [5].

In such cases, the minimum cycle time was given by [5,6] as the sum of:

- the latency ℓ that crosses all devices plus the propagation delay δ for each link and

- and the link transmission time τ for each frame sent by the controller.

Hence, the minimum cycle time is written as:

$$\Gamma = \delta + \ell + n\tau. \tag{4}$$

Equation (4) was developed by analysing the transmission delay τ. The Profinet datagram is encapsulated in the Ethernet frame as shown in Figure 7.

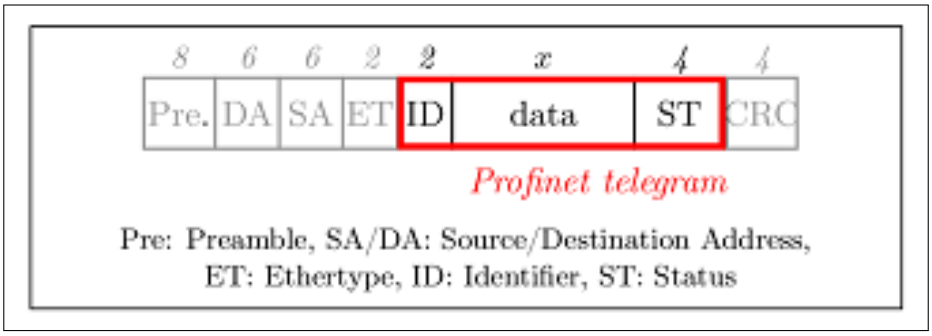


Figure 7: Profinet IRT frame *(field lengths are given in bytes)*

Three fields are added to the Ethernet frame: a data identifier (2 *bytes*), the data value (with the assumption that $x \leq 1494$ *bytes*), and an information status (4 *bytes*). When considering the constraint of the minimal Ethernet frame size, the transmission delay of a Profinet frame was given by:

$$\tau = 8\frac{38 + \max(46, 6 + x)}{C}. \tag{5}$$

The term 38 in the equation (5) corresponds to the size of the Ethernet layer (26 *bytes*) plus the interframe gap (12 *bytes*). As with EtherCAT, the term 46 was added in order to ensure the minimum data size as defined by the Ethernet protocol. If the EtherCAT telegram length was less than 46 *bytes*, an equivalent amount of padding was inserted in the Ethernet frame.

The final form of equation (4) was expressed as:

$$\Gamma = \delta + \ell + n\frac{8}{C}\Big(38 + \max(46, 6 + x)\Big)$$

Comments

The EtherCAT minimum cycle (3) was less than the Profinet IRT one (4), because the Profinet IRT transmission delay was multiplied by the number of devices. The Profinet IRT analysis assumed that all devices had the same clock reference. Figure 6 shows that all devices were synchronized because of the IEEE 1588 protocol. They shared the same clock and periodically sent their messages at the same time. Hence, it may be expected that clock synchronization errors will increase the minimum cycle time. This paper aims only at comparing optimal performances of COTS Ethernet-based protocols, i.e. Profinet IRT nodes sharing the same clock reference. Next studies will hence aim at reporting this synchronization issue.

2.4 Modbus/TCP

Modbus is a serial communication protocol developed by Modicon in 1979. Modbus/TCP is a variant of the Modbus protocol (type 15 in standard IEC 61158, [7,8]), which uses the Ethernet physical and link layers [12]. Modbus/TCP encapsulates a Modbus frame into a TCP frame as

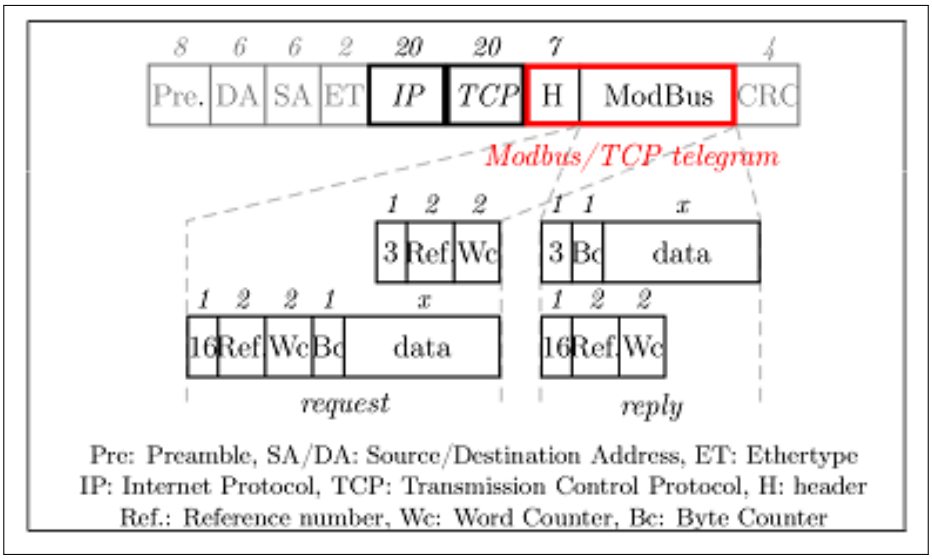


Figure 8: General Modbus/TCP frame and function specific Modbus application frame *(field lengths are given in bytes)*

shown in Figure 8. The Modbus datagram is composed of the ModBus Application Protocol Header (MBAP), the function (read/write), and the data value.

Modbus/TCP is a pragmatic approach that works on several types of configurations that impact on its performance. Parameters, including topology, exchange management in the application layer, and the processing capacity of devices, change the Modbus/TCP time behaviour. Modbus/TCP is based on connection-oriented transactions and it can use different exchange models, such as master/slaves, producer/consumer, or client/server. The objective of this study was to compare several Ethernet products, so it was necessary to use similar contexts for all Ethernet products. Because EtherCat and Profinet IRT were previously analysed using a master/slaves model, this model was also used for Modbus/TCP evaluation. The master/slaves model also simplified the analysis because all the devices were synchronized by network events (none clock synchronization protocol is required).

The communication scheme defined in this study followed these steps. The master sent a frame to one slave and when the slave received this frame, it sent a reply to the master. When the master received the reply, it repeated the same procedure with another slave. All the slaves were processed by the master using a round-robin method. This communication scheme was implemented in the application layer of the OSI model. The request/reply protocol is shown in Figure 9.

Modbus/TCP not only supports application data but also TCP PDU, and it comprises opening and closing TCP connections when acknowledging segment reception. The acknowledgment can be achieved either immediately a segment is received, after the reception of several segments, or inside the next data transmission (piggybacking). In practice, TCP behaviour changes according to the operating system, the TCP configuration, and whether or not the Nagle algorithm is used. In this study, we assumed that the timeout to send the acknowledgement was 0.5 s, meaning that only piggybacking was analysed, as shown in Figure 9. The transient states of the TCP opening and closing steps were not considered in the modelling.

Modbus/TCP supports line, star, and ring topologies. Complex architectures based on switches can be used, especially when the network is shared by several applications, although switches induce additional costs. However, because only the master/slaves communication

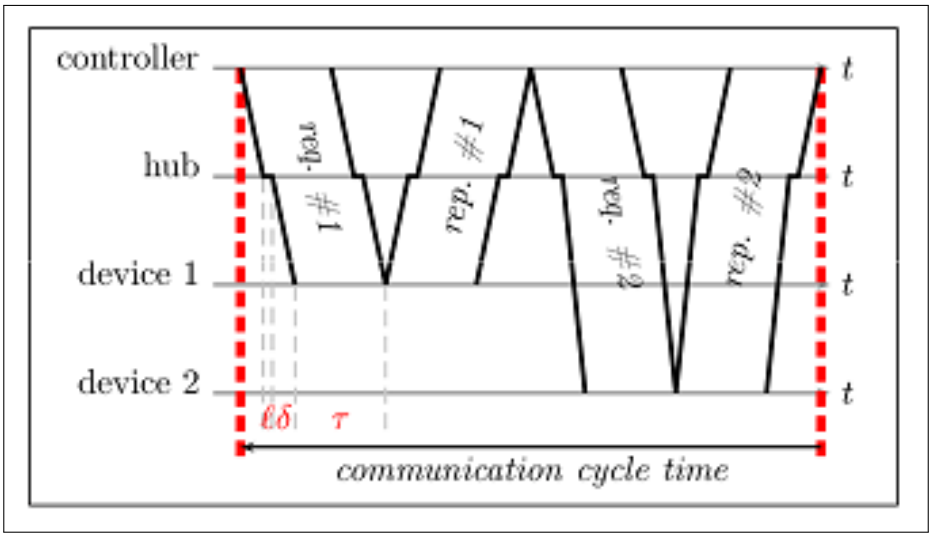


Figure 9: Modbus/TCP space-time diagram

scheme was considered in this study, a bus infrastructure was used.

It was assumed that each device was interconnected through a common hub, and the propagation delay was equal to δ. Figure 9 shows that the cycle time was equal to the number of devices n multiplied by the time required to poll one device.

Given previous assumptions, this can be written as:

$$\Gamma = n\left(\tau_{req} + \tau_{rep} + 2\left(2\delta + \ell\right)\right). \tag{6}$$

In the worst case, where a TCP acknowledgment is sent for each segment, the minimal cycle time would be equal to:$Gamma = n\left(\tau_{req} + \tau_{rep} + 3 \times \left(2\delta + \ell\right) + 2\tau_{ack}\right)$ where $\tau_{ack} = 672/C$ and 672 is the minimal size in bits for an Ethernet frame (interframe gap included). When the architecture was composed of several hubs, the propagation delay increased and (6) was slightly different.

The frame transmission time was composed of:

- a constant part related to the sum of the Ethernet protocol (38 *bytes* with the interframe gap), the IP header (20 *bytes* without options), the TCP header (20 *bytes* without options), and the Modbus/TCP header (7 *bytes* for the ModBus Application Protocol Header),
- a variable part related to the type of Modbus message, with the size changing according to the type of data (function code) and the transaction model state (request or reply).

  Consequently, the analysis of Modbus/TCP was only valid for an application framework. In this study, we considered only write requests.
- a variable part proportional to the payload (because the byte count field was stored as a single byte, this indicates that the date field length was limited to $x \leq 255$ *bytes*).

The sum of IP, TCP, and Modbus header sizes was larger than the minimal Ethernet data length. Thus, no added padding was required. The delay is directly given by:

$$\tau_{req_{10h}} + \tau_{rep_{10h}} = 8\frac{91 + x}{C} + 8\frac{90}{C}$$

such that (6) corresponds to:

$$\Gamma_{10h} = n\left(8\frac{181 + x}{C} + 2\left(2\delta + \ell\right)\right). \tag{7}$$

2.5 EtherNet/IP

Introduction

EtherNet/IP (IP, Industrial Protocol) is a network developed by Rockwell Automation in 2001 and supported by ODVA (Open DeviceNet Vendor Association) [13, 14]. EtherNet/IP (type 2 in standard IEC 61158, [7, 8]) uses the Common Industrial Protocol (CIP) for off-the-shelf Ethernet products and TCP-UDP/IP stack. Ethernet/IP is a connection-based network. A CIP connection defines the type of packet sent to the network. There are two types of connections: the Explicit Messaging connection and the I/O (or Implicit) connection. Explicit Messaging provides generic and multi-purpose communication paths between two nodes, whereas I/O messaging is specific to application I/O data and provides serial purpose communication paths. When the application is time-constrained, I/O Messaging is the preferred mode because it employs UDP rather than TCP sockets. CIP uses the producer/consumer model and requires broadcast exchanges encapsulated in UDP. This study evaluated only I/O connections.

Because EtherNet/IP relies on COTS, no particular topology was specified. Bus- or switch-based architectures are both possible. Switches are interesting because they break the collision domain, allowing the support of VLAN and the classification of service mechanisms [15]. Switched architectures are recommended for exchange management with time-critical (implicit) messaging. However, several switched Ethernet architectures are possible. In this study, a linear switched topology was selected, as shown in Figure 10. This can be viewed as an extension of the experimental set-up considered in [16], where the number of switches varied according to the number of ports per switch.

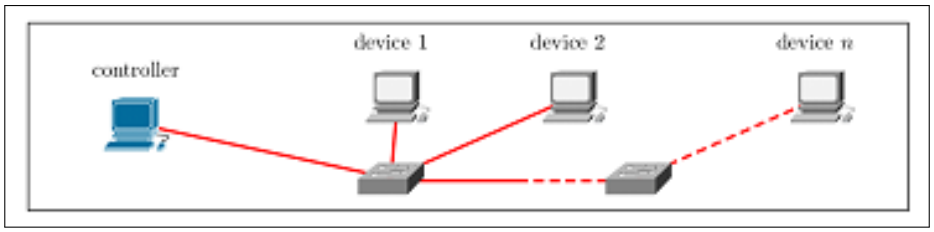


Figure 10: EtherNet/IP switched linear topology

In contrast to Modbus/TCP, utilization of a *slipstreaming effect* on Ethernet/IP does not facilitate the use of a bus topology. Indeed, it is possible that two messages can be on the network at the same time (as shown in Figure 11).

In order to compare the other protocols with Ethernet/IP, a similar exchange scenario was proposed. The controller sent a frame to each slave device and each slave device produced data that were sent to the controller. Figure 11 shows this behaviour.

Initially, EtherNet/IP did not support medium access synchronization points, as found in the master/slaves technique used in Modbus/TCP. This meant that any device could access the network at any time. However, the time synchronization mechanism can be used because of the support of EtherNet/IP by the IEEE 1588 [17] protocol (implemented by the CIPSync profile). Consequently, devices can send messages using the same clock reference. We used the *slipstreaming effect* for the controller and a common departure time for the devices, as shown in Figure 11. Obviously, this profile corresponded to an ideal case related to the minimal cycle time for this architecture, because lock synchronization errors would lead to another profile inducing a longer cycle time and the use of the *slipstreaming effect* requires (see Profinet IRT) that $\tau \geq \delta + \ell$.

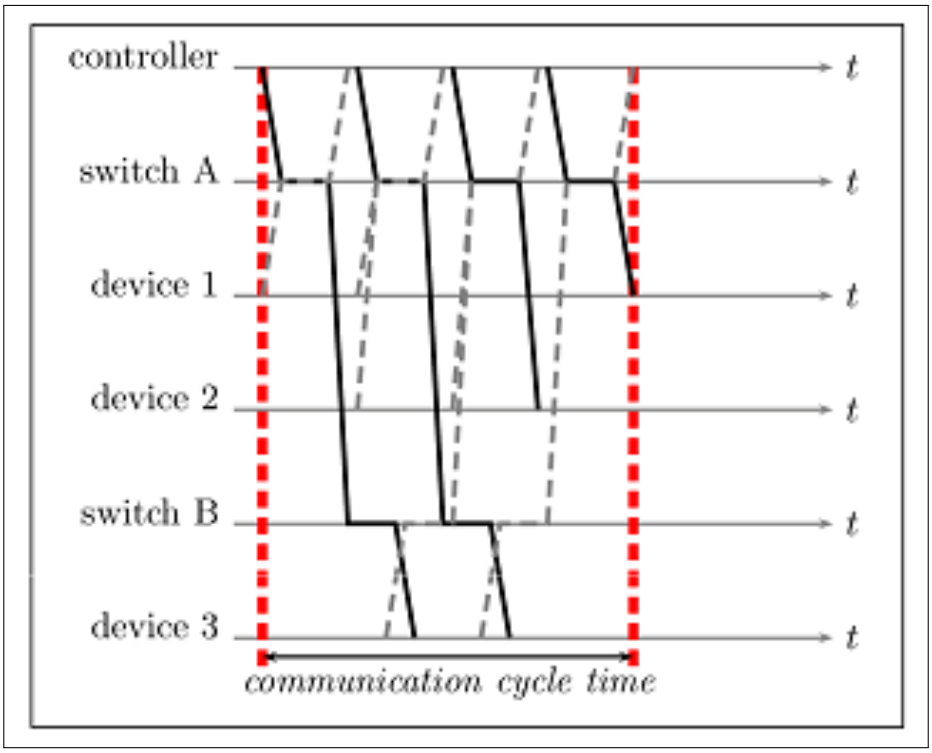


Figure 11: EtherNet/IP space-time diagram for 4-ports switches

Minimum cycle time

Considering the profile given in Figure 11, the minimum cycle time corresponded to the sum of: - the latency ℓ for crossing only one switch plus twice the propagation delay δ, between a controller/device and a switch and - the link transmission time τ for each frame sent by the controller.

Thus, the minimum cycle time can be written as:

$$\Gamma = 2\delta + \ell + n\tau. \tag{8}$$

The term given in (8) was valid only if $\tau \geq \delta + \ell$, meaning that switches did not work in the store-and-forward mode. The link transmission time was computed according to the encapsulation format of I/O messages.

Figure 12 shows that 18 *bytes* were added by the CIP protocol, 8 *bytes* by UDP, 20 *bytes* by IP, and finally 38 *bytes* by Ethernet. In this case, there was no padding. Thus, under the assumption that $x \leq 1454$ *bytes*, the link transmission time was given by:

$$\tau = 8\frac{38 + 20 + 8 + 18 + x}{C}.$$

(8) can be modified as:

$$\Gamma = 2\delta + \ell + 8n\frac{84 + x}{C}. \tag{9}$$

The profile given in Figure 11 can be optimized if several "items" (data) are encapsulated inside one frame, as specified in the CIP protocol. It induces a mitigation of the minimum cycle time.

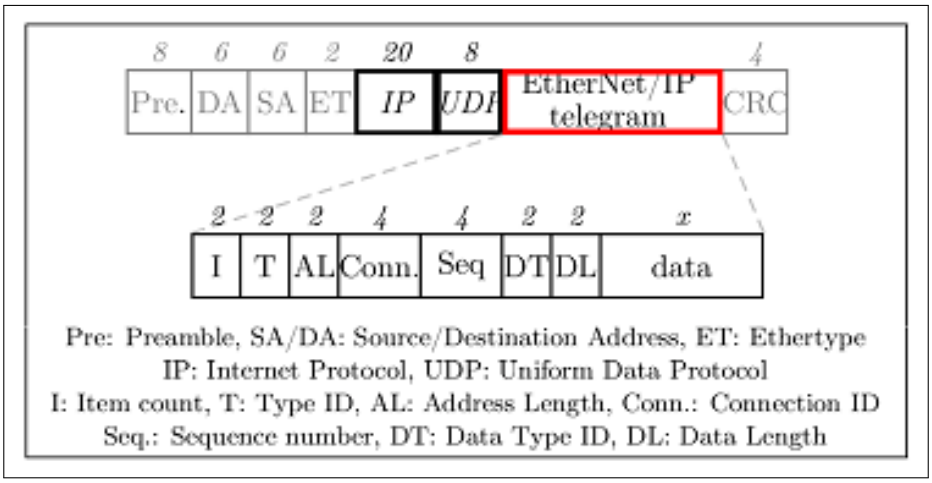


Figure 12: EtherNet/IP I/O messaging frame *(field lengths are given in bytes)*

Since we employed the same approach that for Profinet IRT (in particular the *slipstreaming effect*), next studies will hence aim at reporting the synchronization effect on the minimum cycle time.

In the next section, the results obtained for EtherCAT, Profinet IRT, Modbus/TCP, and EtherNet/IP are compared in different application contexts.

3 Comparisons

The objective was to analyse and compare the behaviour of different Ethernet-based solutions. The topologies used in each assessment were linear. Two bandwidths were studied: 100 Mb/s and 1 Gb/s. We analysed two payload sizes introduced by [6] as being representative of an industrial context: 16 *bytes* and 100 *bytes*. The network device latencies ℓ were defined in [5, 6] and those used in our study are described in Table 2. It was assumed the performances of switches used in EtherNet/IP were the same as those used by Profinet IRT. Clearly, alternative assumed values would yield different results. Thus, the results for the study using EtherNet/IP are only valid when $\tau \geq \delta + \ell$. The link propagation delay was 50 ns and corresponded to a distance of 10 m between two devices. The comparison is related to the minimum cycle time without synchronization errors.

Table 2: Latencies

Protocol	*FastEthernet* (100 *Mb/s*)	*GigaEthernet* (1 *Gb/s*)
EtherCat	1.35 μs	0.85 μs
Profinet IRT	3 μs	0.6 μs
Modbus/TCP	1 μs (*hub*)	
EtherNet/IP	3 μs	0.6 μs

Figure 13 shows the minimum cycle times (in ms), according to the number of slave devices with a payload equal to 100 *bytes* The first observation was that Modbus/TCP provided the worst results, whatever the payload. This was because of its medium access mechanism, which is based on polling at the application level.

In the case of a small payload, EtherCat provided the best results in the FastEthernet mode. The impact of bandwidth on Profinet IRT and Ethernet/IP was very significant, because it

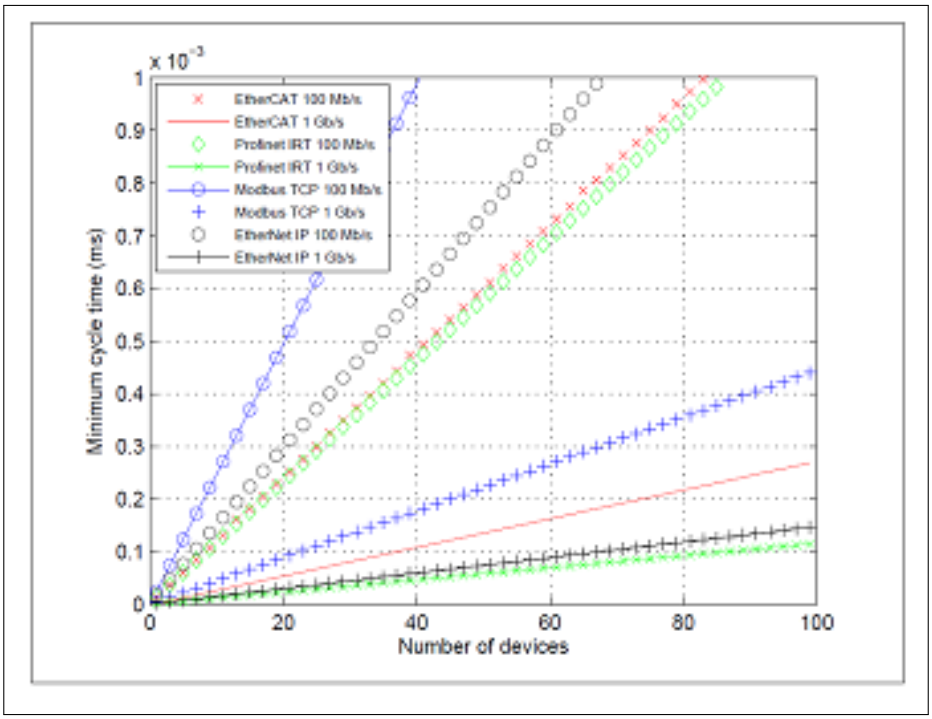


Figure 13: Minimum cycle time as a function of the number of devices with a constant payload of 100 bytes per device and two bit rates of 100 Mb/s and 1 Gb/s

enabled a reduction in the minimum cycle time. This benefit was lower with EtherCat. The explanation is simple: EtherCat sends only one frame for communicating with all its slaves and so EtherCat provides only one link per transmission. In contrast, Profinet IRT and Ethernet/IP sent n frames to dialogue with n slaves, which provides n times the link transmission. The bandwidth was the most crucial parameter in reducing the time cycle. Moreover, $2n$ latencies must be considered with EtherCat because Profinet and EtherNet/IP cycle times were composed of only one switch latency ℓ, because of the *slipstreaming effect*. Profinet IRT provided the best results at 1 Gb/s.

Figure 13 shows that all the Ethernet solutions, with the exception of FastEthernet Modbus/TCP, could interconnect with more than 60 slaves in cycle times of less than 1 ms. This cycle time constraint, and the number of slaves and the frame size, enables adequate coverage for most industrial applications. Performances were similar when the payload was increased. The only difference was that the EtherCat minimal cycle time was larger than Profinet when the number of devices was increased. This was because the EtherCat telegram had to be fragmented, which decreased protocol performance. An increase in frame size means that the choice of the FastEthernet solutions must be carefully considered because the cycle time grows quickly. The GigaEthernet solution eliminated this problem.

4 Conclusion

This study analysed the time performance of Industrial Ethernet protocols. The general conclusion was that all Ethernet protocols perform suitably with a bandwidth of 1 Gb/s for interconnected real-time systems. At 100 Mb/s, special attention is required by engineers in the selection of Ethernet-based protocols. Other considerations have to be taken into account in the selection of Ethernet-based products, including persistence in the market and interoperability

with other industrial equipment and network. Regarding interoperability concerns, Ethernet/IP is a standardized solution and provides acceptable performance for controlling industrial systems. Ethernet/IP can also implement priority mechanisms (IEEE 802.1p), which is important when a network is shared with other applications because this facility enables the differentiation of network services offered by real-time traffic and unconstrained traffic.

Moreover, multicast communications may be interesting in industrial applications. It is hence important to know what is the capacity of the analysed solutions to enhance nodes operation using multicasting techniques. As Ethernet/IP relies on COTS, multicast addresses and VLAN techniques may be used to create multicast communications. Because of its transport protocol (TCP), Modbus/TCP does not seem to be suitable to multicast communications (due to multiple acknowledgement issue). Profinet IRT considers only multicast communications in asynchronous mode. Finally, EtherCAT supports VLAN switches in its topology. As a conclusion, all solutions excepted Modbus/TCP make it possible to support multicast communications.

Future works are related to the clock synchronization issue, especially for solutions based on the *slipstreaming effect*. The objective is to deal with non optimal situations for Profinet IRT and Ethernet/IP where synchronization errors appear.

Bibliography

[1] M. Felser, T. Sauter, The fieldbus war: History or short break between battles?, *4th IEEE International Workshop on Factory Communication Systems*, pp.73-80, 2002.

[2] M. Alves, E. Tovar, and F. Vasques, Ethernet goes real-time: a survey on research and technological developments, *Technical Report HURRAY-TR-2K01, IPP-HURRAY, Polytechnic Institute of Porto (ISEP-IPP)*, 2000.

[3] J.-D. Decotignie, A perspective on ethernet-tcp/ip as a fieldbus, *4th IFAC International Conference on Fieldbus Systems and their Applications*, pp.138-143, 2001.

[4] J.-D. Decotignie, Ethernet-based real-time and industrial communications, *Proceedings of the IEEE*, 93(6):1102-1117, 2005.

[5] J. Jasperneite, M. Schumacher, K. Weber, Limits of increasing the performance of industrial ethernet protocols, *12th IEEE Conference on Emerging Technologies and Factory Automation*, pp.17-24, 2007.

[6] G. Prytz, A performance analysis of ethercat and profinet irt, *13th IEEE Conference on Emerging Technologies and Factory Automation*, pp.408-415, 2008.

[7] IEC, Digital data communications for measurement and control ? fieldbus for use in industrial control systems: Part 3: Data link service definition, *IEC Standard 61158, Part 3*, 2006.

[8] IEC, Digital data communications for measurement and control ? fieldbus for use in industrial control systems: Part 4: Data link protocol specification, *IEC Standard 61158, Part 4*, 2006.

[9] G. Cena, I C. Berlotti, S. Scanzio, On the accuracy of the distributed clock mechanism in ethercat, *4th IEEE International Workshop on Factory Communication Systems*, pp.43-52, 2010.

[10] J. Jasperneite , E. Elsayed, Investigations on a distributed time-triggered ethernet realtime protocol used by profinet, *3rd International Workshop on Real-Time Networks*, 2004.

[11] The Modbus Organization. Modbus.

[12] Modbus-IDA, Modbus application protocol specification v1.1b, 2006.

[13] P. Brooks, Ethernet/IP-industrial protocol, *8th IEEE International Conference on Emerging Technologies and Factory Automation*, Vol.2, pp.505-514, 2001.

[14] V. Schiffer, The common industrial protocol (CIP) and the family of CIP networks, *Technical Report PUB00123R0, Open DeviceNet Vendor Association, Inc. (ODVA)*, 2006.

[15] A. Modlovansky, Utilization of modern switching technology in ethernet/IP networks, *1st International Workshop on Real-Time LANs in the Internet Age*, pp.35-37, 2002.

[16] E. Alessandria, L. Seno, S. Vitturi, Performance analysis of ethernet/ip networks, *7th IFAC International Conference on Fieldbuses and Networks in Industrial and Embedded Systems*, Vol.7, 2007.

[17] IEEE Computer Society, IEEE standard for a precision clock synchronization protocol for networked measurement and control systems, *ANSI/IEEE standard 1588-2002*, 2002.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 759-766

Transmission Control for Future Internet including Error-prone Wireless Region

I. Ryoo, S. Kim

Intae Ryoo, Seokhoon Kim
Dept. of Computer Engineering & RU-IPTV Research Center
Kyunghee Univ., 1, Seocheon-dong, Giheung-gu,
Yongin-si, Gyeonggi-do, Korea
E-mail: itryoo@khu.ac.kr, kimsh@khu.ac.kr

Abstract:
This paper introduces a transmission control scheme which is aimed at enhancing overall transmission capability of internetworks including *error-prone wireless regions*. The proposed scheme can accurately adapt to wireless communication environments by integrating new approaches of *bandwidth estimation*, *loss detection*, and *error recovery* techniques while differentiating data losses due to physical bit error characteristics of *error-prone wireless regions* from those due to network congestion. From the simulations, it has been verified that the proposed scheme shows better throughput performances than the existing major transmission control schemes such as New Reno, Vegas, and Westwood+, while achieving satisfied levels of fairness and friendliness with them.
Keywords: transmission control, error-prone wireless region, bandwidth estimation, loss detection, error recovery

1 Introduction

Wireless and mobile communication environments have already been pervasively deployed, and never-ending communication demand in error-prone wireless regions of upcoming Future Internet makes it difficult to maintain the same level of performances of legacy transmission controls. This is mainly due to the inability of the existing transmission control schemes to differentiate data losses resulting from inevitable bit errors in error-prone wireless regions [1] from those resulting from network congestion. As bad channel conditions and aperiodic disconnections are typically transient phenomena, legacy TCP congestion control responses may be inappropriate and undesirable for the networks with error-prone wireless regions. In order to achieve a higher level of transmission efficiency, we have proposed a new transmission control method that is well suited for error-prone wireless regions as well as wired regions.

Many solutions have been proposed to improve TCP performances in wireless networks and can be categorized into End to End (E2E), Split Connection, and Link Layer solutions [2][3]. The proposed transmission control protocol for error-prone wireless regions (TCP EWR) falls into the E2E category as it measures packet arrival rates and acts according to current network status. It, however, differs from the general E2E solutions in the point that it is aiming to improve its performances in case of random or sporadic data losses by performing a time-stamp based bandwidth estimation at the receiving TCP rather than at the sending TCP, calculating the expected and actual bandwidths, and reacting accordingly to the random wireless packet loss events. It measures bandwidth utilization by using packet arrival information at the receiver rather than by using acknowledgement (ACK) information at the sender like TCP New Jersey [4]. As a result, sending feedback information to the receiver is not necessary in our scheme. Moreover, in order to make the TCP EWR operate in a stable status, we have introduced two thresholds, α and β, which corresponds to having too little and too much data enroute to the destination, respectively. In addition, the TCP EWR performs an enhanced error recovery in

case that the network suffers from timeout expiration due to wireless packet losses. That is, if there is any sign of data loss either by retransmission timeout or by three duplicate ACKs, the TCP EWR sets slow start threshold (*ssthresh*) value to optimal congestion window ($oCwnd$). With these features, the overall performance of the TCP EWR can be increased for the networks including error-prone wireless regions.

In Sect. 2, we discuss the bandwidth estimation, loss detection, enhanced error recovery, and congestion control procedures of the TCP EWR. We also give a pitch of our scheme through performance tests in Sect. 3, and finally bring this paper to a conclusion in Sect. 4.

2 Transmision Control for Future Internet Including Error-Prone Wireless Regions

The proposed transmission control scheme is composed of bandwidth estimation at the receiver, loss detection, enhanced error recovery, and congestion control. The first step of bandwidth estimation is to calculate a sample bandwidth ($Bw_{sample}(n)$) when the n^{th} packet arrives at the receiver at time t_n:

$$Bw_{sample}(n) = d_n/(t_n - t_{n-1}) \tag{1}$$

where d_n is the amount of data currently received by the receiver and t_{n-1} is the previous $(n-1)^{th}$ packet arrival time. That is, differently from the existing schemes such as Westwood+ and New Jersey [5], our scheme uses packet arrival information at the receiver instead of using returning ACK information at the sender. Note that Westwood+ and New Jersey are considered in this work as they are representative transmission control schemes that can distinguish wireless packet losses from congestion packet losses and react accordingly. By incorporating the actual arrival rate information in bandwidth estimation and, at the same time, by including congestion-related packet loss probability into the packet arrival rate calculation [6], the proposed scheme can probe the available bandwidth more accurately. The next step is to estimate the available bandwidth with a time varying coefficient, exponentially weighted moving average filter, which yields to following equation.

$$Bw_{estimated}(n) = Bw_{estimated}(n-1) \times \delta_n + Bw_{sample}(n) \times (1 - \delta_n) \tag{2}$$

where δ_n is a constant filter gain. In our simulation, we set δ_n to 0.7 for n$\geq$2 and zero for n=1 for simplicity. Thus, the optimal congestion window size $oCwnd_n$ is computed as following:

$$oCwnd_n = RTT_{min} \times Bw_{estimated}(n)/seg\_size \tag{3}$$

where RTT_{min} is minimum measured round-trip time (RTT) and seg_size is the length of the TCP segment. Note that we do not consider the available buffer space in the receiver in this work, and $oCwnd_n$ is used as a receiver-advertised congestion window size. Simulation results in Sect. 3 show that this yields good bandwidth estimates for random packet loss scenarios with error-prone wireless region.

In order to proactively detect the incipient stage of congestion and to efficiently figure out the reason for packet losses, the proposed scheme calculates maximum expected throughput and measures actual throughput as followings similar to Vega and Veno approaches [7]:

$$expected\ throughput = WindowSize/RTT_{min} \tag{4}$$

$$actual\ throughput = WindowSize/RTT \tag{5}$$

where RTT is the smoothed round-trip time measured. The difference D between these two throughputs indicates the amount of data that is currently residing on the corresponding path and/or going away due to network congestion or error-prone wireless regions. Also, by using the backlog variable $x = D * RTT_{min}$ and two thresholds, α and β, which correspond to having relatively a few and too many data on the route respectively, we can grasp how to manage the sender's congestion window ($cwnd$) size based on the current network situation. In our simulations, the values of α=1 and β=9 are used, but they can be set to better adapt to Future Internet environment, which is for further study. With these indicators for the current network status, the TCP EWR performs an enhanced error recovery (EER) mechanism. If there is any sign of data loss either by retransmission timeout (RTO) at the sender or by three duplicate $ACKs$ ($DUPACKs$) from the receiver, the proposed scheme sets the $ssthresh$ value to the $oCwnd_n$. Note that, if the connection is restored after physical bit errors, the window size should grow quickly to make up for the previous data loss event. In the proposed scheme, the sender immediately adjusts its congestion window size to the $oCwnd_n$. It is because the receiver has continuously monitored the available bandwidth of the corresponding path as given in Eq. (1) and (2). There is no need to employ any kinds of slow start, additive increase, or fast recovery algorithms whenever the connection is recovered.

With regard to congestion window updates when a packet loss is detected, there are three different cases depending on the values of backlog variable x, and two thresholds α and β. When $x<\alpha$, the proposed scheme updates the congestion window size $cwnd$ to $oCwnd_n$ or follows an additive-increase paradigm (linear increase state) by comparing the current $cwnd$ with a new $ssthresh$ value. The reason why the proposed scheme updates the $cwnd$ like this is that the network condition is not too bad as the backlog $x<\alpha$ implies, although there exits an incident that packets have been lost during transmission. When $\alpha \leq x \leq \beta$, it updates the congestion window size to $oCwnd_n$ or keeps the previous $cwnd$ value unchanged by comparing the current $cwnd$ with a new $ssthresh$ value. The backlog $\alpha \leq x \leq \beta$ implies that there is a high probability that packets will be lost due to either network congestion or wireless bit errors. So, it is desirable to retain the previous $cwnd$ (wait-and-see state) rather than following an additive-increase/adaptive-decrease paradigm. When $x>\beta$, it decreases the congestion window size adaptively. The backlog $x>\beta$ means that the network falls into a congestion state. By adopting this EER approach, we can use the available bandwidth efficiently. The pseudo code of the proposed TCP EWR transmission control algorithm is given below:

a) When packets successfully arrive at the receiver:
   Bw_{sample}(n), $Bw_{estimated}$(n), and $oCwnd_n$ are computed;
b) On ACK reception at the sender:
   $cwnd = oCwnd_n$;
c) When 3 $DUPACKs$ are received or RTO expires:

```
if (x<α)
  ssthresh = oCwnd_n;
  if (cwnd > ssthresh)
    cwnd = oCwnd_n;
  else                        /* linear increase state */
    cwnd=cwnd+2;
 else if (α ≤ x ≤ β)
  ssthresh=oCwnd_n;
  if (cwnd > ssthresh)
    cwnd = oCwnd_n;
  else                        /* wait-and-see state */
    cwnd is retained;
```

```
  end if
else                              /* congestion state */
  ssthresh = oCwnd_n;
  cwnd = 1;
```

Note that the *cwnd* is increased by two in the linear increase state because it is updated after one *RTO* rather than after each *RTT*.

3 Simulation Results

The proposed transmission control scheme has been simulated by using NS-2 and its performances are compared with several representative schemes such as New Reno, Vegas, Westwood+, and New Jersey. Performance metrics used in simulations include throughput, fairness, and friendliness. Although the TCP EWR is proposed for adapting to Future Internet including error-prone wireless regions, it must also operate well for any topology with lossy or congested links. Figure 1 shows a single bottleneck scenario for comparing performance metrics of the TCP EWR with those of well-known schemes under the condition that there exist various link error rates, and background traffics.

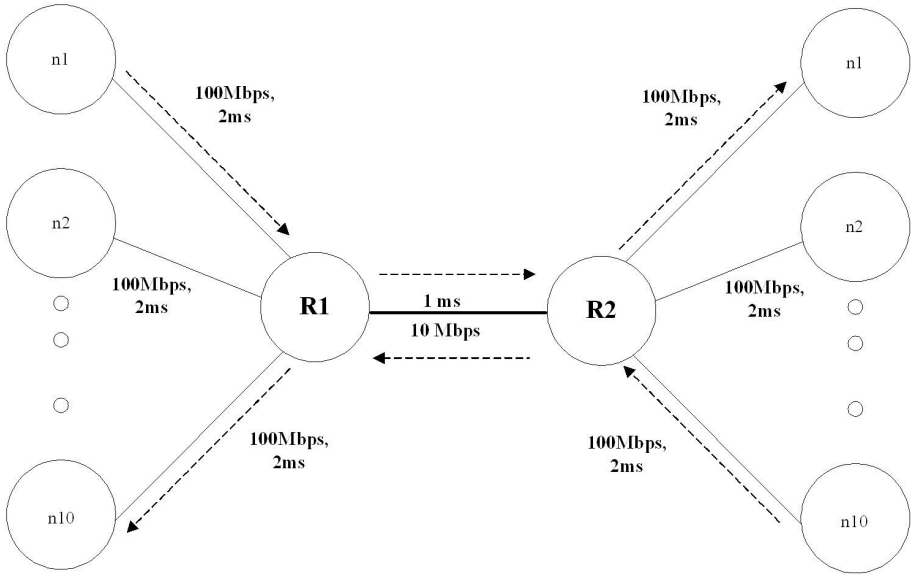


Figure 1: Single Bottleneck Scenario

Figure 2 shows throughput comparison results. The TCP EWR shows almost similar performances when link error rate is low. However, as the link error rate increases, especially from the point of 8 % link error rate, the TCP EWR outperforms other schemes. In addition, fairness index [8] has been compared by considering 10 same TCP flows that share a 10 Mbps bottleneck link. Different TCP schemes have been simulated individually and the corresponding results are summarized in Table 1. A perfect fairness of bandwidth allocation leads to the fair index of 1. It has been verified that, except 0 % link error rate, the TCP EWR's fairness index achieves a satisfactory margin compared to other schemes. The reason why the TCP EWR shows an inferior result under the very low link error condition is that it does not aggressively use the bandwidth but maintain the optimal congestion window size $oCwnd_n$.

Figure 3 shows a mixed wired and error-prone wireless scenario where there exist multiple

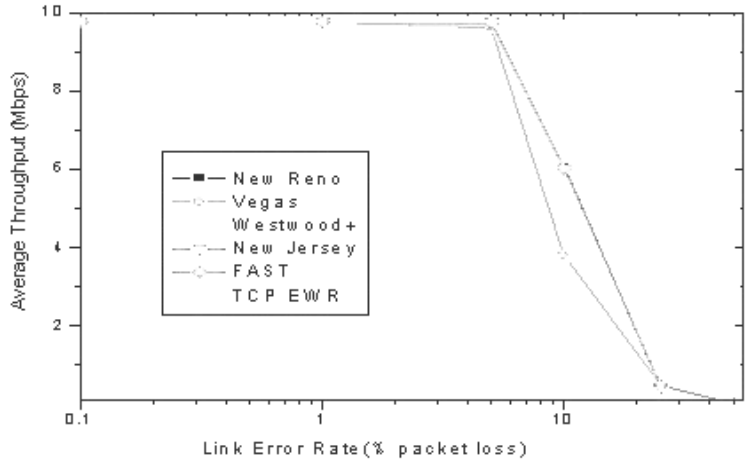


Figure 2: Throughput Comparison for Single Bottleneck Scenario

Table 1: Fairness Comparison for Single Bottleneck Scenario

Error rate(%)	New Reno	Westwood+	EWR
0	0.76	0.80	0.64
0.1	0.31	0.93	0.98
1	0.25	0.74	0.76
10	0.57	0.23	0.62

TCP connections. In this scenario, bidirectional FTP background traffics flow between N1 and N3 and between N2 and N4. The queue sizes of wired links and the wireless link are set to 100 and 10, respectively.

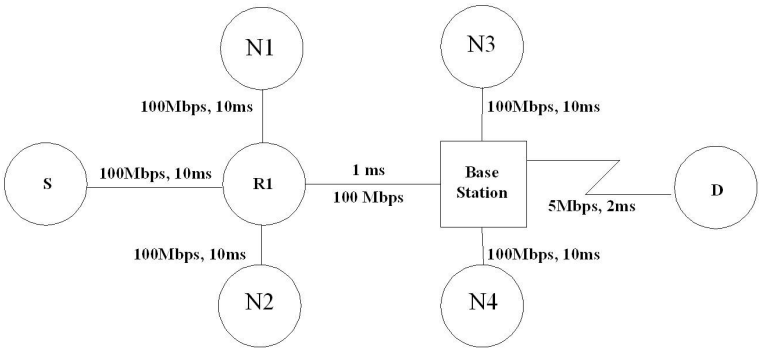


Figure 3: Mixed Wired/Error-prone Wireless Scenario

With these simulation conditions, we have compared the TCP EWR with New Reno, Vegas, and Westwood+. New Reno is considered as it is the leading Internet congestion control scheme. Vegas is considered as it also proposes, as Westwood+, a new mechanism to throttle the congestion window based on measuring the network congestion status. Westwood+ is considered as it remarkably improves utilization of wireless links that are affected by losses not due to congestion. Figure 4 shows the corresponding throughput comparison results. New Reno, Westwood+, and EWR show almost similar throughput results when the link error rate is less than 1 %. When

the link error rate is greater than 1% and less than 10 %, the EWR outperforms Westwood+. Vegas shows poor performance throughout the whole link error range. From the results shown in Figure 2 and Figure 4, we can conclude that the EWR shows better throughput performances than the other schemes for wireless scenario as well as for wired scenario. The reason why the EWR shows again similar result as New Reno and Westwood+ is that the link with more than 10 % error rate makes all the above control schemes fall into congestion state and cannot be logically managed any more.

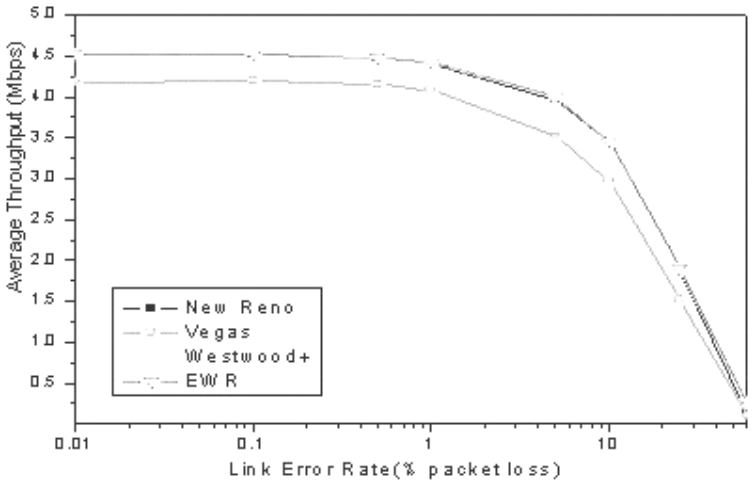


Figure 4: Throughtput Comparison for Mixed Wired/Error-prone Wireless Scenario

For the same topology shown in Figure 3, fairness indices of New Reno, Westwood+, and the EWR are compared. There are seven TCP flows of the same version that share a bottleneck link of which the bandwidth is 5 Mbps. The results are shown in Table 2. From these results, we can verify that the EWR also achieves a satisfactory level of fairness index as New Reno and Westwood+ do.

Table 2: Fairness Comparison for Mixed Wired/Error-prone Wireless Scenario

Error rate(%)	New Reno	Westwood+	EWR
0	0.95	0.94	0.96
0.1	0.94	0.94	0.89
1	0.95	1.00	0.95
10	0.96	0.95	0.96

Table 3: Friendliness Comparison of the EWR with New Reno (Throughputs in Kbps)

Number of Connections(New Reno)	Number of Connections(EWR)	Mean Throughput (New Reno)	Mean Throughput (EWR)
3	7	69.24	59.97
5	5	77.91	77.68
7	3	71.55	71.03

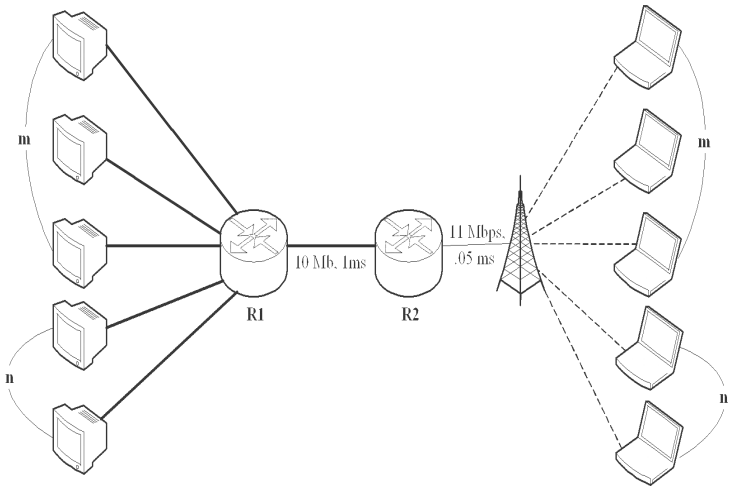


Figure 5: A Network Topology for Verifying Friendliness

Table 4: Friendliness Comparison of the EWR with WestWood+ (Throughputs in Kbps)

Number of Connections(New Reno)	Number of Connections(EWR)	Mean Throughput (New Reno)	Mean Throughput (EWR)
3	7	73.71	81.95
5	5	78.03	97.03
7	3	63.25	70.62

Finally, the friendliness of the proposed scheme with New Reno and Westwood+ has been tested by using a simple network topology shown in Figure 5. In simulations, there are 10 pairs of connections where m is the number of the hosts that communicate based on the proposed scheme and n is the number of hosts that use New Reno or Westwood+. Wireless link error rates have been set to vary between 0.1 and 10, and the corresponding results are summarized in Table 3 and Table 4. From the results, it has been shown that the proposed TCP EWR has good controllable friendliness compared with other two major TCP control schemes in error-prone wireless networking environment.

4 Conclusion

This paper introduces a new transmission control scheme for Future internetworking environment with error-prone wireless regions. The proposed scheme is designed to adjust its congestion window optimally based on the current network situation while estimating the available bandwidth at the receiver. As a result, accurate window update and transmission control can be possible to obtain high throughput while achieving satisfactory levels of fairness and friendliness, which are important indices for the proposed TCP EWR to be a feasible scheme.

Acknowledgement

This research was supported by Kyung Hee University, Korea (20090724).

Bibliography

[1] Sumit Rangwala, Apoorva Jindal, Ki-Young Jang, and Konstantinos Psounis, *Understanding Congestion Control in Multi-hop Wireless Mesh Networks*, Proceedings of the 14th ACM international conference on mobile computing and networking, San Franscisco, California, USA, pp. 291-302, 2008.

[2] Dhiman Barman, Ibrahim Matta, Eitan Altman, and Rachid El Azouzi, *TCP Optimization through FEC, ARQ and Transmission Power Tradeoffs*, Lecture Notes in Computer Science, vol. 2957, pp. 87-98, 2004.

[3] Rajashree Paul and Ljiljana Trajković, *Selective-TCP for Wired/ Wireless Networks*, in Proceedings SPECTS 2006, Calgary, AL, Canada, pp. 339-346, 2006.

[4] Kai Xu, Ye Tian, and Nirwan Ansari, *TCP-Jersey for Wireless IP Communications* ,IEEE Journal on Selected Areas in Communications, vol. 22, no. 4, pp.747-756, 2004.

[5] Luigi A. Grieco and Saverio Mascolo, *Performance Evaluation and Comparison of Westwood+, New Reno and Vegas TCP Congestion Control* , ACM SIGCOMM Computer Communications Review, vol. 34, no. 2, pp. 25-38, 2004.

[6] Jitendra Padhye, Victor Firoiu, Don Towsley, and Jim Kurose, *Modeling TCP Throughput: A Simple Model and its Empirical Validation* , ACM SIGCOMM Computer Communication Review, vol. 28, issue 4, pp. 303-314, 1998.

[7] Cheng Peng Fu and Soung C. Liew, *TCP Veno: TCP Enhancement for Transmission Over Wireless Access Networks* , IEEE Journal on Selected Areas in Communications, vol. 21, no. 2, pp. 216-228, 2003.

[8] Rajendra. K. Jain, Dah-Ming W. Chiu, and William R. Hawe, A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer System, DEC-TR-301, Eastern Research Lab., http://www.cs.wustl.edu/ jain/papers/ftp/fairness.pdf

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 767-775

Use of Reconfigurable IM Regions to Suppress Propagation and Polarization Dependent Losses in a MMI Switch

G. Singh, V. Janyani, P. Yadav

G. Singh, V. Janyani, V. Janyani
Department of Electronics and Communication Engineering
Malaviya National Institute of Technology Jaipur-India
E-mail: gschoudhary75@gmail.com, Tel.: +91-1412713431

Abstract:
With this work, use of reconfigurable index modulated (IM) regions to accelerate the performance of a multimode interference (MMI) based photonic switch is presented. Appropriate dimension for such regions are defined to suppress the transition losses and to optimize the area coverage. It has been noticed that by reconfiguring the IM regions, perfect switching for test wavelengths of 1.3μ m and 1.55μ m with low insertion loss (I.L.) levels, $\leq$ 1.2dB and excess loss (E.L.) levels, $\leq$ 0.17dB can be achieved with vacillation of extremely low polarization dependent losses (PDLs), which are $\leq$ 0.15dB. For either case of input test wavelengths, generated crosstalk (CT) levels are found better than -21.8dB for TE and -20.2dB for TM polarization state.
Keywords: Selfimages in MMI waveguides, Reconfigurable IM regions, I.L. and E.L., CT levels.

1 Introduction

Image replication in the MMI waveguides strongly depends upon the refractive index changing properties of the waveguide material. Various materials such as polymers, SiO_2-SiON, SOI, $LiNbO_3$, III-V group semiconductors and their composites have been used as a backplane to develop multimode waveguides to implement optical couplers and switches. [1–9] Such waveguides are suitable to realize many optical devices utilizing their self-imaging principle and possess various characteristics such as compactness, relaxed fabrication tolerance, large optical bandwidth, and polarization insensitivity. [8–10] Design of ultra-short couplers and switches based on MMI waveguides have been achieved by eliminating photonic confinement to change their behaviors or by alteration in the imaging length with suitable index variance. [5, 6] In recent, researchers has shown a great interest for modeling of polarization independent photonic switches for wide wavelength spectrum, low losses and CT levels using MMI waveguiding structures. This article deals with modeling of a 2×2 MMI-switch, which is tuned by integrated reconfigurable IM regions with its coupling area. The channel waveguides are optimized to reduce overall switch area coverage. The device has been evaluated for its satisfactory performance for operating test wavelengths of 1.3 ÂĺmandÂĺm. The contents of the article are as follows. A MMI-switch structure and its operation with an IM region are described in Section II. The detailed modeling and switching characteristics of the proposed structure are analyzed by finite difference beam propagation method (FD-BPM) in Section III, and then conclusions are summarized in Section IV.

2 MMI-switch based on self-imaging principle

The concept of self-imaging explains the replication of images of random objects in multimode waveguides. [10, 11] The switching in MMI waveguides based switches can be achieved by modifying the refractive index at specific areas within the MMI waveguide, which are collocated

with the occurrence of multiple self-images. [8] This change in the refractive index is the prime factor behind altering the phase relation between the self-images, which ultimately modifies the output image and switches the light between the output waveguides. MMI structures usually realize optical functions by tuning the refractive index of the image modulation region, located within the MMI section. Changes in the refractive indices of the segments (IM regions) inside the MMI section also bring variation in the effective width of the MMI region, thereby changes occurs in the beat length (L_π).

In configuration shown by Fig. 1, the confinement guide regions are shown, which allows the

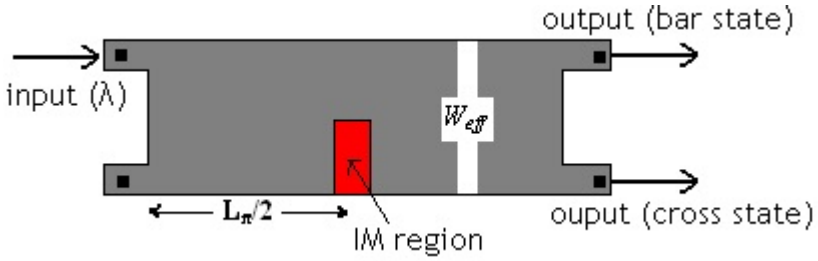


Figure 1: Layout of 2×2 MMI-switch with IM region, ref. [8]

light to pass through the region. Distribution of the optical intensities in the MMI waveguides occurs in accordance to the beat length, defined as follows. [12]

$$L_\pi = \frac{\pi}{\beta_0 - \beta_1} \approx \frac{4 n_{eff} W_{eff}^2}{3\lambda_0} \tag{1}$$

Where β_0 and β_1 are propagation constants of the fundamental and the first-order lateral modes, respectively, λ_0 is a free-space wavelength, n_{eff} is an effective index, and W_{eff} is an effective width of the MMI waveguide. The use of different interference phenomena in a device allows achieving different switching states, namely cross and bar. If the width of the MMI regions is reduced by depressing the refractive index by means of the electro-optic (EO) or thermo-optic (TO) effects, then the imaging locations will be changed. Though, EO-effects are highly effective to use in this type of MMI-switch in comparison with TO-effects., as the power consumption is bit higher which further depends upon the length of IM regions.

3 Modeling of 2×2 MMI-switch with reconfigurable IM regions

This section elaborates the design procedure for 2×2 MMI-switch, which has been tuned by reconfigurable IM regions, while geometries of its various parts are kept very small as compared to. [12–15] Index profile of the channel waveguides is chosen in the range of 2.10-2.20 in agreement with available data and the literature survey to investigate the switching characteristics. Various EO/TO crystals such as Ti-indiffused or proton exchanged lithium niobate (α : $LiNbO_3$), lithium tantalate ($LiTaO_3$), potassium niobate ($KNbO_3$), barium-sodium niobate (BNN), strontium-barium niobate (SBN), bismuth germinates crystals of evlitine structure ($Bi_4Ge_3O_{12}$), lithium iodate (α - $LiIO_3$) etc. can be used to implement such channel waveguides. Perfect dimensions for coupling and IM regions have been obtained using the well-known relationship for general interference in MMI waveguides. [6] Combined effect of geometry and refractive index contrast has been applied for modeling of symmetric waveguides to ensure polarization independent propagation.

The switch characteristics are investigated with a tunable laser to generate an input light beam (output level was fixed at 1mW and wavelengths were tuned in between 1.3 μm and 1.55

Table 1: Tuning strategies (Structure-1)

Region	Refractive index ($n_1 > n_2$)			
	1.55μm		1.3μm	
	cross	bar	cross	bar
Reg.1	n_1	n_2	n_1	n_1
Reg.2	n_1	n_1	n_1	n_2
Reg.3	n_2	n_2	n_1	n_1

μm). The various states of the switch are simulated and analyzed using the FD-BPM. [4] Initially we have evaluated the structure by considering less number of IM regions for its satisfactory operation over a bandwidth of 50nm for either of centre wavelength (1.3 μm and 1.55 μm). Fig.2 illustrates the layout of the proposed 2×2 MMI-switch labeled as structure-1, which represents its x-z slice cut on the wafer of the waveguide with sectional dimensions. The whole device is 500μm long and 8μm wide including IM regions, including 2 μm width for upper and lower straight waveguides. The wafer index assumed to be of same value as that of cladding i.e. 1.49 and the thickness of all the regions has been kept same, which is 15μm. In this structure, the

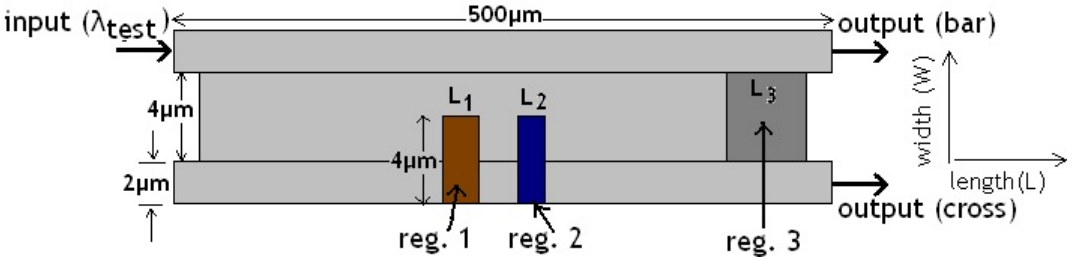


Figure 2: Layout (top view) of the proposed 2×2 MMI-switch with IM regions (structure-1)

index is modulated for three regions (1 to 3) in order to achieve it various switching states in accordance to applied test wavelength. The IM regions represent symmetrical planner multimode waveguides of different area coverage and refractive indices. The size of the index modulated regions are same in terms of their width (4 μm each) but differ in lengths, which are L_1 =12 μm, L_2 = 10 μm and L_3 =71 μm respectively. Switching in both structures has been achieved by varying the number and locations of reconfigurable IM regions. The refractive index, n_1 is kept fixed at 2.22, while n_2 represent reduced index values ($n_2 = n_1$-δn), centered at 2.136. The switch performance parameters (propagation losses, CT levels and PDLs) are calculated by altering the value of δn, which has been varied in the range of 0.05-0.07.

Table 1 depicts the respective index profile for various IM regions, which are used to tune the switch into its various states. By evaluating this structure it has been noticed that the PDLs observed are at high level for most inputs. In particular, the switch response becomes more sensitive to polarization state of the input for high order wavelengths, which causes image-shifting in the specified IM regions. Therefore few more IM regions are introduced, which can tune to get distinct but clear images for higher order wavelength inputs. Figure 3 illustrate the modified version of previous structure (1), which consists of two additional IM regions termed (4 and 5), each of 4ÂĽm wide with a little difference in their length, which are 16.36μm (L_4) and 16.48 μm (L_5) respectively. Due to this modification, the length of the region 3 is also required

Table 2: Tuning strategies for modified MMI-switch (structure 2)

Switch State	Bar state				Cross state			
Mode	TE		TM		TE		TM	
λ (μm)	1.55	1.3	1.55	1.3	1.55	1.3	1.55	1.3
Reg.1	n_2	n_1	n_2	n_1	n_1	n_1	n_1	n_1
Reg.2	n_1	n_2	n_1	n_2	n_1	n_1	n_1	n_1
Reg.3	n_2	n_1	n_2	n_1	n_2	n_1	n_2	n_1
Reg.4	n_1	n_1	n_2	n_1	n_1	n_1	n_2	n_1
Reg.5	n_2	n_1	n_2	n_2	n_2	n_1	n_2	n_2

to reduce by 16.8 μm (L_3 = 54.2 μm).
However both regions can be made of same length, but in that case, observed PDLs are at higher levels ($\geq$ 1dB). Theses extra IM Regions (4 and 5) are defined in order to accommodate TM polarized light images perfectly for either of input wavelength and to suppress the PDLs particularly. The tuning strategies for structure 2 are mentioned in table 2 in details. However

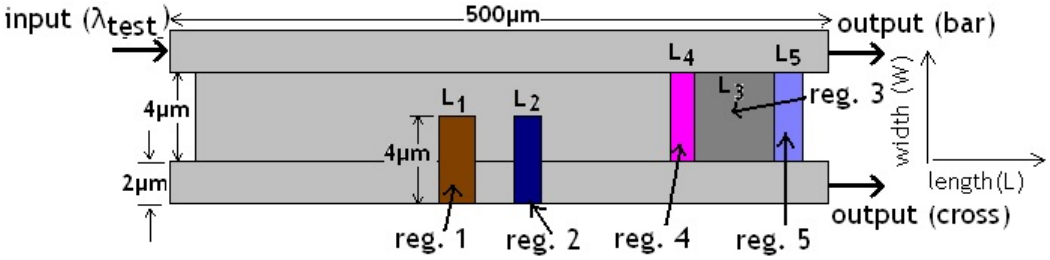


Figure 3: Layout (top view) of the modified 2×2 MMI-switch for wider operating wavelengths (structure 2)

both regions can be made of same length, but in that case, observed PDLs are at higher levels ($\geq$ 1dB). Theses extra IM Regions (4 and 5) are defined in order to accommodate TM polarized light images perfectly for either of input wavelength and to suppress the PDLs particularly. The tuning strategies for structure 2 are mentioned in table 2 in details.

Figures 4(a-c) and 5 (a-c) shows a comparison of calculated E.L., I.L. and corresponding CT levels for both structures. While these are evaluated for TM polarized input of a test wavelengths of 1.55μm and 1.3 μm. From these figures, it is clear that by introducing an extra IM region (Region 5) within the coupling section, as in case of structure 2, switch can be made to operate with reduced losses and better CT levels even for TM polarized optical inputs. Plots of figure 4 clearly indicate that with modified structuring, reduction in the switch losses i.e. E.L. and I.L. of the order of 2.25 dB and 2.50 dB with a reduction in the CT levels up to 5.8 dB can be achieved. Similarly trends in plots of figure 5, gives an indication of suppression in the switch losses (of the order of 1 dB) at a cost of increased CT levels by 1 dB for modified structure case.

Similarly plots in figure 6 (a-c) shows a comparison of calculated E.L., I.L. and corresponding CT levels for TE/TM polarization case for modified structure 2, while the optical test wavelength is 1.55Âľm. From these plots, it is very clear that within modified structure, PDL of less than 0.15 in any case can be maintained, while the switch losses in worst case remains below 0.8 dB and the best CT level of -23.5 dB can be achieved. Again with a slight variation in δn, the switch losses unaffected at large. However CT levels are more sensitive and tends to increase to

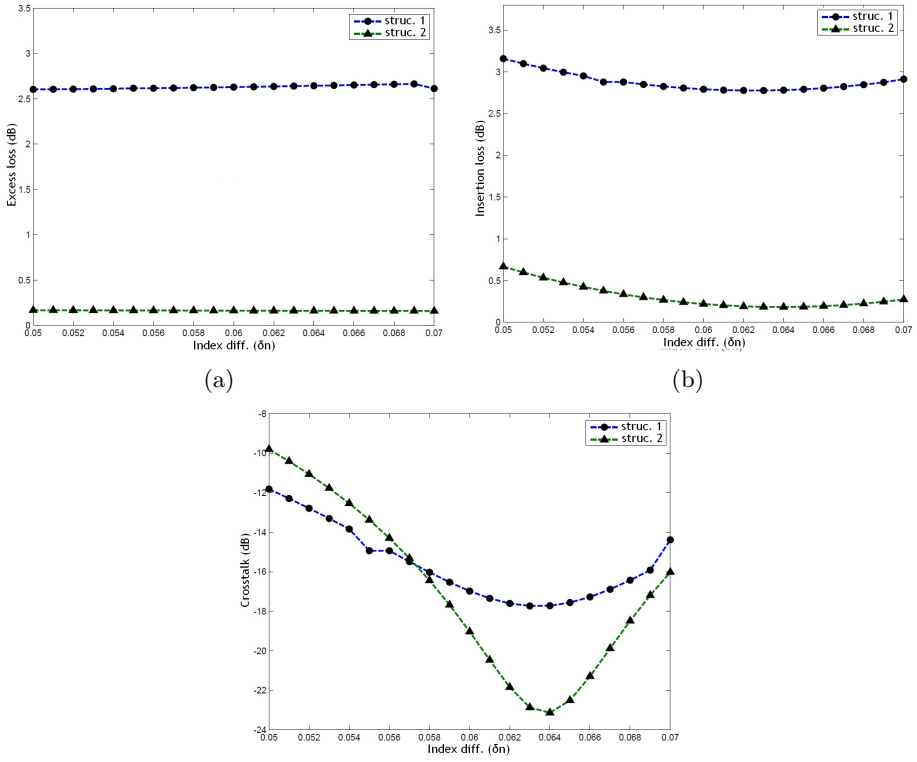


Figure 4: Calculated variation in the (a) E.L. (b) I.L. and (c) CT levels with respect to index variations (δn), while the switches (Struct. 1 and 2) are in bar state, while subjected to a TM polarized optical input of the test wavelength (λ) of 1.55 μm

higher levels in case of TM polarized inputs, thereby affecting the switching performance badly. Similar trends for test input wavelength of 1.3 μm have been noticed depicted by plots of figure 7 (a-c), which are showing a comparison of calculated E.L., I.L. and corresponding CT levels for TE/TM polarization case for modified structure. From these plots, it is clear that with modified structuring, PDLs of less than 0.07 in either case of test wavelength can be maintained, while the structure possess switching losses $\leq$ 1.1 dB with a best value for CT of -23.8 dB.
By evaluating both structures, it has been noticed that by employing extra IM regions, the MMI switch can be made to operate with reduced losses and better CT levels irrespective of polarization state of the inputs. It is also noticed later that for modified structure (2), PDL $\leq$ 0.15 dB in any case can be maintained, while in worst case, the switch possess propagation losses $\leq$ 0.8 dB with a best possible CT level of -23.5 dB. With modified structure, suppression in the switch losses (E.L., I.L.) of 1dB approximately can be achieved. Table 3 summarizes the performance parameters for both considered structures. With this table, one can predict that performance of such switching structure can be optimize further by introducing more IM regions, to generate clear, concentrated and lossless images within the coupling region. However insertion of more IM regions will affect the design complexity, tolerance factors and cost of the operation.

4 Conclusion

MMI waveguides based small size 2×2 photonic switches are realized. Concept of self-imaging in MMI waveguides has been used to introduce reconfigurable IM regions in a high refractive index

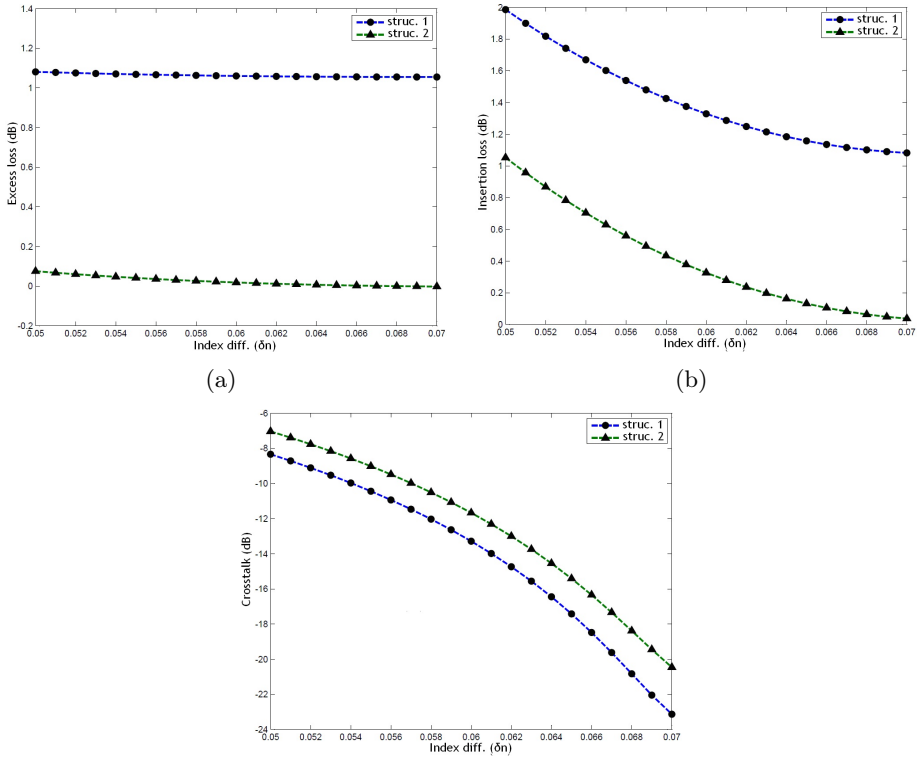


Figure 5: Calculated variation in the (a) E.L. (b) I.L. and (c) CT levels with respect to index variations (δn), while the switches (Struct. 1 and 2) are in bar state, while subjected to a TM polarized optical input of the test wavelength (λ) of 1.3 μm

contrast to tune the switch for different wavelengths. It is observed that with introduction of more IM regions, switch losses, corresponding CT and PDL levels can be suppressed significantly. The smaller dimensions of proposed structures shall provide an opportunity to reducing fabrication costs and increasing the density in optical networks. Therefore such structures can be proven as suitable to realize higher order switches due to their advantages of minimum propagation losses, better CT levels and ease of extendibility into a multi-port configuration. Also their compact nature with higher tolerance to avoid fabrication errors shall makes them promising elements for photonic integrated circuits (PIC) and monolithic functional switches.

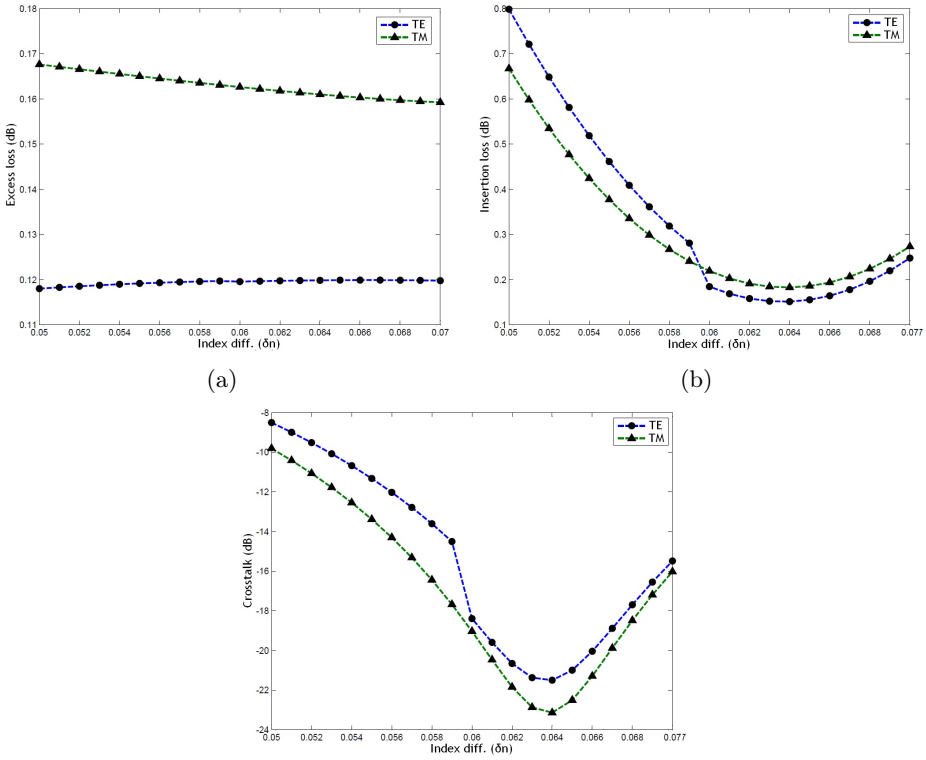


Figure 6: Calculated variation in the (a) E.L. (b) I.L. and (c) CT levels with respect to index variations (δn), while the switch (Structure-2) is in bar state and subjected to a TE/TM polarized optical inputs at test wavelength (λ) of 1.55 μm

Table 3: Comparative summary of performance parameters

Performance parameters (structure 1)	
λ_{test} : 1.3μm	E.L. $\leq$ 1.1 dB (TM) and $\leq$0.07 dB (TE) I.L.$\leq$ 2 dB (TM) and $\leq$ 0.8 dB (TE) CT (best value): -23.4 dB (TE), -23.2 dB (TM) PDL$_{max.}$: $\leq$ 0.1 dB
λ_{test} : 1.55μm	E.L. $\leq$ 2.7 dB (TM) and $\leq$0.12 dB (TE) I.L.$\leq$ 3.3 dB (TM) and $\leq$ 0.8 dB (TE) CT (best value): -21.8 dB (TE) and -17.4 dB (TM) PDL$_{max.}$: $\leq$ 2.6 dB
Performance parameters (structure 2)	
λ_{test} : 1.3μm	E.L. $\leq$ 0.08 dB (TM/TE) I.L.$\leq$ 1.2 dB (TM/TE) CT (best value): -23.8 dB (TE) and -20.2 dB (TM) PDL$_{max.}$: $\leq$ 0.07 dB
λ_{test} : 1.55μm	E.L. $\leq$ 0.17 dB (TM/TE) I.L.$\leq$ 0.8 dB (TM/TE) CT (best value): -21.8 dB (TE) and -23.3 dB (TM) PDL$_{max.}$: $\leq$ 0.15 dB

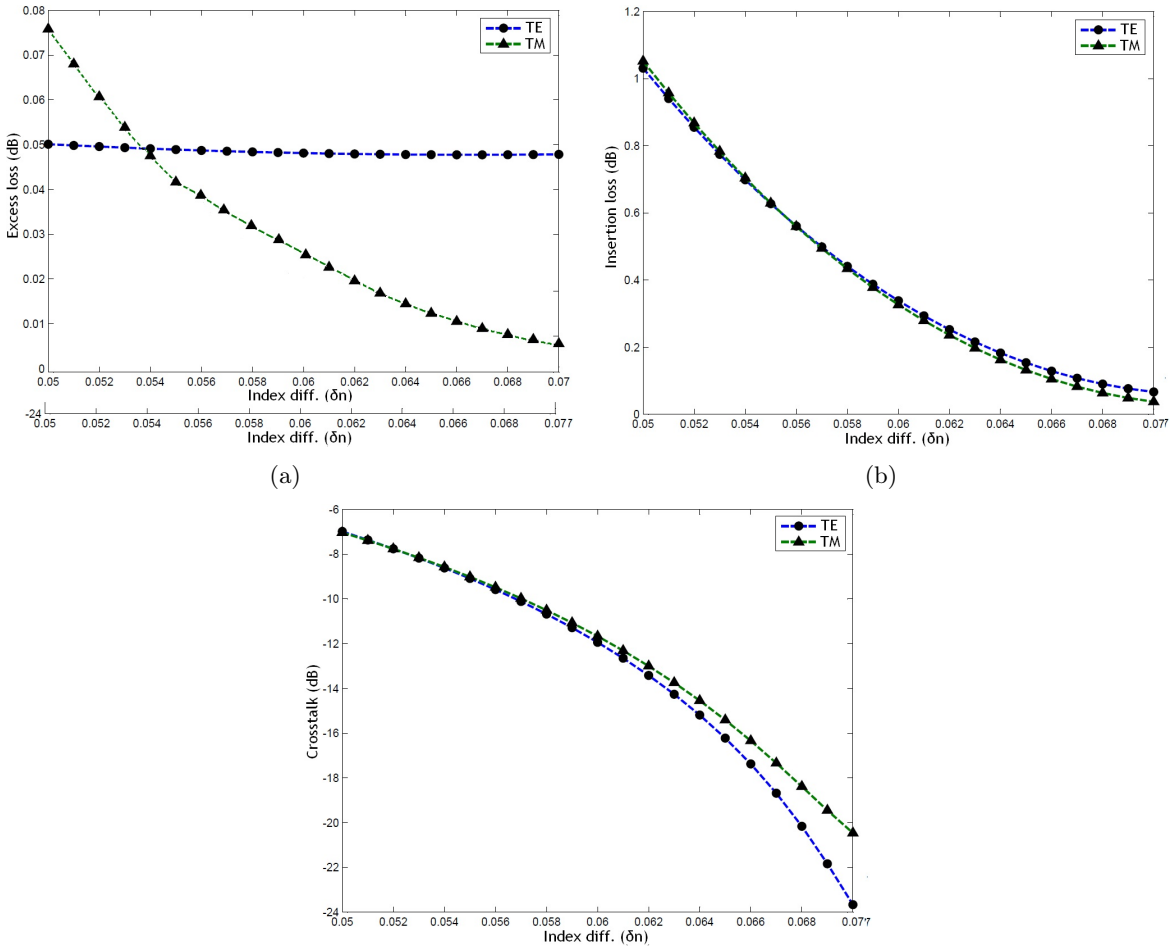


Figure 7: Calculated variation in the (a) E.L. (b) I.L. and (c) CT levels with respect to index variations (δn), while the switch (Structure-2) is in bar state and subjected to a TE/TM polarized optical inputs at test wavelength (λ) of 1.3 μm

Bibliography

[1] S. Kumai, T. Ishikawa, A. Okazaki, K. Utaka, et al., High-speed optical switching of InAlGaAs/InAlAs multimode interference photonic switch with partial index-modulation region (MIPS-P), *IEICE Electron. Express*, 2(23):578-582, 2005.

[2] Z. Jin, G. Peng, Designing optical switches based on silica MMI devices, *Progress in Electromagnetic Research Symposium*, Hangzhou, China, pp. 58-61, August 22-26, 2005.

[3] X. Wu, L. Liu, Y. Zhang, et al., Low electric power drived thermo-optic MMI-switches with tapered heating electrodes, *Optics Communications*, 258: 135-143, 2006.

[4] X.Q. Sun, C.M. Chen, et al., A MMI polymer-silica hybrid waveguide 2×2 thermo-optic switch, *Optica Applicata*, vol. xl, no. 3, 2010.

[5] L. Cahill, The modeling of MMI devices, *Proc. ICTON, ieeexplorer*, vol. 2, pp. 138-141, Nottingham, U.K., June, 2006.

[6] P. P. Sahu, A tapered structure for compact MMI-coupler, *IEEE Photon. Technol. Lett.*, 20(8):638-640, April 15th, 2008.

[7] K. Solehmainen, M. Kapulainen, M. Harjanne, T. Aalto, Adiabatic and MMI-couplers on SOI, *IEEE Photon. Technol. Lett.*, 18(21), Nov. 1st, 2006.

[8] A. M. Al hetar, A. S. M. Supaat, A. B. Mohammad, I. Yulianti, MMI photonic switches, *Optical Engineering*, 47(11), 112001, Nov., 2008.

[9] A. M. Al-Hetar, A. B. Mohammad et al., MI-MZI polymer thermo-optic switch with a high refractive index contrast, *Journal of Lightwave Technology*, 29(2):171-178, January 15th, 2011.

[10] L. B. Soldano, E. C. M. Pennings, Optical multimode interference devices based on self-imaging: Principles and applications, *Journal of Lightwave Tech.*, 13(4):615-627, 1995.

[11] R. Ulrich, Image formation by phase coincidences in optical waveguides, *Opt. Comm.*, 13:259-264, 1975.

[12] S. Nagai, G. Morishima, H. Inayoshi, K. Utaka, Multimode Interference Photonic Switches (MIPS), *Journal of Lightwave Technology*, 20(4):675-681, April, 2002.

[13] Z. Jin, G. Peng, Designing optical switches based on silica MMI devices, *Proc. of progress in Electromagnetic Research Symposium*, Hangzhou, China, pp. 58-61, August 22-26, 2005.

[14] F. Wang, J. Yang, L. Chen, X. Jiang, M. Wang, Optical switch based on MMI-coupler, *IEEE Photon. Technol. Lett.*, 18 (2):421-423, 2006.

[15] D. A. M. Arrioja, N. Bickel, P. Likamwa, Robust 2×2 MMI optical switch, *Optical and Quantum Electronics*, 38:557-566, 2006.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 776-784

Stability of Discrete-Time Systems with Time-Varying Delay: Delay Decomposition Approach

S.B. Stojanovic, D.LJ. Debeljkovic, N. Dimitrijevic

Sreten B. Stojanovic
University of Nis, Faculty of Technology
Serbia, 16000 Leskovac, Bulevar oslobodjenja 124
E-mail: ssreten@ptt.rs

Dragutin LJ. Debeljkovic, Nebojsa Dimitrijevic
University of Belgrade, Faculty of Mechanical Engineering
Serbia, 11120 Beograd, Kraljice Marije 16
E-mail: ddebeljkovic@mas.bg.ac.rs, ndimitri@verat.net

Abstract:
This article deals with the problem of obtaining delay-dependent stability conditions for a class of discrete-time systems with interval time-varying delay. Using the decomposition the delay interval into two unequal subintervals by tuning parameter α, a new interval delay-dependent Lyapunov-Krasovskii functional is constructed to derive novel delay-dependent stability conditions which are expressed in terms of linear matrix inequalities. This leads to reduction of conservatism in terms of the upper bounds of the maximum time-delay. The numerical examples show that the obtained result is less conservative than some existing ones in the literature.
Keywords: time-delay systems, interval time-varying delay, asymptotic stability, delay-dependent stability, Lyapunov-Krasovskii methods.

1 Introduction

Time-delay frequently occurs in many practical systems, such as manufacturing systems, telecommunication and economic systems etc. Since time-delay is an important source of instability and poor performance, considerable attention has been paid to the problem of stability analysis and controller synthesis for continuous time-delay systems (see e.g. [3-5, 10, 11, 17-21, 23-26] and the reference therein). Inversely, less attention has been drawn to the corresponding results for discrete-time delay systems (see e.g. [1, 2, 6-9, 12-15, 22, 24]). This is mainly due to the fact that such systems can be transformed into augmented systems without delay. This augmentation of the system is, however, inappropriate for systems with unknown delays and for systems with time-varying delay which are the subject analysis in this work.

Recently, increasing attention has been devoted to the problem of delay-dependent stability of linear systems with time-varying delay, including continuous-time (see e.g. [5, 10, 11, 18-21, 23], 25, 26]) and discrete-time systems (see e.g. [1, 2, 7-9, 12, 14, 15, 24]) and a great number of delay-dependent stability criteria were derived. The key point for deriving the delay-dependent stability criterions is the choice of an appropriate Lyapunov–Krasovskii functional (LKF). It is known that the existence of a complete quadratic Lyapunov-Krasovskii functional (CQLKF) is a sufficient and necessary condition for asymptotic stability of the time-delay system. Using the CQLKF, one can obtain the maximum allowable upper bound (MAUB) of delay which is very close to the analytical delay limit for stability. However, the CQLKF leads to a complicated system of partial differential equations, yielding infinite dimensional linear matrix inequalities (LMIs). Therefore, to develop simpler stability criteria, many authors have used special forms of LKF rather than CQLKF, which give LMIs with finite order and a reduced value of MAUB.

Further, to reduce the conservativeness of the existing results, some new analysis methods have been proposed, such as descriptor system transformation method [3-5], free weighting matrix

method [8, 11, 23], matrix inequality method [10, 17, 18] and input–output approach [19]. Using these methods, many stability criteria were derived by checking a variation of LKF in a *whole interval* of the time-delay. Contrary to this approach, in [24, 25], in order to obtain some less conservative stability conditions, the interval of the time delay is divided into *multiple equidistant subintervals* and interval delay-dependent LKF (ID-D LKF) is constructed. By checking the variation of the ID-D LKF defined on the subintervals, some new delay-dependent stability criteria are derived. It is worth pointing out that the main difference between LKF and ID-D LKF lies in that the former allows taking different weighing matrices on different subintervals. Therefore, ID-D LKF, as expected, will yield less conservative delay-dependent stability criteria.

Inspired by the idea of Zhu and Yang [26] on splitting the delay of continuous-time systems into *two unequal subintervals*, a new method is developed in this paper for stability analysis for discrete-time systems with time-varying delay. The delay interval $[k-h_M,\ k-1]$ in the ID-D LKF is divided into two unequal subintervals: $[k-h_M,\ k-\alpha-1]$ and $[k-\alpha,\ k-1]$, where $0<\alpha<h_M$ is a tuning parameter. The new ID-D LKF is constructed with different weighing matrices in various subintervals. Free-weighting matrices and model transformation are not used in order to derive delay dependent criterion. It is shown that the presented stability condition is much less conservative than the existing ones [1, 2, 6-9, 13-15, 22, 24], because it has a lower value of MAUB. The derived condition can be seen as an extension of the methods in [24, 25], wherein the whole delay range is divided to $n \geq 2$ equal subintervals. As the number of subintervals in [24, 25] is greater than two, the decomposition approach is more complex and the resulting stability conditions are more conservative and difficult to implement. To demonstrate the effectiveness of the proposed method, numerical examples are given in section 3.

Notation: $\Re^n$ and Z^+ denote the n-dimensional Euclidean space and positive integers. Notation $P>0$ $(P\geq 0)$ means that matrix P is real symmetric and positive definite (semi-definite). For real symmetric matrices P and Q, the notation $P>Q$ $(P\geq Q)$ means that matrix $P-Q$ is positive definite (positive semi-definite). I is an identity matrix with an appropriate dimension. Superscript "T" represents the transpose. In symmetric block matrices or complex matrix expressions, we use an asterisk $(*)$ to represent a term which is induced by symmetry. If dimensions of matrices are not explicitly given, then they are assumed to be compatible for algebraic operations.

2 Main results

Consider the following system with an interval time-varying delay:

$$x(k+1)=Ax(k)+Bx(k-h(k)) \tag{1}$$

where $x(k)\in\Re^n$ is the state at instant k, matrices $A\in\Re^{n\times n}$and $B\in\Re^{n\times n}$ are constant matrices and $h(k)$is the positive integer representing the time delay of the system that we assume to be time dependent and satisfies the following:

$$0\leq h(k)\leq h_M \tag{2}$$

where h_M is known to be a positive and finite integer.

The aim of this article is to establish the sufficient condition that guarantee the delay-dependent stability of the system (1), which is less conservative than the existing results in literature.

We first introduce the following result, which will be used in the proof of our main results.

Lemma 1. *Let $y(k) = x(k+1) - x(k)$. For any matrix $R > 0$ [24]*

$$-(h_M - h_m)\sum_{m=k-h_M}^{k-1-h_m} y^T(m)Ry(m) \leq \begin{bmatrix} x(k-h_m) \\ x(k-h_M) \end{bmatrix}^T \begin{bmatrix} -R & R \\ R & -R \end{bmatrix} \begin{bmatrix} x(k-h_m) \\ x(k-h_M) \end{bmatrix} \tag{3}$$
$$= -\left[x(k-h_m) - x(k-h_M)\right]^T R \left[x(k-h_m) - x(k-h_M)\right]$$

Theorem 2. *For given scalars $h_M (h_M > 0)$ and $\alpha (0 < \alpha < h_M)$, the system described by (1)-(2) is asymptotically stable if there exists matrices $P = P^T > 0$, $Q_i = Q_i^T \geq 0$ and $Z_i = Z_i^T \geq 0$ $(i = 1, 2, 3)$, such that the following LMIs hold:*

$$\Phi = \begin{bmatrix} \Phi_{11} & \Phi_{12} & 0 & 0 & \Phi_{15} \\ * & \Phi_{22} & \Phi_{23} & 0 & \Phi_{25} \\ * & * & \Phi_{33} & \Phi_{34} & 0 \\ * & * & * & \Phi_{44} & 0 \\ * & * & * & * & \Phi_{55} \end{bmatrix} < 0 \tag{4}$$

$$\Psi = \begin{bmatrix} \Psi_{11} & \Psi_{12} & \Psi_{13} & 0 & \Psi_{15} \\ * & \Psi_{22} & \Psi_{23} & \Psi_{24} & \Psi_{25} \\ * & * & \Psi_{33} & 0 & 0 \\ * & * & * & \Psi_{44} & 0 \\ * & * & * & * & \Psi_{55} \end{bmatrix} < 0 \tag{5}$$

where

$$\Phi_{11} = A^T PA - P + Q_1 + Q_3 - \tfrac{1}{\alpha}(Z_1 + Z_3),$$
$$\Phi_{12} = A^T PB + \tfrac{1}{\alpha}(Z_1 + Z_3),\ \Phi_{15} = (A - I)^T U_1,$$
$$\Phi_{22} = B^T PB - Q_3 - \tfrac{1}{\alpha}(2Z_1 + Z_3),\ \Phi_{23} = \tfrac{1}{\alpha}Z_1,\ \Phi_{25} = B^T U_1,$$
$$\Phi_{33} = -Q_1 + Q_2 - \tfrac{1}{\alpha}Z_1 - \tfrac{1}{h_M - \alpha}Z_2,\ \Phi_{34} = \tfrac{1}{h_M - \alpha}Z_2,$$
$$\Phi_{44} = -Q_2 - \tfrac{1}{h_M - \alpha}Z_2,\ \Phi_{55} = -U_1,$$
$$\Psi_{11} = \Phi_{11},\ \Psi_{12} = A^T PB,\ \Psi_{13} = \tfrac{1}{\alpha}(Z_1 + Z_3),\ \Psi_{15} = (A - I)^T U_2,$$
$$\Psi_{22} = B^T PB - Q_3 - \tfrac{1}{h_M - \alpha}(2Z_2 + Z_3),\ \ \Psi_{23} = \tfrac{1}{h_M - \alpha}(Z_2 + Z_3),$$
$$\Psi_{24} = \tfrac{1}{h_M - \alpha}Z_2,\ \Psi_{25} = B^T U_2,$$
$$\Psi_{33} = -Q_1 + Q_2 - \tfrac{1}{\alpha}(Z_1 + Z_3) - \tfrac{1}{h_M - \alpha}(Z_2 + Z_3),$$
$$\Psi_{44} = -Q_2 - \tfrac{1}{h_M - \alpha}Z_2,\ \Psi_{55} = -U_2,$$
$$U_1 = \alpha Z_1 + (h_M - \alpha)Z_2 + \alpha Z_3,\ U_2 = \alpha Z_1 + (h_M - \alpha)Z_2 + h_M Z_3$$

Proof: Construct the interval delay-dependent LKF as

$$V(k) = V_1(k) + V_2(k) + V_3(k) \tag{6}$$

where

$$V_1(k) = x^T(k)Px(k) \tag{7}$$

$$V_2(k) = \sum_{i=k-\alpha}^{k-1} x^T(i)Q_1 x(i) + \sum_{i=k-h_M}^{k-1-\alpha} x^T(i)Q_2 x(i) + \sum_{i=k-h(k)}^{k-1} x^T(i)Q_3 x(i) \tag{8}$$

$$V_3(k) = \sum_{i=-\alpha}^{-1}\sum_{j=k+i}^{k-1} y^T(j)Z_1 y(j) + \sum_{i=-h_M}^{-1-\alpha}\sum_{j=k+i}^{k-1} y^T(j)Z_2 y(j) + \sum_{i=-h(k)}^{-1}\sum_{j=k+i}^{k-1} y^T(j)Z_3 y(j) \tag{9}$$

where $P = P^T > 0$, $Q_i = Q_i^T \geq 0$ and $Z_i = Z_i^T > 0$ $(i = 1, 2, 3)$. Note, the delay interval $[k - h_M,\ k - 1]$ in the LKF is divided into two unequal subintervals: $[k - h_M,\ k - \alpha - 1]$ and $[k - \alpha,\ k - 1]$, where $0 < \alpha < h_M$ is a tuning parameter.

Taking the difference of $\Delta V_i(k) = V_i(k+1) - V_i(k)$, we can obtain

$$\begin{aligned}\Delta V_1(k) = {} & x^T(k)\left(A^T P A - P\right)x(k) + 2x^T(k)A^T P B x\left(k - h(k)\right) \\ & + x^T\left(k - h(k)\right)B^T P B x\left(k - h(k)\right)\end{aligned} \tag{10}$$

$$\begin{aligned}\Delta V_2(k) = {} & x^T(k)Q_1 x(k) - x^T(k-\alpha)Q_1 x(k-\alpha) \\ & + x^T(k-\alpha)Q_2 x(k-\alpha) - x^T(k-h_M)Q_2 x(k-h_M) \\ & + x^T(k)Q_3 x(k) - x^T\left(k - h(k)\right)Q_3 x\left(k - h(k)\right)\end{aligned} \tag{11}$$

$$\begin{aligned}\Delta V_3(k) = {} & \sum_{i=-\alpha}^{-1}\left[y^T(k)Z_1 y(k) - y^T(k+i)Z_1 y(k+i)\right] \\ & + \sum_{i=-h_M}^{-1-\alpha}\left[y^T(k)Z_2 y(k) - y^T(k+i)Z_2 y(k+i)\right] \\ & + \sum_{i=-h(k)}^{-1}\left[y^T(k)Z_3 y(k) - y^T(k+i)Z_3 y(k+i)\right] \\ = {} & \alpha y^T(k)Z_1 y(k) - \sum_{i=-\alpha}^{-1} y^T(k+i)Z_1 y(k+i) \\ & + \left(h_M - \alpha\right)y^T(k)Z_2 y(k) - \sum_{i=-h_M}^{-1-\alpha} y^T(k+i)Z_2 y(k+i) \\ & + h(k)y^T(k)Z_3 y(k) - \sum_{i=-h(k)}^{-1} y^T(k+i)Z_3 y(k+i) \\ = {} & y^T(k)\left[\alpha Z_1 + \left(h_M - \alpha\right)Z_2 + h(k)Z_3\right]y(k) \\ & - \sum_{m=k-\alpha}^{k-1} y^T(m)Z_1 y(m) - \sum_{m=k-h_M}^{k-1-\alpha} y^T(m)Z_2 y(m) \\ & - \sum_{m=k-h(k)}^{k-1} y^T(m)Z_3 y(m) \\ = {} & y^T(k)\left[\alpha Z_1 + \left(h_M - \alpha\right)Z_2 + h(k)Z_3\right]y(k) \\ & - \sum_{m=k-\alpha}^{k-1} y^T(m)Z_1 y(m) - \sum_{m=k-h_M}^{k-1-\alpha} y^T(m)Z_2 y(m) \\ & - \sum_{m=k-h(k)}^{k-1} y^T(m)Z_3 y(m)\end{aligned} \tag{12}$$

It is know from (2) that, for any $k \in Z^+$, $h(k) \in [0, \alpha - 1]$ or $h(k) \in [\alpha, h_M]$. Define two sets

$$\Omega_1 = \left\{k :\ h(k) \in [0, \alpha]\,,\ k \in Z^+\right\} \tag{13}$$

$$\Omega_2 = \left\{k :\ h(k) \in [\alpha + 1, h_M]\,,\ k \in Z^+\right\} \tag{14}$$

In the following, we will discuss the variation of $\Delta V(k)$ for two cases ($k \in \Omega_1$ and $k \in \Omega_2$).

Case 1. For $k \in \Omega_1$, i.e. $0 \leq h(k) \leq \alpha$.

$$\sum_{m=k-\alpha}^{k-1} y^T(m)Z_1 y(m) = \sum_{m=k-\alpha}^{k-1-h(k)} y^T(m)Z_1 y(m) + \sum_{m=k-h(k)}^{k-1} y^T(m)Z_1 y(m) \tag{15}$$

$$\begin{aligned}\Delta V_3(k) = y^T(k)\left[\alpha Z_1 + (h_M-\alpha)\,Z_2 + h(k)Z_3\right]y(k) - \sum_{m=k-\alpha}^{k-1-h(k)} y^T(m)Z_1y(m)\\ - \sum_{m=k-h(k)}^{k-1} y^T(m)\left(Z_1+Z_3\right)y(m) - \sum_{m=k-h_M}^{k-1-\alpha} y^T(m)Z_2y(m)\end{aligned} \tag{16}$$

Because $Z_1 + Z_3 > 0$, $h(k) \le \alpha$ and $\alpha - h(k) \le \alpha$, using Lemma 1, it follows

$$\begin{aligned}&-\sum_{m=k-h(k)}^{k-1} y^T(m)\left(Z_1+Z_3\right)y(m)\\ &\quad\le -\tfrac{1}{h(k)}\left[x(k)-x\left(k-h(k)\right)\right]^T\left(Z_1+Z_3\right)\left[x(k)-x\left(k-h(k)\right)\right]\\ &\quad\le \tfrac{1}{\alpha}x^T(k)\left(-Z_1-Z_3\right)x(k) + \tfrac{1}{\alpha}2x^T(k)\left(Z_1+Z_3\right)x\left(k-h(k)\right)\\ &\qquad+\tfrac{1}{\alpha}x^T\left(k-h(k)\right)\left(-Z_1-Z_3\right)x\left(k-h(k)\right)\end{aligned} \tag{17}$$

$$\begin{aligned}&-\sum_{m=k-\alpha}^{k-1-h(k)} y^T(m)Z_1y(m)\\ &\quad\le -\tfrac{1}{\alpha-h(k)}\left[x\left(k-h(k)\right)-x\left(k-\alpha\right)\right]^T Z_1\left[x\left(k-h(k)\right)-x\left(k-\alpha\right)\right]\\ &\quad\le \tfrac{1}{\alpha}x^T\left(k-h(k)\right)\left(-Z_1\right)x\left(k-h(k)\right) + \tfrac{1}{\alpha}2x^T\left(k-h(k)\right)Z_1x\left(k-\alpha\right)\\ &\qquad+\tfrac{1}{\alpha}x^T\left(k-\alpha\right)\left(-Z_1\right)x\left(k-\alpha\right)\end{aligned} \tag{18}$$

$$\begin{aligned}&-\sum_{m=k-h_M}^{k-1-\alpha} y^T(m)Z_2y(m)\\ &\quad\le -\tfrac{1}{h_M-\alpha}\left[x\left(k-\alpha\right)-x(k-h_M)\right]^T Z_2\left[x\left(k-\alpha\right)-x(k-h_M)\right]\\ &\quad\le \tfrac{1}{h_M-\alpha}x^T\left(k-\alpha\right)\left(-Z_2\right)x\left(k-\alpha\right) + \tfrac{1}{h_M-\alpha}2x^T\left(k-\alpha\right)Z_2x\left(k-h_M\right)\\ &\qquad+\tfrac{1}{h_M-\alpha}x^T\left(k-h_M\right)\left(-Z_2\right)x\left(k-h_M\right)\end{aligned} \tag{19}$$

Combining (10)-(19), it yields

$$\begin{aligned}\Delta V(k) &\le \xi^T(k)\hat{\Phi}\xi(k)\\ \hat{\Phi} &= \begin{bmatrix} \Phi_{11}+(A-I)^TU_1(A-I) & \Phi_{12}+(A-I)^TU_1B & 0 & 0\\ * & \Phi_{22}+B^TU_1B & \Phi_{23} & 0\\ * & * & \Phi_{33} & \Phi_{34}\\ * & * & * & \Phi_{44}\end{bmatrix}\\ \xi(k) &= \begin{bmatrix} x^T(k) & x^T\left(k-h(k)\right) & x^T(k-\alpha) & x^T(k-h_M)\end{bmatrix}^T\end{aligned} \tag{20}$$

Obviously, $\Delta V(k) < 0$ for $k \in \Omega_1$ if $\hat{\Phi} < 0$. Using the Schur complement, it is easy to see that $\Delta V(k) < 0$ holds if $\Phi < 0$ and $h(k) \in [0, \alpha]$.

Case 2. Similarly, for $k \in \Omega_2$, i.e. $\alpha + 1 \le h(k) \le h_M$, using the Schur complement, it is easy to see that $\Delta V(k) < 0$ holds if $\Psi < 0$.

From the above discussions, we can see that for all $k \in Z^+$ if (4)-(5) hold, $\Delta V(k) < 0$, which completes the proof. □

Remark 3. Theorem 2 presents a stability result which depends on the maximum delay bound h_M. The conditions in Theorem 2 are expressed in terms of LMIs, and therefore, they can be easily checked by using standard numerical software.

Remark 4. The delay interval $[k - h_M,\ k - 1]$ in the ID-D LFK is divided into two unequal subintervals: $[k - h_M,\ k - \alpha - 1]$ and $[k - \alpha,\ k - 1]$, where $0 < \alpha < h_M$ is a tuning parameter.

Consequently, the different weighing matrices in the Lyapunov functional are used in various subintervals and the information of delayed state $x(k-\alpha)$ can be taken into full consideration. Further, using the subintervals and Lemma 1, the upper bounds of some terms in $\Delta V_3(k)$ are more accurately estimated than by using the previous methods since the upper bound h_M of delay $h(k)$ on the interval $0 \le h(k) \le h_M$ is substituted with two less conservative upper bounds α and h_M on the subintervals $0 \le h(k) \le \alpha$ and $\alpha < h(k) \le h_M$, respectively. So the decomposition method presented in Theorem 2 can reduce the value of MAUB.

An algorithm for seeking a corresponding values of α $(0 < \alpha < h_M)$ subject to (4)-(5), such that the MAUB of h_M has maximal value, can easily be obtained.

Algorithm 5. *Step 1. Let* $h = 0$ *and* $\alpha = 0$.

Step 2. $h = h + 1$.

Step 3. $\alpha = \alpha + 1$.

Step 4. If inequalities (4)-(5) is feasible, then $\alpha_m = \alpha$, $\alpha = 0$ *and go to step 2; otherwise, go to step 5.*

Step 5. If $\alpha = h - 1$, *go to step 6; otherwise, go to step 3.*

Step 6. The maximal delay is $h_M = h - 1$ *and the minimal value of tuning parameter*α*is* α_m.

3 Numerical examples

In this section, two examples are presented. The obtained results have been compared with several existing criteria in the literature.

Example 1. Consider system (1) with the time-varying delay $h(k)$ satisfying (2) and

$$A = \begin{bmatrix} 0.8 & 0 \\ 0 & \lambda \end{bmatrix}, \quad B = \begin{bmatrix} -0.1 & 0 \\ -0.1 & -0.1 \end{bmatrix}, \quad \lambda \in \{0.91,\ 0.97\}$$

Case 1 ($\lambda = 0.91$). This system was considered in [13] and [22]. Table 1 lists the MAUB of delay obtained from Theorem 2 of this paper. For comparison, results from [13, 22] are also listed in the table. It is clear that Theorem 2 leads to better results than those in [13, 22].

Table 1. Comparison of MAUB of delay based on different existing methods for $\lambda = 0.91$	
Method	h_M
[13]	41
[22, Corollary 1]	42
Theorem 2 in this paper	46 for $\alpha = 19, \cdots, 30$

Case 2 ($\lambda = 0.97$). For comparison, the results from [2, 6, 7, 9] and this paper are listed in Table 2. It is clear that Theorem 2 gives much better results than the existing delay-dependent criteria.

Example 2. Consider system (1) with the time-varying delay $h(k)$ satisfying (2) and

Case 1. [1, 14, 15]

$$A = \begin{bmatrix} 0.6 & 0 \\ 0.35 & 0.7 \end{bmatrix}, \quad B = \begin{bmatrix} 0.1 & 0 \\ 0.2 & 0.1 \end{bmatrix}$$

Case 2. [8, 9, 15, 24]

$$A = \begin{bmatrix} 0.8 & 0 \\ 0.05 & 0.9 \end{bmatrix}, \quad B = \begin{bmatrix} -0.1 & 0 \\ -0.2 & -0.1 \end{bmatrix}$$

Table 2. Comparison of MAUB of delay based on different existing methods for $\lambda = 0.97$	
Method	h_M
[9, Theorem 1]	4
[7, Theorem 3]	8
[6, Lemma 2]	8
[2, Theorem 1]	10
Theorem 2 in this paper	17 for $\alpha = 9, 10, 11$

For the above systems, the LMIs conditions for delay-independent stability [16]

$$P = P^T > 0, \quad Q = Q^T > 0, \quad \begin{bmatrix} -P+Q & 0 & A^TP \\ * & -Q & B^TP \\ * & * & -P \end{bmatrix} < 0$$

are feasible, from which we deduce that both systems are stable for $0 \leq h(k) < \infty$. Using the existing delay-dependent criteria, it can be obtained only the finite value of MAUB, which will guarantee the stability of the given systems. Tables 3 (Case 1) and 4 (Case 2) list the intervals of time delay for different methods. Based on Theorem 2 in this paper, large numerical values of MAUB are obtained ($h_M \to \infty$). Hence, using of Theorem 2 leads to better results than those in [1, 8, 9, 14, 15, 25].

Table 3. Interval of time delay for Case 1	
Method	Interval
[1, Th. 3.1]	$2 \leq h(k) \leq 10$
[15, Theorem 1], [15, Theorem 2]	$2 \leq h(k) \leq 13$
[14, Theorem 3.2]	$0 \leq h(k) \leq 12$
[2, Theorem 1]	$2 \leq h(k) \leq 15$
Theorem 2 in this paper	$0 \leq h(k) \leq 10 \cdot 10^{21}$

Table 4. Interval of time delay for Case 2	
Method	Interval stability
[9, Theorem 1]	$0 \leq h(k) \leq 6$
[15, Theorem 2]	$0 \leq h(k) \leq 10$
[8, Theorem 1]	$0 \leq h(k) \leq 12$
[24, Theorem 5]	$2 \leq h(k) \leq 19$
[24, Theorem 7]	$2 \leq h(k) \leq 20$
Theorem 2 in this paper	$0 \leq h(k) \leq 9.61 \cdot 10^8$

4 Conclusion

In this paper, the problem of obtaining delay dependent stability conditions for a class of systems with interval time-varying delay is discussed. A new interval delay-dependent Lyapunov–Krasovskii functional is constructed by splitting the delay interval into two unequal intervals by tuning parameter α. The free-weighting matrices and model transformation are not used in order to derive delay-dependent criteria. Numerical examples show that the results proposed in this paper are much less conservative while comparing the maximum allowable upper bound of delay with the existing results in the literature.

Bibliography

[1] E.K. Boukas, Discrete-time systems with time-varying time delay: stability and stabilizability, *Mathematical Problems in Engineering*, Article ID 42489:1-10, 2006.

[2] K.F. Chen and I-K Fong, Stability of discrete-time uncertain systems with a time-varying state delay, Proc. IMechE, Part I: *J. Systems and Control Engineering*, 222: 493-500, 2008.

[3] E. Fridman and U. Shaked, A descriptor system approach to $H\infty$ control of linear time-delay systems, *IEEE Transactions on Automatic Control*, 47(2): 253–270, 2002.

[4] E. Fridman and U. Shaked, H_∞ control of linear state delay descriptor systems: an LMI approach, *Linear Algebra and its Applications*, 351–352: 271–302, 2002.

[5] E. Fridman, New Lyapunov–Krasovskii functionals for stability of linear retarded and neutral type systems, *Systems and Control Letters*, 43: 309–319, 2001.

[6] E. Fridman and U. Shaked, Delay-Dependent H_∞ Control of Uncertain Discrete Delay Systems, *European Journal of Control*, 11: 29-37, 2005.

[7] E. Fridman and U. Shaked, Stability and guaranteed cost control of uncertain discrete delay systems, *International Journal of Control*, 78(4): 235-246, 2005.

[8] H. Gao and T. Chen, New results on stability of discrete-time systems with time-varying state delay, *IEEE Transactions on Automatic Control*, 52: 328–334, 2007.

[9] H. Gao, J. Lam and Y. Wang, Delay-dependent output-feedback stabilization of discrete-time systems with time-varying state delay, *IEE Proc.: Control Theory Applications*, 151(6): 691-698, 2004.

[10] Q.L. Han and D. Yue, Absolute stability of Lur'e systems with time-varying delay, *IET Control Theory*, 1(3): 854–859, 2007.

[11] Y. He, M. Wu, J.H. She and G.P. Liu, Parameter-dependent Lyapunov functional for stability of time-delay systems with polytopic-type uncertainties, *IEEE Transactions on Automatic Control*, 49: 828–832, 2004.

[12] X. Jiang, Q.L. Han and X.H. Yu, Stability criteria for linear discrete-time systems with interval-like time-varying delay, *Proc. American Control Conference*, New Orleans, USA, 2817–2822, 2005.

[13] Y.S. Lee and W.H. Kwon, Delay-dependent robust stabilization of uncertain discrete-time state-delayed systems, *Proc. 15^{th} IFAC World Congr.*, 15(1): Barcelona, Spain 2002.

[14] V. Leite and M. Miranda, Robust Stabilization of Discrete-Time Systems with Time-Varying Delay: An LMI Approach, *Mathematical Problems in Engineering*, 2008: Article ID 876509, 15 pages, 2008.

[15] X.G. Liu, R.R. Martin, M. Wu and M.L. Tang, Delay-dependent robust stabilization of discrete-time systems with time-varying delay, *IEE Proc.: Control Theory and Applications*, 153(6): 689–702, 2006.

[16] M.S. Mahmood, *Robust control and filtering for time-delay systems*, Marcel-Dekker, New York, 2000.

[17] Y.S. Moon, P. Park and W.H. Kwon, Robust stabilization of uncertain input-delayed systems using reduction method, *Automatica*, 37: 307–312, 2001.

[18] P. Park and J.W. Ko Stability and robust stability for systems with a time-varying delay, *Automatica*, 43: 1855–1858, 2007.

[19] E. Shustin and E. Fridman, On delay-derivative-dependent stability of systems with fast-varying delays, *Automatica*, 43: 1649–1655, 2007.

[20] M. Wu, Y. He, J.H. She and G.P. Liu, Delay-dependent criteria for robust stability of time-varying delay systems, *Automatica*, 40: 1435–1439, 2004.

[21] S. Xu and J. Lam, Improved Delay-dependent Stability Criteria for time-delay systems, *IEEE Transactions on Automatic Control*, 50(3): 384-387, 2005.

[22] S. Xu, J. Lam and Y. Zou, Improved conditions for delay-dependent robust stability and stabilization of uncertain discrete time-delay systems, *Asian Journal of Control*, 7(3): 344-348, 2005.

[23] D. Yue and Q.L. Han, A delay-dependent stability criterion of neutral systems and its application to a partial element equivalent circuit model, *IEEE Transactions on Circuits and Systems-II*, 51(12): 685–689, 2004.

[24] D. Yue, E. Tian and Y. Zhang, A piecewise analysis method to stability analysis of linear continuous/discrete systems with time-varying delay, *International Journal of Robust and Nonlinear Control*, 19: 1493–1518, 2009.

[25] X.M. Zhang and Q.L. Han, A delay decomposition to delay-dependent stability for linear systems with time-varying delays, *International Journal of robust and nonlinear control*, 19: 1922–1930, 2009.

[26] X.L. Zhu and G.H. Yang, New results of stability analysis for systems with time-varying delay, *International Journal of robust and nonlinear control* 20: 596–606, 2010.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 4 (November), pp. 785-790

Clustering-Based Energy-Efficient Broadcast Tree in Wireless Networks

J. Yu, H. Jiang, G. Wang, Q. Guo

Jiguo Yu, Honglu Jiang
School of Computer Science,
Qufu Normal University, Rizhao, 276826, P.R. China
E-mail: jiguoyu@sina.com; jianghonglu88@163.com

Guanghui Wang
School of Mathematics, Shandong University,
Jinan, 250100, P.R. China
E-mail: ghwang@sdu.edu.cn

Qiang Guo
Key Laboratory for Computer Networks of Shandong Province
Shandong Computer Science Center, Jinan, 250014, P.R. China
E-mail: guoq@keylab.net

Abstract: The characteristics of wireless networks present formidable challenges to the study of broadcasting problem. A crucial issue in wireless networks is the energy consumption, because of the nonlinear attenuation properties of radio signals. Another crucial issue is the trade-off between reaching more nodes in a single hop by using higher power versus reaching fewer nodes in that single hop by using lower power. Given a wireless network with a specified source node that broadcasts messages to all other nodes in the network, the minimum energy broadcast (MEB) problem is NP-hard. In this paper, we propose a hybrid approach CBEEB(clustering-based energy-efficient broadcast) for the MEB problem based on clustering. Theoretical analysis indicates the efficiency and effectiveness of CBEEB. Simulation results show that CBEEB has better performance compared with the existing heuristic approaches.
Keywords: Wireless Network, Energy-Efficient, Broadcast, Clustering

1 Introduction

Nodes in wireless networks are usually powered by batteries with limited capacity. Therefore, energy efficiency is one of the most important design issues for wireless networks. The energy consumption in transmission can be further reduced by using one or more intermediate nodes instead of transmitting directly.

Broadcasting in wireless networks is different from that in wired networks, since all nodes that are within the transmission range of the sender can receive the transmission without any additional cost at the sender. This characteristic of wireless transmission is known as wireless multicast advantage (WMA) [1]. This allows us to seek power optimal range assignment for the energy-efficient broadcast problem. The minimum energy broadcast (MEB) problem is to find a broadcast scheme with minimum energy consumption. The problem is also known as minimum power broadcast (MPB) problem or minimum energy consumption broadcast subgraph (MECBS) problem [2]. We use energy and power alternatively throughout this paper.

In this paper, we study the problem of broadcasting in wireless networks, where every node is equipped with omni directional antennas. We propose a hybrid algorithm based on clustering and prove its correctness. Theoretical analysis shows the effectiveness of our approach. We compare it with the Broadcast Incremental Power (BIP) [1] algorithm and the best heuristic approaches known in [2–4] by simulation.

2 Related work

Most previously developed models for broadcasting and multicasting problems were link-based models, which does not properly reflect the properties of the all-wireless network environment. To solve the MPB problem, Wieselthier et al. first proposed the node-based approach (BIP)which is more suitable for wireless environment than the link-based algorithms [1]. It was subsequently shown in [5] that the BIP algorithm has an approximation ratio between 13/3 and 12. Das et al. gave an improvement procedure called r-shrink for MEB problem [6]. Cagalj et al. [7] proposed another improvement procedure called embedded wireless multicast advantage (EWMA). Kang et al. [8] generalized the EWMA into another heuristic called largest expanding sweep search(LESS). In [9], an ant-colony based approach for the problem was presented. An algorithm called CM using ant colony optimization approach was also presented in [3]. Furthermore, in [4] a simple simulated annealing algorithm with Metropolis chains of dynamic length was presented. The algorithm is in fact a probabilistic version of the 1-shrink tree-improvement algorithm described in [9]. Kang et al. [10] gave a perturbation based iterated local optimization method that uses LESS as local search. In [11], Montemanni et al. presented a simulated annealing approach based on r-shrink. They used sweep procedure to improve the solution obtained through simulated annealing. Al-Shihabi et al. [12] developed a hybrid approach combining nested partitioning with LP relaxation and r-shrink method. In [13], Wolf et al. proposed an evolutionary local search, which uses modified r-shrink as local search and random increase in transmission power of some nodes as mutation operator. Recently, Wu et al. [14] developed a generational genetic algorithm that uses permutation encoding to represent a broadcast scheme. Hashemi et al. presented a simulated annealing algorithm using a special node selection mechanism in its neighborhood structure [2]. In [15], Singh et al. proposed a hybrid approach to the MEB problem combining a genetic algorithm with a local search heuristic, which is a modified version of r-shrink improvement procedure [6].

3 Preliminaries and model

We consider static multi-hop ad hoc wireless networks with omni-directional transmitters, and each node can adjust its transmitting power based on the distance to the receiving node. In the most common power attenuation model, received signal power varies as $d^{-\alpha}$, where d is the distance from the transmitter and α is an environment-dependent constant typically between 2 and 5. Therefore, the transmitter power required to support the direct communication is proportional to d^{α}.

To represent one wireless ad hoc network, let $G = (V, E)$ be a directed graph, where V denotes the set of nodes and E denotes the set of edges, and a special node $s \in V$ that broadcasts messages to all other nodes of v. The transmission energy required by a node in an arborescence is determined by the longest edge among all edges of that arborescence. Leaf nodes do not relay messages to any other node. Therefore, the transmission energy required by leaf nodes is zero. The total transmission energy required for the broadcast can be computed by adding the energy which is required by each node in that arborescence.

Definition 1. Min-Energy Broadcast Problem(MEB): Let $G = (V, E)$ be a directed graph, find a broadcast tree $T \subseteq E$ rooted at s, with the minimum energy cost $\sum_{u \in V} max_{(u,v) \in T} d(u, v)^{\alpha}$, where $d(u, v)$ is the distance between the nodes u and v.

4 Hybrid approach–CBEEB

We now propose a hybrid approach CBEEB (clustering-based energy-efficient broadcast) to MEB problem, which includes a clustering algorithm and the IBIP (improved BIP) algorithm.

A special node s broadcasts messages to all other nodes of V, and each node except s has one unique identifier ID. For convenience, the notation and terminology used in the rest of this paper are summarized in Table 1.

Definition 2. Priority of nodes: For two nodes $i, j \in V\backslash\{s\}$,$i.prio > j.prio \Leftrightarrow$ $d(i) > d(j)$, or $d(i) = d(j)$ & &$ID(i) > ID(j)$

Table 1. Notation and Terminology

$ID(i)$	The unique identifier of node i
$i.prio$	The priority of node i
$d(i)$	The degree of node i
N_i^1	The set of one-hop neighbors of node i
$i.ch$	The clusterhead elected by node i
$i.type$	The type of node i, which is a clusterhead or a leaf node

The clustering algorithm is as follows:

Step 1. Each node except s collects the information of all its one-hop neighbors N_i^1. Such information can be acquired by each node exchanging message including its ID and priority to its one-hop neighbors. Then each node sorts the priority for all nodes within its one-hop neighborhood.

Step 2. A node i decides to become a clusterhead if either one of the following criteria is satisfied:

(1)Node i has the highest priority in its one-hop neighborhood.

(2) Node i is an one-hop neighbor of a node b. Furthermore, i has the highest priority in the one-hop neighborhood of b.

Step 3. All the clusterheads are added to the set D. Assign to each clusterhead $v \in D$ the transmission range $r_v = max_{v\prime \in N_i^1} d(v, v\prime)$. Then, all the cluster members are the leaf nodes of the broadcast tree.

The following Figure 1 gives an example of the clustering algorithm in a twenty-node network.

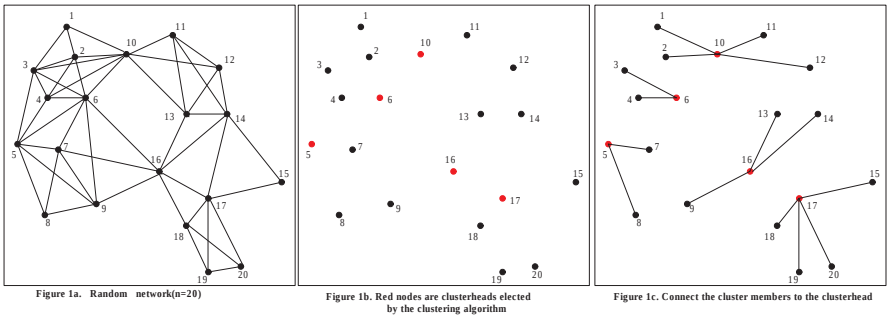


Figure 1. An example of clustering in a twenty-node network.

We give the IBIP(Improved BIP) algorithm.

The IBIP algorithm runs on the clusterhead set D. The cluster members elected by the clustering algorithm consist of the leaf nodes of the broadcast tree.

Step 1. Initially, the tree consists of only the source node s. We begin by determining the node that node s can reach with minimum expenditure of power.

Step 2. Determine which new clusterhead can be added to the tree at minimum additional power in the set of D. If node i is already in the tree, and node j is not yet in the tree, $(i, j \in D)$.

If $P_{i,j} < r_i^{\alpha}$, then we add j to the tree, where $P_{i,j}$ is the power required to support the transmission between node i and j, and r_i is the transmission range of node i which is assigned in the clustering algorithm.

Otherwise let $P\prime_{i,j} = P_{i,j} - max\{r_i^{\alpha}, P(i)\}$, where $P\prime_{i,j}$ represents the incremental power associated with adding j to the set of nodes to which node i is already transmitting. The pair $\{i, j\}$ that causes the minimum value of $P\prime_{i,j}$ is selected. Node i transmits at a power level sufficient to reach node j. Thus, one new node is added to the tree.

Step 3. This process continues until all nodes are included in the tree. The total power required to maintain this tree is the sum of the transmitted powers at each of the transmitting nodes.

The following Figure2 gives an example of the IBIP algorithm.

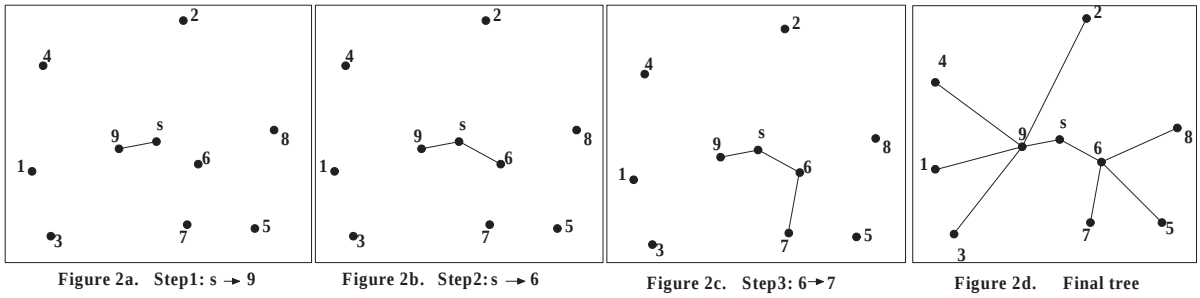


Figure 2a. Step1: s → 9 Figure 2b. Step2: s → 6 Figure 2c. Step3: 6→7 Figure 2d. Final tree

Figure 2. An example of IBIP algorithm in a nine-clusterhead network with source node s.

5 Algorithm Analysis and Simulation

Theorem 3. *The set of clusterheads elected by the clustering algorithm cover all network nodes.*

Proof: Because each node either has the highest priority among its one-hop neighbors or has a neighbor with the highest priority among one-hop neighbors of the node. By the definition of the dominating set, a node either becomes a dominator itself or elects the neighbor with the highest priority among one-hop neighbors of the node as a dominator. Thus the network has a dominating set elected by the algorithm. Therefore, we conclude that the sets of clusterheads elected by the clustering algorithm can cover all network nodes. □

Theorem 4. *The CBEEB algorithm has time complexity of $O(M^2)$, where $M \leq \frac{N}{d_{min}+1}$ is the number of clusterhead elected by the clustering algorithm and d_{min} is the smallest degree of nodes in the network.*

Proof: The clustering algorithm adopts a distributed clustering strategy. Thus, the time complexity of the entire network equals that of a single node $O(1)$. In the algorithm IBIP, it runs on the set of clusterheads. Obviously, the number of the clusterheads is smaller than $\frac{N}{d_{min}+1}$. Moreover, the IBIP algorithm is based on the Prim's algorithm. Thus the time complexity is $O((\frac{N}{d_{min}+1})^2)$. Therefore, the total time complexity of the CBEEB algorithm is $O(1 + M^2)$, that is $O(M^2)$. □

Theorem 5. *The overhead complexity of control messages in the network of the CBEEB algorithm is $O(N + M^3)$, where $M \leq \frac{N}{d_{min}+1}$ is the number of clusterhead elected by the clustering algorithm and d_{min} is the smallest degree of nodes in the network.*

Proof: At the beginning of the clustering algorithm, each node broadcasts a message to collect its information of its one-hop neighbors. Therefore, the overhead complexity of control messages in the network is $O(N)$. The IBIP runs on the set of clusterheads elected by the clustering algorithm based on the Prim's algorithm. Because we need to update the power $P\prime_{i,j}$ at each step of the algorithm, the message complexity is $O(M^3)$ when a straightforward implementation is used [1]. Therefore, we can conclude that the overhead complexity of control messages in the network of the CBEEB algorithm is $O(N + M^3)$. □

Comparing with the time complexity of the BIP algorithm is $O(N^2)$, and the message complexity is $O(N^3)$ [1]. Obviously, our algorithm CBEEB shows better effectiveness.

We now compare the performance of CBEEB algorithm with other six broadcast algorithms BLiMST [1], BIP [1], CM [3], SA [4], ESA [2]and GA [15] by simulations. The simulations are carried out in an ideal network generated over a $100 \times 100m^2$ area without consideration for packet lost. The number of network nodes changes from 25 to 250 with each increment of 25.

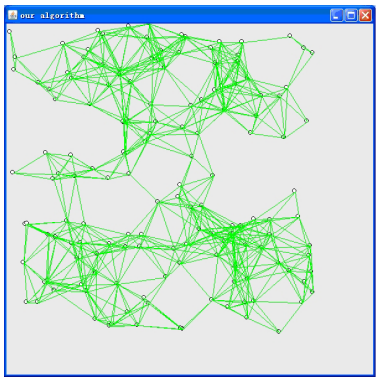


Fig 3a. The original network topology with 150 nodes

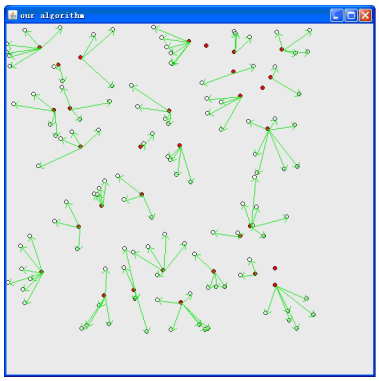


Figure 3b.The topology after running clustering algorithm

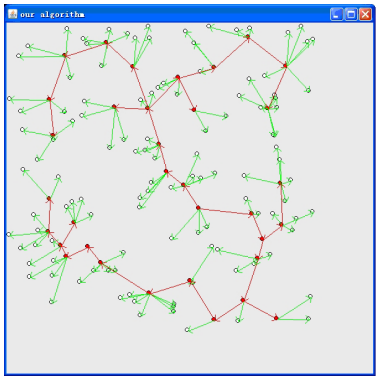


Figure 3c.The topology after running IBIP

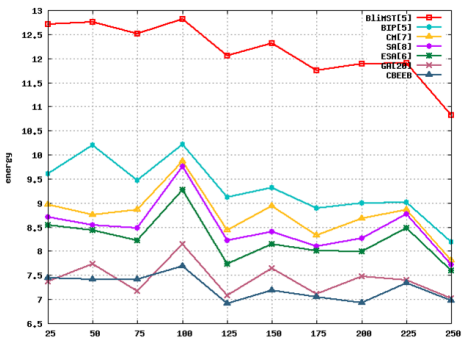


Figure 4.Energy cost comparison among seven algorithms

Figure 3 illustrates the backbone topology construction process using CBEEB. In Figure 4, we compare the energy cost of our algorithm with other six algorithms. The simulation results show that our algorithm makes further improvements on the energy cost compared with the BliMST and BIP algorithm. CM, SA, ESA and GA algorithm can improve the energy cost over the BIP algorithm. Moreover, we can see CBEEB performs better than the GA algorithm especially when the number of nodes is large.

6 Conclusions

In this paper, we propose a hybrid approach for the MEB problem in wireless networks. Compared with existing approaches, theoretical and experimental results show the superiority. As further work, we will develop fully distributed algorithms for MEB problem.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China for contract (11101243, 60373012), Natural Science Foundation of Shandong Province for contract(ZR2009GM009, ZR2009AM013), STPU of Shandong Province for contract (J10LG09) and TKP of Shandong Province for contract (2009GG10001014).

Bibliography

[1] J. E. Wieselthier, G. D. Nguyen, A. Ephremides, On the construction of energy efficient broadcast and multicast trees in wireless networks, *in: Proc. of INFOCOM 2000,* 585-594, 2000.

[2] S. M. Hashemi, Mohsen Rezapour, Ahmad Moradi, Two new algorithms for the Min-power Broadcast problem in static ad hoc networks, *Applied Math. and Compu.*, Vol.190, pp. 1657-1668, 2007.

[3] A. K. Das, R. J. Marks, M. EI-Sharkawi, P. Arabshi, A. Gray, A cluster-merge algorithm for solving the minimum power broadcast problem in large scale wireless networks, *in Proc. of the Milcom conference*, 13-16, 2003.

[4] A. K. Das, R. J. Marks, M. EI-Sharkawi, P. Arabshi, A. Gray, The minimum power broadcast problem in wireless networks: a simulated annealing approach, *in Proc. of WCNC*, 2057-2062, 2005.

[5] P. J. Wan, G. Calinescu, X. Y. Li, O. Frieder, Minimum-energy broadcast routing in static ad hoc wireless networks, *in Proc. of INFOCOM*, 1162-1171, 2001.

[6] A. K. Das, R. J. Marks, M. EI-Sharkawi, P. Arabshai, A. Gray, *r*-shrink: a heuristic for improving minimum power broadcast trees in wireless networks, *in Proc. of GLOBECOM* 523-527, 2003.

[7] M. Cagalj, J. P. Hubaux, C. Enz, Minimum-energy broadcast in all-wireless networks: NP-completeness and distribution issues, *in Proc. of MobiCom'02,* 172-182, 2002.

[8] I. Kang, R. Poovendran, Broadcast with heterogeneous node capability, *in Proc. of GLOBECOM* 4114-4119, 2004.

[9] A. K. Das, R. J. Marks, M. EI-Sharkawi, P. Arabshi, A. Gray, The minimum power broadcast problem in wireless networks: an ant colony system approach, *in Proc. of the IEEE CAS Workshop on Wireless Communications and Networking*, 5-6, 2002.

[10] I. Kang, R. Poovendran, Iterated local optimization for minimum energy broadcast, *in: Proc. of WiOpt*, 332-341, 2005.

[11] R. Montemanni, L. M. Gambardella, A. K. Das, The minimum power broadcast in wireless networks: a simulated annealing approach, *in Proc. of WCNC,* 2057-2062, 2005.

[12] S. Al-Shihabi, P. Merz, S. Wolf, Nested portioning for the minimum energy broadcast problem,*in Proc. of LION 2, LNCS 5313*, 1-11, 2008.

[13] S. Wolf, P. Merz, Evolutionary local search for the minimum energy broadcast problem,*in Proc. of EvoCOP'08, LNCS 4972*, 61-72, 2008.

[14] X. Wu, X. Wang, R. Liu, Solving minimum power broadcast problem in wireless ad hoc networks using genetic algorithm,*in Proc. of CNSR*, 203-207, 2008.

[15] Alok Singh, Wilson Naik Bhukya, A hybrid genetic algorithm for the minimum energy broadcast problem in wireless ad hoc networks, *Applied Soft Computing*, Vol.11, pp. 667-674, 2011.

Author index