Studies in Computational Intelligence 365

Roger Lee (Ed.)

Computers, Networks, Systems, and Industrial Engineering 2011

Springer

Roger Lee (Ed.)

Computers, Networks, Systems, and Industrial Engineering 2011

Studies in Computational Intelligence, Volume 365

Further volumes of this series can be found on our homepage: springer.com

Vol. 346. Weisi Lin, Dacheng Tao, Janusz Kacprzyk, Zhu Li, Ebroul Izquierdo, and Haohong Wang (Eds.)
Multimedia Analysis, Processing and Communications, 2011
ISBN 978-3-642-19550-1

Vol. 347. Sven Helmer, Alexandra Poulovassilis, and Fatos Xhafa
Reasoning in Event-Based Distributed Systems, 2011
ISBN 978-3-642-19723-9

Vol. 348. Beniamino Murgante, Giuseppe Borruso, and Alessandra Lapucci (Eds.)
Geocomputation, Sustainability and Environmental Planning, 2011
ISBN 978-3-642-19732-1

Vol. 349. Vitor R. Carvalho
Modeling Intention in Email, 2011
ISBN 978-3-642-19955-4

Vol. 350. Thanasis Daradoumis, Santi Caballé, Angel A. Juan, and Fatos Xhafa (Eds.)
Technology-Enhanced Systems and Tools for Collaborative Learning Scaffolding, 2011
ISBN 978-3-642-19813-7

Vol. 351. Ngoc Thanh Nguyen, Bogdan Trawiński, and Jason J. Jung (Eds.)
New Challenges for Intelligent Information and Database Systems, 2011
ISBN 978-3-642-19952-3

Vol. 352. Nik Bessis and Fatos Xhafa (Eds.)
Next Generation Data Technologies for Collective Computational Intelligence, 2011
ISBN 978-3-642-20343-5

Vol. 353. Igor Aizenberg
Complex-Valued Neural Networks with Multi-Valued Neurons, 2011
ISBN 978-3-642-20352-7

Vol. 354. Ljupco Kocarev and Shiguo Lian (Eds.)
Chaos-Based Cryptography, 2011
ISBN 978-3-642-20541-5

Vol. 355. Yan Meng and Yaochu Jin (Eds.)
Bio-Inspired Self-Organizing Robotic Systems, 2011
ISBN 978-3-642-20759-4

Vol. 356. Slawomir Koziel and Xin-She Yang (Eds.)
Computational Optimization, Methods and Algorithms, 2011
ISBN 978-3-642-20858-4

Vol. 357. Nadia Nedjah, Leandro Santos Coelho, Viviana Cocco Mariani, and Luiza de Macedo Mourelle (Eds.)
Innovative Computing Methods and their Applications to Engineering Problems, 2011
ISBN 978-3-642-20957-4

Vol. 358. Norbert Jankowski, Włodzisław Duch, and Krzysztof Grąbczewski (Eds.)
Meta-Learning in Computational Intelligence, 2011
ISBN 978-3-642-20979-6

Vol. 359. Xin-She Yang, and Slawomir Koziel (Eds.)
Computational Optimization and Applications in Engineering and Industry, 2011
ISBN 978-3-642-20985-7

Vol. 360. Mikhail Moshkov and Beata Zielosko
Combinatorial Machine Learning, 2011
ISBN 978-3-642-20994-9

Vol. 361. Vincenzo Pallotta, Alessandro Soro, and Eloisa Vargiu (Eds.)
Advances in Distributed Agent-Based Retrieval Tools, 2011
ISBN 978-3-642-21383-0

Vol. 362. Pascal Bouvry, Horacio González-Vélez, and Joanna Kolodziej (Eds.)
Intelligent Decision Systems in Large-Scale Distributed Environments, 2011
ISBN 978-3-642-21270-3

Vol. 363. Kishan G. Mehrotra, Chilukuri Mohan, Jae C. Oh, Pramod K. Varshney, and Moonis Ali (Eds.)
Developing Concepts in Applied Intelligence, 2011
ISBN 978-3-642-21331-1

Vol. 364. Roger Lee (Ed.)
Computer and Information Science, 2011
ISBN 978-3-642-21377-9

Vol. 365. Roger Lee (Ed.)
Computers, Networks, Systems, and Industrial Engineering 2011, 2011
ISBN 978-3-642-21374-8

Roger Lee (Ed.)

Computers, Networks, Systems, and Industrial Engineering 2011

Springer

Editor

Prof. Roger Lee
Central Michigan University
Computer Science Department
Software Engineering & Information
Technology Institute
Mt. Pleasant, MI 48859
U.S.A.
E-mail: lee1ry@cmich.edu

Guest Editors

Yung-Cheol Byun
Kiumi Akingbehin
Petr Hnetynka

ISBN 978-3-642-21374-8 e-ISBN 978-3-642-21375-5

DOI 10.1007/978-3-642-21375-5

Studies in Computational Intelligence ISSN 1860-949X

Library of Congress Control Number: 2011927931

Typeset & *Cover Design:* Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

The purpose of the 1st International Conference on Computer and Information Science (CNSI 2011) held on May 23–25, 2011 Jeju Island, Korea was to bring together researchers and scientists, businessmen and entrepreneurs, teachers and students to discuss the numerous fields of computer science, and to share ideas and information in a meaningful way. Our conference officers selected the best 22 papers from those papers accepted for presentation at the conference on order to publish them in this volume. The papers were chosen based on review scores submitted by members of the program committee, and underwent further rounds of rigorous review.

In Chapter 1, Eswar Bathini et al. Our results show the performance of dominating sets with two-hop neighborhood information which have improved the broadcast latency, packet delivery, and load on network.

In Chapter 2, HS Johal et al. In this paper, we study the effect of changing the retransmission mechanism on QoS parameters and propose a mechanism which would dynamically adjust the retry limits (which controls the retransmission policy) based upon network state (feedback force effect) and applications (driving force effect). We start by explaining the brief theory behind retransmission mechanism followed by some basic analysis which provides the motivation for the need to dynamically control the retransmission mechanism.

In Chapter 3, Soojin Park et al. This study proposes the mechanism in which the traceability between analysis model, design model and pattern elements is set in real time while applying the architecture pattern simultaneously to minimize the costly change of non-functional requirements. *NFR Tracer*, a tool that supports the automation of the above mentioned process, will also be introduced. NFR Tracer proposes the quantitative evaluation results on correctness and completeness automatically set on trace link, showing performance improvement.

In Chapter 4, Purushothaman Surendran et al. Here the objective of this paper is to analyze the detection performance of non-coherent detectors such as square law detector, linear detector and logarithmic detector in a background of white Gaussian noise (WGN) for Ultra Wide Band Short Range Radar in Automotive applications. It is assumed that the target is stationary and energy reflected from the target is distributed. The non-coherent detectors have almost similar detection probability in a background of white Gaussian noise. The performance of the detector is analyzed and simulation has been done in order to verify.

In Chapter 5, Yuji Hashiura et al. To solve the problems, we propose the following new modeling. First, we define a modeling method regarding relationship and transmission of effect between nodes. In addition, we consider a strength of cause and effect to analyze complex condition and situation. Second we define a new unified modeling language to express a complex graph-based qualitative simulation model. Our proposed method can give an analyzed result of an nonautonomous system where the parameter depends on time-pass.

In Chapter 6, Yeong-tae Song et al. Goal-oriented approach is a requirement analysis technique that supports early requirements analysis, which can be used to define a process for bridging EA requirements to ArchiMate. Therefore, we propose guidelines using goal-oriented approach and then apply them to the interoperability of prescriptions in a healthcare system.

In Chapter 7, Nandeeshkumar Kumaravelu et al. This paper analyzes the detection performance of non-coherent logarithmic detector in log normal and weibull clutter environment. The detection probability of the detector is obtained with different mean and variance value in log normal clutter environment, and with different shape parameter and scale parameter values in weibull clutter environment at different system bandwidth of 1GHz, 500MHz and 100MHz.

In Chapter 8, Eunyoung Cheon et al. To cope with such problems this paper suggests a regional matchmaking technique which can redistribute works in consideration of the characteristics of the works and the participant resources.

In Chapter 9, Myoung Jin Kim et al. In this paper, we propose an intelligent multi-agent model for resource virtualization (IMAV) to automatically allocate service resources suitable for mobile devices in cloud computing environment supporting social media services. Our model can recommend optimized resource allocation suitable for mobile devices through virtualization rules and multi-agent. IMAV consists of user agent, distributed agent, gathering agent, virtualization register agent manager and system resource manager.

In Chapter 10, Sung Ho Park et al. Therefore, with this technique, the maximum stack usage of each thread can be precisely determined during testing, and thus allowing the stack of each thread to be accurately optimized. Unlike the solutions proposed in previous research, this technique does not have problems such as limited availability, the possibility of undetectable stack usage, and memory fragmentation. Although this technique adds some overhead, the overhead is added only during the stack optimization process in the development phase. Also, despite the necessity for modification of the compiler and operating system, this technique is easy to implement. It can be implemented by slight modification of the existing compiler and operating system. To verify this technique, it was implemented and evaluated on the ARM platform by modifying the GNU ARM C compiler and the Ubinos, which is an operating system for memory-limited embedded systems.

In Chapter 11, Dong-Gil Kim et al. This paper presents a fault-tolerant clock synchronization method that can be used for a time-triggered wireless network. The effectiveness of the proposed method is demonstrated through a set of experiments.

In Chapter 12, Naoki Fukuta et al. In this paper, I present a preliminary empirical analysis of P2P-based semantic file sharing and retrieval mechanism. The mechanism enables us to utilize private ontologies for flexible concept-oriented se-mantic searches without loss of privacy in processing semantic matching among private metadata of files and the requested semantic queries. The private ontolo-gies are formed. I show the effectiveness of the use of private ontologies in meta-data-based file retrieval. Also I show that the mobile agent approach has rather less overheads in execution time when the network latency is relatively high, while it is small enough even when the network is ideally fast.

In Chapter 13, Takayuki Ito et al. The main problem is such iterative and agile process might need the tremendous number of communications and documents for collaboration because of iteration. Thus, software collaboration tools will be valuable for such situations in global software developments.

In Chapter 14, Cheoljong Yang et al. This paper addresses the gesture mode of interface and proposes an effective gesture language set capable of providing automotive control via hand gesture as natural but safe human-vehicle interface. Gesture language set is designed based on practical motions of single hand gesture. Proposed language set is optimized for in-vehicle imaging environment. Feature mapping for recognition is achieved using hidden Markov model which effectively captures the hand motion descriptors. Representative experimental results indicate that the recognition performance of proposed language set is over 99%, which makes it promising for real vehicle application.

In Chapter 15, Nga Dinh et al. This paper focuses on minimizing energy consumption under a guaranteed delay performance using a Power Efficiency based Delay Constraint (PEDC) mechanism which finds jointly-optimal parameters Tmin and Tmax. Simulation results validate that our proposed PEDC mechanism indeed minimizes consumed power for a specified frame response delay. In addition, in some cases, it can save up to 60% of consumed power compared to the original power saving mechanism in Mobile WiMAX.

In Chapter 16, Longzhe Han et al. In this paper, an adaptive scheduling algorithm has been proposed to overcome these limitations. This algorithm is based on H.264 Scalable Video Coding (H.264/SVC) and P2P paradigm to provide scalable video streaming services for heterogeneous users and self-adapt the dynamic network environment.

In Chapter 17, Namje Park et al. This paper is about enterprise DLP system having a function of coping with civil suits. We loaded various functions needed at discovery process of civil suit to conventional enterprise DLP system. The proposed system can reduce enterprise's damage by coping spontaneously about enterprise's litigation dispute.

In Chapter 18, Jong Hwan Lim et al. This study presents a methodology to perform the optimal sizing of a new and renewable hybrid generation system. The methodology aims at finding the configuration, among sets of system components, that meets the desired system requirements, with the lowest value of the energy cost. The availability of the methodology is demonstrated with the field data acquired from sets of experiments.

In Chapter 19, Sang-Goog Lee et al. This paper proposes a new smart web sensor model. Since the proposed smart web sensor is based on IEEE 1451.0, most of the existing sensor interfaces may be used, and the smart web sensor can be achieved using TEDS information. In addition, as XML is used, the web service is user friendly and a remote user can easily handle all kinds of information related to the sensor. This research presents a reference model for a smart web sensor and, to prove how valuable it is, a web-service using a gas sensor is utilized.

In Chapter 20, Sun-Myung Hwang et al. This paper is to propose evaluation scope computation model related with operating system evaluation to be enforced in the future.

In Chapter 21, Daniel Leonardo Niko et al. This paper presents a framework in per-forming geoprocessing to validate geographic references submitted by users and integrate them with the existing spatial database, i.e., parcel, hydrology, facility. The proposed framework will surely lead us to create a more complete spatial data of the damage that are to be utilized in a web based GIS application for effective disaster management.

In Chapter 22, Haeng-kon Kim et al. In this paper, we approach the EMG signals from electrodes placed on the forearm and recognizes the four kinds of motion. We also develops the prototype of a u-health device controller that controls hardware devices using EMG signal.

It is our sincere hope that this volume provides stimulation and inspiration, and that it will be used as a foundation for works yet to come.

May 2011 Guest Editors

Yung Cheol Byun
Kiumi Akingbehin
Petr Hnetynka

Contents

List of Contributors

Young-Hwan Bang
Korea Institute of Industrial Technology,
Korea

Hyunsuk Hwang
National University, Korea
Email: hhs@pknu.ac.kr

Eswar Bathini
Central Michigan University, USA
E-mail: bathi1e@cmich.edu

Jounghoon Beh
University of Maryland, USA
E-mail: jhbeh@umics.umd.edu

Hyung-Gi Byun
Kangwon National Univ., Korea,
byun@kangwon.ac.kr

Yung-Cheol Byun
National University, Korea,
Corresponding Author
E-mail: ycb@jejunu.ac.kr

Eunyoung Cheon
Chungnam National University, Korea
E-mail: eycheon@cnu.ac.kr

Nga T. Dinh
Bell Labs Seoul, Seoul, Korea
E-mail: Dinh.Nga@alcatel-lucent.com

Naoki Fukuta
Shizuoka University, Japan
E-mail: fukuta@cs.inf.shizuoka.ac.jp

Longzhe Han
Korea University, South Korea
E-mail: lzhan@korea.ac.kr

Yuji Hashiura
Yamagata University, Japan
E-mail: hashiura2009@e-activity.org

Chul-Ho Hong
Hoseo University, Korea
Email: chhong@hoseo.edu

Dowon Hong
Electronics & Telecommunications
Research Institute, Korea
Email: dwhong@etri.re.kr

Sun-Myung Hwang
Korea Institute of Industrial Technology,
Korea
E-mail: sunhwang@dju.ac.kr

Takayuki Ito
Nagoya Institute of Technology, Gokiso,
Nagoya
E-mail: ito.takayuki@nitech.ac.jp

Hoh Peter In
Korea University, South Korea
Email: hoh_in@korea.ac.kr

Hartinder Singh Johal
Lovely Professional University, India
E-mail: hs.johal@lpu.co.in

Chul Ung Kang
Jeju national University, Korea
cukang@jejunu.ac.kr

Soon Ju Kang
Kyungpook National University –
EECS, Korea

Dong-Gil Kim
National University, Korea
Haeng-Kon Kim
Catholic University of Daegu,
Korea
Email: hangkon@cu.ac.kr

Mikyoung Kim
Chungnam National University, Korea
E-mail: lpgrane@cnu.ac.kr

Myoung Jin Kim
Konkuk University, Korea
E-mail: tough105@konkuk.ac.kr

Changsoo Kim
National University, Korea,
Corresponding Author
cskim@pknu.ac.kr

Hyeon Soo Kim
Chungnam National University, Korea
hskim401@cnu.ac.kr

Younglok Kim
Sogang University, Korea
ylkim@sogang.ac.kr

Youngsoo Kim
Electronics & Telecommunications
Research Institute, Korea
E-mail: blitzkrieg@etri.re.kr

Jung-Ju Kim
Hoseo University, Korea
Email: ichromosome@nate.com

Jeong-Do Kim
Hoseo University, Korea
E-mail: jdkim@hoseo.edu

Mark Klein
MIT Sloan School of Management,
USA
E-mail: mklein@mit.edu

Hanseok Ko
Korea University, Korea
E-mail: hsko@ispl.korea.ac.kr

Seok-Jun Ko
Jeju National University, Korea
E-mail: sjko@jejunu.ac.kr

Seunghak Kuk
Chungnam National University, Korea
E-mail: triple888@cnu.ac.kr

Nandeeshkumar Kumaravelu
Jeju National University, Korea
E-mail: nandeeshforu@gmail.com

Kewal Krishan
Lovely Professional University, India
E-mail: kewal.krishan}@lpu.co.in

Dong Kyu Lee
Kyungpook National University –
EECS, Korea

Dongik Lee
National University, Korea
E-mail: dilee@ee.knu.ac.kr

Han Ku Lee
Konkuk University, Korea
Email: hlee@konkuk.ac.kr

Hyeoncheol Lee
Towson University, USA
E-mail: hlee23@students.towson.edu

Jong-Hun Lee
Daegu Gyeongbuk Institute of Science
& Technology, Korea

Roger Lee
Central Michigan University, USA
E-mail: lee1ry@cmich.edu

Sang-Goog Lee
The catholic University of Korea, Korea
E-mail: sg.lee@catholic.ac.kr

Yugyung Lee
University of Missouri at Kansas City,
USA
Email: leeyu@umkc.edu

Jong Hwan Lim
Jeju national University, Korea
Email: jhlim@jejunu.ac.kr

Tokuro Matsuo
Yamagata University, Japan
E-mail: matsuo@yz.yamagata-u.ac.jp

Amandeep Nagpal
Lovely Professional University,
E-mail: amandeep.nagpal@lpu.co.in

Daniel Leonardo Niko
National University, Korea
E-mail: daniel.gultom@gmail.com

Kun Hyun Park
Jeju national University, Korea
E-mail: 3313park@jejunu.ac.kr

Namje Park
National University, Korea
E-mail: namjepark@jejunu.ac.kr

Soojin Park
Sogang University, Korea
Email: psjdream@sogang.ac.kr

Sooyong Park
Sogang University, Korea
Email: sypark@sogang.ac.kr

Sung-Dae Park
Hoseo University, Korea
E-mail: sdpark@control.hoseo.ac.kr

Sung Ho Park,
National University – EECS, Korea
E-mail: slblue@ee.knu.ac.kr

Balraj Singh
Lovely Professional University, India
E-mail: balraj.13075@lpu.co.in

Sookyeong Song
Sogang University, Korea
E-mail: tamnacs@gmail.com

Yeong-tae Song
Towson University, USA
E-mail:ysong@towson.edu

Purushothaman Surendran
National University, Jeju, Korea

Cheoljong Yang
Vision Information Processing, Korea
University, Seoul, Korea
E-mail: cjyang@ispl.korea.ac.kr

Jongsung Yoon
Korea University, Korea
E-mail: jsyoon@ispl.korea.ac.kr

Hyo Gun Yoon
Konkuk University, Korea
E-mail: hgyun007@gmail.com

Using Dominating Sets with 2-Hop Neighborhood Information to Improve the Ad-Hoc On-Demand Distance Vector Routing

Eswar Bathini and Roger Lee

Abstract. Mobile Ad-Hoc network is a wireless network whose nodes in the network must coordinate among them to determine connectivity and routing. Many improvements were done to Ad-Hoc on-demand distance vector routing by using dominating sets to decrease the network load and flooding. But further with the increase of nodes at higher terrain dimensions the performance of Ad-Hoc on-demand distance vector routing decreases. To overcome this problem, we apply dominating sets with two-hop neighborhood information to Ad-Hoc on-demand distance vector routing protocol using GloMoSim-2.03 simulator, where route to the destination can be identified using the two-hop information at the dominating nodes where the packet to destination is present. Our results show the performance of dominating sets with two-hop neighborhood information which have improved the broadcast latency, packet delivery, and load on network.

Keywords: Total Dominant Pruning, Dominating Sets, two-hop Neighbor Information.

1 Introduction

Mobile IP and wireless networks accessing the fixed networks have provided support for the mobility. But it is still restrictive in forcing the connectivity at least to the core network. It puts impediments on supporting the true mobility in the network. In this connection, one area which is getting much attention in last couple of years is Mobile Ad Hoc Networks (MANETs). **A MANET (*Mobile Ad-Hoc***

Eswar Bathini
Software Engineering & Information Technology Institute,
MSIS Department, Central Michigan University
e-mail: `bathi1e@cmich.edu`

Roger Lee
Software Engineering & Information Technology Institute,
Computer Science Department, Central Michigan University,
e-mail: `lee1ry@cmich.edu`

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 1–9.
springerlink.com

Network) is a type of Ad-Hoc network with rapidly changing topology [2]**.** Since the nodes in a MANET are highly mobile, the topology changes frequently and the nodes are dynamically connected in an arbitrary manner.

There are two broad categories of unicast routing protocols for MANETs, proactive and reactive. With *proactive routing* (e.g., OLSR), nodes keep routing information to all nodes in the network, not subject to any existing data flow. OLSR is a link state protocol using an optimized broadcast mechanism for the dissemination of link state information. In *reactive routing* (e.g., AODV [3]), routes are found on demand and nodes find routes to their destinations as they are needed. Route discovery starts by broadcasting a *route request* (RREQ) message throughout the network. This message is relayed until it reaches a node with a valid route to the destination, or the destination itself. Once this happens, a *routereply* (RREP) message is sent back to the source by reversing the path traversed by the RREQ message. Only after receiving the corresponding RREP message can the source start sending packets to the destination. Reactive and proactive routing can be combined, resulting in *hybrid protocols* (e.g.,ZRP). In this case, routes to some nodes (usually the nearest ones) are kept proactively, while routes to the remaining nodes are found on-demand as in [3].

In this paper we evaluated the performance improvement of the Ad-Hoc on-demand distance vector routing algorithm by using the dominating sets and 2-hop neighborhood information i.e. Total Dominant Pruning [1] which increases performance at higher terrain ranges and improves broadcast latency, packet delivery, and load on network.

The rest of the paper is organized as follows. Section 2 is the related work. Section 3 deals with Dominating sets Section 4 Dominant pruning with 2-Hop neighborhood information. Section 5 Route Request in AODV using dominating sets with 2-hop Neighbor Information. Section 6 Shows the Simulations using GloMoSim 2.03 Section 7 Presents Conclusion.

2 Related Work

Powerful Broadcasting techniques in Ad-Hoc wireless networks have been extensively reviewed in [4]. A subset of nodes is called a dominating set if every node in the network is either in the set or a neighbor of a node in the set. In [6] it discussed about dominant pruning algorithm which uses 2-hop neighborhood information. The forward node list is selected in such a way that covers all the nodes within two hops.

Dominant Pruning [6] is a neighbor-designated method (i.e the sending node decides which adjacent nodes should relay the packet). The relaying nodes are selected using a distributed Connected Dominating Set algorithm, and the identifiers (IDs) of the selected nodes are piggy-backed in the packet as the forwarder list. A receiving node that is requested to forward the packet again determines the forwarder list.

Williams and Camp [8] have shown that neighbor information methods are preferred over other types of broadcast protocols. Between the two classes of neighbor information methods, Lim and Kim [6] show that the simplest form of neighbor-designated algorithm outperforms the simplest form of self-pruning, and Wu and Dai [9] show that an improved self-pruning technique outperforms the most efficient neighbor-designated algorithm (both algorithms based on the two-hop neighborhood information).

In Neighbor Information Methods [6], a node has partial topology information, which typically consists of the topology within two hops from the node (two-hop neighborhood). There are two main classes of methods in the category. In a neighbor designated method a node that transmits a packet to be flooded specifies which one-hop neighbors should forward the packet. In a self pruning method a node simply broadcasts its packet, and each neighbor that receives the packet decides whether or not to forward the packet.

Some of the improvements have been reported recently for dominant pruning [1, 11]. Lou and Wu [1] propose two enhancements to DP: total dominant pruning (TDP), and partial dominant pruning (PDP). Simulation results assuming an ideal MAC layer with which no contention or collisions occur show that both TDP and PDP. Improve dominant pruning (DP) in a static environment. A dynamic scenario is also evaluated, and DP is shown to perform better than both TDP and PDP. Spohn and Garcia-Luna-Aceves proposed enhanced dominant pruning (EDP) [12], which they applied to AODV to show its improvements compared to DP. Spohn and Garcia-Luna-Aceves also showed that EDP improves the performance of AODV in the context of directional antennas [12].

In our proposed method, we used dominating sets model and two hop neighborhood information to identify the Route Request forwarding nodes within the dominating set nodes with two hop neighborhood information to the destination. This approach decreases the overhead on route requests of the AODV by eliminating the redundant RREQ forwarding towards destination and also improves the performance of AODV at higher terrains with increase of number of nodes.

3 Dominating Sets

We use a undirected graph G= (V, E) consists of a set of vertices V represents set of wireless mobile nodes and E represents a set of edges. A set D $\subseteq$ V of vertices in a graph G is called a dominating set (DS) if every vertex $n_i \in$ V is either an element of D or is adjacent to an element of D [7]. If the graph induced by the nodes in D is connected, we have a connected dominating set (CDS). The problem of computing the minimum cardinality DS or CDS of any arbitrary graph is known to be NP-complete.

3.1 Identifying a Dominating Set from a Set Covering

Let (X, U) be an instance of the set cover problem with the universe U and the family of subsets X= {X_i: i $\in$ I}; we assume that U and the index set I are disjoint. Construct a graph G = (V, E) as follows: the set of vertices is V= I $\cup$ U, there is an edge {i, j} $\in$ E between each pair I, j $\in$ I, and there is also an edge {i, u} for each i ∈ I and u ∈ X_i. That is, G is a split graph: I is a clique and U is an independent set.

Now id C ={X_i : I ∈ D} is a feasible solution of the set cover problem for some subset D ⊆ I, then D is a dominating set for G, with ǀDǀ=ǀCǀ: First, for each u ∈ U there is an i ∈ D such that u ∈ X_i, and by construction, u and I are adjacent in G; hence u is dominated by i. Second, since D must be nonempty, each i ∈ I is adjacent to a vertex in D.

Conversely, let D be a dominating set for G. Then it is possible to construct another dominating set X such that ǀX1ǀ≤ǀDǀ and X1 ⊆ I : simply replace each u ∈ D ∩ U by a neighbor I ∈ I of u. Then C ={ X_i : I ∈ X1} is a feasible solution of the set cover problem, with ǀCǀ = ǀX1ǀ≤ǀDǀ.

The following Figure 1 shows the construction for U= {a, b, c, d, e}, I={1, 2, 3, 4}, X_1={a, b, c}, X_2={a. b}, X_3={b, c, d}, and X_4={c, d, e}.

In this figure 1 C= {X_1, X_4} is a set cover; this corresponds to the dominating set D={1, 4}.

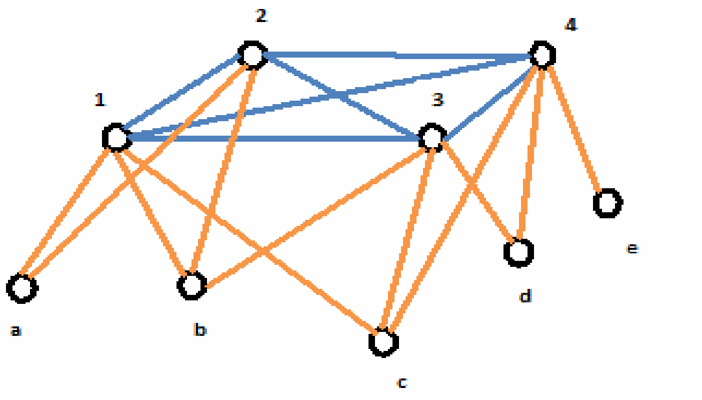


Fig. 1 Showing the nodes connected in sets.

D= {a, 3, 4} is another dominating set for the graph G. Given D, we can construct a dominating set X1={ 1, 3, 4} which is not larger than D and which is a subset of I. The dominating set X1 corresponds to the set cover C={X_1, X_3, X_4}.

4 Dominant Pruning Method with 2-Hop Neighborhood Information

We use a simple graph, G=(V,E) to represent an ad hoc wireless network, where V represents a set of wireless mobile hosts(nodes) and E represents set of edges(Links). The network is seen as unit desk graph [13], i.e. the nodes within the circle around node *v* (corresponding to its radio range) are considered its neighbors.

In dominant pruning methods sending node decides which adjacent nodes should relay the packet. The relaying nodes are selected using a selected connected dominating set algorithm, and the identifiers of the selected nodes are piggy backed in the packet as the forwarder list. A receiving node that is requested to forward the packet again determines the forwarder list. Nodes keep information about their two-hop neighbors, which can be obtained by the nodes exchanging their adjacent node list with their neighbors. Dominant Pruning is a distributed algorithm that determines a set cover based on the partial knowledge of the two-hop neighborhood. To decrease the transmissions the number of forwarding nodes must be minimized. Anyhow, the optimal solution is NP-complete and requires that nodes know the entire topology of the network.

Total Dominant Pruning which always makes the most efficient way to use neighborhood information. With total dominant pruning two-hop neighborhood information from the immediate sender is included in the header of every broadcast packet a node which receives that packet builds its forward list based on that information.

4.1 Total Dominant Pruning Algorithm

If node v can receive a packet piggybacked with N(N(u)) from node u, the 2-hop neighbor set that needs to be covered by n's forward node list F is reduced to U= N(N(v)) – N(N(u)). The total dominant pruning algorithm uses the above method to reduce the size of U and hence to reduce the size of F [1].

Total Dominant Pruning algorithm [1].

1. Nodes v uses $N(N(v))$, $N(N(u))$, $N(v)$ to obtain

   $U = N(N(v)) - N(N(u))$

   And

   $B = N(v) - N(u)$.

2. Node *v then calls the selection process to determine F.*

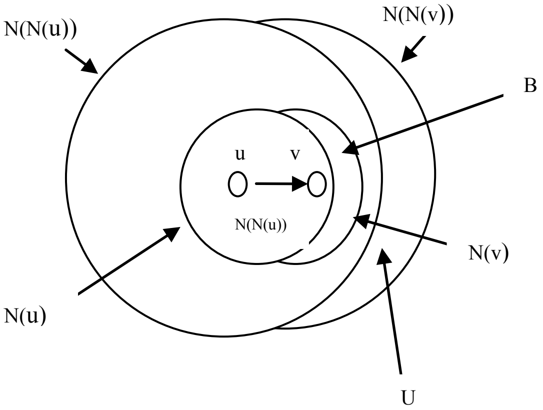


Fig. 2 Showing Total Dominant pruning

5 Route Request Forwarding Algorithm Using Dominating Sets with Two-Hop Information

```
Data: Nodes $n_i$, Source S, Destination D, Dominating set $D_{s,}$ $F_S$ Forwarder list,
  Begin:
        If valid route to D Then
                send RREQ to the Destination (D)
        Else
                Send RREQ to the dominating sets
                If dominating set nodes have valid route to destination then
                (Using TDP [10] shown in section 4.1 to reduce the sending of packets to the same nodes twice)
                Send RREQ to the destination
                End If
        End
```

6 Simulation Results and Parameters

The GloMoSim-2.03 [14] simulator is used to run the simulation and PARSEC [15] is used for simulation coding for the below table which summarizes the simulation parameters that we have used. The simulation parameters are declared in the glomosim-2.03 simulator code which is in the network module of the config.in file where the configuration settings are done. The simulation time was 15 minutes according to the simulation clock. A total of 60 and above nodes were randomly placed in the field of 1500 X 2000 m^2 and 2500 X 2000 m^2. The power range of each node is 250m.

Table 1 Simulation Parameters

Number of nodes	60
Terrain Size	1500 X 2000 m^2
Power range	250 m
Mac protocol	IEEE 802.11
Network protocol	AODV & Rough AODV
Transport Layer Protocol	UDP
Propagation function	FREE space
Node placement	Random
Simulation time	15M
Mobility interval	10-30 sec

The network consists of 60 nodes and they are spread over an area of 2000 X 1000 m^2 and also in 2500 X 2000 m^2 area. The source-destination pairs are chosen randomly among the nodes that are distributed in the network. The source nodes are always active flows during each simulation time and the destinations are randomly selected as needed. The performance is tested at every interval increasing the nodes and the pause time variations.

The simulations at different nodes and the terrain dimensions are shown in the graphs calculating various issues such as number of route requests, number of route replies, packet delivery ratio, number of control packets, end to end delay.

6.1 *Simulations of Aodv at Range (1500, 2000) with 60 to 120 Nodes in Network with an Increment of 10 Nodes at Each Point*

Table 2 Simulations of conventional Aodv

Number of nodes	Route Requests	Throughput	Collisions	End to End Delay	Control packets
60	1413	7751	5967	0.01661	1417
70	1662	7753	10686	0.01783	1666
80	1906	7753	14316	0.017970	1910
90	2083	7747	17586	0.015716	2087
100	2471	7758	28831	0.022295	2475
110	2691	7752	35972	0.023368	2695
120	2949	7765	46102	0.028067	2953

6.2 *Simulation of Dominating Sets with Two-Hop Neighbor's Information in Aodv at Range (1500, 2000) with 60 to 120 Nodes in the Network with an Increment of 10 Nodes at Each Point*

Table 3 Simulations of dominating sets with two-hop neighbors in Aodv

Number of nodes	Route Requests	Throughput	Collisions	End to End delay	Control packets
60	1081	3691	8391	0.034304	1088
70	1280	3611	10708	0.033615	1287
80	1654	3610	17902	0.038677	1661
90	1822	3610	24282	0.043407	1829
100	1921	3732	29601	0.163151	1928
110	2199	3738	29866	0.053361	2120
120	2441	3612	45531	0.041761	2448

At this point of simulation when we compare the results of the Conventional Aodv with the improved Aodv the results shows that with 60 nodes till 120 nodes. When the results are compared-

1) The route requests have been decreased in the in each and every increment of the nodes showing the decrease of Network Load while transmission of packets from source to destination.
2) The Throughput between number of packets originated by the application layer sources and the number of packets received by the sinks at final destinations was improved.
3) Packet collisions in both the conventional and improved AODV for some point of nodes collisions in Aodv are decreased and for some point of nodes collisions in improved Aodv decreases and for increasing of nodes in improved Aodv collisions decreased.
4) End-to End Delay in both the conventional Aodv and the improved Aodv shows approximately equal performances at each node levels.
5) Total number of control packets at each level decreased in the improved Aodv.

7 Conclusions and Future Work

In this paper we used dominating sets with two hop information to send the route request to the dominating nodes and hence the requests are forwarded to the destinations through the two hop information available at the dominating node to the destination which improved the packet latency and decreased the load on the network by reducing the route requests and collisions at some intervals. We used the dominating sets and total dominant pruning algorithm for the two hop information which regulates in sending the packets to the nodes which already received the packets. Our results show that Aodv has been improved in various aspects of decreasing network load, packet latency and throughput. Further research can be done using the rough set theory concepts in Aodv to reduce the unnecessary broadcasting in the network.

References

[1] Lou, W., Wu, J.: On reducing broadcast redundancy in adhoc wireless networks. IEEE Transactions on Mobile Computing 1(2) (April-June 2002)

[2] Mobile Ad Hoc Networking: Routing Protocol Performance Issues and Evaluation Considerations. Request For Comments 2501, `http://www.ietf.org`

[3] Perkins, C.E., Royer, E.M.: Ad-hoc on-demand distance vector routing. Technical report, Sun Micro Systems Laboratories, Advacnced Developmenet Group, USA

[4] Calinescu, G., Mandoiu, I., Wan, P.J., Zelikovsky, A.: Selecting Forwarding Neighbors in Wireless Ad Hoc Networks. In: Proc. ACM Int'l Workshop Discrete Algorithms and Methods for Mobile Computing (DIALM 2001), pp. 34–33 (December 2001)

[5] Perkins, C., Royer, E.M.: Ad-Hoc O-Demand Distance Vector Routing. In: Proc. Second IEEE Workshop Mobile Computing Systems and Applications (WMCSA), pp. 90–100 (February 1999)

[6] Lim, H., Kim, C.: Flooding in Wireless Ad hoc Networks. Computer Comm. J. 24(3-4), 353–363 (2001)

[7] Haynes, T.W., Hedetniemi, S.T., Slater, P.J. (eds.): Fundamentals of Domination in Graphs. Marcel Dekker, Inc., New York (1998)

[8] Williams, B., Camp, T.: Comparison of broadcasting techniques for mobile ad hoc networks. In: Proceedings of MOBIHOC, pp. 194–205 (2002)

[9] Wu, J., Dai, F.: Broadcasting in ad hoc networks based on self-pruning. In: INFOCOM (2003)

[10] Stojmenovic, I., Seddigh, S., Zunic, J.: Dominating Sets and Neighbor Elimination Based Broadcasting Algorithms in Wireless Networks. IEEE Trans. Parallel and Distributed Systems 13(1), 14–25 (2002)

[11] Spohn, M.A., Garcia-Luna-Aceves, J.J.: Enhanced dominant pruning applied to the route discovery process of on demand routing protocols. In: Proceedings of the 12th IEEE ICCCN (2003)

[12] Spohn, M.A., Garcia-Luna-Aceves, J.J.: Enhancing the route discovery process of on-demand routing in networks with directional antennas. In: Proceedings of The Third IFIP-TC6 Networking Conference (2004)

[13] Clark, B.N., Colbourn, C.J., Johnson, D.S.: Unit disk graphs. Discrete Math. 86, 165–177 (1990)

[14] GloMoSim. Global Mobile Information Systems Simulation Library, `http://pcl.cs.ucla.edu/projects/glomosim/`

[15] PARSEC (Parallel Simulation Environment for complex systems), `http://pcl.cs.ucla.edu/projects/parsec/`

Dynamically Controlling Retransmission Mechanism for Analysing QoS Parameters of IEEE 802.11 Networks

Hartinder Singh Johal, Balraj Singh, Amandeep Nagpal, and Kewal Krishan

Abstract. Wireless medium being error prone, the task of handling retransmissions is very important and it is assigned to the MAC subtype of IEEE 802.11. In this paper, we study the effect of changing the retransmission mechanism on QoS parameters and propose a mechanism which would dynamically adjust the retry limits (which controls the retransmission policy) based upon network state (feedback force effect) and applications (driving force effect). We start by explaining the brief theory behind retransmission mechanism followed by some basic analysis which provides the motivation for the need to dynamically control the retransmission mechanism.

Keywords: RTS/CTS, retry limits, threshold, jitter, delay, throughput.

1 Introduction

The use of long and short retry limits provide a limit to number of retransmissions, thereby enabling frames to reach their destination in case there are earliest attempts results in frames getting lost or waiting timer expires [1]. However, there is a cost associated with retry count values. As we increase the values of retry counts, losses may decrease but on other hand, delays may increase significantly [2]. Therefore, we perform a basic analysis as done previously for RTS Threshold [3]. We use network topology as depicted in fig 1. TCP and UDP data frames are sent with application level sizes of 512 bytes and 80 bytes, respectively. RTS Threshold value is kept as 250 throughout all the experiments, since we are mainly interested in evaluating retry limits. This also means that frames of TCP traffic will follow retransmissions according to long retry count while frames of UDP traffic will work according to short retry count [4][5]. QoS parameters for the following cases are analysed. Case 1: Short Retry Count = 1, Long Retry Count = 10. We study effect of reducing SRL to 1 and increasing LRC to 10 on QoS parameters for both

Hartinder Singh Johal · Balraj Singh · Amandeep Nagpal · Kewal Krishan
Department of Computer Science and Engineering, Lovely Professional University
144402 Phagwara, India
e-mail: {hs.johal,balraj.13075,amandeep.nagpal,
kewal.krishan}@lpu.co.in

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 11–23.
springerlink.com

TCP and UDP traffic. Case 2: Short Retry Count = 10, Long Retry Count = 1. We study effect of increasing SRL to 1 and decreasing LRC to 1 on QoS parameters for both TCP and UDP traffic. Case 3: Short Retry Count = 10, Long Retry Count = 10 [6]. We study effect of keeping both SRL and LRC to 10 on QoS parameters for both TCP and UDP traffic. By observing the delay, jitter, loss and throughput graphs, it is evident that delay and jitter reduces when we use a small SRL for small frames i.e. UDP traffic [7]. However, loss for such traffic increases significantly when compared with the case when SRL is large [8]. Increase loss makes the throughput to decrease and throughput fluctuates very frequently.

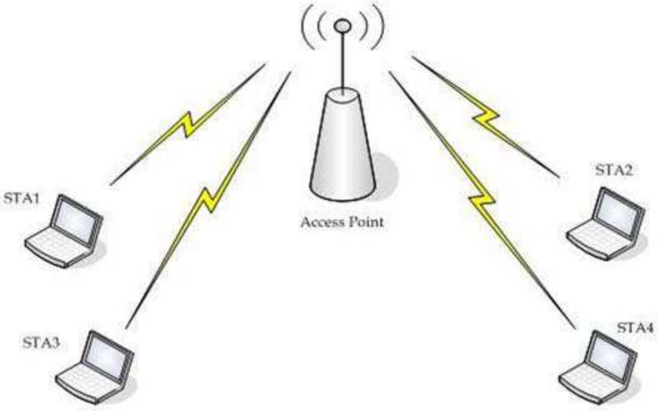


Fig. 1 Simple infrastructure Based Wireless Network

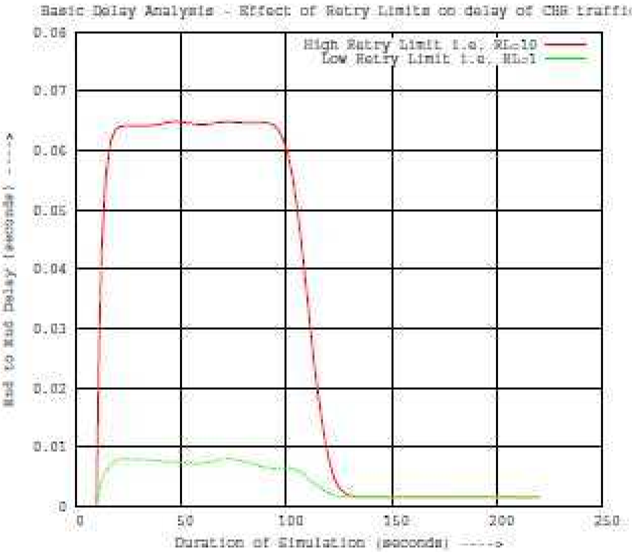


Fig. 2 Delay vs Time (Basic Analysis)

2 Proposed Scheme

Proper setting of SRC/LRC values reduces loss in frames, but if its values are poorly configured, then there is always a possibility that we are unnecessarily paying a cost [9]. For instance, for real time UDP traffic, a frame which has finally reached its destination after too many retransmissions and in the process its delay becoming too large to be of any use for the application [10]. Like the RTS Threshold, the dilemma here is to decide an appropriate value for SRC/LRC which would neither increase the delay nor increase the loss. As in the case with RTS/CTS, the only differentiation provided is the frame size which obviously is not a effective criteria as it neither captures the network state (driving force effect) nor reflects the application requirements (driving force effect) [11]. In its present form, retransmission policy based on statically pre-configured retry limits fails miserably against the proposed Progressive QoS Enhancement model. Henceforth, we propose a very simple scheme for QoS enhancement by efficiently adjusting the values of SRC/LRC [12]. So, accordingly, we incorporate the effects of driving forces (application) and feedback forces (network state). The driving force effect is implemented by using different SRL/LRL values for different applications or in other words, different application traffic is controlled by different SRC/LRC values for retransmission behaviour [13]. The feedback force effect is captured by incrementing or decrementing the values of SRC/LRC depending upon retransmission and successful transmission after a retry attempt counts, respectively as shall be explained head [14]. The process of incrementing and decrementing is done dynamically over the complete network operation [15]. As explained before, we follow a step by step approach by first providing the necessary feedback mechanism and thereafter employing the effect from the application.

2.1 Feedback Force Effect

In this step, we control the retransmission mechanism based on the current network conditions. Unlike the legacy mode of IEEE 802.11 operation, now the network runs in three phases namely Retry Increment, Retry Decrement and No Change [16]. The basic principle is that do not keep retry limits very high, rather increase them dynamically only when there is a frame drop due to RETRY LIMIT EXCEEDED reason code. At the same time, when frames are getting transmitted successfully during retransmission attempts, then decrease the value of retry limits. The above intuition can be very easily explained. If there is a frame drop due to RETRY LIMIT EXCEEDED reason code error, then it means that we have configured retry limits with an underestimated value. On other hand if frames are getting transmitted in first attempt then there is no point keeping the retry limit large as it may increase delay for real time traffic as explained above [17]. During the network operation, the state of the network changes dynamically from one state to another based on the retransmission attempts and successful sends on retry attempt. Short retry counter is changed when there is a retransmission attempt for either RTS frame or DATA frame, since RTS frame is of small size while Long retry counter is changed only in case of retransmission attempt for DATA frame [18]. Instead of incrementing and/or decrementing retry limits by 1, we use different values for increment and decrement namely incrementLevel and decrementLevel. This is useful for two reasons. Firstly, successive retry attempts and successful sends in retransmissions may cause retry limits to fluctuate between two values. Secondly and more importantly, more control is provided over retransmission behavior. For instance, increasing/decreasing retry limits by 1 will change the retry limit marginally which may not be significantly useful in a wireless network which is experiencing large number of retransmissions. Presently, the two variables incrementLevel and decrementLevel are pre-configured with values 3 and 1, respectively [19]. However, an extension of the proposed scheme could be to make them dynamic. As mentioned before with the proposed scheme, network operates in one of the following three phases at any given time:1. Retry Increment - Network switches to this phase when there is either RTS frame or data frame is lost due to RETRY LIMIT EXCEEDED reason code. The associated retry limit is incremented by incrementLevel value. 2. Retry Decrement Network switches to this phase when either retransmitted RTS frame or retransmitted data frame is successfully sent. The associated retry limit is decremented by decrementLevel value. 3. No change as long as there is no retransmission attempts and all frames are getting transmitted in their first transmission attempt, network operates in this phase.

The assumption in above proposed scheme is that with every successful sending of frame during retransmissions, the network has overcome from a state which was the cause for the successive retransmissions. Similarly, when there is a frame drop due to RETRY LIMIT EXCEEDED reason code, then we assume that network has entered into a state where likelihood of retransmissions would be more in future, so we increase the retry limit. Although, these assumptions may

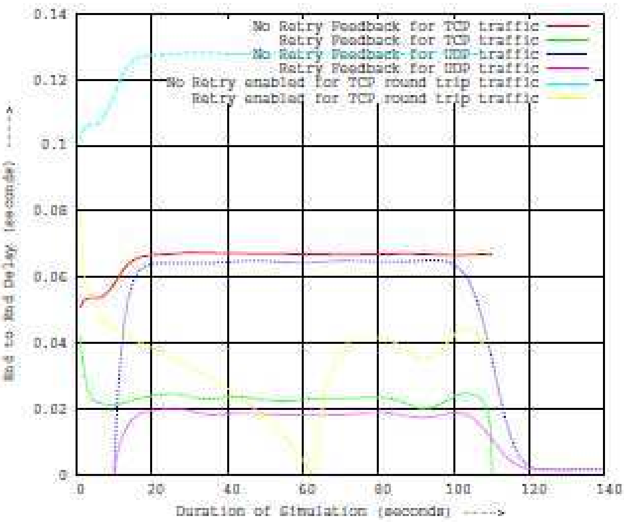


Fig. 3 Delay vs Time (SRC, LRC) = 0

not hold true at all times, but is fairly good enough for a scheme which intends to be simple. The feedback from the network accounts for feedback force effect according to the Progressive QoS Enhancement model [20]. Having explained the proposed mechanism, next we analyze its effect on QoS parameters in the 802.11 wireless networks. For performing analysis of above proposed scheme, we took two sets of plots. 1. In first set, we plot QoS parameter for all traffic types (TCP, UDP, ack) keeping SRC/LRC values constant. The aim here was to study to effect of proposed scheme on various traffic flows keeping SRC/LRC value fixed. 2. In second set, proposed scheme was studied for all SRC/LRC values keeping the traffic flow same [21][22]. The aim here was to study the combined effect of proposed scheme with varying SRC/LRC values. The fig 3 shows that the proposed scheme reduces delay for all traffic types while when SRC/LRC is set to 10. It can be observed that the proposed scheme reduces the delay substantially. The proposed scheme stabilizes the network and makes it to operate at an optimum level of delay. By stabilization, we mean that for (SRC, LRC) = (10, 1) delay is reduced but for (SRC, LRC) = (1, 10), delay increased slightly. The increase is due to the fact that proposed scheme increases the SRC value during the network operation causing delay to increase for UDP traffic. We analysed Loss graphs for all traffic types for (SRC, LRC) = (1, 10) and for all UDP (CBR) traffic for all values of (SRC, LRC) = (10, 10; 10, 1; 1, 10). Unlike previously, we have chosen (SRC, LRC) = (1, 10) instead of (SRC, LRC) = (10, 10) since more errors are expected when SRC value is less. The proposed scheme reduces loss when operated with (SRC, LRC) = (1, 10), although delay and jitter for this case is slightly increased. This is quite obvious since our proposed scheme will increase the SRC value which accounts for reduced loss. The increased loss is due to the fact that network operation increases SRC value for reducing the delay as shown in delay analysis. We analysed throughput graphs for all traffic types for (SRC,

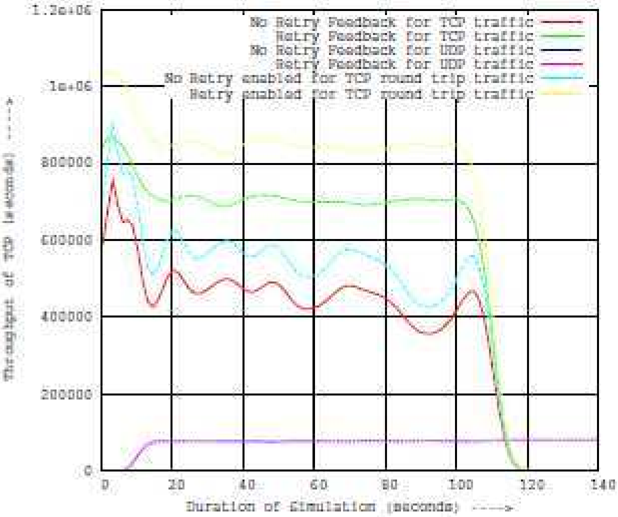


Fig. 4 Throughput vs Time, (SRC, LRC) = (1, 10)

LRC) = (1, 10) and for all TCP (FTP) traffic for all values of (SRC, LRC) = (10, 10; 10, 1; 1, 10). The fig 4 shows that the proposed scheme enhances the throughput for TCP and its corresponding ack traffic which fluctuates quite often when scheme is not used. This is quite obvious since our proposed scheme will switch the network operation from increase SRC value, thereby stabilizing the throughput.

2.2 Driving Force Effect

Having proposed a scheme for capturing the network state and controlling retransmission mechanism accordingly, we now further enhance the scheme by incorporating requirements of the application (driving force effect). On similar lines to that for RTS Threshold, we replace the classical concept that same SRC/LRC parameters control retransmission behavior for all frames.

Using same SRC/LRC value for all types of traffic would mean that there is no differentiation among various classes of traffic. Considering the fact that applications with varying requirements would be transmitted, the above classical concept is poorly designed. Henceforth, we propose that different applications with varying requirements shall have their own values for SRC/LRC values. So, frames of i^{th} application will have SRC_{th} (i) and LRC_{th} (i) as its SRC/LRC values. Since we are running two types of application namely the TCP based FTP application and UDP based CBR application, so we have four SRC/LRC values namely SRC_{th} (TCP FLOW), SRC_{th} (UDP FLOW), LRC_{th} (TCP FLOW) and LRC_{th} (UDP FLOW). Using such a scheme means that frames from each application can be controlled for retransmission mechanism independent of other applications. For instance, taking example of application flows which we have used, the TCP packet size is 512 bytes and UDP packet size is 80 bytes. Now, we can have an RTS_{th} = 250 fixed for all application types and study effect of application dependent SRC/LRC values on QoS parameters.

Additionally, we incorporate the feedback mechanism as discussed above along with the proposed application based differentiation scheme. The feedback mechanism is designed for each application flows and is selectively applied for different application flows. Since we have two kinds of applications each running over TCP and UDP, so like the case with RTS Threshold we have the following four cases: Case 0: No Feedback applied. In this case there is no feedback mechanism being applied. However, the application level differentiation (i.e. different SRC/LRC for different applications) is present and is studied without any influence from network. Case 1: Feedback only applied to TCP traffic. In this case there is feedback mechanism being applied only for TCP traffic. The application level differentiation (i.e. different SRC/LRC values for different applications) is present. Study is done to explore behavior of different traffic flows in this setup. Case 2: Feedback only applied to UDP traffic. In this case there is feedback mechanism being applied only for UDP traffic. The application level differentiation (i.e. different SRC/LRC values for different applications) is present. Study is done to explore behavior of different traffic flows in this setup. Case 3: Feedback only applied to both TCP and UDP traffic. In this case there is feedback mechanism being applied for both TCP and UDP traffic. The application level differentiation (i.e. different SRC/LRC values for different applications) is present. Study is done to explore behavior of different traffic flows in this setup.

Network feedback follows the transition similar to that discussed before, the only difference being that now application effect has been also introduced. In order to implement the four cases discussed above, we introduce a new operating mode variable (useRetryAlgorithmMode) which can be configured by application user with values 0, 1, 2, , n and n+1, considering that there are n applications. Here, 0 means feedback is not applied for any traffic, n means that feedback is applied for nth application and n+1 means that feedback is applied for all traffic.

Before the transfer of any frame, each frame undergoes our proposed network feedback algorithm according to the value of useRetryAlgorithmMode. The value of SRC/LRC incremented or decremented based on the decision made by the proposed algorithm. Thereafter, all associated frames uses updated values of SRC/LRC. The algorithm for the proposed scheme is briefed below:

```
  reTransmitRTS(packet *pktRTS )
  /* About to resend RTS packet to PHY layer */
  {
  /* Current retry limit is more than configured SRC */
  if( src [pktRTS →type] ≥ macmib.getSRC(pktRTS →type)
)
  /* Increment the SRC value by increaseLevel*/
  macmib .setSRC(macmib.getSRC(pktRTS →type) +
increaseLevel, pktRTS →type);
  else
  /* Decrement the SRC value by decreaseLevel*/
  macmib setSRC(macmib.getSRC(pktRTS →type) -
decreaseLevel, pktRTS →type);
  if( macmib .getSRC(pktRTS → type) ≤ MIN RETRY
THRESHOLD)

  macmib .setSRC(MIN RETRY THRESHOLD , pktRTS →type);
  if( macmib .getSRC(pktRTS → type) ≥ MAX RETRY
THRESHOLD)
  macmib .setSRC(MAX RETRY THRESHOLD , pktRTS →type);
  }
  .
  .
  reTransmitDATA(packet *pkt )
  /* About to resend DATA packet to PHY layer */
  {
  if( pkt →size() ≥ macmib .getRTSThreshold() )
  {
  /* Current retry limit is more than configured LRC */
  if( lrc [pkt →type] ≥ macmib.getLRC(pkt →type) )
  /* Increment the LRC value by increaseLevel*/
  macmib .setSRC(macmib.getSRC(pktRTS →type) +
increaseLevel, pktRTS →type);
  else
  /* Decrement the SRC value by decreaseLevel*/
  macmib .setLRC(macmib.getLRC(pkt →type) -
decreaseLevel, pkt →type);
  if( macmib .getLRC(pkt → type) ≤ MIN RETRY THRESHOLD)
  macmib .setLRC(MIN RETRY THRESHOLD , pkt → type);
  if( macmib .getLRC(pkt → type) ≥ MAX RETRY THRESHOLD)
  macmib .setLRC(MAX RETRY THRESHOLD , pkt → type);
  }
  else
  {
  /* Current retry limit is more than configured SRC */
  if( src [pktRTS → type] ≥ macmib.getSRC(pktRTS →
type) )
  /* Increment the SRC value by increaseLevel*/
  macmib setSRC(macmib.getSRC(pktRTS → type) +
increaseLevel, pktRTS →
  type);
  else
  /* Decrement the SRC value by decreaseLevel*/
  macmib setSRC(macmib.getSRC(pktRTS → type)
decreaseLevel, pktRTS → type);
  if( macmib .getSRC(pktRTS →type) ≤ MIN RETRY
THRESHOLD)
  macmib .setSRC(MIN RETRY THRESHOLD , pktRTS →type);
  if( macmib .getSRC(pktRTS →type) ≥ MAX RETRY
THRESHOLD)
  macmib .setSRC(MAX RETRY THRESHOLD , pktRTS →type);
  }
  }
```

The application requirements consideration accounts for the driving force effects and the feedback from the network accounts for feedback force effect according to our proposed QoS Enhancement model. The decision for adjusting retransmission mechanism based on the network feedback as well as the application. Having explained the proposed mechanism for adjusting retransmission mechanism, next we analyze its effect on QoS parameters in the 802.11 wireless networks. For performing analysis of above proposed scheme, we follow the approach as: 1. Studying feedback force effect: In first case we keep the driving force effect i.e. SRC/LRC for various applications constant and study the effect of differentiating network feedback (four cases discussed above).

The aim is to study the network feedback when application dependent SRC/LRC values are kept same. 2. Studying driving force effect: In second case, we keep the network feedback effect i.e. the four differentiating cases as constant and making the driving force effect varying by setting different values of SRC/LRC for different applications.

The aim is to study the application dependent differentiating scheme when same feedback is applied Delay Analysis: Graphs for all traffic types for the case when SRC_{th} [TCP FLOW, UDP FLOW] = (5, 5) & LRC_{th} [TCP FLOW, UDP FLOW] = (5, 5) and feedback is applied to all traffic types. Next graphs are for all UDP (CBR) traffic for all values of (SRC_{th}[TCP FLOW,UDP FLOW], LRC_{th}[TCP FLOW,UDP FLOW]) = (10,1, 10,1 ; 1,10, 1,10 ; 5,5, 5,5) with different cases of applying feedback effect. It is quite interesting to observe that delay is not reduced when feedback is applied only for UDP traffic. This is due to the fact that competing TCP traffic makes the above feedback ineffective. But, when feedback is applied to only TCP traffic, delays for UDP traffic is reduced, almost same reduction when feedback is applied for both TCP & UDP traffic. With no feedback is applied and we study the effect of using different values of (SRC_{th} [TCP FLOW,UDP FLOW], LRC_{th} [TCP FLOW,UDP FLOW]) for UDP traffic. It is quite interesting to observe here that delay is not reduced even when SRC is reduced because competing TCP traffic SRC/LRC is set to 10.

Decreasing SRC/LRC for TCP traffic reduces delay for UDP traffic. Further we study the effect of feedback being applied for only TCP traffic. Delay for all cases is found to be stabilized i.e. between the upper and lower delays. Lastly, we study the effect of feedback being applied for only UDP traffic and it is observed that delay is less for the case when SRC/LRC for TCP traffic is less. Jitter graphs are analysed for all traffic types for the case when SRC_{th} [TCP FLOW, UDP FLOW] = (5, 5) & LRC_{th} [TCP FLOW, UDP FLOW] = (5, 5) and feedback is applied to all traffic types. Next graphs are for all UDP (CBR) traffic for all values of (SRC_{th} [TCP FLOW,UDP FLOW], LRC_{th} [TCP FLOW,UDP FLOW]) = (10,1, 10,1 ; 1,10, 1,10 ; 5,5, 5,5) with different cases of applying feedback effect.

We record the effect of proposed scheme when it is applied for all four cases discussed above, keeping SRC_{th} [TCP FLOW,UDP FLOW] = (5,5) & LRC_{th} [TCP FLOW,UDP FLOW] = (5, 5) for all UDP traffic. Although the improvement is less but a closer look tells us that jitter is reduced when feedback is applied only for both UDP & TCP traffic when compared with the case when no feedback is applied. With no feedback applied and we study the effect of using different values of (SRC_{th} [TCP FLOW, UDP FLOW], LRC_{th} [TCP FLOW, UDP FLOW]) for UDP traffic.

We observe a similar trend as observed in delay analysis that jitter is not reduced even when SRC is reduced because competing TCP traffics SRC/LRC is set to 10. Decreasing SRC/LRC for TCP traffic reduces jitter for UDP traffic, even if SRC/LRC is high for UDP traffic. In fig 18 we study effect of feedback being applied for only TCP traffic. Jitter for the case when SRC/LRC for TCP is high is found to be less because of the applying of feedback. Lastly, we study the effect of feedback being applied for only UDP traffic and it is observed that jitter is less for the case when SRC/LRC for TCP traffic is less, even though SRC/LRC for UDP is more because feedback will anyway reduce this value during network operation. Loss Analysis: Below are the loss graphs for all traffic types for the case when SRC_{th} [TCP FLOW,UDP FLOW] = (10,1) & LRC_{th} [TCP FLOW,UDP FLOW] =

(10,1) and feedback is applied to all traffic types. Next graphs are for all UDP (CBR) traffic for all values of (SRC$_{th}$ [TCP FLOW,UDP FLOW], LRC$_{th}$ [TCP FLOW,UDP FLOW]) = (10,1, 10,1 ; 1,10, 1,10 ; 5,5, 5,5) with different cases of applying feedback effect. The fig 20 shows the effect of proposed scheme on frame loss when feedback is applied for all four cases discussed above, keeping SRC$_{th}$ [TCP FLOW,UDP FLOW] = (10,1) & LRC$_{th}$ [TCP FLOW,UDP FLOW] = (10,1) constant for all UDP traffic. It can be observed that scheme significantly reduces loss for UDP traffic when feedback is applied to UDP traffic and both UDP & TCP traffic. It can also be observed that loss of UDP traffic is not reduced when feedback is applied only for TCP traffic.

We study effect of differentiating SRC/LRC values keeping no feedback for all UDP traffic. Quite interestingly, we observe that loss of UDP traffic is less when SRC/LRC for TCP traffic is less. On contrarily, loss is more even if SRC/LRC values for UDP is less if competing TCP traffics SRC/LRC is high. We observe loss on UDP traffic for different values of RTS$_{th}$ when there is feedback only for TCP traffic. It is observed that frame losses are less when SRC/LRC of competing TCP traffic is less.

Lastly, we observe loss on UDP traffic for different values of SRC/LRC when there is feedback for only UDP traffic. It is observed that losses get reduced for the case which has higher value for SRC/LRC for UDP traffic due to applying of the proposed scheme. Throughput analysis is performed for all traffic types for the case when SRC$_{th}$ [TCP FLOW, UDP FLOW] = (1, 10) & LRC$_{th}$ [TCP FLOW, UDP FLOW] = (1, 10) and feedback is applied to all traffic types. Next graphs are for all TCP (FTP) traffic for all values of (SRC$_{th}$[TCP FLOW,UDP FLOW], LRC$_{th}$[TCP FLOW,UDP FLOW]) = (10,1, 10,1 ; 1,10, 1,10 ; 5,5, 5,5) with different cases of applying feedback effect.

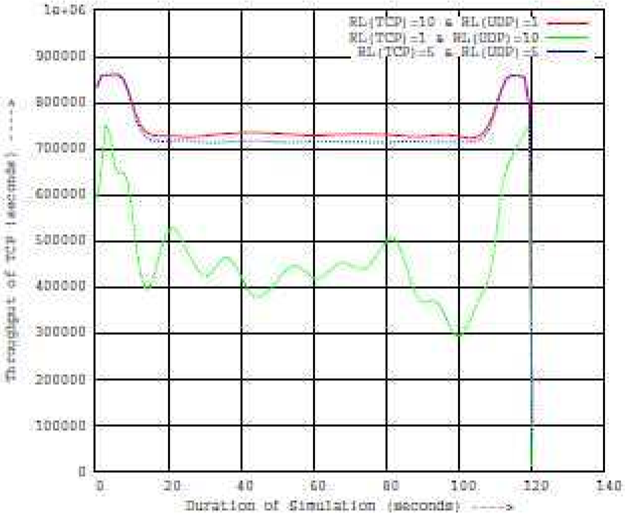


Fig. 5 UseRetryAlgorithm = 0

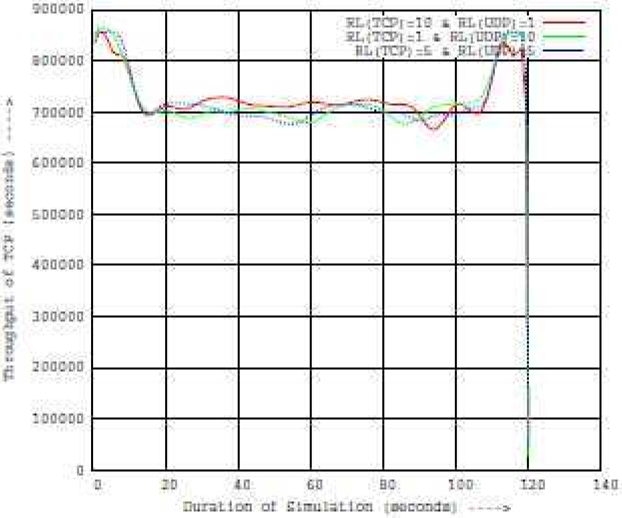


Fig. 6 UseRTSAlgorithm = 1

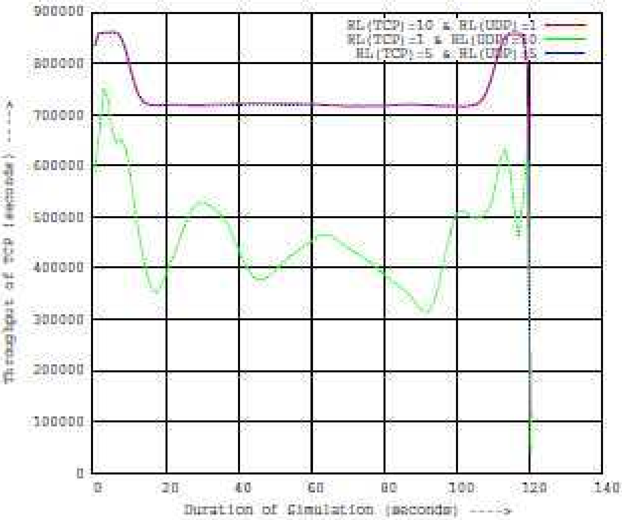


Fig. 7 UseRTSAlgorithm = 2

3 Conclusion

It can be observed that scheme increases the throughput in the cases when feedback is applied for TCP traffic and when applied for both TCP & UDP traffic. In fig 5 we study the throughput for different values of SRC/LRC when no feedback is applied. The aim is to study the application level differentiation on throughput of TCP traffic. It can be observed that throughput of TCP traffic decreases when competing UDP traffic is having higher SRC/LRC values. In fig 6, we study the throughput for different values of SRC/LRC when feedback is applied only to TCP traffic. We observe that throughput is enhanced for almost all the cases even for the case when competing UDP traffic is having higher SRC/LRC values. Lastly, we observe in fig 7 the throughput on TCP traffic for different values of SRL/LRC when there is feedback only UDP traffic. It is observed that throughput for the case when UDP traffic has higher SRC/LRC values is not enhanced even on applying the proposed scheme. The probable reason could be that scheme needs to be more stabilizing by increasing the value of increaseLevel.

References

1. Mico, P.C.F., Luis, O.: Qos in ieee 802.11 wireless lan: Current research activities. In: Electrical and Computer Engineering, Canadian Conference, May 2004, vol. 1, pp. 447–452 (2004)
2. Qiang, L.R.N., Turletti, T.: A survey of qos enhancements for ieee 802.11 wireless lan: Research articles. Wireless Communications and Mobile Computing 4, 547–566 (2004)
3. Zhai, X.C.H., Fang, y.: A call admission and rate control scheme for multimedia support over ieee 802.11 wireless lans. Wireless Networks 12, 451–463 (2006)
4. Pau, D.D., Maniezzo, G.: A cross layer framework for wireless lan qos support. In: Proceedings, International Conference on Information Technology: Research and Education, August 2003 pp. 331–334 (2003)
5. Raisinghani, V.T., Iyer, S.: Cross layer design optimizations in wireless protocol stacks. In: Computer Communications, vol. 27, pp. 720–725. Elsevier, Amsterdam (2004)
6. Ieee 802.11 wireless lan medium access control and physical layer specifications. IEEE Standard (June 2003)
7. Khurana, S., Kahol, A., Jayasumana, A.P.: Effect of Hidden Terminals on the Performance of IEEE 802.11 MAC Protocol. In: Proc. IEEE LCN 1998 (October 1998)
8. Shiann-Tsong, Chen, S., Jenhui, T., Ye, C.F.: The impact of RTS threshold on IEEE 802.11 MAC protocol. In: Proc. of the Ninth International Conference on Parallel and Distributed Systems (December 2002)
9. Choi, W.-Y., Lee, S.-K.: A real-time updating algorithm of RTSCTS threshold to enhance EDCA MAC performance in IEEE 802.11e wireless LANs. In: Proc. IEEE 60th Vehicular Technology Conference VTC 2004-Fall (September 2004)
10. Zhang, L., Shu, Y.: RTS threshold self-tuning algorithm based on delay analysis on 802.11 DCF.In: Proc. of the International Conference on Wireless Communications, Networking and Mobile Computing (September 2005)
11. Liu, J., Guo, W., Xiao, B.l., Huang, F.: RTS Threshold Adjustment Algorithm for IEEE 802.11 DCF. In: Proc. of the 6th International Conference on ITS Telecommunications (2006)

12. Sobrinho, J.L., de Haan, R., Brazio, J.M.: Why RTS-CTS is not your ideal wireless LAN multiple access protocol. In: Proc. IEEE Wireless Communications and Networking Conference (March 2005)
13. Rahman, A., Gburzynski, P.: Hidden Problems with the Hidden Node Problem. In: Proc. of the 23rd Biennial Symposium on Communications (June 2006)
14. Raptis, P., Banchs, A., Vitsas, V., Paparrizos, K., Chatzimisios, P.: Delay Distribution Analysis of the RTS/CTS mechanism of IEEE 802.11. In: Proc. 31st IEEE LCN, Tampa, Florida, USA (November 2006)
15. Ho, C.K., Jean-Paul, M., Linnartz, G.: Analysis of the RTS/CTS Multiple Access Scheme with Capture Effect. In: Proc. 17th IEEE PIMRC 2006 (2006)
16. Ray, S., Starobinski, D.: On False Blocking in RTS/CTS Based Multihop Wireless Networks. IEEE Transactions on Vehicular Technology (March 2007)
17. Heusse, M., Rousseau, F., Berge-Dabbatel, G., Duda, A.: Performance Anomaly of 802.11b. In: Proc. IEEE INFOCOM (March 2003)
18. NCTUns Simulator, http://nsl.csie.nctu.edu.tw/nctuns.html
19. CRAWDAD, http://crawdad.cs.dartmouth.edu/
20. CoralReef, http://www.caida.org/tools/measurement/coralreef/
21. Wiethlter, C.H.S.: Design and verification of ieee 802.11e edcf simulation model for ns2.26 (technical report). Telecommunication Networks Groups, Technische Universitt Berlin (2003)
22. Network simulation using network-simulator 2.29

A Non-functional Requirements Traceability Management Method Based on Architectural Patterns

Sookyeong Song, Younglok Kim, Sooyong Park, and Soojin Park*

Abstract. Unlike the functional requirements that support a certain degree of locality in the system, non-functional requirements, being related to the system quality, apply to the overall qualities of the system. In most cases, non-functional requirements provide solutions in pattern and are applied to the system in the design phase. If the traceability between the analysis model elements, design model elements and the elements involved in the architectural pattern is not maintained in the application process, it may be very costly to reflect the changes of the non-functional requirements in the system. This study proposes the mechanism in which the traceability between analysis model, design model and pattern elements is set in real time while applying the architecture pattern simultaneously to minimize the costly change of non-functional requirements. NFR Tracer, a tool that supports the automation of the above mentioned process, will also be introduced. NFR Tracer proposes the quantitative evaluation results on correctness and completeness automatically set on trace link, showing performance improvement.

1 Introduction

Non-functional requirements (‘*NFRs*’) are the requirements that define how the service is provided to the users, rather than what. While functional requirements focus on the individual services provided by the system, NFRs define the certain expectations of the quality of the software, such as “the response time must be within 1 second in all services” and thus are also referred to as the quality attributes[4]. Considering its pervasiveness, NFRs, once enabled, require substantial cost to make changes in relative terms. However, NFRs can change for various reasons such as business opportunities or changes in technical trend without functional requirements. Even if there were no specific change requirements, there could be multiple changes due to the experimental prototyping until the software is

Sookyeong Song · Younglok Kim · Sooyong Park · Soojin Park
Sogang University
1 Shinsoo-Dong Mapo-Gu, Seoul, Korea
e-mail: `tamnacs@gmail.com`, `{ylkim,sypark,psjdream}@sogang.ac.kr`

* Corresponding author.

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 25–35.
springerlink.com

stabilized overall. Thus it is necessary to identify the correlation between NFRs and various artifacts from the early stage of building the system, which is referred to as requirements traceability management.

There are many studies on the efficient requirements traceability management, but the focus on NFRs is relatively less than that of functional requirements. There are two flows in the current practice of NFRs: i) the model-driven approach based on the perspective that NFRs can be reflected in the system by enabling the design pattern [8, 12] and ii) the information retrieval based approach that views the trace link setting as the mapping of relevant information[2, 5, 10, 14, 16]. The NFR tracing technique suggested in this study is based on the former of the two flows. Most literatures to date seek the automation of trace link of the requirements based on their own technique but cannot suggest any quantitatively verified results. Even if there are quantitative results on performance, the automated trace links do not seem to reach the credible expectations of the developers. It seems that these results come from taking the traceability recovery approach after the artifacts are created.

Therefore this study proposed the mechanism in which *the NFRs trace link is synchronized with the creation of objective artifact,* which is the destination of the trace, and in which the number of design elements that does not have any trace-from link to the analysis or deliverables in the architecture stage. The ultimate goal of this study is to use such result to build *the NFR traceability management framework that has the precision and recall value of '1'.*

This study will focus on the following: Chapter 2 analyzes the problems of the existing studies by discussing the current practices. Chapter 3 introduces the NFRs trace model established based on the above analysis. Chapter 4 explains the sub-mechanism for the auto setting of the proposed trace model and NFR Tracer that supports automation. In Chapter 5, the evaluation is provided on the proposed techniques and tools by utilizing the actual examples of applying NFR Tracer when generating the design model. Lastly, Chapter 6 discusses the conclusion and future research directions.

2 Current Practices

Currently, major literatures on NFRs include:[4, 15, 17]. [4] and [15] deal with the continued trace of the changes by setting the traceability link from the architectural element to the code element, while [17] suggests the technique of expressing NFRs through safety, timing, hardware and performance for the safety-critical system, and introduces the technique to verify and trace each NFR based on EAST-ADL2 and MARTE. [4] proposes the event-based methodology that manages traceability between certain NFR, design deliverable and source codes based on the design pattern. However, it is limited in that the managed traceability link is confined to the architectural element and code element, excluding the relationship between the non-functional requirement and functional requirements in the code.

Other literatures mostly concentrate on classifying the correlation between artifacts and maintaining the trace link while generalizing all software artifacts without distinguishing between the traceability maintenance of functional requirements or requirements type. The major techniques used in the current practices to classify the requirements traceability link are as follows: *Information Retrieval (IR)* [2, 10, 11, 14, 16]; *rule-based approach* [2, 13, 18, 20]; *model-based approach* [1, 9]; and Special Integrator[19].

The analysis of the current practices listed above showed several issues. First of all, the studies on automatic requirements traceability management that present the quantitative evaluation data on performance were rare, limited to [2, 3, 11, 16, 21]. As for the literatures that include the quantitative evaluation results, most studies show a wide gap between precision and recall. The reason for such big difference is that the trace link in these studies was set after the trace source and destination artifacts were generated. In respect to the timing of the setting the link, the techniques used by those studies are technically more inclined toward recovering the missing trace than creating the requirements trace. Fundamentally, the mechanism cannot ensure that both precision and recall at '1' due to the limitations. In this study, the timing of setting the requirement traceability link is synchronized to the creation of the deliverables, the destination of the link, thereby trying to build the mechanism in which all artifacts, except for the deliverables from the very first requirements phase, secure the trace-from link.

3 Non-functional Requirements Trace Model

The elements of the NFRs related trace model are trace node and trace link. Trace node refers to the artifact for which to maintain traceability and trace link expresses the traceability relationship of each artifact. The link to the traceability between the deliverables in different development phases was defined as vertical trace link (*V_TR*), while the traceability between the deliverables in the same

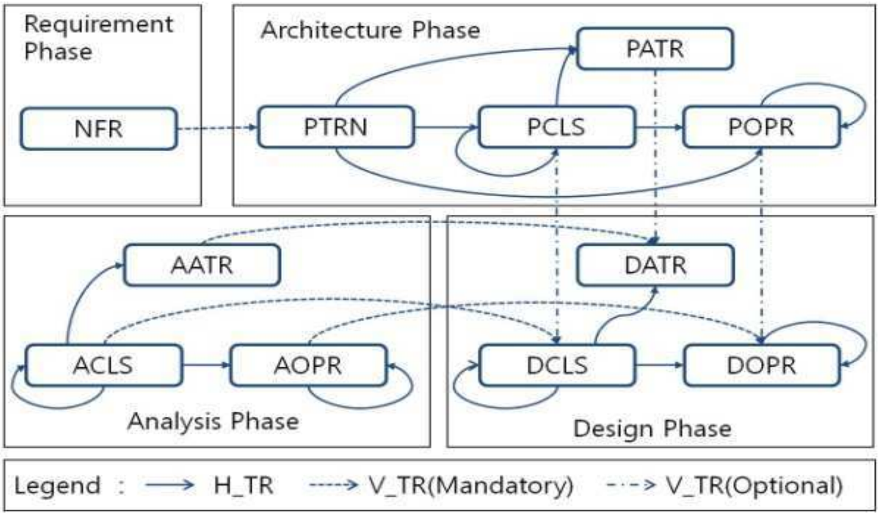


Fig. 1 A Metamodel for NFR Trace

phases was defined as horizontal trace link (*H_TR*). *H_TR* expresses the classes in each model and their relationship, the inclusive relationship between class, attribute and operation, and the call relationship between operations.

As can be seen in Fig. 1., the *NFR node* classified in the requirement stage is traced to the node of the architectural pattern (*PTRN*) which is selected to satisfy the requirement. Since the NFR node and *PTRN* node are artifacts that belong to the requirement phase and architecture phase, *V_TR* link is set. The pattern (*PTRN*) constructed in the architecture phase is formed by the collaboration of different classes (*PCLS*), and the classes (*PCLS*) in the pattern have multiple operations (*POPR*) and attribute (*PATR*), according to the NFR Trace Model. Moreover, the pattern behavior comes from the mutual call between the operations of the classes that comprise the pattern, and *POPR* and *PCLS* node in Fig. 1. will have *H_TR* link to themselves.

Design elements from the design phase are no different from the elements that comprise the architectural pattern. However, design elements can be divided into two: the design class that retains the analysis class classified by the functional requirements and the design class that is changed or newly classified due to the application of the architectural pattern in the course of shifting from the analysis to design phase. Therefore, the trace node of all *DCLS, DATR* and *DOPR* types in Fig. 1. needs to have more than one *ACLS, AATR* and *AOPR* nodes and *V_TR* link, but there may or may not be *V_TR* link with the *PCLS, PATR* and *POPR* nodes in the architecture phase. Such relationship is expressed in Fig. 1. as *'optional'* and *'mandatory'* under *V_TR* link.

4 Automatic Trace Link Generation between Analysis and Design Artifacts

By using the interface with the UML based modeling tool that supports the user-defined pattern automation including the GoF Pattern [6], we developed *NFR Tracer*, the tool that automatically sets the elements of trace models defined in Chapter 3. The NFR Tracer runs as *the add-in application embedded in the IBM Rational Software 7.0 ('RSA').* RSA was selected for the interface because RSA has flexible extensibility since it runs in the eclipse environment. Fig. 2. depicts the structure of the NFR Tracer. As shown in Fig. 2., the pattern instantiation is done on RSA environment and the derived changes from the instantiation are reflected to the trace model which is managed by the NFR Tracer. How the NFR Tracer can update the traceability model according to the changes of the design model will be explained at the end of this chapter.

There is an example to help understand the process of the automatic trace link generation: the fraction of the trace model generated when a simple GoF pattern (*Proxy Pattern*) is applied to the analysis model. The trace model defined in Fig. 1. of Chapter 3 includes all operations and attributes of each class, but Fig. 3. only shows the class information in the trace model in order to focus only on the mapping relationship between the artifacts and trace node. Fig. 3. shows the process of setting the trace link for the classes that are changed or newly generated as the

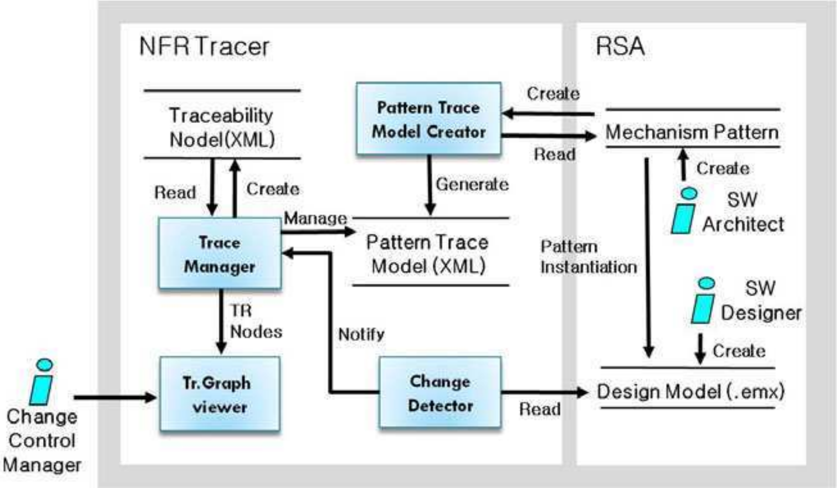


Fig. 2 Structure of NFR Tracer

result of applying the proxy pattern, which is selected to satisfy the modifiability related requirements among the quality attributes of the course registration system, to CRC class, one of the classes of <<*control*>> type.

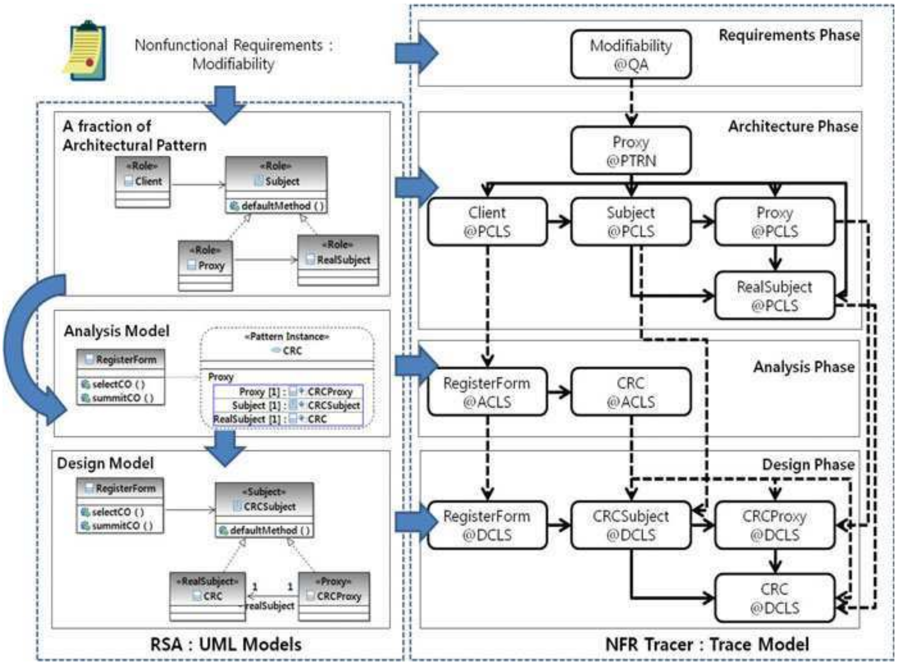


Fig. 3 Relationship between Artifacts and Trace Nodes

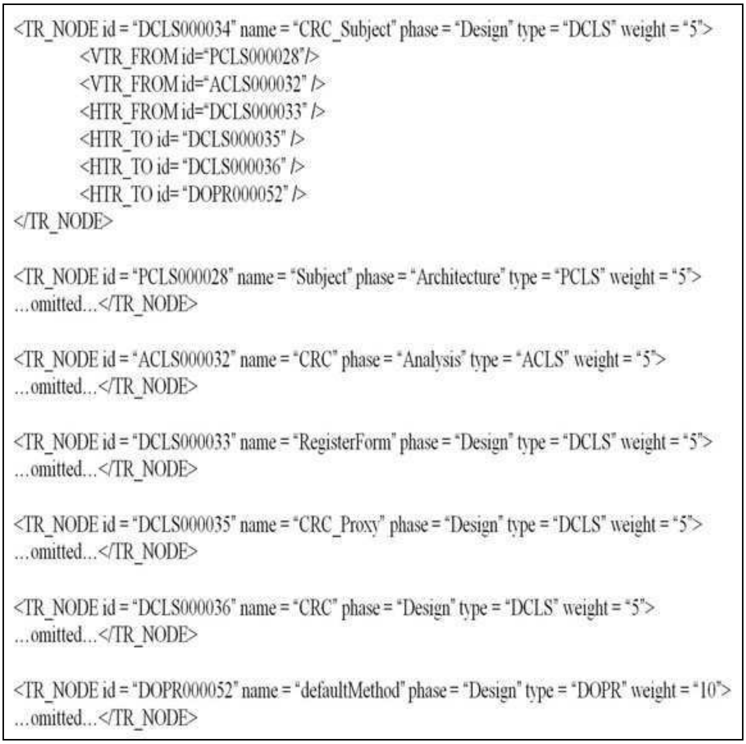
All information about trace node and link is converted to XML and managed by the NFR Tracer. The XML fragment that expresses the information of *CRCSubject@DCLS* from the nodes in Fig. 3. is described in Fig. 4.. <VTR_FROM> shows the vertical trace link data from the analysis model's *CRC (ACL00032)* and *Subject (PCL000028)* node from the architectural pattern connected to *CRC_Subject*. The relationship between *RegisterForm (DCL000033)* and *CRC_Subject* in the same data model are the horizontal trace link and *CRC_Subject* is the destination of the relationship, showing the reverse link as it is with <HTR_FROM>. As for *CRCProxy (DCLS000035)* and *CRC (DCLS000036)* that exist in the same design phase, they are the classes that implement *CRC_Subject*. Because they are the artifacts that are affected by the changes of *CRC_Subject*, it is linked to *HTR_TO* link. Lastly, it can be seen that there is <HT R_TO> link with *defaultMethod (DOPR000052),* the operation of *CRC_Subject*, which was omitted in the trace model on Fig. 1. The name of *'defaultMethod'* will be changed as *'summitCO'* during design model refinement. Then, the change of the UML model will be automatically reflected to the trace model by NFR Tracer.

The weight included in each node data in Fig. 3. is used to calculate the change impact degree inside the system when there is a certain trace node change requirement. It can be adjusted according to the application characteristics. For now, 5 is given to class, 2 to attribute and 10 to operation that includes the enabling logic.

The source for all data in Fig. 4. is *.emx,* the model file managed by RSA. The biggest issue for enabling the NFR Tracer was how to reflect the changes in the models on RSA, the COTS tool, in real time. It was because the greatest merit and advantage of the approach proposed by this study is to detect the changes in the COTS tool and reflect them in the trace model so as to ensure synchronicity of the trace link and software artifacts, which serve as the destination of the trace links, as well as to maintain the data consistency of the two models.

Since the source codes for the COTS tool are not open to public generally, it is impossible to change the COTS tool itself to link to the requirements traceability management tool. To solve such problems, *Instrumentation Technique* [7] was used in this study. Instrumentation refers to detecting the impact to the COTS software from outside. The enabling techniques include API based instrumentation and machine code based instrumentation. The former observes and controls the interaction between OS and COTS software, and is also called *'hooking'* [7]. The latter literally is to alter the binary representation to change the COTS software itself. Hooking was chosen among the two since all it takes is to detect the action that changes the UML model in the RSA in real time.

Fig. 5. shows the RSA hooking algorithm of the NFR Tracer and those steps are the necessary parts selected by referring to [7]. Construction of the initial design model should precede running of the NFR Tracer and hooking the RSA model data is done explicitly on the NFR Tracer. The NFR Tracer generates the trace node and link that reflect the relevant analysis model at the point when the evaluation on the analysis model ends on the RSA and is managed as the deliverable of Version 1.0. At the same time, all *ACLS, AATR* and *AOPR* nodes as well as the links between the nodes are copied, and the link between the *DCLS, DATR* and *DOPR* nodes and the links between the nodes.



```
<TR_NODE id = “DCLS000034” name = “CRC_Subject” phase = “Design” type = “DCLS” weight = “5”>
        <VTR_FROM id=“PCLS000028”/>
        <VTR_FROM id=“ACLS000032” />
        <HTR_FROM id=“DCLS000033” />
        <HTR_TO id= “DCLS000035” />
        <HTR_TO id= “DCLS000036” />
        <HTR_TO id= “DOPR000052” />
</TR_NODE>

<TR_NODE id = “PCLS000028” name = “Subject” phase = “Architecture” type = “PCLS” weight = “5”>
...omitted...</TR_NODE>

<TR_NODE id = “ACLS000032” name = “CRC” phase = “Analysis” type = “ACLS” weight = “5”>
...omitted...</TR_NODE>

<TR_NODE id = “DCLS000033” name = “RegisterForm” phase = “Design” type = “DCLS” weight = “5”>
...omitted...</TR_NODE>

<TR_NODE id = “DCLS000035” name = “CRC_Proxy” phase = “Design” type = “DCLS” weight = “5”>
...omitted...</TR_NODE>

<TR_NODE id = “DCLS000036” name = “CRC” phase = “Design” type = “DCLS” weight = “5”>
...omitted...</TR_NODE>

<TR_NODE id = “DOPR000052” name = “defaultMethod” phase = “Design” type = “DOPR” weight = “10”>
...omitted...</TR_NODE>
```

Fig. 4 XML fragments for "CRC_Subject" node and related nodes

The NFR Tracer hooks all the modifications of the designer, including the application of the architecture pattern after the initial design model is generated. The main components of the NFR Tracer are *Change Detector* and *Trace Manager*. While Change Detector monitors the changes of the *.emx* files, which are UML models managed by RSA, Trace Manager manages the trace models generated and managed by NFR Tracer.

The steps taken by the NFR Tracer in Fig. 5. are as follows: ① First, Trace Manager detects the location of RSA in the system, and ② hook RSA. If there are any changes made to the UML model on RSA, ③ Change Detector detects the interaction between RSA and OS. If the changes made to the UML model caused by the manipulation inside RSA and need to be reflected on the trace model, ④ the data on such changes are read from the UML model, and ⑤ are delivered to Trace Manager. ⑥ Trace Manger directly reflects the UML model changes received from Change Detector to the trace model. To prevent any UML model changes while the NFR Tracer hooks a change, RSA is locked during hooking in order to maintain consistency between the trace model and UML models managed by RSA.

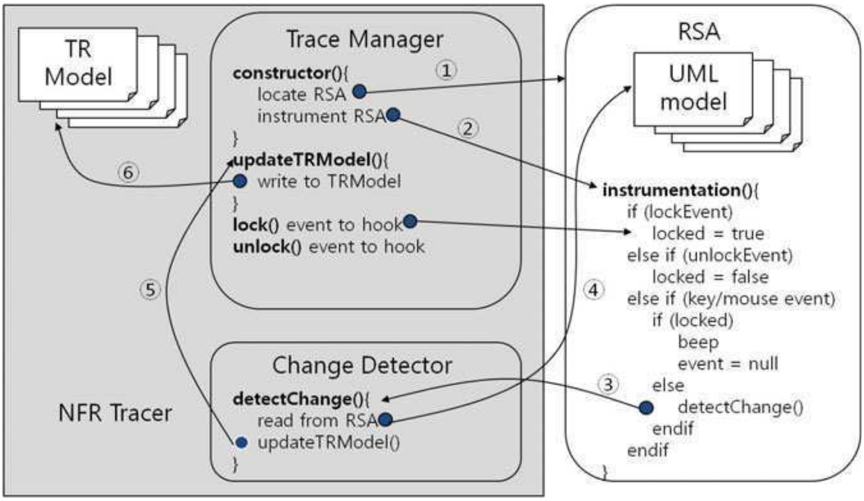


Fig. 5 Hooking Mechanism of NFR Tracer

5 Evaluation

In order to measure the correctness and completeness of the techniques for requirements traceability proposed by this study, two were selected from the patterns that form the architecture in the analysis model for two applications and the transformation into the design model was made on RSA to apply the NFR Tracer.

- Target Application
  Course Registration System, Payroll System
- Applied Patterns
- RMI (Remote Method Invocation) pattern which is supporting distribution mechanism
- JDBC (Java DataBase Connectivity) pattern which is supporting persistency mechanism

Precision and *recall* were the metrics used to measure the accuracy and completeness of the trace link automatically generated by the NFR Tracer. The definitions of precision and recall were used as suggested by [1], but the definition of information retrieval was excluded, giving a re-interpretation to precision and recall as follows:

$$\text{Precision} = |\text{ Correct Links }| / (|\text{ Correct Links }| + |\text{ Incorrect Links }|)$$

$$\text{Recall} = |\text{ Correct Links }| / (|\text{ Correct Links}| + |\text{Missed Links }|)$$

Precision refers to the correct link, which is one of the trace links set by the NFR Tracer, and recall refers to how much of the total trace links were set that were supposed to be set by the NFR Tracer. Precision being ‘1’ means that all trace set by the NFR Tracer is 100% credible. Recall being '1' means that the NFR Tracer can set the incorrect link but that there is no missed link.

Table 1. shows the study of the automatically generated trace-link by using the NFR Tracer. In both payroll system and course registration system, *precision was '1'* when the NFR Tracer automatically generated the trace, *which means that there was no incorrect link in the trace-link found by the NFR Tracer*. On the other hand, recall was not '1'(average value was 0.85), because the missed link was caused by the incompleteness issue of RSA API, the COTS tool interfaced by the NFR Tracer, not the theoretical error from the trace model. The NFR Tracer application of patterns in the *.emx* files of RSA that can be detected from outside RSA. This is why the NFR Trace could not complete hooking. Recall in Table 1. is lower for JDBC than for RMI pattern because there are more operation changes in the JDBC pattern application, causing more missed links during hooking.

Table 1 Precision and Recall of NFR Tracer

Application	Pattern Name	Correct Links	Incorrect Links	Missed Links	Precision	Recall
Payroll System	RMI	239	0	21	1	0.92
	JDBC	274	0	91	1	0.75
Course Registration System	RMI	205	0	16	1	0.93
	JDBC	250	0	72	1	0.78

Considering that there were cases in which precision exceeds 0.8 and recall is below 0.5[3,16], whereas recall is almost 1 but precision is below 0.1[21] in some cases, the results described in Table 1. are balanced ones between correctness and completeness in setting trace links.

6 Conclusion

The approach proposed in this study seeks to *synchronize timing of the creation the trace link and its destination artifacts*. It can be done by synchronizing the trace link set between analysis model elements and design model elements to the application of the architectural pattern that takes place during the shift from the analysis model to the design model in order to secure the trace link of NFRs. Refinement from an analysis model to a design model is not fulfilled by adoption of only one architectural pattern. Although the case which is described in chapter 3 is just a snapshot for showing the progress of the construction of some requirements traceability model fragments by adoption of an architectural pattern, we show the proposed general approach can work on different applications and different patterns through the experiment of chapter 4. The result says that recall could not reach 1 due to the RSA API, the interface modeling tool, being open limitedly to the outside. However, there is a possibility for securing correctness and completeness of the automatic setting of the requirements traceability by showing that *the average precision* reached *1* and *the average recall* also approached *0.85* from experiments with two different system cases and four different patterns.

In the future, the study will refine hooking as the modeling tool in the NFR Tracer and build the automatic environment in which recall can reach 1. Moreover, the tools will be applied to various applications to refine their performance.

Acknowledgment. This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)(NIPA-2011-C1090-1131-0008).

References

1. Anquetil, N., et al.: A model-driven traceability framework for software product lines. Software Systems Modeling 9(4), 427–451 (2009)
2. Antoniol, G., et al.: Recovering traceability links between code and documentation. IEEE Transactions on Software Engineering 28(10), 970–983 (2002)
3. Asuncion, H., et al.: Software traceability with topic modeling. In: Proceedings of the 32nd ACMIEEE International Conference on Software Engineering ICSE 2010, vol. 0, p. 95 (2010)
4. Cleland-Huang, J., Schmelzer, D.: Dynamically Tracing Non-Functional Requirements through Design Pattern Invariants. In: Proceedings of the 2nd International Workshop on Traceability in Emerging Forms of Software Engineering TEFSE (2003)
5. De Lucia, A., et al.: Assessing IR-based traceability recovery tools through controlled experiments. Empirical Software Engineering 14(1), 57–92 (2009)
6. Gamma, E., et al.: Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley, Reading (1995)
7. Egyed, A., Balzer, R.: Integrating COTS Software into Systems through Instrumentation and Reasoning. Automated Software Engineering 13(1), 1–39 (2006)
8. Gross, D., Yu, E.: From Non-Functional Requirements to Design through Patterns. Requirements Engineering 6(1), 18–36 (2001)
9. Dubois, M., et al.: A Model for Requirements Traceability in a Heterogeneous Model-Based Design Process: Application to Automotive Embedded Systems. 15th IEEE International ConferenceEngineering of Complex Computer Systems (ICECCS) 0, 233–242 (2010)
10. Hayes, J., et al.: Improving requirements tracing via information retrieval. In: Proceedings 11th IEEE International Requirements Engineering Conference, vol. 0, pp. 138–147 (2003)
11. Kim, S., et al.: Quality-driven architecture development using architectural tactics. Journal of Systems and Software 82(8), 1211–1231 (2009)
12. Leveson, N.: Safeware: System Safety And Computers. Society. Addison-Wesley, Reading (1995)
13. Mader, P., et al.: Semi-automated Traceability Maintenance: An Architectural Overview of traceMaintainer. Automated Software Engineering 0, 425–426 (2009)

14. Marcus, A., Maletic, J.: Recovering documentation-to-source-code traceability links using latent semantic indexing. In: Proceedings of 25th International Conference on Software Engineering, vol. 6(0), pp. 125–135 (2003)
15. Murta, L.P., et al.: Continuous and automated evolution of architecture-to-implementation traceability links. Automated Software Engineering 15(1), 75–107 (2008)
16. Oliveto, R., et al.: On the Equivalence of Information Retrieval Methods for Automated Traceability Link Recovery. In: Proc. ICPC(International Conference on Program Comprehension) Short, vol. 0, pp. 68–71 (2010)
17. Peraldi-Frati, M., Albinet, A.: Requirement traceability in safety critical systems. In: Proceedings of the 1st Workshop on Critical Automotive applications Robustness Safety CARS(Critical Automotive applications Robustness Safety) 2010, vol. 0, pp. 11–14 (2010)
18. Ramesh, B., Jarke, M.: Toward reference models for requirements traceability. IEEE Transactions on Software Engineering 27(1), 58–93 (2001)
19. Sherba, S., et al.: A Framework for Mapping Traceability Relationships. In: Proceedings of the 2nd International Workshop on Traceability in Emerging Forms of Software Engineering, vol. 0, pp. 32–39 (2003)
20. Spanoudakis, G., et al.: Rule-based generation of requirements traceability relations. Journal of Systems and Software 72(2), 105–127 (2004)
21. Sundaram, S., et al.: Assessing traceability of software engineering artifacts. Requirements Engineering 15(3), 313–335 (2010)

Performance of Non-coherent Detectors for Ultra Wide Band Short Range Radar in Automobile Applications

Purushothaman Surendran, Jong-Hun Lee, and Seok Jun Ko*

Abstract. A detector is said to be superior if the information is extracted in a best way for some particular purpose. Here the objective of this paper is to analyze the detection performance of non-coherent detectors such as square law detector, linear detector and logarithmic detector in a background of white Gaussian noise (WGN) for Ultra Wide Band Short Range Radar in Automotive applications. It is assumed that the target is stationary and energy reflected from the target is distributed. The non-coherent detectors have almost similar detection probability in a background of white Gaussian noise. The performance of the detector is analyzed and simulation has been done in order to verify.

Keywords: UWB Radar, Coherent Integration, Target Detection.

1 Introduction

The demand for short range radar sensors which are used for target detection has increased fabulously in automotive sector [2, 3]. The key ideas behind the development of Short Range Radar (SRR) in automotive sector are collision avoidance and reduce traffic fatality. The source for target detection is the radar signals reflected by the target, the received radar signal is a mixture of noise and varied signals. The designed system must provide the optimal technique to obtain the desired target detections, preferred detection can be determined by using specific algorithm for measuring the energy of the signals received in the receiver side. Decision is made on the basis of the received echo signal which is determined by target geometry. The detection algorithm is used to make a decision on the target present and to measure the range of the target. In order to predict the range more exactly, a larger

Purushothaman Surendran · Seok Jun Ko
Department of Electronics Engineering, Jeju National University, Jeju, Korea

Jong-Hun Lee
DGIST(Daegu Gyeongbuk Institute of Science & Technology), Daegu, Korea
e-mail: sjko@jejunu.ac.kr

[*] Corresponding author.

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 37–47.
springerlink.com

signal bandwidth is required. This can be accomplished by reducing the range resolution cell of the radar so that the range accuracy is increased. Consequently, the range resolution cell size of the radar is smaller than the target size and target is surrounded by homogeneous distributed noise [4].

Previous work [4, 5] has been mainly focused on the influence of increasing range resolution on the detection ability of targets with dimensions greater than the resolution cell. Also mathematical analysis of these circumstances has been resolved.

In this paper we analyze the performance of various detectors for Ultra Wide Band Short Range Radar (UWB-SRR) in automotive applications. It is assumed that the target is stationary and the energy reflected from the target is assumed to be 1. The performance of the detectors is shown.

The organization of this paper is as follows. In Section 2, the system model is described. In section 3, description about non-coherent detector. In section 4, the probability of detection and false alarm is expressed. In Section 5, simulation results for various detectors. Finally, conclusions are presented in Section 6.

2 System Description

The block diagram of a UWB radar system as shown in fig. 1 is split into two parts, the transmitter part and the receiver part.

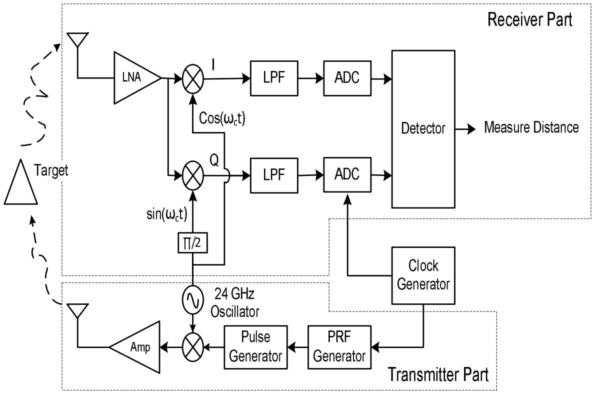


Fig. 1 Block Diagram of a UWB radar system

In the transmitter part, the pulses are initiated by the Pulse Repetition Frequency (PRF) generator which triggers the pulse generator which in turn generates Gaussian pulses with sub-nano second duration as shown in fig. 2. The Pulse Repetition Interval (PRI) is controlled by the maximum range of the radar. The maximum range for unambiguous range depends on the pulse repetition frequency and can be written as follows

$$R_{\max} = \frac{c}{2 \cdot f_{PRF}} \tag{1}$$

where f_{PRF} is pulse repetition frequency and c is the velocity of light. And the range resolution can be written as

$$\Delta R = \frac{c \cdot T_P}{2} \tag{2}$$

where T_P is pulse width and c is the velocity of light. And then the transmitted signal can be written as follows

$$s(t) = A_T \cdot \sin(2\pi f_c t + \varphi_0) \cdot \sum_{n=-\infty}^{+\infty} p(t - n \cdot T_{PRI}) \tag{3}$$

where $p(t)$ is the Gaussian pulse as follows

$$p(t) = \exp\left[-2\pi\left(\frac{t}{\tau_p}\right)^2\right] \tag{4}$$

where τ_p represents the time normalization factor, A_T is the amplitude of single transmit pulse, φ_0 is the phase of the transmit signal, f_c is the transmit frequency, T_{PRI} is the pulse repetition interval obtained from pulse repetition frequency given as $T_{PRI}=1/f_{PRF}$.

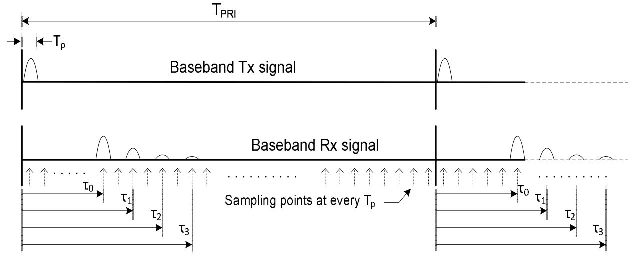


Fig. 2 Transmitted signal and received baseband signal

The form of the transmitted signal in this system is known, but the received signal usually is not completely known. Since the range resolution of this UWB radar system is much less than the extent of the target it must detect, the echo signal is the summation of the time-spaced echoes from the individual scattering centers that constitute the target [3]. In this paper, we assume that the target is stationary and the target has L independent reflecting cells. Then the target model is written as

$$h(t) = \sum_{l=0}^{L-1} \alpha_l \cdot \delta(t - \tau_l) \tag{5}$$

where the number of scatters L, the amplitude of the scatters α_l, and the time delays of the scatters τ_l are all unknown. We assume that the τ_0 in fig. 2 indicates the target range.

The radiated electromagnetic signals generated by the transmit antenna is reflected by the target and they are received in the receiver antenna. First, the received signal is pre -amplified by the Low Noise Amplifier (LNA). Then the signal is multiplied with carrier signal and divided between the in-phase and quadrature-phase baseband signal as shown in fig. 1. We assume that the low pass filter will match the pulse envelope spectral shape as close as possible to provide a fully matched optimum filter. Then the baseband received signal $r(t)$ is written as

$$r(t) = A_T \sum_{n=-\infty}^{+\infty} \sum_{l=0}^{L-1} \alpha_l \cdot e^{j\theta_l} p(t - nT_{PRI} - \tau_l) + n(t) \tag{6}$$

where $n(t)$ is the white Gaussian noise (WGN) with two-sided power spectral density $N_0/2$ and θ_l is the arbitrary phase of l-th scatter that can be written as $\theta_l = -2\pi f_c \tau_l + \varphi_0$. The sampling rate of the A/D converters is same to the pulse width. And we assume that the baseband received signal is sampled at peak point of $p(t)$ as like the fig. 2. When the target size is greater than the radar resolution, then the echo consists of a signal characterized by eq. (6) and fig. 2; the received echo signal will be composed of individual short duration signals in a pulse train. A gain rather than a loss can be obtained when a high-resolution waveform resolves the individual scatters of a distributed target such as the UWB radar system. Because there will be no signal addition from close scattering centers, detection will be depend on the reflected strength of individual centers for weak returns [3].

3 Non-coherent Detectors

The detector of the UWB radar receiver must determine that a signal of interest is present or absent. And then the UWB radar processes it for some useful purpose such as range determination, movement, and etc [3]. In this paper, we analyze the performance of non-coherent detectors against a background of white Gaussian noise for range determination, as shown in fig. 3. The non-coherent detectors consist of coherent integration, non-coherent detector and non-coherent integration.

The in-phase (I) and quadrature (Q) sampled values at every T_p are used as the input of the detector. It is assumed that the sampling rate (T_p) is same to the pulse width of 2 ns and the range resolution can be 30cm from (2). Also it is assumed that the maximum target range is 30m and by using (1) we can get T_{PRI} of 200 ns. From the above-mentioned range resolution and maximum target range, the range gates of at least 100 are required to detect the target. It is equal to the number of memory in the coherent and non-coherent integration. The sampled value at every T_p is applied to the switch I of the coherent integration. The switch I is shifted at every T_p sample, i.e., the range gate. It takes $N{\cdot}T_{PRI}$ to coherently integrate and dump for all of the range gates.

The coherent integration for the *i*-th range gate in I branch can be expressed as follows

$$X^{I}(i)=\frac{1}{N}\sum_{n=1}^{N}\mathrm{Re}\left\{r_n(iT_P)\right\} \tag{7}$$

where

$$r_n(iT_P)=A_T\sum_{l=0}^{L-1}\alpha_l\cdot e^{j\theta_l}p((nT_{PRI}+iT_p)-nT_{PRI}-\tau_l)+n^{'}(nT_{PRI}+iT_P) \tag{8}$$

and where $n'(nT_{PRI}+iT_p)$ is a sampled value of $n(t)$ at $nT_{PRI}+iT_p$. Then the summations for each gate are stored in each memory of the coherent integration. Therefore it is possible to achieve an improvement of signal-to-noise ratio (SNR) as much as 10·log(*N*) [dB].

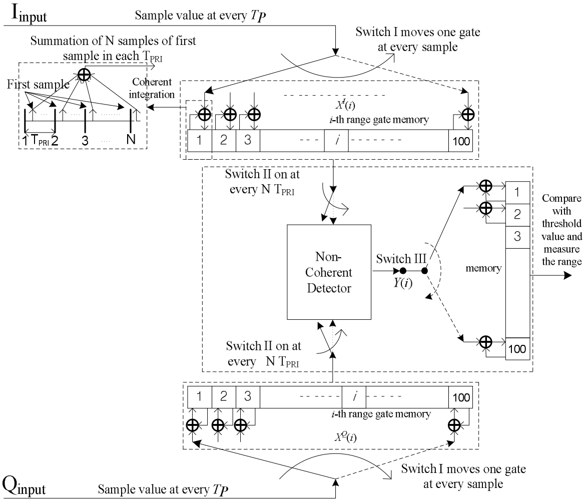
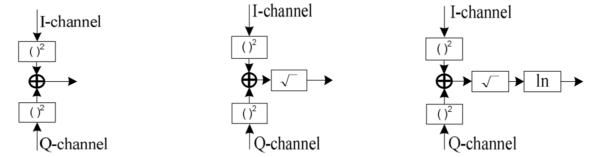


Fig. 3 Block diagram of receiver with detector



(a)Square Law (b) Linear (c) Logarithmic Detector

Fig. 4 Non-coherent Detector

The sample value received from the coherent integration is squared and operates at every $N \cdot T_{PRI}$. The squared range gate samples are combined and then both I and Q branch values are summed as shown in fig. 4a. The output after squaring $Y(i)$ is known as square law detector can be represented as

$$Y(i) = \{X^{I}(i)\}^{2} + \{X^{Q}(i)\}^{2} \tag{9}$$

In the case of a linear detector as shown in fig. 4b the sample value received from the coherent integration is squared and operates at every $N \cdot T_{PRI}$. The squared range gate samples are combined and then both I and Q branch values are summed and square root is applied to the summed value. The output of the linear detector $Y(i)$ can be represented as

$$Y(i) = \sqrt{\{X^{I}(i)\}^{2} + \{X^{Q}(i)\}^{2}} \tag{10}$$

In Logarithmic detector as shown in fig. 4c the sample value received from the coherent integration is squared and operates at every $N \cdot T_{PRI}$. The squared range gate samples are combined and then both I and Q branch values are summed and square root is applied to the summed value and natural logarithm is taken. The output of the logarithmic detector $Y(i)$ can be represented as

$$Y(i) = \ln\left(\sqrt{\{X^{I}(i)\}^{2} + \{X^{Q}(i)\}^{2}}\right) \tag{11}$$

All of the reflected signals from the target can be added non-coherently.

The value $Y(i)$ is stored in the i-th register of the non-coherent integration at every $N \cdot T_{PRI}$ for $N \cdot M \cdot T_{PRI}$. The output of the non-coherent integration $Z(i)$ can be written as

$$Z(i) = \frac{1}{M}\sum_{m=1}^{M} Y_{m}(i) \tag{12}$$

where $Y_m(i)$ is the output of the squared window at $m \cdot N \cdot T_{PRI}$. If the above result is greater than the defined threshold, then we can determine that a target is present. And the index i represents the position of the target; the target range indicates $i \cdot 30$ cm. It takes $N \cdot M \cdot T_{PRI}$ to decide the target range.

4 Detection and False Alarm Probability for Square Law Detector

To calculate the detection characteristics, the probability density functions $p_0(z)$ and $p_1(z)$ of $Z(i)$ in (12) should be determined. Here we consider the detection and false alarm probability for square law detector. If the echo signal is absent, then the probability density function $p_0(z)$ is determined at the detector output by using the following expression [6]

$$P_0(Z) = \frac{1}{\sigma^2 2^{n/2}\Gamma(n/2)} Z^{(n/2)-1} e^{-z/2\sigma^2} \tag{13}$$

where the central chi-square distribution $p_0(z)$ with n degree of freedom has zero mean and variance σ^2. And if the echo signal is present, then the probability density function $p_1(z)$ is determined as following [6]

$$P_1(Z) = \frac{1}{2\sigma^2}\left(\frac{z}{s^2}\right)^{(n-2)/4} e^{-(s^2+z)/2\sigma^2} I_{(n/2)-1}\left(\sqrt{z}\frac{s}{\sigma^2}\right) \tag{14}$$

where $I_{n/2\text{-}1}$ is the n-th order modified Bessel function of the first kind. The non-central chi-square distribution $p_1(z)$ with n degree of freedom has mean s^2 and variance σ^2.

To get the detection and false alarm probability, we must calculate the following integral; the probability of false alarm can be written as [6]

$$P_{FA}(Z) = \int_{u_{th}}^{\infty} \frac{1}{\sigma^2 2^{n/2}\Gamma(n/2)} Z^{(n/2)-1} e^{-z/2\sigma^2} dz \tag{15}$$

and the probability of detection is given as

$$P_d(Z) = \int_{u_{th}}^{\infty} \frac{1}{2\sigma^2}\left(\frac{z}{s^2}\right)^{(n-2)/4} e^{-(s^2+z)/2\sigma^2} I_{(n/2)-1}\left(\sqrt{z}\frac{s}{\sigma^2}\right) dz \tag{16}$$

where u_{th} is the threshold value. On the basis of the above mentioned formulas, the detection characteristics of the received echo signal is determined for proposed detector.

Table 1 System parameters

Parameter	Notation	Value
Pulse Repetition Interval	T_{PRI}	200ns
Pulse Width	T_P	2ns
Maximum Target Range	R_{max}	30m
Range Resolution	R	30cm
Number of coherent integration	N	10
Number of non-coherent integration	M	variable

5 Computer Simulation Results

The purpose of the simulation is to assess the performance of the non-coherent detectors. First, we compare theoretical results with computer simulation results by using the probability density function. And then the probabilities of detection and false alarm are evaluated. In the simulations, we use the percentage of total energy reflected from each flare point as 1. We simulate the probability density functions using the system parameters as given in table I. A large enough number of trials are used to obtain each point of the probability density functions. The number of trials is about 1000000 times. The signal-to-noise ratio (SNR) is defined as $\bar{E}/N_0$, where Ē represents the total average energy reflected from a target.

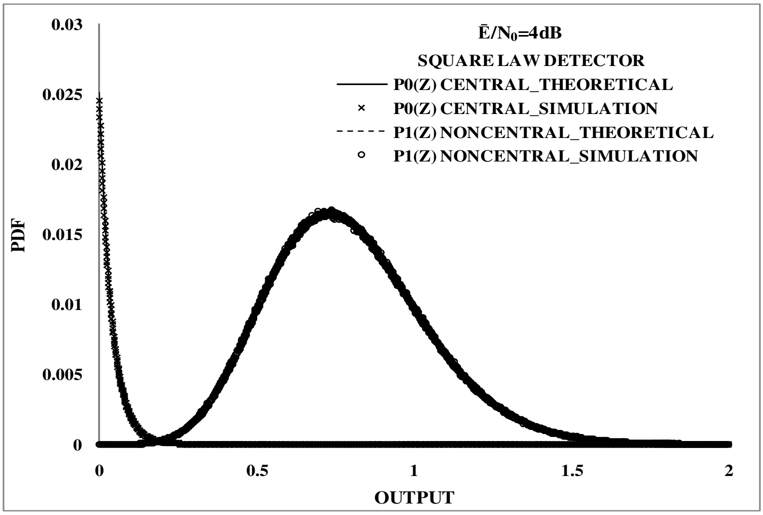


Fig. 5 The probability density function

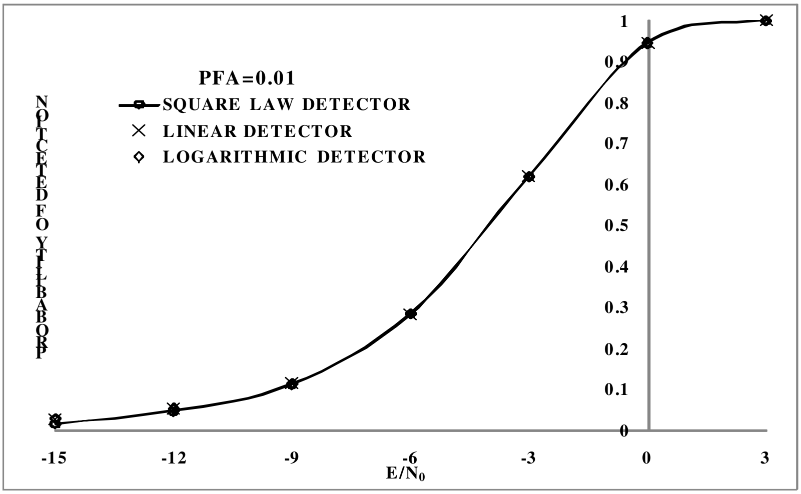


Fig. 6 The detection probability vs. $\bar{E}/N_0$ at P_{FA}=0.01

Fig. 5 shows the result of the probability density functions (PDF) $p_0(z)$ and $p_1(z)$ Ē/N$_0$=4dB for non-coherent detector. The simulation result is compared with the theoretical result. It shows that the simulation result and the theoretical result are in excellent agreement. And the PDF has 2 degrees of freedom and the average signal energy is 1. By using the probability density functions, the detection characteristics of the detector can be plotted.

Fig. 6 shows the detection probability versus Ē/N$_0$ for stationary target at P_{FA}=0.01. The non-coherent detectors have detection probability of 1 at Ē/N$_0$=3dB and the performance of all the three detectors is same.

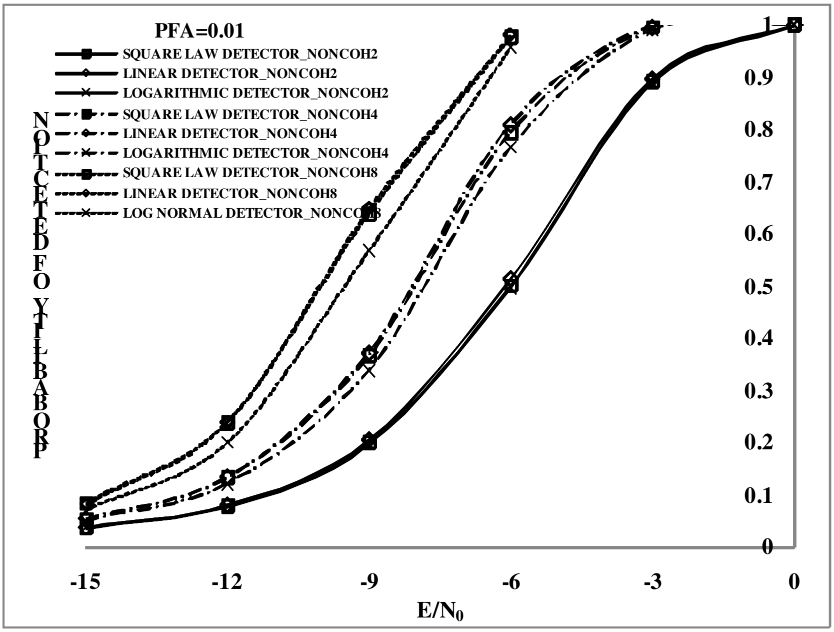


Fig. 7 The detection probability vs. Ē/N$_0$ at P_{FA}=0.01

Fig. 7 shows the detection probability versus Ē/N$_0$ for various stationary target at P_{FA}=0.01. The non-coherent detectors have better detection probability as the number of non-coherent integration increases. The simulation result shows that at Ē/N$_0$=-6dB the probability of detection is approximately 1 for non-coherent integration number (NONCOH) of 8 for all the detectors, On the other hand the probability of detection reduces by 0.2 when the number of non-coherent integration reduces to 4.We can also predict that the performance of all the detectors is almost same when we use the same number of non-coherent integration. Therefore we can use any of the three detectors for automobile applications.

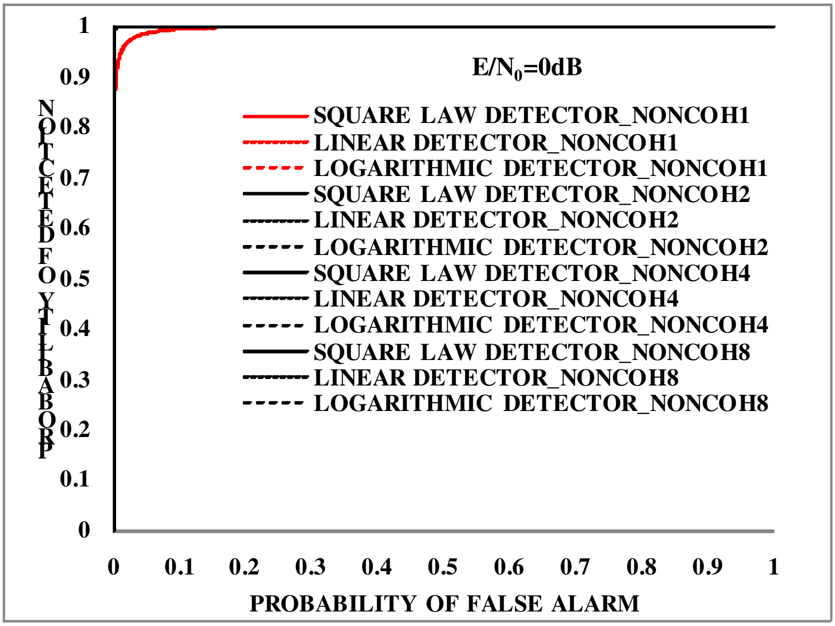


Fig. 8 The detection probability vs. false alarm probability for various detectors

Fig. 8 shows the detection probability versus the false alarm probability for various detectors at $\bar{E}/N_0$=0dB. The result shows that the performance of the detectors is increased as the number of non-coherent integration increases from 1 to 8. We can predict that all the non-coherent detectors mentioned in this paper have similar performance in a background of white Gaussian noise.

6 Conclusion

In this paper, we have analyzed the performance of non-coherent detection algorithm for Ultra Wide Band Short Range Radar (UWB-SRR) signals in automotive applications. The detection probability is found to be same for all the non-coherent detectors such as square law detector, linear detector and logarithmic detector in various SNR. Also, in order to get the detection probability to be above 0.9 for P_{FA}=0.01, $\bar{E}/N_0$ is required to be more than 0dB. Therefore it is necessary that the number of coherent integration must be increased.

Acknowledgments. This work was supported by Daegu Gyeongbuk Institute of Science and Technology and funded by the Korean Ministry of education & Science Technology.

References

1. Strohm, K.M., et al.: Development of Future Short Range Radar Technology. Radar Conference (2005)
2. Taylor, J.D. (ed.): Ultra-Wideband Radars Technology: Main Features Ultra-Wideband (UWB) Radars and differences from common Narrowband Radars. CRC Press, Boca Raton (2001)
3. Taylor, J.D.: Introduction to Ultra-Wideband (UWB) Radars systems, Boca Raton, FL (1995)
4. Hughes, P.K.: A High-Resolution Radar Detection Strategy. IEEE Transaction on Aerospace and Electronic Systems ASE 19(5), 663–667 (1983)
5. Van Der Spek, G.A.: Detection of a Distributed Target. IEEE Transaction on Aerospace and Electronic Systems ASE-7(5), 922–931 (1971)
6. Proakis, J.G.: Digital Communications. McGraw-Hill, New York (2001)
7. Surendran, P., Ko, S.J., Kim, S.-D., Lee, J.-H.: A Novel Detection Algorithm for Ultra Wide Band Short Range Radar in Automobile Application. In: IEEE VTC2010-Spring (May 2010)

Flexible Modeling Language in Qualitative Simulation

Yuji Hashiura and Tokuro Matuo

Abstract. Qualitative Simulation is one of research areas in Artificial Intelligence. It is a strong tool to analyze various types of dynamics, but simulation model used in the existing qualitative simulation has a strong limitation where the complex condition and nonautonomous model can not be expressed. To solve the problems, we propose the following new modeling. First, we define a modeling method regarding relationship and transmission of effect between nodes. In addition, we consider a strength of cause and effect to analyze complex condition and situation. Second we define a new unified modeling language to express a complex graph-based qualitative simulation model. Our proposed method can give an analyzed result of an nonautonomous system where the parameter depends on time-pass.

1 Introduction

Qualitative Reasoning/Simulation is one of research fields in Artificial Intelligence and is applied to analyze various types of dynamics. And also, qualitative simulation is one of effective tools to simulate complex and dynamic systems[1][2][3][4]. In qualitative simulation, model to run a simulation is expressed by qualitative differential equations. Thus, the simulator provides a result of simulation on whole of the dynamics. Users can not find the result of simulation, but also view the process of the simulation, because the qualitative simulation-based method shows the condition and state in each time-step. To make a decision and determine a strategy, causal

Yuji Hashiura
Yamagata University 4-3-16, Jonan, Yonezawa, Yamagata, 992-0051, Japan
e-mail: hashiura2009@e-activity.org

Tokuro Matsuo
Yamagata University 4-3-16, Jonan, Yonezawa, Yamagata, 992-0051, Japan
e-mail: matsuo@yz.yamagata-u.ac.jp

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 49–62.
springerlink.com

graph-based simulation is one of promising field to analyze without qualitative differential equation[5][6]. Causal graph model has nodes showing factors in a system and arcs that is connecting with related nodes[7]. When there are related nodes, they are connected by arcs. Each node has a qualitative state value and qualitative state trend. Each arc has rules about a changing trend of effect and transmission speed. The characteristics on arcs are given initially and qualitative simulation can be conducted by just characteristics and initial value on nodes. A node give an influence by combination of qualitative value on nodes and characteristics on arcs. Time passes of simulation are iterative and set of qualitative set on node is a result of the simulation.

In recent years, a lot of contributions are provided in application of qualitative simulation[8][9][10][11]. However, simulation model used in the existing qualitative simulation has a strong limitation where the complex condition and dynamical model can not be expressed by existing tool to make qualitative simulation model[12][13]. To solve the problem, we propose a new unified modeling language to make a qualitative simulation model and explain the way to be used. We propose a modeling method regarding a strength of cause and effect. In actual dynamic systems, there are a lot of characteristics between causes and effects. We give a definition of changing state of strength of relationship between nodes. We also propose a conditional transmission of effect between nodes. By using our proposed methods, in addition to static and non-autonomous system like existing researches, simulation model regarding dynamic and autonomous system can be described. This paper includes three following contributions; first, we proposed a unified modeling language to make a qualitative simulation model; second, users can make more precise simulation graph model; finally, in qualitative simulation research, using our proposed methods, autonomous systems can be analyzed where the system is changed with time passes.

The rest of this paper shows as follows. In Section 2, we explain about definitions of qualitative simulation. Section 3 proposes a strength of relationship and transmission feature of effect between nodes. Then, in Section 4, we propose new description methods to make a graph-based qualitative simulation model including conditional influence transmission, nodes division/integration, and sequential flow of cyclic graph. After that, Section 5 gives short discussion of our proposed method. Finally, we summarize our study and recall the contribution of this paper in Section 6.

2 Qualitative Simulation

In this section, we give some definitions and assumptions for graph-based qualitative simulation. A graph has nodes and arcs that have qualitative state values and effect degrees. Node indicates a factor that is contained in a dynamics. Arc indicates an effect between each adjacent node.

2.1 Definitions on Node

2.1.1 Qualitative State on Nodes

Each node has a state value related with a time passes. Generally, the value is quantitatively defined between $-\infty$ and ∞. When qualitative method is employed, intervals divided by some landmarks are defined as qualitative value. For example, when there is a quantitative space between $-\infty$ and ∞, The qualitative values in the space are defined as follows if we set a landmark on zero.

- $(-\infty,0) \rightarrow [-]$
- $0 \rightarrow [0]$
- $(0,\infty) \rightarrow [+]$

These 3 sort of qualitative values include a value on the landmark.

In qualitative simulation regarding analysis of social dynamics, it is not easy to decide and define the value of landmark because there is not conception about landmark to distinguish a state on node. Each node has a qualitative state on each time step. Thus, we prepare a qualitative state value on node without defined landmark. We define a qualitative state value $[x(t)]$ of node x on time t shown in Table 1[14][15][16]. Normally, the qualitative state value can be shown by H, M, and L. When the qualitative simulation should be conducted more precise, the qualitative state value can be defined as shown Table 1. $H+$ and $L-$ is respectively larger value than H and L. $M+$ and $M-$ indicates probabilistic expression that is respectively never decreased and increased at the next step.

Table 1 Qualitative State Value of $[x(t)]$

$[x(t)]$	Qualitative State
$H+$	Qualitative value of $[x(t)]$ in the next time step is increased rather than the value on the current step.
H	Qualitative value of $[x(t)]$ in the next time step is weakly increased rather than the value on the current step.
$M+$	Qualitative value of $[x(t)]$ in the next time step may be increased rather than the value on the current step.
M	Qualitative value of $[x(t)]$ in the next time step is not increased and decreased rather than the value on the current step.
$M-$	Qualitative value of $[x(t)]$ in the next time step may be decreased rather than the value on the current step.
L	Qualitative value of $[x(t)]$ in the next time step is weakly decreased rather than the value on the current step.
$L-$	Qualitative value of $[x(t)]$ in the next time step is decreased rather than the value on the current step.

2.1.2 Trend of State Change on Nodes

In qualitative simulation, qualitative value of change trend is defined by qualitative differential equation. Generally, qualitative values to be shown change trend are defined that is shown in Table 2. In this paper, we define, in Table 3 , a qualitative value $[\delta x(t)]$ to be shown a trend of state change of node by temporal differentiation.

Table 2 Qualitative State Trend on Node *x*

$[\delta x/\delta t]$	Change Trend of Qualitative Value
$+$	$[x(t)]$ is an increase trend.
0	$[x(t)]$ is stable.
$-$	$[x(t)]$ is a decrease trend.

Table 3 Qualitative State Trend on Node *x*

$[\delta x/\delta t]$	Change Trend of Qualitative Value
I	$[x(t)]$ is an increase trend.
S	$[x(t)]$ is stable.
D	$[x(t)]$ is a decrease trend.

2.2 *Definitions on Arcs*

2.2.1 Transmission of Effects on Arcs

Transmission of effect from arcs is defined by a trend of state change of the arc. In this paper, we define a qualitative value of change trend of nodes effected by adjacent causal nodes shown in Table 4. $D(x,y)$ is a qualitative value of transmission trend where a causal node *x* effects result node *y*.

Table 4 Transmission of Effects

$D(x,y)$	Transmission of Effects
$+$	When qualitative value on node *x* increases (decreases), the qualitative value on node *y* increases (decreases).
$-$	When qualitative value on node *x* decreases (increases), the qualitative value on node *y* increases (decreases).

2.2.2 Transmission Speed of Effects on Arcs

We define a transmission speed where node *x* influences node *y*. Transmission speed depends on a causal node. Transmission speed is defined by V_n. When n=0, the effect is transmitted simultaneously from node *x* to node *y*. When $n \geq 1$, the effect is transmitted with n time step delay. Namely, n shows a time step of transmission delay and has a feature $V_{n-1} < V_n < V_{n+1}$. When the dynamics system is a closed system, n is decided by a law and rule. On the other hands, when the dynamics system is an open system, n is decided by statistical and objective data. Simply, the definition of transmission speed is shown in Table 5. $V(x,y)$ is a qualitative value of transmission speed where node x influences node y.

Table 5 Transmission Speed of Effects

$V(x,y)$ & Transmission Speed
V_0 & Node x influences simultaneously node y, when the node x changes.
V_1 & Node x influences node y with one time step delay after the node x changes.
: & :
V_n & Node x influences node y with n time step delay after the node x changes.
$V_?$ & Transmission speed is unknown.

For example, let us consider the qualitative value V_0 about transmission speed on arc from node x to node y. We consider x is quantity of tasks to do in eight hours and y is quality of the finished task. When the amount of tasks are increased, the quality of finished task goes down. In this relationship between x and y, the effect from node x is influenced simultaneously to node y. On the other hands, for example, let us consider the qualitative value V_1 about transmission speed on arc from node x to node y. We consider x is amount of consumption and y is amount of production in economics dynamics. In the case there is no market expectation, when the amount consumption increases, the amount of production increases slowly. This is important feature of relationship between nodes to make users understand the condition and situation of the system.

2.2.3 Integration of Effects from Multiple Adjacent Nodes

When, a node has multiple adjacent nodes connected by arcs, the qualitative value on the node is changed by multiple influences from the connected nodes.The integration of effects is defined as follows.

- $[\delta z] = [\delta x] + [\delta y]$

Figure 1 shows examples that show integration of effects. Figure 1(a) shows integration of $[\delta x]$ and $[\delta y]$ that has same types of change trend. Figure 1(b) shows integration of $[\delta x]$ and $[\delta y]$ that has different types of change trend. Figure 1(c) shows integration of $[\delta x]$ and $[\delta y]$ that has different types of transmission speed. Figure 1(d) shows an integration of $[\delta x]$, $[\delta y]$, and $[\delta w]$. In Figure 1(a), when qualitative value of $[\delta x]$ and $[\delta y]$ are respectively both $[I]$, integrated qualitative value $[\delta z]$ is also $[I]$. In Figure 1(b), when qualitative value of $[\delta x]$ and $[\delta y]$ are respectively $[D]$ and $[D]$, integrated qualitative value of $[\delta z]$ is not simply determined. In Figure 1(c), when transmission speed is different, integrated qualitative value of $[\delta z]$

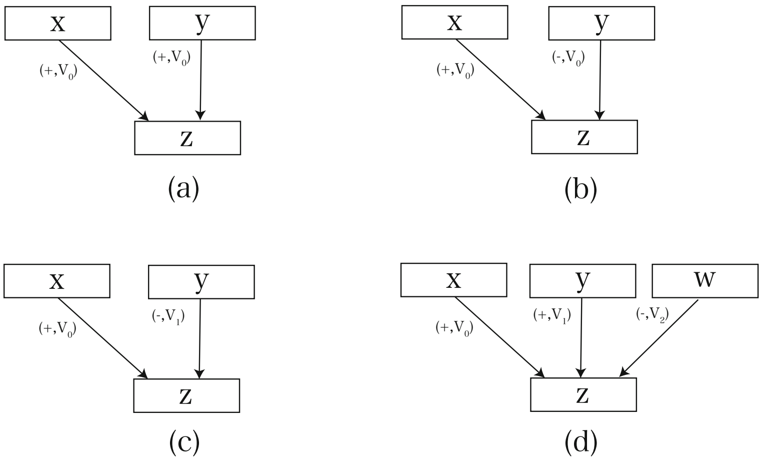


Fig. 1 Integration

is determined by n time step delayed effects of node x and y. Because the delay is n=1from node y to node z in Figure 1(c), the integrated effect at node z at n=2 is determined by the node $x's$ qualitative value at n=1 and the node $y's$ qualitative value at n=0. In Figure 1(d), the calculation rule of integration is employed from simple types of integration that is shown in Figure 1(a) and (b). The calculation rule of integration is shown in Table 6. ”?” shows unknown value.

Table 6 Integration of Effects

$+$	I	S	D	$-$	I	S	D
I	I	I	?	I	D	D	?
S	I	S	D	S	D	S	I
D	?	D	D	D	?	I	I

3 Strength of Cause and Effect

In social dynamics, there is a strength of cause and effect, that is connected with each node by arcs. For example, when a real estate company has a lot of lands and the market price is going up, the company has a chance to get a lot of money selling the lands. This is strong relationship of cause and effect. On the other hands, although a real estate company has a lot of lands and the market price is going down, people may not have a strong motivation to by the lands. This is a weak relationship of cause and effect. Table 7 shows qualitative values $R(x,y)$ showing the strength of cause and effect from nodes x and y. In the rule to conduct a simulation, when a node has influences from multiple nodes that have different strength of effect, the effect of weak relationship is discounted based on the threshold value s. For example, Figure 2 (a), (b), (c) and (d) shows connections with different strength of

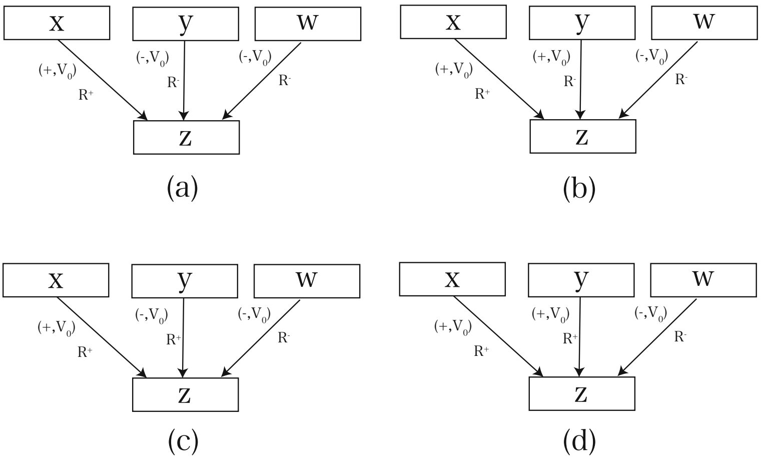


Fig. 2 Strength of Influence

relationships. When the qualitative trends on nodes are given like left of the Table 8, the transmission of effect is influenced like right of the Table 8 (a), (b), (c), and (d) respectively if threshold value s is 2. When $s=1$ and the strength of multiple effect is same, the influences are transmitted by the rule shown in Table 6.

Table 7 Qualitative State Trend on Node $[\delta x(t)]$

$R(x,y)$ & Strength of Cause and Effect
R^+ & Strong relationship between nodes x and y.
R^* & Medium relationship between nodes x and y.
R^- & Weak relationship between nodes x and y.

3.1 Change State of Strength of Cause and Effect

Strength $R(x,y)$ of cause and effect is given as initial value for simulation. However, the strength may be changed with time passes or changed by external impact. In this paper, we define change states of a strength $R(x,y)$ of cause and effect as follows.

- Normal: $R(x,y)$ is fixed. It never changes from the initial value.
- Super Local Effect Nodes $R(x,y) \leftarrow z$: Strength of effect is changed by local effect. It changes based on condition of the adjacent node z connecting with node x or y.
- Local Effect Nodes $R(x,y) \leftarrow (w,v,z)$: Strength of effect is changed by local effect. It changes based on condition of node w.
- Global Effect Nodes $R(x,y) \leftarrow P$: Strength of effect is changed by global effect. It changes based on trend of the system.

Table 8 Strength of Influence

(a)	$x \rightarrow I$, $y \rightarrow I$ and $w \rightarrow I$	$z \rightarrow ?$
(b)	$x \rightarrow D$, $y \rightarrow I$ and $w \rightarrow D$	$z \rightarrow D$
(c)	$x \rightarrow I$, $y \rightarrow I$ and $w \rightarrow D$	$z \rightarrow ?$
(d)	$x \rightarrow I$, $y \rightarrow I$ and $w \rightarrow ?$	$z \rightarrow I$

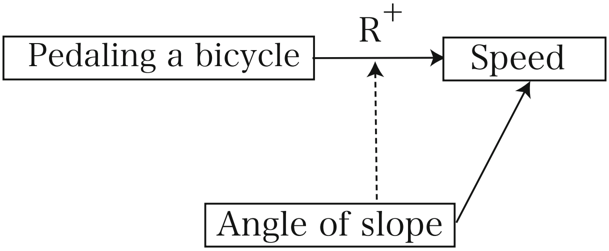


Fig. 3 Super Local Effect Nodes

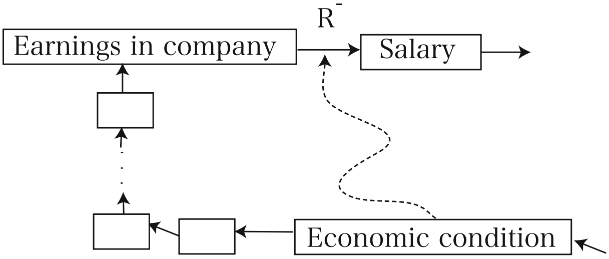


Fig. 4 Global Effect Nodes

We call a super local effect nodes, local effect nodes and a global effect nodes that gives an effect to $R(x,y)$. Normally, $R(x,y)$ is defined initially their characteristics.

Super local effect nodes are connected nodes with node x or y. When a qualitative state/trend on node z is changed, $R(x,y)$ is changed. The super effect nodes can be multiple nodes. Local effect nodes are a partial graph which consists of a set of nodes that makes $R(x,y)$ change. Global effect nodes are a node or larger partial graph which is important in the simulation model. For example, in a macro economics simulation, assumption of food can not be a global effect node but economic condition can be a global effect. When the global effect nodes consist multiple nodes, they are a large partial graph like agricultural industry. Figure 3 and Figure 4 are examples of a part of graph including a super local effect node and a global effect node. In Figure 3, strength of cause and effect between a power to pedal a bicycle and speed is changed by the angle of slope. When pedaling uphill, the pedaling power and speed have a strong relationship with each other in order to pick up speed. Contrary, when going downhill by bicycle, the relationship is not strong because the bicycle does not need a pedaling power to pick up speed. Figure 4 shows an example of global effect nodes showing a salary. The figure shows a relationship between economic condition and salary. Salary is increased by result of the economic condition. There is a relationship between business performance and salary, but they do not have directly relationship with economic condition.

Table 9 Strength of Influence

$+$	R^+	R^*	R^-
R^+	R^+	R^+	?
R^*	R^+	R^*	R^-
R^-	?	R^-	R^-

3.2 Combining of Strength of Cause and Effect

When $R(x,y)$ is influenced by multiple causal nodes, the effect is combined. Table 9 shows a definition of combining strength of cause and effect. ”?” shows a symbol that can not be defined.

4 Partial Graph Functions

4.1 Integration and Division of Nodes

When a simulator is conducted, nodes are divided and integrated based on condition and state value. For example, in order to analyze economical dynamics, when amount of commission fee of the Internet auction changes, the total number of transaction by traders is changed. Generally, node N' is defined as a meta/abstracted

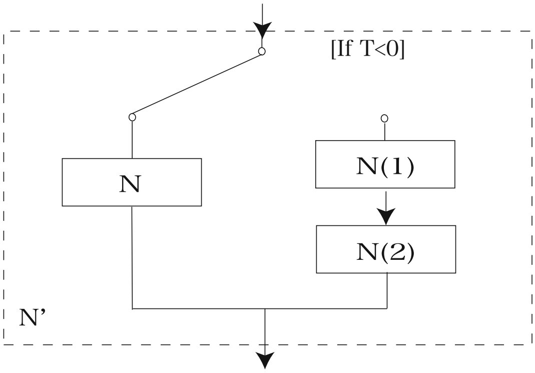


Fig. 5 Integration and Division of Nodes

node. We consider node N is divided in m nodes. When node N is divided, new partial graph that has m modes is generated. Meta-graph N' has node N and divided m nodes. Contrary, there is a case that multiple nodes are merged in one node. Figure 5 shows models of node division and node merge. This is shown as circuit diagram and the switch chooses and control the path based on condition. Condition is shown at right side of the switch. In the Figure 5, when the change trend is lower than zero, the switch changes to integration path. The condition is set by time, qualitative value of other node, trend of the dynamics system, stochastic and several other factors. Nodes are divided in the switch box and it can be recognized as structural model.

4.2 Conditional Nodes

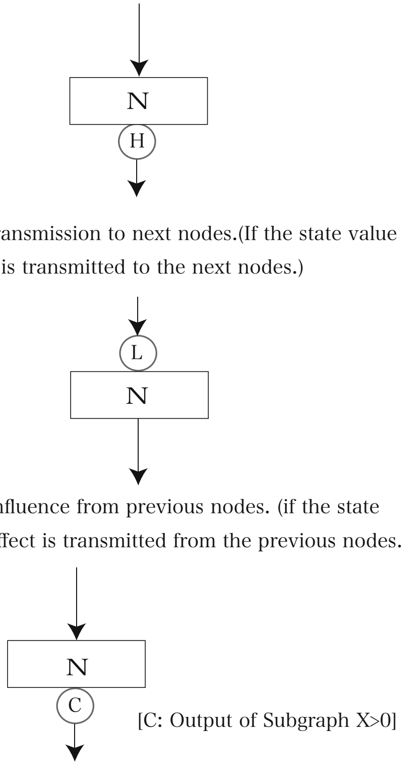
As same as division and integration of nodes, there is a node that has a relationship with other connected nodes in a condition. For example, when the simulation model contains a node showing a decision making, the effect is influenced by a condition. When a company trades with other companies, she makes a decision to trade if items are dealt in at a low price. In this subsection, we define two definitions to express and understand conditional nodes and arcs. They also can be used as a partial graph.

4.2.1 Conditional Transmission Based on State on Nodes

Effect from node sometimes influences to other nodes by a special condition. For example, a company staff gets an incentive salary if he/she make the company increase earnings of himself/herself. This condition is based on qualitative state on node about earnings in qualitative simulation model in company management. Namely, when the condition is set, it does not matter whether the effect is influenced or not. To clarify a conditional transmission, we define a modeling method shown in Figure 6 (a) and (b). Figure 6(a) shows an example of conditional model where a node give an influence to other nodes, when qualitative value on node meet a condition to transmission. When the qualitative state value on node N is H', the node can influence to other connected node. On the other hands, Figure 6(b) shows an example where a node can get effect from previous node, when the When the condition is met, when qualitative value on node meet a condition. When the qualitative state value on node N is L', the node can get an influence from other connected node.

4.2.2 Conditional Transmission Based on Global Dynamics

In addition to a conditional transmission shown in the previous definitions, the effect is sometimes transmitted by multiple nodes like global effect. Concretely, Figure 6(c) shows a modeling where this transmission is decided by qualitative states in multiple nodes. Instead of qualitative state value on node, the condition to transmission is shown by a symbol C. In the Figure 6(c), when a synthetic qualitative states



(a) Condition of transmission to next nodes.(If the state value is high, the effect is transmitted to the next nodes.)

(b) Condition of influence from previous nodes. (if the state value is low,the effect is transmitted from the previous nodes.)

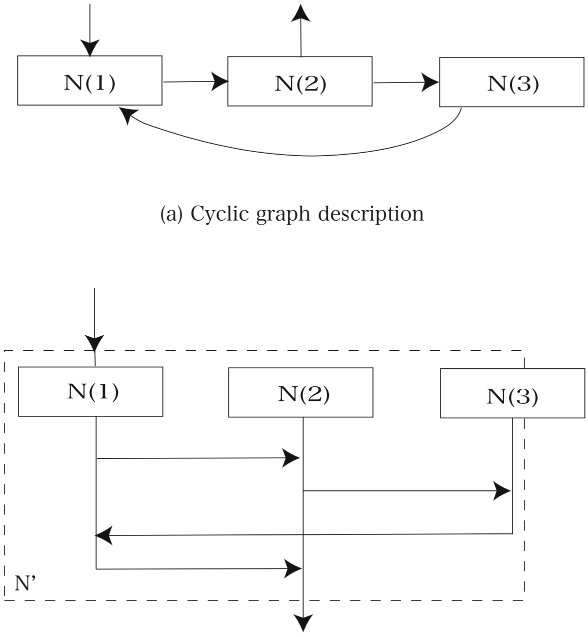
(c) Condition of transmission based on global effect. (If the partial gtaph state is positive, the effect is transmitted to the next nodes.)

Fig. 6 Conditional Transmission

on partial graph is positive, the effect is influenced to connected node from node *N*. For example, when the company makes a decision to trade based on synthetic economic condition, this mode can be employed. This model also can be combined with the strength of transmission shown in the previous section.

4.3 Cycles of Transmission of Effects

The correct order of transmission is not easy to be understood and expressed, if there is a cyclic graph in the model. Generally, it is difficult to express a complicated relationships in a partial graph. For example, Figure 7(a) shows a cyclic graph connecting with same nodes and is not easy to be understood. In order to show the precise flow of effects with time passes in a partial graph, we define an modeling method based on sequential model shown in Figure 7(b). Figure 7(a) and (b) are logically same meaning, but it is not easy to understand how many times the effects are



(a) Cyclic graph description

(b) Sequential description

Fig. 7 Sequential Description of Cyclic Graph

influenced in the cycle in Figure 7(a). In Figure 7(b), using this modeling method, users can easily to understand the start point and number of cycle in the graph.

5 Discussion

In existing research regarding graph-based qualitative simulation, stochastic expression and strength of cause and effect are not referred[17][18][19]. We also can view a lot of probabilistic phenomena and various types of relationship between issues. To solve the problem where we can not make more precise qualitative simulation model, this paper proposed new types of unified modeling set to make a simulation model including integration/division of nodes, visual expression of path, and branch on condition. In expression of causal graph model. We can view and understand a state of dynamic system and structural relationship between multiple nodes. And also, we can grasp a transmission type and its flow by using our proposed model. In text-based qualitative simulation systems like QSIM, it is difficult to make a causal graph model with complex condition and expression. Also, in existing graph-based qualitative simulation systems, we can not conduct a simulation based on complicated model with condition and stochastic operation.

6 Conclusion

In this paper, we proposed unified modeling set to make a graph-based qualitative simulation model. Particularly, the strength of relationship between cause and effect can be useful to make a more precise graph model in qualitative simulation. The strength model can be used by combining with stochastic and integrated definition. This paper also defined three describing methods including division/integration of nodes, transmission with conditions, and sequential flow of influence. By using these description methods, users can view and understand easily the simulation model. Contribution of this paper is (1) we proposed a unified modeling language to make a qualitative simulation model, and (2) users can make more precise simulation graph model, and (3) in qualitative simulation research, using our proposed methods, autonomous systems can be analyzed where the system is changed with time passes.

References

1. Kuipers, B.: Qualitative Reasoning. The MIT Press, Cambridge (1994)
2. Bredeweg, B., Forbus, K.: Qualitative Modeling in Education. AI Magazine (2004)
3. Forbus, K., Carney, K., Sherin, B., Ureel, L.: Vmodel: A visual qualitative modeling environment for middle-school students. In: Proc. Innovative Applications of Artificial Intelligence Conference (2004)
4. de Kleer, J.: Modeling when connections are the problem. In: Proc. International Workshop on Qualitative Reasoning (2006)
5. Bredeweg, B., et al.: Towards a structured approach to building qualitative reasoning models and simulations. Ecological Informatics 3(1-1), 1–12 (2008)
6. Friedman, E.S., Forbus, D.K.: Learning Qualitative Causal Models via Generalization & Quantity Analysis. In: Proc. International Workshop on Qualitative Reasoning (2008)
7. Bouwer, A., Bredeweg, B.: VisiGarp: Graphical Representation of Qualitative Simulation Models. In: Proc. International Workshop on Qualitative Reasoning (2001)
8. Bredeweg, B., Salles, P.: Qualitative models of ecological systems - Editorial introduction. Ecological Informatics 4, 261–262 (2009)
9. Miyasaka, F., Yamasaki, T., Yumoto, M., Ohkawa, T., Komoda, N.: Real-Time Simulation for Fault Detection and Diagnosis using Stochastic Qualitative Reasoning. In: Proc, IEEE International Conference on Emerging Technologies and Factory Automation, vol. 1, pp. 391–398 (2001)
10. Tomas, R.V., Garcia, A.L.: Freeway Traffic Qualitative Simulation, Innovative in Applied Artificial Intelligence. Lecture Notes in Computer Science 3533, 360–362 (2005)
11. Zhang, H., Huo, M., Kitchenham, B., Jeffery, R.: Qualitative Simulation Model for Software Engineering Process. Proc. Australian Software Engineering Conference, 391–400 (2006)
12. Agell, N., Aguado, C.J.: A hybrid qualitative- quantitative classification technique applied to aid marketing decisions. Proc. International Workshop on Qualitative Reasoning (2000)
13. Bredeweg, B., Forbus, D.K.: Qualitative Modeling in Education, vol. AI magazine 24(4), 35–46 (2004)

14. Hiramatsu, A., Hata, S., Ohkawa, T., Komoda, N.: Scenario generator for qualitative simulation system. In: Proc. of, IEEE International Conference on Systems, Man and Cybernetics, pp. 143–148 (1995)
15. Samejima, M., Akiyoshi, M., Komoda, N., Sasaki, R.: Social Consensus Making Support System by Qualitative and Quantitative Hybrid Simulation. In: Proc, IEEE International Conference on Systems, Man, and Cybernetics, pp. 896–901. IEEE Computer Society Press, Los Alamitos (2010)
16. Samejima, M., Akiyoshi, M., Mitsukuni, K., Komoda, N.: Business scenario design support by qualitative-quantitative hybrid simulation. In: Proc. of The IEEE International Conference on e-Technology, e- Commerce, and e-Services and the IEEE International Conference on Electronic Commerce, pp. 401–408 (2007)
17. Bredeweg, B., et al.: Garp3: A New Workbench for Qualitative Reasoning and Modelling. In: Proc. International Workshop on Qualitative Reasoning (2006)
18. Shults, B., Kuipers, B.: Proving properties of continuous systems: qualitative simulation and temporal logic. Artificial Intelligence 92, 91–129 (1997)
19. Liem, J., Bredeweg, B.: OWL and qualitative reasoning models. In: Lecture Notes in Computer Science, vol. 4314, pp. 33–48. Springer, Heidelberg (2007)

Bridging Enterprise Architecture Requirements to ArchiMate

Hyeoncheol Lee and Yeong-tae Song

Abstract. Properly aligned business and Information Technology (IT) can provide competitive edge to an organization. In order to align business with IT, IT should fully support business operations. Enterprise Architecture (EA) can be used to support business and IT alignment. To help design such systems using EA, Enterprise Architecture Frameworks (EAF) may be used. There are frameworks to support EA, such as Zachman Framework, the Department of Defense Architecture Framework (DoDAF), and The Open Group Architecture Framework (TOGAF). They help to design, evaluate, and build the right architecture and reduce the costs of planning, designing, and implementing [8]. ArchiMate is an open and independent architecture modeling language that complements EAF for modeling and visualizing EA. Regardless of chosen EAF, conversion processes from requirements to EA using ArchiMate are not well-defined to the best of our knowledge. Goal-oriented approach is a requirement analysis technique that supports early requirements analysis, which can be used to define a process for bridging EA requirements to ArchiMate. Therefore, we propose guidelines using goal-oriented approach and then apply them to the interoperability of prescriptions in a healthcare system.

1 Introduction

Information technology (IT) supporting business strategies and processes are key elements of successful organizations [1], which refers to as business and IT alignment[6].

Hyeoncheol Lee
Towson University, 8000 York Road, Towson, 21252-0001 Maryland, USA
e-mail: hlee23@students.towson.edu

Yeong-tae Song
Towson University, 8000 York Road, Towson, 21252-0001 Maryland, USA
e-mail: ysong@towson.edu

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 63–78.
springerlink.com

Enterprise Architecture (EA) is understood as a blueprint for the optimal and target-conformant placement of resources in the IT environment for the ultimate support of business functions. Therefore, EA may be used as an approach to solve business and IT alignment problems. Some of the advantages of the alignment of business and IT include more efficient IT operations, cost reduction, and faster, simpler and cheaper procurement [8]. KAM et. al. suggested that a framework and architecture approach can be used in any business scenario [9]. There are many EA frameworks available, such as the Zachman Framework [10][11], Department of Defense Architecture Framework (DoDAF) [12], and The Open Group Architecture Framework (TOGAF) [8].

ArchiMate, developed by the Open Group, is an open and independent architecture modeling language [13]. It is able to describe business processes, organizational structures, information flows, IT systems, and technical infrastructure in an architectural level. ArchiMate can be used with an architecture framework to structure the concept and relationships among architectural items of an enterprise. The ArchiMate language defines three main layers: Business layer, Application Layer and Technology Layer [13]. They are inter-related with cross-layer dependencies. The ArchiMate language complements EA frameworks, such as TOGAF in producing TOGAF artifacts. It is supported by many tools, such as BiZZ design Architect [16] and ARIS ArchiMate Modeler [17].

However, to the best of our knowledge, modeling guidelines for bridging requirements to each layer of ArchiMate do not exist. For that reason, we incorporate goal-oriented approach [20] that is a requirement analysis technique, where goals can be decomposed into sub goals and relationships among goals are specified recursively until it reaches leave operations [20]. It can be used to transform given requirements into the form that can be used by ArchiMate to produce architectural models. This paper focuses on proposing guidelines for the transformation using a goal- oriented approach.

The remainder of the paper is organized as follows: Section 2 provides the background of this research: business and IT alignment, EA, ArchiMate, and goal-oriented approach; Section 3 shows specific guidelines based on the background; Section 4 shows how guidelines are applied to real EA modeling, and Section 5 describes available future research topics; the paper concludes with Section 6.

2 Background

2.1 Business and IT Alignment

Business and IT Alignment is the capacity to demonstrate a positive relationship between information technologies and the accepted financial measures of performance [5]. It also shows a desired scenario where information technology (IT) is used by a business organization to achieve business objectives.

According to Silvus's research, the business and IT alignment are essential elements of a company's concern [1]. A survey by Synstar [2] shows problems about business and IT alignment, specifically on the rising stress levels in IT. It shows that 79% of IT managers in Europe think business and IT Alignment is one of serious problems within their organization. Cumps et.al argued that Business and IT Alignment is a complex and multidimensional problem that remains among the top-10 issues for many organizations [3].

2.2 Enterprise Architecture

Open Group [8] defines enterprise as ”any collection of organization that has a common set of goals”. According to their definition, government, corporation, and a department can be an example of enterprise. It includes information technology services, processes, and infrastructure, as well as partners, suppliers, customers and internal business units. EA covers all functional groups, inforamtion, technology, processes and infrastructure within enterprise [7].

The biggest advantage of EA is the alignment between IT and a business. That motivates practitioners to develop as high-level management in organizations concerning IT system that address organizations and business functions and problems [8]. Therefore, EA aims to align IT with business in these days [4]. Architects address this problems associated with stakeholders’ concerns and problems. To solve the problem, architects should consider the requirements and concerns of stakeholders with EA.

Framework and architecture approach can be used in any business scenario [9]. The Open Group Architecture Framework (TOGAF) is an architecture framework that provides the methods and tools for development, usage, and maintenance of EA [8]. It includes an iterative model process called Architecture Development Method (ADM), which is supported by the best practices and a re-usable set of existing company-specific existing assets.

TOGAF ADM, the core of TOGAF, describes a method for developing an EA. It is used to populate the foundation architecture for some enterprise. It provides guidelines and techniques that support application of the enterprise. Left side of Figure 3 shows the basic structure of the ADM.

EA frameworks, such as Zachman Framework, DoDAF and TOGAF, help to design, evaluate, and build the right architecture, and reduce the costs of planning, designing, and implementing architectures [13]. These kinds of enterprise framework need modeling language to describe, analyze, and visualize architecture.

2.3 ArchiMate

ArchiMate is an open and independent architecture modeling language [13]. ArchiMate is hosted by the Open Group as an open standard. Consulting organizations and tool vendors support ArchiMate. It is able to describe business process, organizational structures, information flows, IT systems, and technical infrastructure. ArchiMate can be used with an architecture framework to structure the concept and relationships of it. It provides a structuring mechanism for architecture domains, layers, and aspects. The Open Group accepted ArhciMate meta-model as a part of TOGAF in 2009 [14].

A key challenge in the development of a general meta-model for EA is to strike a balance between the specificity of languages for individual architecture domains, and a very general set of architecture concepts, which reflects a view of systems as a mere set of inter-related entities. Figure 1 illustrates different levels of the specialization concept. The most general meta-model for system architectures is at the top of the triangle. The design of the ArchiMate language started from a set of relatively generic concepts. These were then specialized towards application at different architecture layers. The meta-model of the architecture modeling concepts used by specific organizations and a variety of existing modeling languages and standards are at the base of the triangle. The language consists of active structure, behavioral elements and passive structure elements. The active structure elements are the business actors, application components and devices that display actual behavior. The active structure concepts are assigned to behavioral concepts to show who or what performs the behavior. The passive structure elements are the objects on which behavior is performed. These are usually information or data objects. The passive structure is also used to represent physical objects. These associations are the core concepts of ArchiMate. Figure 2 shows the concepts. The ArchiMate language defines three main layers based on specialization of the core concepts. The business layer offers products and services to external customers, which is realized in the organization through a business processes performed by business actors. The application layer supports the business layer with application services that are realized by any (software) applications. The technology layer offers infrastructure services (e.g., processing, storage, and communication services) needed to run applications, realized by computer and communication hardware and system software. The general structure of models within the different layers is similar. The same type of concepts and relations are used, although their exact nature and granularity differ.

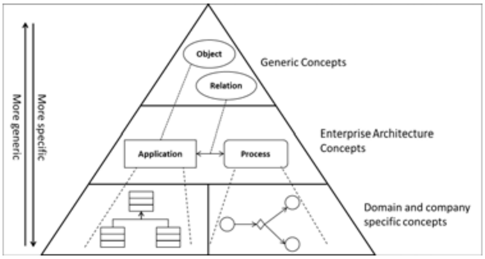


Fig. 1 Different Levels of Specialization in ArchiMate [13]

According to identified core concepts and layers, a framework of nine ”cells” can be illustrated in right side of Figure 3. ArchiMate complements EA framework,

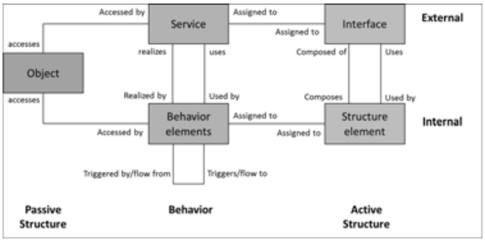


Fig. 2 Core Concept of ArchiMate [13]

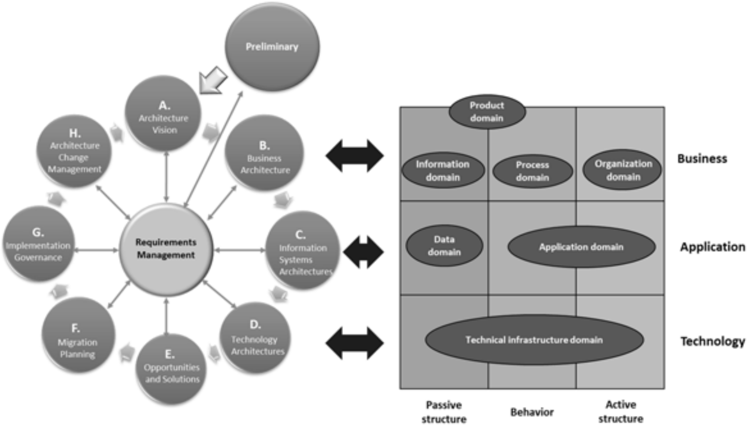


Fig. 3 Framework of Nine Cells in ArchiMate [13] and its relationship with TOGAF ADM

such as TOGAF, in that it provides a vendor-independent set of concepts, including a graphical representation that helps to create a consistent, integrated model, which can be depicted in the form of TOGAF's views. The structure of the ArchiMate language neatly corresponds with the three main architectures as addressed in the TOGAF ADM. Figure 3 shows relationship between TOGAF and ArchiMate. A central issue in EA is business-IT alignment. For this reason, how to match these layers is a very important issue. In ArchiMate, each layer has its concepts that are relevant in their domain and their cross layer dependencies. Each layer's concepts and their cross layer dependencies between each other are illustrated in Figure 4. This business layer includes a number of business concepts, which are relevant in business domain. The application layer and technology layer also include a number of applications and technology concepts.

Tools certificated by the ArchiMate Foundation for compliance with the standard are used to design EA using ArchiMate. BiZZ design Architect is one of EA modeling tools that enables enterprise architects to visualize and analyze EA [16]. It provides the functionality for creating and maintaining consistent EA. It supports modeling EA elements, such as services, functions, processes, applications

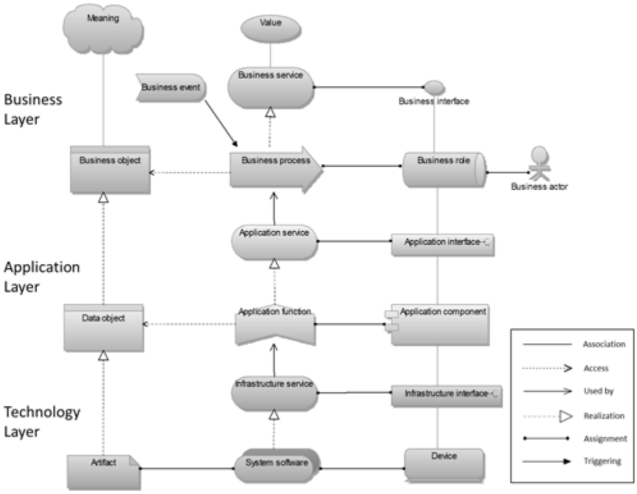


Fig. 4 Layer Concept and Cross Layer Dependencies in ArchiMate[13]

and infrastructure based on ArchiMate. ARIS ArchiMate Modeler is also one of the EA modeling tools [17]. It provides complete integration of ArchiMate framework into ARIS. It uses web-based process and IT design. It enables IT architecture to combine with process strategy, design, and implementation phases. It also provides comprehensive analysis options via queries, reports, and exports of result to various formats, such as, XML and Excel. There are other EA modeling tools, such as Metis, Corporate Modeler and System Architect. These tools have been certified by the ArchiMate Foundation for compliance with the standard [13].

Typically, simple and understandable architecture is more useful in an enterprise environment [9]. ArchiMate defines three domains: business layer, application layer and technology layer. These enable the complexity of architectural domain analysis with the defined meta-model, cross layer dependencies and visualization techniques. In this respect, the ArchiMate is suitable for modeling complex EA because of its features [15].

Even though there are many EA modeling tools available for ArchiMate, modeling guidelines for bridging requirements to each layer of ArchiMate do not exist at the time of this writing.

2.4 Requirements and Goal Oriented Apporach

Software requirements express the needs and constraints placed on a software product that contributes to the solution of some real-world problem [18]. It includes the elicitation, analysis, specification, and validation of software requirements [19].

Goal-oriented approach is a requirement analysis technique. Goals are decomposed into sub goals and specific relationships between goals are specified [20]. It supports early requirements analysis [21] that allows for modeling and analyzing processes that involve multiple participants and intensions. Goals represent strategic interests of actors who model an entity that has strategic goals and intentionality within system or organizational setting [22]. Goal modeling rests on the analysis of an actor goals, conducted from the point of view of the actor, by using three basic reasoning techniques: means-end analysis, contribution analysis, and AND/OR decomposition. In particular, means-end analysis identifies plans, resources and goals that provide means for achieving a goal. Contribution analysis identifies goals that can contribute positively or negatively in the fulfillment of the goal to be analyzed. Therefore, in can be considered as an extension of means-end analysis, with goals as means. AND/OR decomposition combines AND and OR decompositions of a root goal into sub-goals, modeling a finer goal structure. Goal modeling is applied to early and late requirement models in order to refine them and to elicit new dependencies.

3 Guidelines of Bridging for EA Requirements to ArchiMate

3.1 Identify a Top Goal, Sub Goals, and Business Service Based on a Goal Oriented Approach

3.1.1 Identify the Top Goal

In the first step, the top goal that concerns business strategies in terms of business and IT alignment is identified. Therefore, a business strategy for an organization drives the top goal. In other words, the top goal represents the business strategies of an organization.

Before the goal is identified, however, strategic concerns for an organization should be identified. Systems functions and behaviors should support the top goal.

3.1.2 Examine What Is Needed to Satisfy the Top Goal

We examine what is needed to satisfy the top goal in the second step. Once the top goal is identified, the goal is decomposed into sub goals. To decompose the sub goals, a variety of sub goals are concerned in this step. The most important thing in this step is that the sub goals should satisfy the top goal. Figure 5 shows the relationship between the top goal and the sub goals.

3.1.3 Incremental Expansion and Identifying Business Services to Meet Upper Level Goals

After the first level of sub goals is identified, sub goals can be decomposed into lower level of sub goals. Lower level sub goals also must satisfy upper level sub goals. When all relevant goals have been analyzed, the business services reflecting as-is business services or to-be business services are derived. Business services must satisfy upper level sub goals. The services are going to be business services and roles in the business layer of ArchiMate. Figure 5 shows the relationship among upper level sub goals, lower level sub goals, and business services.

3.2 *Elicit Required Elements for ArchiMate and Model EA Using ArchiMate*

3.2.1 Specify Elements of Business Layer in ArchiMate

In this step, we specify elements of the business layer in ArchiMate based on elicited goals and services. First of all, the business service(s) identified in the previous step are decomposed into two parts: a business service and a business role in ArchiMate, because a business service exposes the functionality of a business role. According to ArchiMate specification rules, the name of a business service should be a verb ending with ”-ing” and the name of a business role should be a noun. The names of a business service and a business role are transformed by the rule. A business interface defines how the functionality of a business role can be used by other business roles or environments. Therefore, a business interface is identified by asking how a business role can connect with its environment, such as a business service. The identified business interface connects a business service and a business role in the next step. Business collaboration is a configuration of two or more business roles. It can be identified based on the business roles when they need collaborative functions. Business behavior elements, events, functions, actors, data objects and values are identified and specified by analyzing business services. These can be specified through WHO, HOW, WHAT, and WHY questions. WHO questions help architects identify actors who play a role in the business services. HOW questions help architects identify business behavior elements, such as business processes, business functions, business interaction and business events, which specify how business services proceed. WHAT questions help architects identify the kind of data used in the business service. A value is what a part gets by selling or making available some product or service. WHY questions help architects identify value for a business services. Since a value represents the reasoning for a business service or product, values can be considered sub goals. A business representation is a perceptible form of information carried by a business object. It can be specified in terms of medium (electronic, paper, audio, etc.) or document format (HTML, ASCII, PDF, etc.) Therefore, business representation is derived from business objects by analyzing needed form to represent business objects or what is needed to realize business objects. A meaning is a representation-related counterpart of a value. Therefore, a meaning is identified by analyzing specific contents of a representation, and also documents or information related to business object and business service can be analyzed for identifying a meaning. A contract is a formal or informal specification of an agreement. Therefore, a contract is derived from requirements specifications, proposals and agreements. A product is a coherent collection of services, accompanied by a contract or set of agreements. Therefore, a product is identified by analyzing a business service and a contract. Moreover, a product can be a business service for upper level sub goals or a top goal. Each element will be arranged and designed

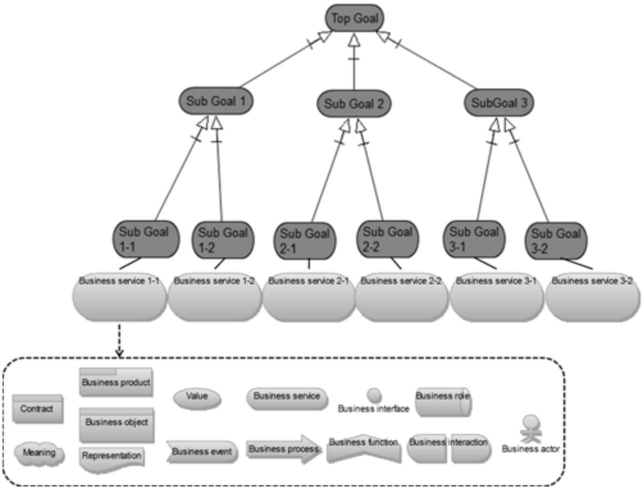


Fig. 5 Relationship between Goals, Business Services and derived Elements

into ArchiMate model in the next step from different viewpoints. Figure 5 shows relationship among goals, business services and each element derived in this step.

3.2.2 Arrange Each Elements to Business Layer of ArchiMate

In the final step, we arrange each element derived from the previous step to the business layer of ArchiMate. The business layer of ArchiMate can be arranged and designed based on the meta-model of ArchiMate, and the artifacts derived from the previous step. In addition, each element can be presented in different architecture viewpoints, such as function viewpoint, process viewpoint and service realization viewpoint. Figure 6 shows the relationship among each element derived in the previous step and the business layer of ArchiMate. After the business layer of ArchiMate is modeled, the application layer and the technology layer can be identified based on the business layer. However, this paper does not cover the modeling application layer or the technology layer because it focuses on bridging the enterprise requirements to the business layer of ArchiMate.

3.3 Contribution of Bridging EA Requirements to ArchiMate

There are significant advantages from the proposed guidelines that bridges EA requirements to ArchiMate based on the goal oriented approach. Firstly, the goals in the goal oriented approach are derived from business strategies and concerns. The business services reflect and support the goals. Therefore, we can model EA based on business services that reflect and support business strategies and concerns. Through this, we can achieve business and IT alignment, which concerns many IT

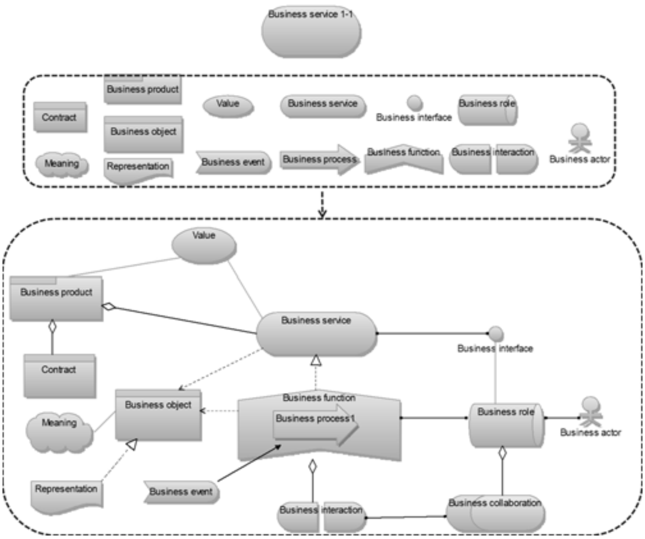


Fig. 6 Matching Each Element Derived from Figure 5. to ArchiMate

managers. Secondly, the guidelines help architects model and design EA based on EA requirements because the proposed guidelines include specific and well-defined steps for the early ArchiMate design phase.

4 Case Study

4.1 Specification of Interoperability for Healthcare

In a healthcare context, EA has the potential to facilitate integrating healthcare units with business architecture. The healthcare domain is considered significant because of its complexity, sensitivity of operations and human involvement. Adapting appropriate EA for healthcare management has a significant impact on healthcare organizations as discussed in [9]. Information sharing within a company or cross enterprise is the core need of any enterprise in order to be effective in business operations. Utilizing resources efficiently expedites the business processes of enterprise. Hussain et al. suggested that healthcare information is more complex and has more diverse dimensions than any other enterprise domains, and sharing information by integrating various healthcare information systems is a great challenge [24]. One of the key aspects in this challenge lies on their interoperability. For this reason, we have chosen interoperability for healthcare domain as our case study.

Healthcare Information Technology Standards Panel (HITSP) presented the Electronic Healthcare Record (EHR) Centric Interoperability specification [23]. The Interoperability Specifications consolidate all information exchanges that involve an EHR System. The Interoperability Specifications are organized as a set of HITSP

Capabilities. Each capability specifies a business service that an EHR system addresses. In this section a capability that describes communicate ambulatory, and long term care prescription is analyzed by guidelines proposed in the previous section.

Communicate Ambulatory and Long Term Care Prescription addresses interoperability requirements that support electronic prescribing in the ambulatory and long term care environment. The capability supports the transmittal of new or modified prescriptions, transmittal of prescription refills and renewals, communication of dispensing status, and accessing to formulary and benefit information. This capability has a five-construct list that must be implemented to satisfy the capability. Table 1 shows list of constructs for the capability.

Table 1 List of Constructs for the Capability [23]

Constructs	Description
Administrative Transport to Health Plan	Provides the transport mechanism for conducting administrative transactions with health plans
Patient Health Plan Eligibility Verification	Provides the status of a health plan covering the individual, along with details regarding patient liability for deductible, co-insurance amounts for a defined base set of generic benefits or services
Medication Dispensing Status	Provides a medication prescriber the dispensing status of an ordered prescription This Transaction is used for original prescriptions, refills and renewals
Medication Orders	Defines transactions between prescribers and dispensers. It is used for new prescriptions, refill requests, prescription changes requests and prescription cancellations
Medication Formulary and Benefits Information	Performs an eligibility check for a specific patient's pharmacy benefits and obtains the medication formulary and benefit information

4.2 Identify a Top Goal, Sub Goals, and Business Service Based on Goal Oriented Approach

4.2.1 Identify a Top Goal

HITSP's first concern is the capability of interoperability. It is trying to provide the ability for exchanging information about prescriptions. Therefore, providing interoperability for prescriptions is identified as a top goal.

4.2.2 Examine What Is Needed to Satisfy the Top Goal

The sub goals are examined to satisfy the top goal in this step. HITPS had presented what the capability supports. Those can be sub goals of providing interoperability for prescriptions. Therefore, the top goal is decomposed into four sub goals: Transmittal of New or Modified Prescriptions, Transmittal of Prescription Refills and Renewals, Communication of Dispensing Status, and Accessing to Formulary and Benefit Information. Figure 7 shows the relationship between the top goal and its sub goals.

4.2.3 Incremental Expansion and Identifying Business Services to Meet Upper Level Goals

Since HITPS has already identified what the capability supports, which expose all sub goals, the sub goals identified in the previous step cannot be decomposed into another lower level of sub goals. After all relevant goals have been analyzed, the business services reflecting as-is business services or to-be business services are derived. In the specification, the list of capability exposes business services. Therefore, Ordering medication, Checking Medication Dispensing Status, Verifying Patient Health Plan Eligibility, Checking Medication Formulary and Benefit Information are identified as services to meet upper level goals. Note that a business service is able to be associated with two or more upper level goals. For example, Ordering medication is used to define transactions between prescribers (who write prescriptions) and dispensers (who fill prescriptions). It is used for new prescriptions, refill requests, prescription changes requests, and prescription cancellations. Therefore Ordering medication should support Transmittal of New or Modified Prescription, as well as Transmittal of Prescription Refills and Renewals. Moreover, a sub goal can be supported by two business services. For example, Accessing to Formulary and Benefit Information needs two business services: Verifying Patient Health Plan Eligibility and Checking Medication Formulary and Benefit Information. Figure 7 shows the relationship between goals and business services.

4.3 Elicit Required Elements for ArchiMate and Model EA Using ArchiMate

4.3.1 Specify Elements of Business Layer in ArchiMate

In this section, we specify the elements of business layers in ArchiMate based on elicited goals and services. ArchiMate provides many different views on designing EA. We identify key elements of the business layer in ArchiMate for the business function viewpoint and the business process viewpoint. The business function viewpoint requires a business role and function. The business process viewpoint requires a business process, events, data object and representation. To help understand the overall concept of ArchiMate, we identify more additional elements: business service, interface, meaning and actor. First of all, the business service, Verifying Patient Health Plan Eligibility, is decomposed into business service and its role in ArchiMate because a business service exposes functionality of business roles. According to the ArchiMate specification rules, the business service will be Patient Health Plan Eligibility Verifying and the business role will be Patient Health Plan Verifier.

A business interface is identified by asking how a business role can connect with its environment, such as a business service. Since web forms are used to connect the business service with the business role in this case, Web Form is identified as an interface. The actor is identified by asking a WHO question. One answer to ”who is responsible for Verifying Patient Health Plan Eligibility” question can be a receptionist. Receptionists play the role when they accept payment for services or for registering patients. The data object is identified by asking a WHAT question. All data about patients’ information is saved in Patient Health Plan Information in EHR. Receptionists access Patient Health Plan Information in EHR and verify the patient’s health plan. Therefore, Patient Health Plan Information in EHR can be a data object for Verifying Patient Health Plan Eligibility. The representation that describes perceptible form of information carried by a business object will be derived from the business object by analyzing what kind of form is needed to represent the business object or what is needed to realize the business object. Therefore, business representation will be a form that include Patient’ Name, SSN, Insurance Company, Health Plan. A meaning is identified by analyzing specific contents of a representation. Therefore, a meaning for the identified representation is Policy Explanation and Coverage Description. The behavior elements, such as business function, business processes and events can be identified by asking a HOW question. Patient Health Plan Getting, Patient Health Plan Evaluating and Health Plan Applying can be identified as business functions. Access to EHR, Get Patient Information from EHR, Evaluate Patient Health Plan, and Apply Health Plan to Payment can be identified as certain processes of Verifying Patient Health Plan Eligibility. Request for Health Plan Verifying is identified as events. Figure 7 shows the relationship between goals, business services and the identified elements of the business layer.

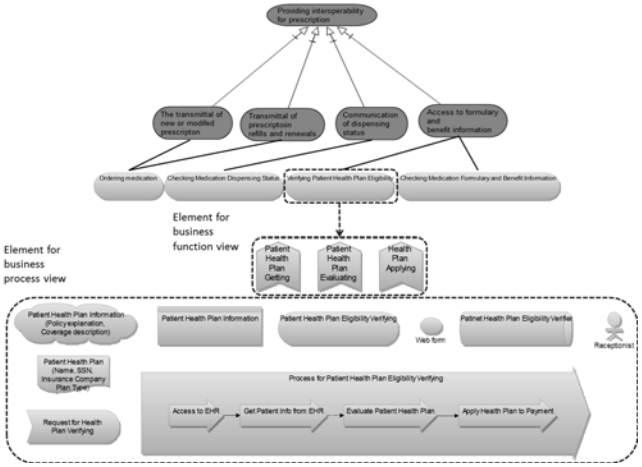

Providing interoperability for prescription
The transmittal of new or modifed prescripton
Transmittal of prescriptoiin refills and renewals
Communication of dispensing status
Access to formulary and benefit information
Ordering medication
Checking Medication Dispensing Status
Verifying Patient Health Plan Eligibility
Checking Medication Formulary and Benefit Information
Element for business function view
Element for business process view
Patient Health Plan Getting
Patient Health Plan Evaluating
Health Plan Applying
Patient Health Plan Information (Policy explanation, Coverage description)
Patient Health Plan Information
Patient Health Plan Eligibility Verifying
Web form
Patinet Health Plan Eligibility Verifier
Receptionist
Patient Health Plan (Name, SSN, Insurance Company Plan Type)
Process for Patient Health Plan Eligibility Verifying
Request for Health Plan Verifying
Access to EHR
Get Patient Info from EHR
Evaluate Patient Health Plan
Apply Health Plan to Payment

Fig. 7 Relationship between Goals, Business Services and Other Elements of Business Layer for Prescription Interoperability

4.3.2 Arrange Each Element to Business Layer of ArchiMate

In the previous step, required elements for business function view and business process view are identified. Each element is directly arranged and designed based on the meta-model of ArchiMate and specific viewpoints. In addition, other more elements are added and designed in the model to understand the relationship among elements in the business layer. Figure 8 and 9 show the resulting ArchiMate models in the business function view and in the business process view.

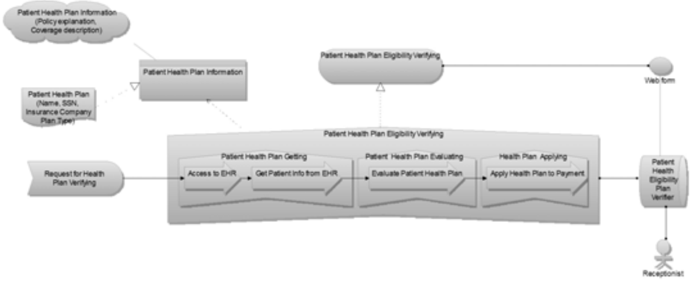

Patient Health Plan Information (Policy explanation, Coverage description)
Patient Health Plan Information
Patient Health Plan (Name, SSN, Insurance Company Plan Type)
Patient Health Plan Eligibility Verifying
Web form
Patient Health Plan Eligibility Verifying
Patient Health Plan Getting
Patient Health Plan Evaluating
Health Plan Applying
Request for Health Plan Verifying
Access to EHR
Get Patient Info from EHR
Evaluate Patient Health Plan
Apply Health Plan to Payment
Patient Health Eligibility Plan Verifier
Receptionist

Fig. 8 Business Function Viewpoint for Verifying Patient Health Plan Eligibility

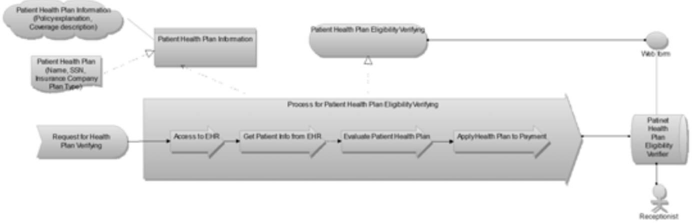

Patient Health Plan Information (Policy explanation, Coverage description)
Patient Health Plan Information
Patient Health Plan (Name, SSN, Insurance Company Plan Type)
Patient Health Plan Eligibility Verifying
Web form
Process for Patient Health Plan Eligibility Verifying
Request for Health Plan Verifying
Access to EHR
Get Patient Info from EHR
Evaluate Patient Health Plan
Apply Health Plan to Payment
Patinet Health Plan Eligibility Verifier
Receptionist

Fig. 9 Business Process Viewpoint for Verifying Patient Health Plan Eligibility

5 Future Study

We have proposed guidelines for bridging EA requirements to ArchiMate. However, it is restricted to the business layer of ArchiMate. Since ArchiMate consists of three layers: Business Layer, Application Layer, and Technology Layer, Overall guideline, design principle, and process to design overall layers are needed.

Moreover, ArchiMate can be extended by adding new attributes to concepts and relations [13]. The core of ArchiMate contains only the concepts and relationships that are necessary for general architecture modeling. However, users might want to be able to, for example, perform model-based performance or cost calculations, or to attach supplementary information (textual, numerical, etc.) to the model elements.

6 Conslusion

In this paper, we have proposed the guidelines on how to bridge EA requirements to ArchiMate based on a goal-oriented approach. The guidelines describe specific steps for bridging EA requirements to ArchiMate. They consist of two big steps: 1. Identify a top goal, sub goals and business service based on goal-oriented approach and 2. Elicit required elements for ArchiMate and model EA using ArchiMate. In the first step, we identified a top goal, sub goals, and business services. In the second step, we identified required elements for designing and modeling ArchiMate, and then modeled EA using ArchiMate based on identified elements. In the case study, the interoperability for prescriptions from healthcare system was modeled by the guidelines in two viewpoints, business function viewpoint and business process viewpoint. We showed that these guidelines can be applied to the interoperability for prescriptions from the healthcare system. Through the guidelines and case study, we showed that requirements can be transformed to the business layer of ArchiMate, which can help architects design and model EA using ArchiMate.

References

1. Gilbert Silvus, A.J.: Business & IT Alignment in theory and practice. In: HICSS 2007.IEEE, Los Alamitos (2007)
2. SYNSTAR, The Pressure Point Index:V,2004 sysntar(2004)
3. Cumps, B., Viaene, S., Dedne, G.: Managing for Better Business-IT Alignment. IT Pro (2006)
4. Brown, E.J., Yarbery Jr., W.A.: The effective CIO: How to achieve outstanding success through strategic alighment, finacial management, and IT governance. An Auerbach Book (2009)
5. Strassmann, P.A.: What is Alignment: Alignment is the delivery of the required results. Cutter IT Journal (1998)
6. Wegmann, A., Regev, G., Rychkova, I., Le, L.-S., Cruz, J.D.D.L., Julia, P.: Business and IT Alignment with SEAM for EA. In: EDOC 2007, IEEE, Los Alamitos (2007)
7. IEEE Standard 1471-2000. IEEE Recommended Practice for Architectural Description of Software-Intensive Systems.IEEE, Los Alamitos (2000)
8. The OpenGroup, TOGAF (2009), http://www.opengroup.org/togaf/
9. Ahsan, K., Shah, H., Kingston, P.: Healthcare Modelling thorugh EA: A Hosptal case. In: International Conference on Infromation Technology (2010)
10. The Zachman Frameworks: The Official Concise Definition Zachman International (2008)
11. Zachman, J.A.: The Zachman Framework for EA: Primer for Enterprise Engineering and Manufacturing. In: Zachman International, electronic book (2003)
12. United States Department of Defense, DoD Architecture Framework 2.02 (2010)
13. The OpenGroup, ArchiMate 1.0 Specification (2009), http://www.archimate.org/
14. Koing, J., Zhu, K., Nordstorm, L., Ekstedt, M., Lagerstrom, R.: Maaping the Substation Configuration Language of IEC 61850 to Archimate. In: EDOCW (2010)
15. Steen, M.W.A., Akehurst, D.H., ter Doest, H.W.L., Lankhorst, M.M.: Supporting Viewpoint-Oriented Enterprise Archtiecture. In: EDOC (2004)

16. BiZZdesign Architect (2010), http://www.bizzdesign.com
17. ARIS ArchiMate Modeler (2010), http://www.ids-scheer.com/en/ARIS/ARIS-Platform/AIRS-ArchiMate_Modeler
18. Kotonya, G., Sommerville, I.: Requiremetns Engineering: Processes and Techniques. John Wiley & Sons, Chichester (2000)
19. Abran, A., Pierre, Tripp, L.L.: SWEBOK(Software Engineering Body of Knowledge). IEEE, Los Alamitos (2004)
20. Giorgini, P., Rizzi, S., Garzetti, M.: Goal-Oriented Ruqirement Analysis for Data Warehouse Design. In: DOLAP (2005)
21. Dardenne, A., van Lamsweerde, A., Fickas, S.: Goal-directed requirements acquisition. Science of Computer Programming (1993)
22. Bresciani, P., Perini, A., Giorgini, P., Giunchgia, F., Mylopoulos, J.: Tropos: An Agent Oriented Software Development Methodology. Kluwer Academic, Dordrecht (2004)
23. Healthcare Information Technology Standards Panel. HITSP HER-Centric Interoperability Specification (2009)
24. Hussain, M., Afzal, M., Farooq Ahmad, H., Khalid, N., Ali, A.: Healthcare Applications Interoperability through Implementation of HL7 Web Service Basic Profile. Sixth international Conference on Information Technology: New Generations (2009)

Non-coherent Logarithmic Detector Performance in Log-Normal and Weibull Clutter Environment (UWB Automotive Application)

Nandeeshkumar Kumaravelu, Jong-Hun Lee, Seok-Jun Ko,
and Yung-Cheol Byun*

Abstract. High range resolution ultra wideband radars attract considerable attention as short range automotive radar for target detection and ranging. Radar signal reflected from a target often contains unwanted echoes called as clutter, so the detection of target is difficult with clutter echoes. Therefore, it is important to investigate the radar detector performance for the better detection of the reflected signals. This paper analyzes the detection performance of non-coherent logarithmic detector in log normal and weibull clutter environment. The detection probability of the detector is obtained with different mean and variance value in log normal clutter environment, and with different shape parameter and scale parameter values in weibull clutter environment at different system bandwidth of 1GHz, 500MHz and 100MHz.

Keywords: Logarithmic detector, clutter, log-normal clutter, weibull clutter, Coherent and non-coherent integration.

1 Introduction

UWB impulse radar is used for short range measurements in the Intelligent Transport System (ITS) [1]. Since Short-Range Radar (SRR) [2, 3] is used in vehicles, the Federal Communications Commission (FCC) has confirmed the spectrum

Nandeeshkumar Kumaravelu · Seok-Jun Ko
Dept. of Electronics Engineering, Jeju National University, Korea

Jong-Hun Lee
Daegu Gyeongbuk Institute of Science & Technology, Daegu, Korea

Yung-Cheol Byun
Dept. of Computer Engineering, Jeju National University, Korea
e-mail: nandeeshforu@gmail.com,{sjko,ycb}@jejunu.ac.kr

* Corresponding author.

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 79–88.
springerlink.com

from 22 to 29GHz for UWB radar with a limit power of –41.3dBm/MHz [4-5].In Intelligent Transport System automotive radar facilitates various functions which increase the driver's safety and convenience. Exact measurement of distance and relative speed of objects in front, beside, or behind the car allows the realization of systems which improve the driver's ability to perceive objects during bad optical visibility or objects hidden in the blind spot during parking or changing lanes [6]. Using radar technology it is possible to detect a target more accurately by the high resolution range profile because the radar resolution is smaller than the vehicle size.

In this paper, we use the system bandwidth of 1GHz, 500MHz and 100MHz centered at 24.125GHz for analyzing the performance of the logarithmic detector. In the UWB automotive short range radar the clutter echoes are the echoes from the objects in the road environment. The road clutter resembles log-normal distribution for a bandwidth of 500MHz or more and it resembles weibull distribution for 100MHz bandwidth [7]. Here we discussed the performance of the non-coherent logarithmic detectors in log normal and weibull clutter environment.

The organization of this paper is as follows. In Section 2, the system model is described. In Section 3, clutter characteristics are described. In Section 4, detector model is described. In Section 5, results were shown in plots. In Section 6, conclusion is given.

2 UWB Radar System

UWB radar system is split into two parts: the transmitter and the receiver as shown in the figure 1. First, in the transmitter, the Gaussian pulse is generated at each time that the Pulse Repetition Frequency (PRF) generator triggers the pulse generator. The Gaussian pulse (T_P) has a sub-nano second duration.

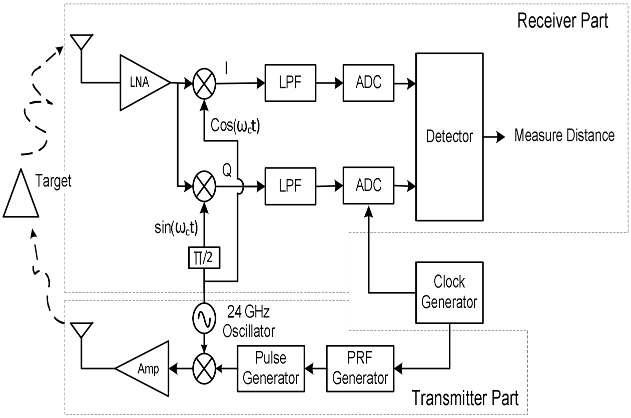


Fig. 1 Block Diagram of a UWB radar system

Therefore we can write the transmitted signal as follows,

$$s(t) = A_T \cdot \cos(2\pi f_c t + \varphi_0) \cdot p_n(t) \qquad (1)$$

$$p_n(t) = \sum_{n=-\infty}^{+\infty} p(t\text{-}n \cdot T_{PRI}) \qquad (2)$$

where $p_n(t)$ Gaussian pulse train.

The parameters employed in this UWB radar system are described as follows;
A_T is the amplitude of single transmit pulse,
φ_0 is the phase of the transmit signal,
f_c is the carrier frequency, and T_{PRI} is the pulse repetition time.

Since the range resolution of the UWB radar system is much less than the extent of the target, the echo signal is the summation of the time-spaced echoes from the individual scattering centers that constitute the target [8]. Therefore, in this paper, we can assume that the target has L independent reflecting cells. The target model is written as,

$$h(t) = \sum_{l=0}^{L-1} \alpha_l \cdot \delta(t - \tau_l) \qquad (3)$$

where the number of scatters L, the amplitude of the scatters α_l, and the time delays of the scatters τ_l are all unknown, the baseband complex received signal reflected from the target is given by

$$\bar{r}(t) = A_T \sum_{n=-\infty}^{+\infty} \sum_{l=0}^{L-1} \alpha_l \cdot e^{j\theta_l} p(t - nT_{PRI} - \tau_l) + \bar{n}_e(t) \qquad (4)$$

where $\bar{n}_e(t)$ is the reflected clutter signal from the clutter.

3 Clutter Characteristics

a) Log Normal Distribution

A log normal distribution is a probability distribution of a random variable whose logarithm is normally distributed. If X is a random variable with a normal distribution, then Y=*exp*(X) has a lognormal distribution.

The probability density function of the lognormal distribution is;

$$f_X(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, x > 0 \qquad (5)$$

where μ and σ are the mean and standard deviation of the variables of natural logarithm [9] .

b) Weibull Distribution

The weibull distribution is a continuous probability distribution. The probability density function of a weibull random variable *X* is,

$$f(x;\lambda,k)=\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^{k}}, x\geq 0 \qquad (6)$$

where λ>0 is scale parameter and k >0 is shape parameter of the distribution [9]. The Probability density function of the log-normal distribution and weibull distribution is shown below,

Fig. 2 Log Normal Probability Density Function

In the figure 2, the probability density funtion of the log normal distribution at different mean μ and standard deviation σ is shown.

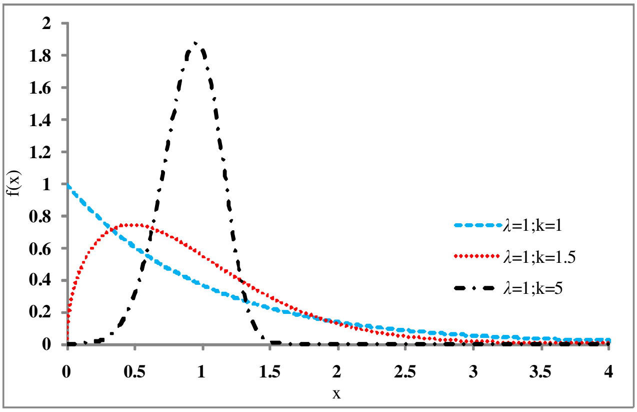


Fig. 3 Weibull Probability Density Function

In the figure 3, the probability density function of the weibull distribution at different λ scale parameter and k shape parameter is shown.

4 Non-coherent Logarithmic Detector

First, in the receiver, the signal detector of the UWB radar must determine that a signal of interest is present or absent. And then the UWB radar processes it for range determination and measuring velocity.

In this paper, we use non-coherent logarithmic detector. The detector consists of coherent range gate's memory, non-coherent range gate's memory, coherent integrator, and non-coherent integrator. The coherent and non-coherent range gate's memory size (*M*) is less than maximum range and indicates the total number of target range to be tested. These are used as buffer to coherently and non-coherently integrate.

Therefore, at every T_{PRI}, we use the samples as much as the range gate's memory size (*M*). At every T_p, the in-phase (*I*) and quadrature (*Q*) sampled values are used as the input of the detector. The switch-I is shifted at every sampling time T_p and the samples at each range gate are coherently integrated. It takes $N_c \cdot T_{PRI}$ time to coherently integrate and dump for all range gates; N_c indicates the coherent integration length. If the round trip delay (τ) from target is equal to the time position of *i*-th range gate ($i \cdot T_p$), then the target range can be expressed as $i \cdot T_p/2 = i \cdot \Delta R$ where the range resolution ΔR is given by following formula $\Delta R = c \cdot T_p/2$. From the above assumption and to find whether the target is present or not, the output of the coherent integrator can be distinguished between the two hypotheses.

$$H_1: \ \overline{X}_m(i) = \frac{A_T \alpha}{N_c} \sum_{n=mN_c}^{(m+1)N_c-1} e^{j\theta_l} p(t - nT_{PRI} - \tau_l) + \overline{n}_e(i) \tag{7}$$

$$H_0: \ \overline{X}_m(i) = \frac{1}{N_c} \sum_{n=mN_c}^{(m+1)N_c-1} \overline{n}_e(i) \tag{8}$$

where m indicates the m-th coherent integration and H_1 is for $\tau = i \cdot T_p$ and H_0 for $\tau \neq i \cdot T_p$. Also we assume that the sampling rate of the ADC is equal to the pulse width. The baseband received signal is sampled at peak point of $p(t)$.Then the values of the coherent integration for each range gate ($\overline{X}(i)$, i=1, 2, , *M*) are stored in the coherent range gate's memory.

After the coherent integration, the power $Y(i)$ is given by

$$Y(i) = \ln\sqrt{(X^I(i))^2 + (X^Q(i))^2} \tag{9}$$

Then power $Y(i)$ is integrated at every $N_c \cdot T_{PRI}$.

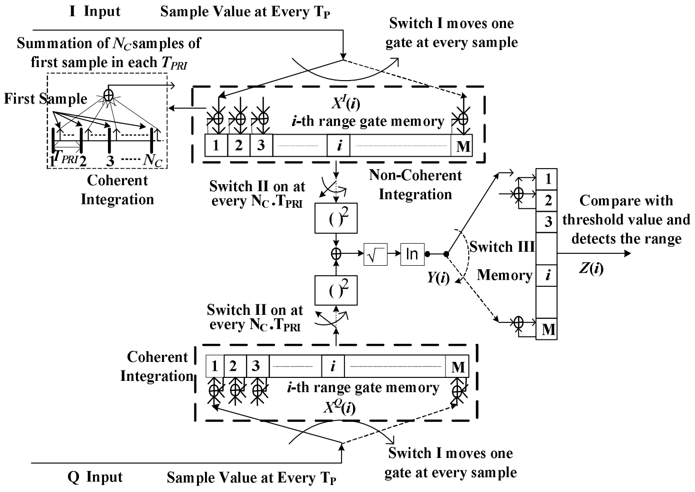

I Input
Sample Value at Every T_P
Switch I moves one gate at every sample
Summation of N_C samples of first sample in each T_{PRI}
First Sample
T_{PRI} 2 3 ····· N_C
Coherent Integration
$X^I(i)$
i-th range gate memory
1 2 3 i M
Switch II on at every $N_C \cdot T_{PRI}$
Non-Coherent Integration
$()^2$
ln
Switch III
$Y(i)$
Memory
1 2 3 i M
Compare with threshold value and detects the range
$Z(i)$
$()^2$
Switch II on at every $N_C \cdot T_{PRI}$
Coherent Integration
1 2 3 i M
i-th range gate memory
$X^Q(i)$
Switch I moves one gate at every sample
Q Input
Sample Value at Every T_P

Fig. 4 Block diagram of the receiver with logarithmic detector

The total number of the non-coherent integration is N_n that means $N_n \cdot N_c \cdot T_{PRI}$ time duration. When the power is stored in the i-th non-coherent range gate's memory at every $N_c \cdot T_{PRI}$, then the output of the non-coherent integration can be written as

$$Z(i) = \frac{1}{N_n} \sum_{m=1}^{N_n} Y_m(i) \tag{10}$$

where $Y_m(i)$ is the power at $m \cdot N_c \cdot T_{PRI}$.

5 Computer Simulation Result

In this paper, we assume that each clutter is independent and uncorrelated. The parameters we used are, coherent integration number N_c is 200 and the non-coherent integration number N_n is 100. The empirical data's are available for the 24 GHz UWB automotive short range radar clutters. The experiment has been carried out in the University of Kitakyushu at different clutter environment and the values are tabulated [7]. The tabulated values are used for the checking the performance of the logarithmic detector in simulation.

Table 1 Clutter Characteristics

BW (Band Width)	Log normal		Weibull	
	μ	σ	λ	k
1GHz	5.0	0.8	1.4	6.7
500MHz	5.2	0.8	1.6	6.9
100MHz	5.7	0.7	2.5	8.7

In the figure 5, the performance of the non-coherent logarithmic signal detectors is analyzed for log normal clutter with coherent integration number of 200 and non-coherent integration number of 100. The detection probability of logarithmic detector is optimum at the system bandwidth of 100MHz than 1GHz and 500MHz, because of the small variance value.

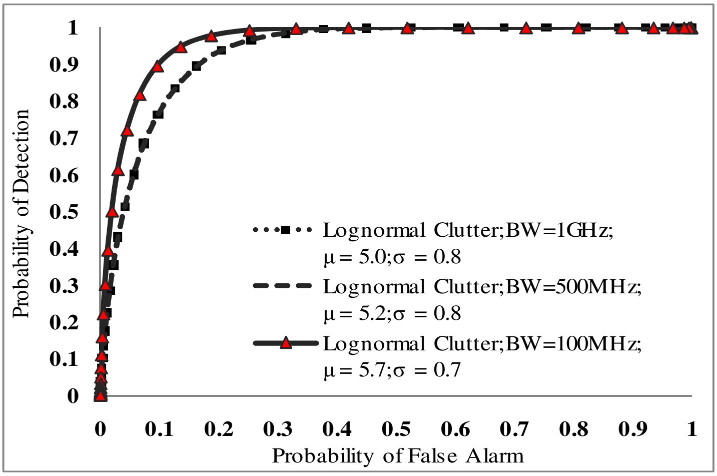


Fig. 5 Performance of non-coherent Logarithmic detector in log normal clutter at 1 GHz, 500MHz and 100MHz bandwidth (BW)

The clutter power decreases as the variance value decreases. So the performance of the detector increases if the clutter power decreases.

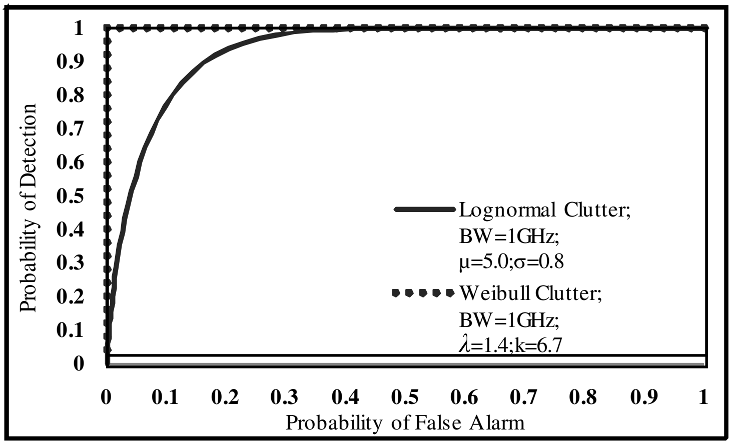


Fig. 6 Performance of non-coherent Logarithmic detector in log normal and weibull clutter at 1 GHz bandwidth (BW)

In the figure 6, the performance of the non-coherent logarithmic signal detectors is analyzed for log normal clutter and weibull clutter environment. At 1GHz bandwidth the logarithmic detector with coherent integration number of 200 and non-coherent integration number of 100, gives better performance in lognormal clutter environment with mean μ=5.0 and standard deviation σ=0.8 and also in weibull clutter environment with scale parameter λ=1.4 and shape parameter k =6.7. In weibull clutter environment the logarithmic detector gives maximum detection probability because the clutter power is very low compared with the log normal clutter power.

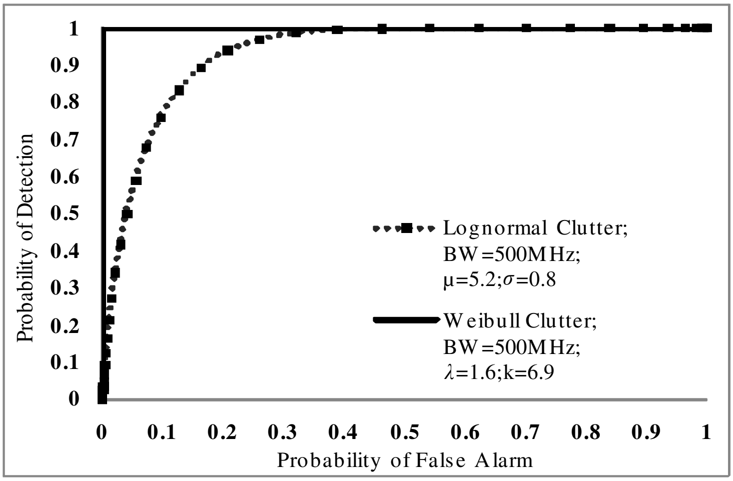


Fig. 7 Performance of non-coherent Logarithmic detector in log normal and weibull clutter at 500MHz bandwidth (BW)

In the figure 7, the performance of the non-coherent logarithmic signal detectors is analyzed for log normal clutter and weibull clutter environment. At 500MHz bandwidth the logarithmic detector with coherent integration number of 200 and non-coherent integration number of 100, gives better performance in lognormal clutter environment with mean μ=5.2 and standard deviation σ=0.8 and also in weibull clutter environment with scale parameter λ=1.6 and shape parameter k =6.9. In weibull clutter environment the logarithmic detector gives maximum detection probability because the clutter power is very low compared with the log normal clutter power.

In the figure 8, the performance of the non-coherent logarithmic signal detectors is analyzed for log normal clutter and weibull clutter environment. At 1GHz bandwidth the logarithmic detector with coherent integration number of 200 and non-coherent integration number of 100, gives better performance in lognormal clutter environment with mean μ=5.7 and standard deviation σ=0.7 and also in weibull clutter environment with scale parameter λ=2.5 and shape parameter k =8.7. In weibull clutter environment the logarithmic detector gives maximum

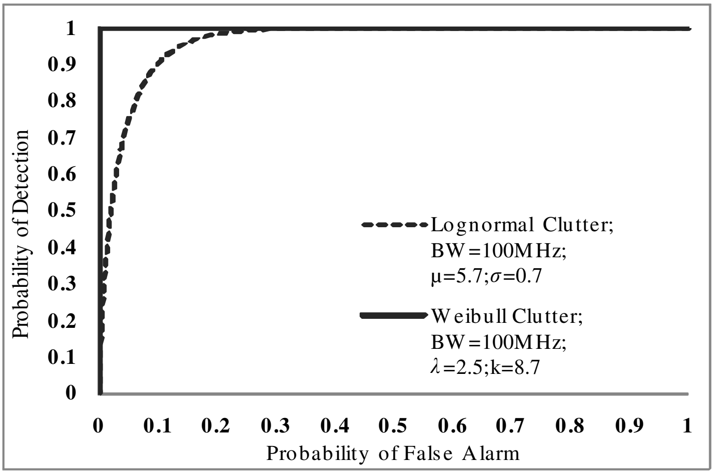


Fig. 8 Performance of non-coherent Logarithmic detector in log normal and weibull clutter at 100MHz bandwidth (BW)

detection probability because the clutter power is very low compared with the log normal clutter power.

6 Conclusion

In the UWB Automotive Short Range Radar the clutter echoes are the echoes from the objects in the road environment. For the system bandwidth of 500MHz the road clutter resembles log-normal clutter distribution and for 100 MHz the clutter resembles weibull distribution. Considering log normal clutter environment and weibull clutter environment the performance of the logarithmic detector is discussed.

The optimum performance of the non-coherent logarithmic detector is analyzed at different system bandwidth (1GHz, 500MHz, and 100 MHz) using different sets of mean and variance value in lognormal clutter environment and in weibull clutter environment. In all the cases the performance of the logarithmic detector in log normal clutter and weibull clutter varies as the function of their distributional parameters. The mean and variance values of the distribution decreases, the clutter power decreases. If the clutter power decreases the performance of the detector increases. So, logarithmic detector gives better performance in weibull clutter environment at the system bandwidth of 1GHz, 500MHz and 100MHz because of the very less clutter power value compared with log-normal clutter power.

Acknowledgement. This work was supported by the research grant of the Jeju National University in 2006, and also supported by the Daegu Gyeongbuk Institute of science and Technology and funded by the Korean Ministry of Education & Science Technology.

References

1. Matsumura, T., Eshima, K., Mizutani, K., Kohno, R.: Inter-Vehicle Communication and Ranging System Using UWB. In: The 25th Synposium on Information Theory and Its Applications (SITA 2002), Gunma, Japan, pp. 291–294 (2002)
2. Takeshi, F., Noboru, N., Shuichi, U., Masaaki, N., Hiroyuki, S., Tsuyoshi, T., Daisuke, U.: A 26GHz Short-Range UWB Vehicular-Radar Using 2.5Gcps Spread Spectrum Modulation. In: IEEE/MTT-S International Microwave Symposium (2007)
3. Gresham, I., Jenkins, A., Egri, R., Eswarappa, C., Kinayman, N., Jain, N., Anderson, R., Kolak, F., Wohlert, R., Bawell, S., Bennett, J., Lanteri, J.: Ultra-wideband radar sensors for short-range vehicular applications. IEEE Transactions on Theory and Techniques 52(9) (2004)
4. Win, M.Z., Scholtz, R.A.: Impulse radio: how it works. IEEE Commun. Lett. 2(2), 36–38 (1998)
5. Bloecher, H.L., Sailer, A., Rollmann, G., Dickmann, J.: 79 GHz UWB automotive short range radar – Spectrum allocation and technology trends. Adv. Radio Sci. 7, 61–65 (2009)
6. Matsunami, I., Nakahata, Y., Ono, K., Noguchi, Y., Kajiwara, A.: Empirical Study on Ultra-Wideband Vehicle Radar. In: IEEE Vehicular Technology Conference, VTC2008-Fall (2008)
7. Taylor, J.D.: Ultra-Wideband Radars Technology: Main Features Ultra-Wideband (UWB) Radars and differences from common Narrowband Radars. J.D.Taylor. CRC Press, Boca Raton (2001)
8. Minkler, G., Minkler, J.: CFAR: The principles of Automatic Radar detection in clutter
9. Taylor, J.D. (ed.): Introduction to Ultra-Wideband (UWB) Radars systems, Boca Raton, FL (1995)
10. Proakis, J.G.: Digital Communications. McGraw-Hill, New York (2001)
11. Gradshteyn, I.S., Ryzhik, I.M.: Table of Integrals, Series, and Products. Academic Press, London (1980)
12. Surendran, P., Ko, S., Kim, S.-D., Lee, J.-H.: A Novel Detection Algorithm for Ultra Wide Band Short Range Radar in Automobile Application. In: IEEE VTC2010-Spring (2010)

A Regional Matchmaking Technique for Improving Efficiency in Volunteer Computing Environment*

Eunyoung Cheon, Mikyoung Kim, Seunghak Kuk, and Hyeon Soo Kim

Summary. Volunteer computing is a computing paradigm in which the operations of a large-scale application are processed by idle resources of Internet-connected computers. The redundant-work distribution techniques are mainly used to exclude the malicious participants in the volunteer computing environment. However, the techniques cause some problems like delay of the work completion time or inefficiency of the work execution due to continuous requests of the work redistribution which arise from the reason that the techniques do not consider the characteristics of works and dynamic resources. To cope with such problems this paper suggests a regional matchmaking technique which can redistribute works in consideration of the characteristics of the works and the participant resources.

1 Introduction

Volunteer computing is a high-performance parallel processing system built through voluntary idle resources of Internet-connected computers [1]. However, the participating computers are volatile. They join freely in the volunteer computing environment and also leave freely. They have no responsibility to offer the reliability of the operation. Therefore the volunteer computing environments lack reliability of results, in addition, it is difficult to guarantee completion of operations on time. Distributing redundant-work has been proposed to cope with this problem. However, it does not resolve the issue about inefficient use of tasking time and delays in task performance because it does not consider characteristics of resource. These problems require the continuous work redistribution on the part of the server and eventually become obstacles to performance.

This paper suggests a regional matchmaking technique to improve efficiently the redundant-work distribution in the volunteer computing environment. Generally, matchmaking techniques are used for allocating resources between the zsuppliers and the buyers in the web services environment or in the distributed resource management systems. With this matchmaking technique, the most

Eunyoung Cheon · Mikyoung Kim · Seunghak Kuk · Hyeon Soo Kim
Department of Computer Science and Engineering
Chungnam National University, Daejeon, Korea
e-mail: {eycheon,lpgrane,triple888,hskim401}@cnu.ac.kr

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 89–97.
springerlink.com

appropriate task is assigned to the resource provider. Reducing the rate of redistribution of the redundant-work can shorten total completion time.

This paper consists of as follows: Section 2 describes the volunteer computing environments and the matchmaking for efficient distribution of the tasks on the resources. Section 3 suggests a matchmaking technique for distributing redundant-works in the P2P-based volunteer computing environments. Section 4 presents the concluding remarks and the future challenges.

2 Background

2.1 Volunteer Computing Environments

In the volunteer computing environments works are processed with parallel by a number of voluntary idle resources in Internet-connected computers. The volunteer computing environments are organized as follows:

(1) Client: A presenter of problems, she/he wants to solve her/his problems on the volunteer computing environment. In order to solve such problems the large distributed-processing is usually required. The client uploads her/his problem to the server and waits for the results from the server.
(2) Resource Provider: The volunteers who provide their computing resources being idle currently. The computing resource executes problem tasks assigned from the server, and then submits the results back to the server.
(3) Central Management Server: It assigns problem tasks to the resource providers. It plays a role as a manager.

Resource providers in the volunteer computing environments have the following difficulties. It is difficult to predict task completion time due to lack of information about the characteristics of resources as well as that of problem tasks. This causes instability in the computing environment and may lead to network disruptions. In such cases, it may not be able to collect the results of the distributed tasks because of unreported escapes. Because resource providers are not able to guarantee completion of the entire tasks, it may take a long time to accomplish the entire tasks and result in performance degradation.

2.2 Matchmaking

Matchmaking is a method that connects a service user and the most appropriate service provider according to the specifications which describe the characteristics of the service provider and the needs of the service user, respectively. The method is applied for utilization of the network-connected resources to the stable distributed computing environment such as Grid [4].

P2P-based matchmaker is used to perform matchmaking for distributing the redundant works among the resource providers [5]. The larger the scale of an application is, the higher the load of the P2P-based server is. As a result, the performance of the matchmaker is drastically degraded, a problem so called the scalability problem happens.

3 Regional Matchmaking Techniques

The P2P based regional matchmaking technique can reduce the server load and guarantee the reliability of the results. Even in the case of a server collapse it can increase vitality of works and fault tolerance [4]. When, however, redundant works are distributed, new problems occur such that the tasks are continuously redistributed from the server, consequently the task completion delays. Redistribution of tasks may appear because there is no any consideration about the resource requirement of the task and the capability of the resource provider. If the mismatch between tasks and the resource providers remains, it results in eventually performance degradation.

3.1 Matchmaking Technique

The matchmaking algorithm proposed in this paper considers dynamic capability of the resource providers. It identifies the capabilities of space and time, respectively. Capability of time is the amount of available time of the used resources, which is expressed as a continuous time function.

We use the elapsed time of the resource provider which attends to the volunteer computing to perform the actual work. The time is determined the available time of the resource provider.

Space availability means a measurement unit of physical space capacity. We consider the dynamic change of CPU and memory capacities when performing a given task.

3.1.1 Dynamic Capability-Based Matchmaking

As seen from Condor [6] or web services, in order to apply matchmaking to volunteer computing, the nature of the environment - the volatility of resources – should be considered. In the matchmaking technique, as resources take part in the system, static information of resources is made known to a matchmaker. Based on that information, the matchmaker matches the resources with the most appropriate works. In the distributed computing environment, time to complete a work depends on the performances of individual resource providers taking part in computation, as well as their availability. The autonomy and volatility of resource providers mean stable processing of work is not guaranteed. In the volunteer computing environment that uses only idle resources, considering only static resource information (memory capacity or CPU performance) leads to redistribution of works caused by processing delays arising from the dynamic nature of the resource providers' environment. As a result, the overall time it takes to complete the work may be delayed. Therefore, a redistribution algorithm that makes use of the dynamic availability information of resources when redistributing works is needed.

3.1.2 Locality-Based Matchmaking for Distributing Redundant Work among Resource Providers

In locality-based matchmaking, the type of peers in the P2P environment is the superpeer, and resource providers that are super peers perform the role of matchmaker. There are many factors to consider when choosing super peers, such as the maximum number of leaf nodes and localities. In the volunteer computing environment, resources are often stopped from participating even if an overlay network is built through the selection of super peers because of the unstable nature of the environment. This results in much overhead from rebuilding the network, making it difficult for application. Therefore, the matchmaking technique proposed in this paper is a locality-based matchmaking technique that involves peers of the same level. As shown in Fig. 1, all resource providers have the same functionality - work request, processing and response and selection of appropriate works - so matchmaking is done through organic communication among peers.

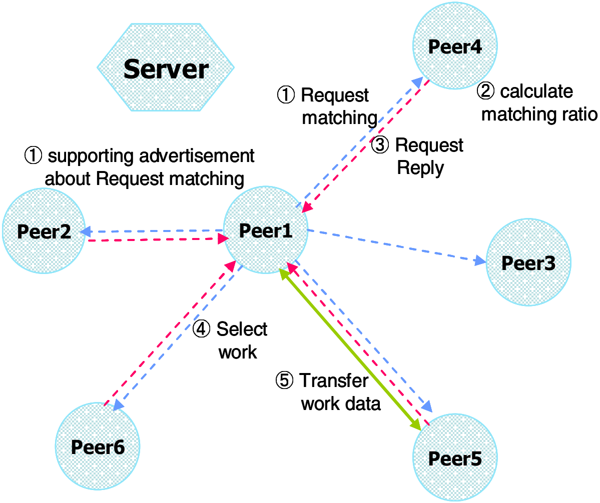


Fig. 1 Locality-based matchmaking peers of the same level

3.2 Components of the Proposed Matchmaking Technique

The following describes the components of our matchmaking technique.

3.2.1 Work/Resource Information Specification

Before conducting matchmaking, resources and works need to be specified.

(1) Work specification: Work specifications are transmitted to resource providers along with work data from the server. They are written when an administrator

uploads a work to the server, and are stored in a database. Tags specifying the weights, which are dependent on the amount of resources that should be dedicated to the work, were added to the work specification, which are memory weight, flops weight and average processing time weight (Fig. 2).

(2) Resource information specification of resource providers: The resource information specification of resource providers describes the expected time and space capacity at the time of work processing, which is used in match ratio calculations (Fig. 3).

```
<joblist>
<appid></appid>
<projectid></projectid>
<policy>
<platform></platform>
<max_runtime></max_runtime>
<os></os>
<mem></mem>
<lib></lib>
<hdd></hdd>
<RATE_mem></RATE_mem>
<RATE_flops></RATE_flops>
<RATE_time></RATE_time>
</policy>
<commonfiles>..</commonfiles>
<workid></workid>
</joblist>
```

Fig. 2 Specification of work

```
<cominfo>
<cpu num="" cat="" clock=""></cpu>
<memory physical=""></memory>
<os cat="" ver=""></os>
<flops></flops>
<A_mem></A_mem>
<A_flops></A_flops>
<avg_times></avg_times>
</cominfo>
```

Fig. 3 Specification of Resource

3.2.2 Resource Dynamic Availability Prediction Module

For matchmaking based on dynamic availability of resources, availability is predicted with more weight put on recent values of availability of resource systems. The predicted availability is used for match ratio calculations as a dynamic resource factor.

(1) Monitoring of expected available space of resources: The matchmaking technique that takes into account availability of resources (memory, CPU) checks the current availability of resources at a specified interval based on users' previous use patterns, and available space is predicted based on that result.

① T_n : Check frequency for resource #n.

② A_n : T_n resource available space

③ PA_n : Predicted resource available space at T_n

④ Predicted available resource space calculation for T_n

$$PA_n = Avg(PA_{n-1} + P_n) = \frac{(PA_{n-1} + A_n)}{2}$$

$$= \frac{A_0 + 2^0 A_1 + 2^1 A_2 + 2^2 A_3 + \cdots + 2^{n-1} A_n}{2^n}$$

As time goes on, for finding predicted available space for A_0 - the initial available space - the effect becomes extremely small, at $\frac{A_0}{2^n}$. However, A_n, the most recent update of available space, is applied using the weight of $\frac{2^{n-1}}{2^n} A_n$, and used as predicted availability information.

3.2.3 Match Ratio Calculation

The steps of the final calculation that determine the match ratios between resources and works based on resource dependence ratio and resource availability, given by the characteristics of the particular work specified in the work specification, are described below:

(1) Redundant work distribution match ratio algorithm: In the redundant work distribution, after the work of an application given by the server has been processed, distribution of the redundant work is requested to the near resource providers.

① Check minimum requirements using static information.

② Calculate work/resource match ratio using dynamic information.

③ Select works by sorting out with the match ratios.

(2) Check minimum requirements using static information: Check if the minimum requirements are met by a resource provider that requests a work.

(3) Availability based work/resource match ratio calculation: Calculated with time/space availability of resources of resource providers.

$$
\begin{aligned}
&Matching_{ratio} = \\
&SMem_{pos}(Mem_{pspace}(peer_x))*Mem_{appsfactor}(apps_y) \\
&+ SFlops_{pos}(Flops_{pspace}(peer_x))*Flops_{appsfactor}(apps_y) \\
&+ Time_{avail}(peer_x)*Time_{appsfactor}(apps_y)
\end{aligned}
$$

- $Mem_{pspace}(peer_x)$: Prediction of the available space of memory at the time a peer performs the work, calculated with current conditions.
- $Flops_{pspace}(peer_x)$: Prediction of the available CPU (flops)
- $SMem_{pos}(\ )$, $SFlops_{pos}(\ )$: represents the location of resources in the system of the predicted available space with the entire volunteer computing environment's memory resources and CPU capacity values from 0 to max (resource space) equally divided. Since the basic units of memory and CPU are different when doing the matching calculation, the same representation ways are required, so the availability within the system is expressed as a percentage. This indicates the value of the resource from the system's perspective.
- $Mem_{appsfactor}(apps_y)$, $Flops_{appsfactor}(apps_y)$: Weight on the level of resource dependence for a work.

3.3 Suggested Matchmaking Scenario

The proposed matchmaking flow is as follows (Fig. 4).

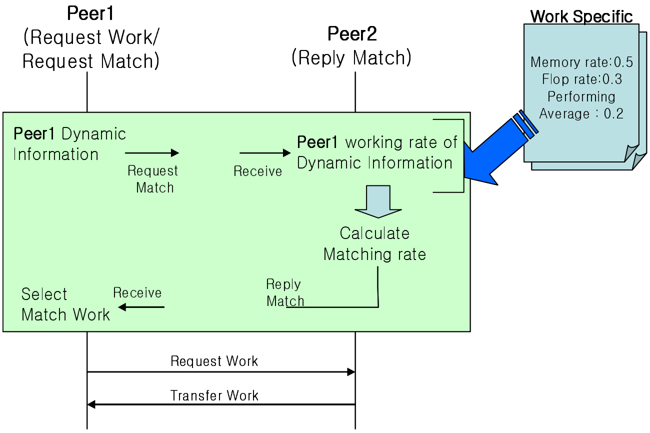


Fig. 4 Task request / response flows

(1) Performing the task exhausted the resources of the resource providers around the availability of providers to predict their dynamic information and redundant-work together to match the request.

(2) Matching demand with available information received dynamic prediction task defined in the specification of the task queue exists as a resource dependency ratio will save the match ratio.

(3) A "match the ratio" request is sent to the resource provider.
(4) The resource provider which requested the transfer of tasks selects the highest match ratio.
(5) The selected resource providers receive the redundant work.

This flow is shown in Fig. 5, which shows the relationship between the request and the reply work.

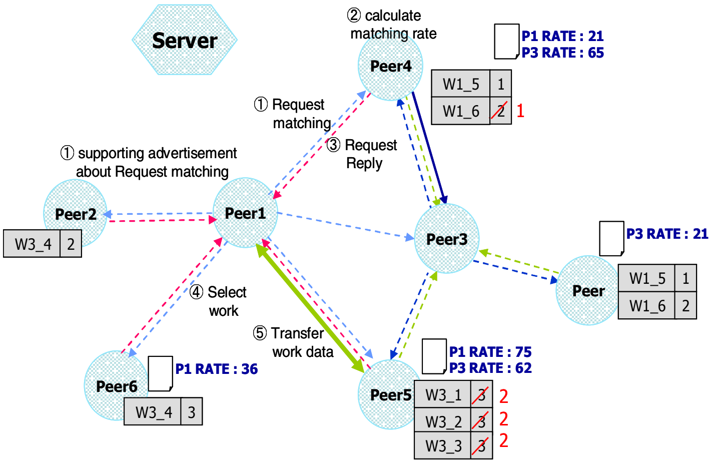


Fig. 5 Distribution of redundant-work for regional matchmaking scenarios

4 Conclusion

Because volunteer computing resource providers use variable idle resources, matching by utilizing only static metadata causes completion time delays and work redistribution from the server. To cope with performance degradation due to this problem, this paper introduces organic matchmaking techniques through communication between neighboring Peers to redistribute works according to the dynamic resource spaces and characteristics of the works. Operations under the attribute specify the resource dependence, and considering the information available about the task of matching resources, considering the suitability of selecting the most appropriate resources to minimize the execution time and reduce the server's rates were redistributed. Through the experiments, we identified that the work distribution method proposed in this paper reduces the rate of work redistribution and the total execution time for the works. But when peers in the same level communicate with each other for regional matchmaking, it can cause overhead and communication expenses. If characterization of the application was incorrect, it can also draw the wrong results which can't be distributed well. Therefore, further study is needed on ways to characterize the applications and quantify the dependence.

References

1. Luis, F.G.S., Satoshi, H.: Bayanihan: building and studying web-based vol-unteer computing systems using Java, pp. 675–686. Elsevier Science Publishers B. V, Amsterdam (1999)
2. Choi, J.H., et al.: Non-disturbance Scheduling based on User Usage Pattern for Volunteer Computing. In: Proc. of International Conference on Convergence Technology and Information Convergence (2007)
3. Anderson, D.P., Korpela, E., Walton, R.: High-performance task distribu-tion for volunteer computing. In: Proc. of First International Conference on e-Science and Grid Computing (2005)
4. Uppuluri, P., et al.: P2P grid: service oriented framework for distributed re-source management. In: Proc. of IEEE International Conference on Services Computing (2005)
5. Chakravarti, A.J., Baumgartner, G., Lauria, M.: The organic grid: self-organizing computation on a peer-to-peer network. IEEE Transactions on Sys-tems, Man and Cybernetics, Part A 35(3), 373–384 (2005)
6. Alain Roy, M.L.: Condor and preemptive resume scheduling, pp. 135–144 (2004)

An Intelligent Multi-Agent Model for Resource Virtualization: Supporting Social Media Service in Cloud Computing

Myoung Jin Kim, Hyo Gun Yoon, and Han Ku Lee*

Summary. With the rapid growth Internet, communication technologies and hand devices, there are a lot of communication and social activities of users in various information services based on SNS. As the availability of social media contents generated from users increases dramatically, the task of providing high-quality media contents in social media sites that need user contributions is emerging as unimportant issue in IT field. In this paper, we propose an intelligent multi-agent model for resource virtualization (IMAV) to automatically allocate service resources suitable for mobile devices in cloud computing environment supporting social media services. Our model can recommend optimized resource allocation suitable for mobile devices through virtualization rules and multi-agent. IMAV consists of user agent, distributed agent, gathering agent, virtualization register agent manager and system resource manager.

1 Introduction

Social media means media for social interaction, using highly accessible and scalable communication techniques [1]. In recent years, Social Network Service (SNS) based on social media contents has obtained a lot of interests from numerous users. In fact, 75% of Internet surfers used social media contents in the second quarter of 2008 by joining social networks and reading blogs according to Forrester Research [2].Thus, SNS have played a significant role in revitalizing

Myoung Jin Kim · Han Ku Lee
Division of Internet and Multimedia Engineering
Konkuk University, Seoul, Korea
e-mail: {tough105,hlee}@konkuk.ac.kr

Hyo Gun Yoon
Center for Social Media Cloud Computing
Konkuk University, Seoul, Korea
e-mail: hgyun007@gmail.com

* Corresponding author.

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 99–111.
springerlink.com

communication and social activities of participants[3][4]. In particular, SNS based on mobile and hand devices such as Facebook and Twitter is used a lot by users because of the advancement of Internet and communication techniques as well as the proliferation of mobile network infrastructure. In these service environments, users wish to receive media information or contents directly and smoothly from providers anytime and anywhere [5].

However, users have not gained reliable and high quality SNS services including social media contents in mobile environment due to the mobile data explosion and the limitation of hand device performance. To overcome this situation, most service providers are applying cloud computing techniques, especially virtualization techniques to SNS [6]. However, social media services in cloud computing environment allocate service resources depending on only user grades and limitations of service resource. In general, this traditional approach to resource allocation has two problems. The first one is that Constantinople service resources is required in order to allocate available resources to users. That is to say, administrators have to control and monitor resource allocation. The second problem is that it causes the increase observer load or network load.

In this paper, to overcome these problems, we suggest an Intelligent Multi-Agent for Resource Virtualization (IMAV) based on virtualization rules [7] that automatically allocates service resources suitable for mobile and hand devices in cloud computing environment supporting social media services. The proposed model can monitor service resources in real time and learn serviced context information that is basic data for resource virtualization. IMAV learning user behavior infers user on-demand and readjusts service resources for conceptualizations that service resources with high reliability can be provided to users. In addition, as idle resources of systems providing SNS are utilized as much as possible in cloud computing systems, system availability can increases. In the section of performance evaluation, we evaluated our IMAV under the criteria of virtualization result by multi-agent model. We also start to discuss performance of IMAV quantitatively. Our system has not quite reached the point where we can provide accurate performance of fully functional IMAV running on mobile cloud computing environments, whereas we address some of the performance questions that seem most urgent in the particular case of IMAV.

This paper is structured as follows. In section 2, we introduce virtualization of cloud computing and multi-agent for context aware. The following section explains the structure of IMAV. Section 4 shows our evaluation whose result is described. In the last section, we conclude our research with future work.

2 Related Works

2.1 Virtualization of Cloud Computing

By abstracting physical computing resources to logical resources, virtualization in cloud computing is able to provide flexibility in the use of computing resources [8].Therefore, it can simplify complicated computing environments and improve distributing job process and management efficiency.

Virtualization is broadly divided into two types: Hypervisor and Hosted. Furthermore, server virtualization, desktop virtualization, embedded virtualization, and mobile virtualizations are divided by system size. According to the purpose of virtualization, it is also classified into processor virtualization, memory virtualization and devices virtualization.

The most useful and significant technique in cloud computing fields, especially mobile cloud, is virtualization. In particular, mobile virtualization in cloud computing environment requires technologies that support various services and applications including real time characteristics. In addition, it has to guarantee reliability in order to satisfy diverse needs that users want. The purpose of mobile virtualization is to provide suitable services to users by using resources such as processor, memory, storage and applications offered from server due to the fact that computing power related to physical resources of mobile devices is inadequate.

The representative researches in mobile virtualization are focusing on virtualization to support platforms running on Android, Windows Mobile, and iPhone OS.

2.2 Multi-Agent for Context-Aware

Multi-agent is being used actively in the fields correlated with the development of automation systems Multi-agent has a knowledge base for learning users behavior as well as the function to infer purposes according to services.

Multi-agent has four features: autonomy, intelligence, mobility and social ability. Multi-agent conducts message passing or shared memory techniques using ACL (Agent Communication Language) and transmits messages and protocol using QML (Knowledge Query and Manipulation Language).

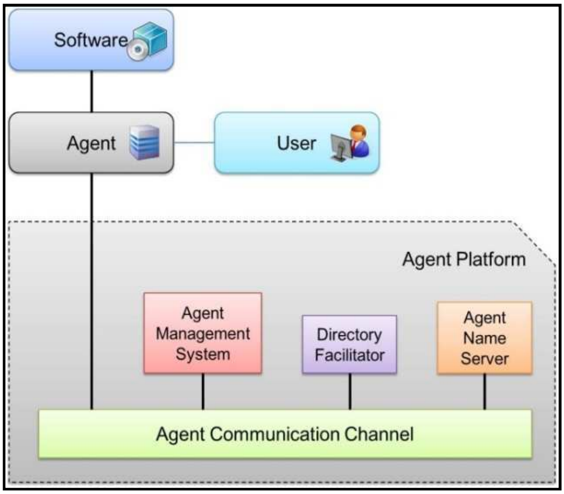


Fig. 1 Agent platform of FIPA

Recently, many researchers have proposed multi-agent models combined with mobile computing environments using knowledge base that contains context information including social data and user's location information [9][10][11].

Figure 1 shows the agent platform of FIPA (The Foundation for Intelligent Physical Agents) [12]. FIPA is a platform for developing and setting computer software standards for heterogeneous and interacting agents and agent-based systems.

3 The Proposed IMAV Model

This section describes the proposed IMAV model. IMAV (Intelligent Multi-Agent for Resource Virtualization)is intelligent virtualization rules based multi-agent model supporting social media service. Our model is able to configure service applications and resources appropriate for users' situation via multi agents and virtualization rules. Moreover, IMAV can manage resources of cloud computing in real time and reconfigure resources according to user behavior.

We assumed that service circumstance is based on the system offering mobile cloud service in order to suggest our model. To address not only the problem with recording user context data,but also the mechanical problem with processing large amounts of data in hand devices in real time, collaborative agents except for mobile agents are managed at server. Mobile agents perform to record and manage location information of users as well as log files which contains personal information, service history and request signal for accessing cloud service system. That is to say, when users execute cloud applications, the mobile agent checks service information in log files. These log files is the basic information to recommend new services to users.

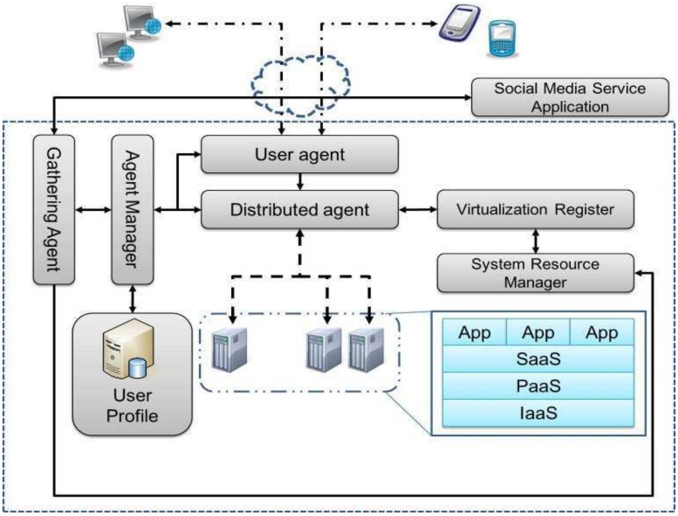


Fig. 2 IMAV (Intelligent Multi-Agent for Resource Virtualization)

Figure 2 depicts the basic model of IMAV. The proposed model is divided 6 domains: user agent, distributed agent, gathering agent, virtualization resister, agent manager and system resource manager. In the next step, 6 parts will be explained in detail.

4 Structure of IMAV

4.1 User Agent

The main role of user agent is to receive information generated from mobile agent. User agent analyzes service types and access types of services. Access type can be analyzed via user location and data transfer rate and service types can be seized through applications of social media service. The analyzed information above is utilized as user context.

The integrated information with system state information created from agent manager is applied to intelligence virtualization rules. In order to appropriately allocate system resources according to intelligence virtualization rules, the results are sent to distributed agent. User agent also offers scope of services to users. In other words, user agent generates users' context that integrates devices information connected by users and user connection information. The purpose of access, user behavior and service access patterns can be analyzed though the created context data. Therefore, service resources can be virtualized according to the purpose of service access and service usage patterns.

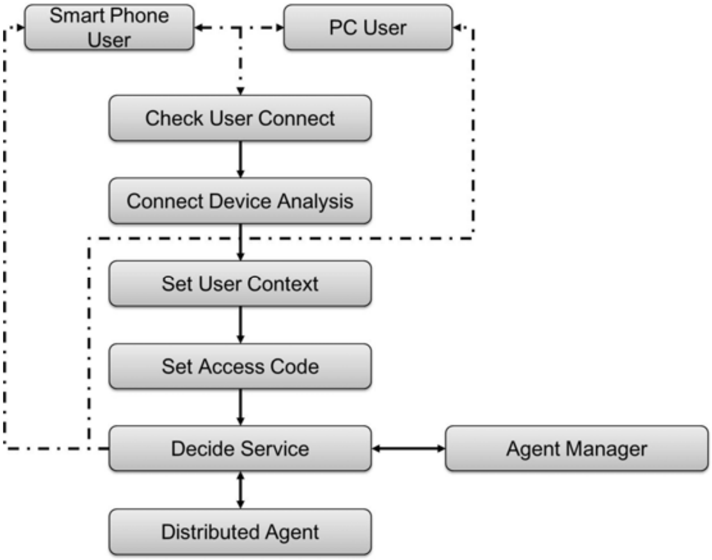


Fig. 3 User Agent

4.2 Distributed Agent

Distributed agent can distributes social cloud service resource via correlation between users and services. In addition, distributed agent carry out intelligent virtualization of service resources by constantly learning services requested by users and state information.

To virtualize services resources of system,distributed agent adopts MLP (Multi-Layer Perceptron) and SVM (Support vector machine). It includes virtualization module that verifies and processes information regarding service resources.

The information used by distributed agent to virtualize service resources consists of user context information transmitted by user agent and system context information transmitted by system resource manager. The user context information contains coefficient of determination that decides service to be provided to users.

If physical resources of system are exceptional, distributed agent recommends replaceable systems and resources. To support and maintain reliable services, lists of service resources concentrated on specific users with relocation rank are notified to system administrators. Relocation of system resources can expand physical resources and perform clustering through system context information. It is possible to adequately perform virtualization by transmitting configured system resources per user to virtualization module. Virtualization information including virtual ID is registered in virtualization resister. Distributed agent sends user states and monitoring information with respect to virtualization depending to time line to agent managers.

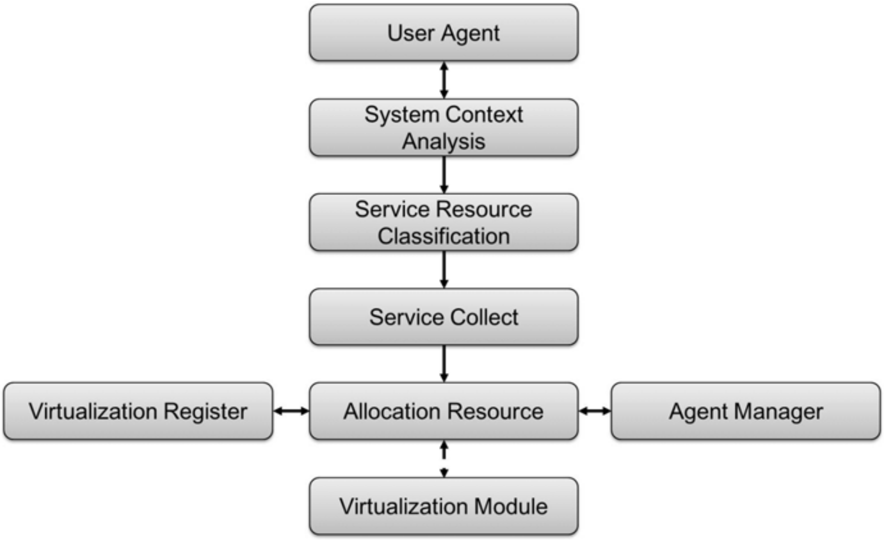


Fig. 4 Distributed Agent

4.3 Gathering Agent

Gathering agent collects service information and social media data needed for systems. Service information is collected at the content sites serving social media service as well as service data connected with applications registered in cloud system. The scope of gathering data is restricted within social sites such as UCC, Blog & Micro-Blog, SNS and News.

Gathering agent adopts Trend Watching (Media tablet) as searching engine of gathering agent. In addition, we also apply MapReduce based on Hadoop to process large amount of data in parallel.

4.4 Agent Manager

Agent manager manage and controlthe stages of creation, registration, event and deletion of each agent. Moreover, agent manager provides knowledge-base to each agent, monitoring the whole agents according to use types of social service resources. It contains the ability to control activity of each agent. The main role of monitoring in agent manager is to record event state and values between agents as well as to provide fault data generated from system to administrators.

Agent manger obtains only relational information from user agent via log information and context information. The relational information determines the event of creation of distributed agent and service items to be provided to users. Agent manager creates the event of creation, activity and deletion of distributed agentdepending on the system time line. The structure of control signal is divided into agents' ID, control information, MAC, TAG and Trap.

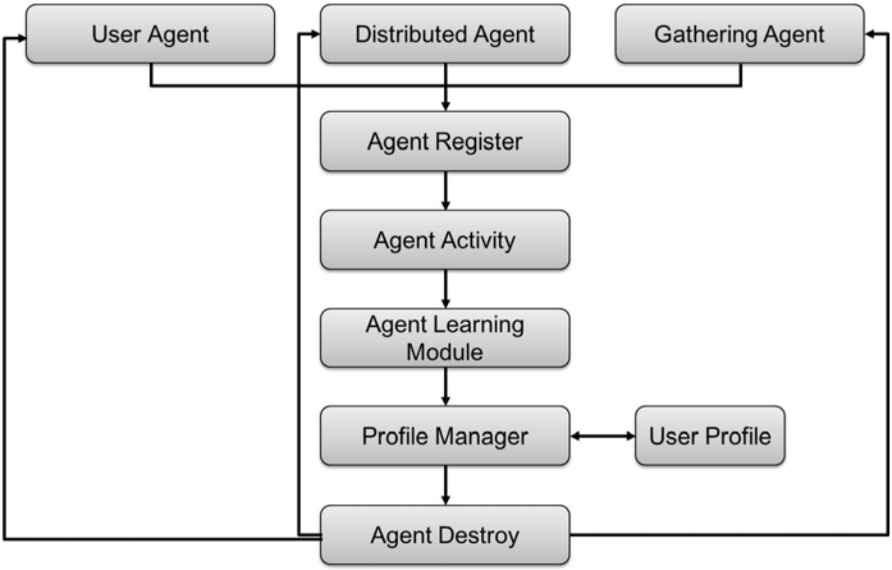


Fig. 5 Agent Manager

Control signal of user agent consists of SIP, RequestID and Trap. SIP is an index including the occurrence information of user agent. RequestID is the request information for the management of context information and for services. Trap controls the action state of user agent. Trap is made up of the creation and deletion of agents and signals that distinguish types of events.

Control signal of distributed agent is composed of UA (User Agent) index, RequestID and TRAP. UA index includes virtualization ID offered to users and information that matches it. RequestID is provided by user agent. Finally, Trap contains the action state value of distributed agent as well as state value regarding the information event received the system.

4.5 Virtualization Register

Virtualization register registers and manages virtualization information of social resources that are distributed by distributed agent. In addition, the utilization rate of service resources and the state of system resources are managed and provided to administrators. It regularly analyzes log information of virtualization, supporting efficient management of system resources.

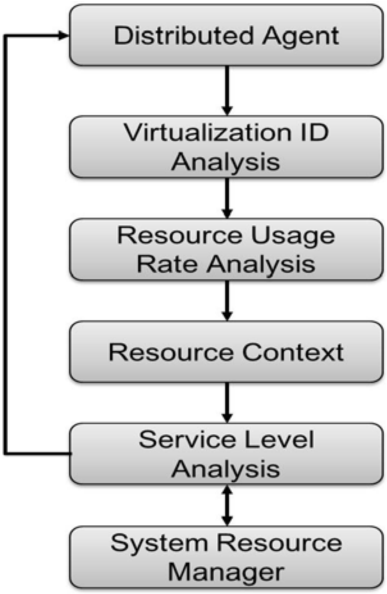


Fig. 6. Virtualization Register

Virtualization Registrants synchronized with distributed agent manage log data from the creation of virtualization to its destruction. Log data of virtualization register consists of virtualization resource ID allocated to users, lists of service resource, priority information and correlation weight. The correlation weight can decide the priority of system resources provided to users when system is virtualized. The utilization rate (rs) is calculated through the following formula. UC is user context. SR is the system resources configured by UC.

$$r_s = \frac{\sum_{i=1}^{n}(SR_i \times UC_i)}{\sum_{i=1}^{m} MSE(UC_i)}$$

The updated correlation weight is used as learning weight of multi-agent. Therefore, multi-agent can predict the way of virtualization in advance when service resources are reallocated. It also can provide reconfigured service resources to users. In addition, the reconfigured services and the whole users are recorded in system service history. Multi-agent is able to set service levels by comparing system service history with service resources provided to other users.

4.6 System Resource Manager

The information recorded in the lists of virtualization register is controlled by system resource manager. System resource manager have a function that controls the management of resources provided to users according to the correlation information in the lists of virtualization resister so that administrators can obtain distributed state monitored by system resource manager and can directly or indirectly control system resources.

The structure of system resource manager is composed of system usage rate analysis domain, system context management domain, system level analysis domain, system resource classification domain, system management information DB. It is able to monitor and manage physical and logical resources of the cloud computing system. The system usage rate analysis domain uses MAXMIN algorithm for its analysis. The following formula is for system usage rate.

$$v_s = \min\left(\max\left[\frac{(\sum_{i=1}^{n} SC(r)\frac{Y}{Z} \times r_s \times n_Y)}{\sum_{i=1}^{n} SC^{Y}}\right]\right)$$

System context management domain makes fundamental data to analyze system levels. The formed context information is data that reconfigures system state depending on the user situation. System level analysis domain sets the level or rank of each resources offered to users. System level is the result to analyze amounts of specific resources provided to users among the whole resources. That is to say, by analyzing the availability of system resources, system level analysis domain decides system grades according to whether resources have high or low availability. The determined level which is provided to users supports to decide additional decision making of system resources. System management information DB is managed by timeline. Furthermore, it provides regular reports and is monitored by administrators.

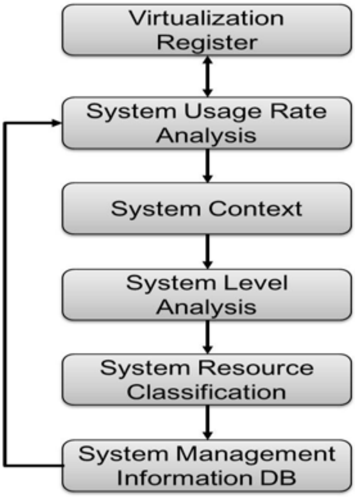

Virtualization Register
System Usage Rate Analysis
System Context
System Level Analysis
System Resource Classification
System Management Information DB

Fig. 7 System Resource Manger

5 Performance Evaluations

In this section we discuss performance evaluations of IMAV. For purposes of this paper- and to understand what directions are most profitable for future development- the MovieLens datasets that are used a lot as most recommendation systems have been applied to IMAV. Table 1 shows the condition of evaluations for the training of multi-agent.

Table 1 Condition of Evaluation

Contents	Value
Times	1000
Like Frequency	0.6
Data VS (Learning Data: Test Data)	4:1
Input Node	12
Hidden Node	6
Learning Rate	0.15

Table 2 Result of Evaluation with UC and SC

Contents	UC	SC	DR
Accuracy	78.1	71.3	81.2
Precision	82.9	69.2	86.0
Recall	83.5	72.6	88.1
F-measure	83.0	71.3	87.4

In our experiment, user correlation between User Context (UC) and System Context (SC) has been tested in terms of accuracy, precision, recall and F-measure respectively. Result of Evaluation is shown in table2.

In the case of UC, the result is reflected by the direct choice of users. Although users have good relationship with system resources, it is a choice without considering other users or the state of the system. In the case of SC, the system directly provides resources to users. Although system efficiency is good, user preference is low. DR shows the result reflected service recommendation through our model after learning correlation between users and system resources. Figure 8 shows recommended resource virtualization using IMAV.

Multi-agent recommends service resources to users according to user context. The first level is service resources that cloud system have. The second level is resources to be virtualized by user request. As shown figure, our multi-agent model recommends H3, H5 and H7 for virtualization of service resources. As a result, it is possible for our system to recommend virtualization services suitable for user requirements and system environments in cloud system.

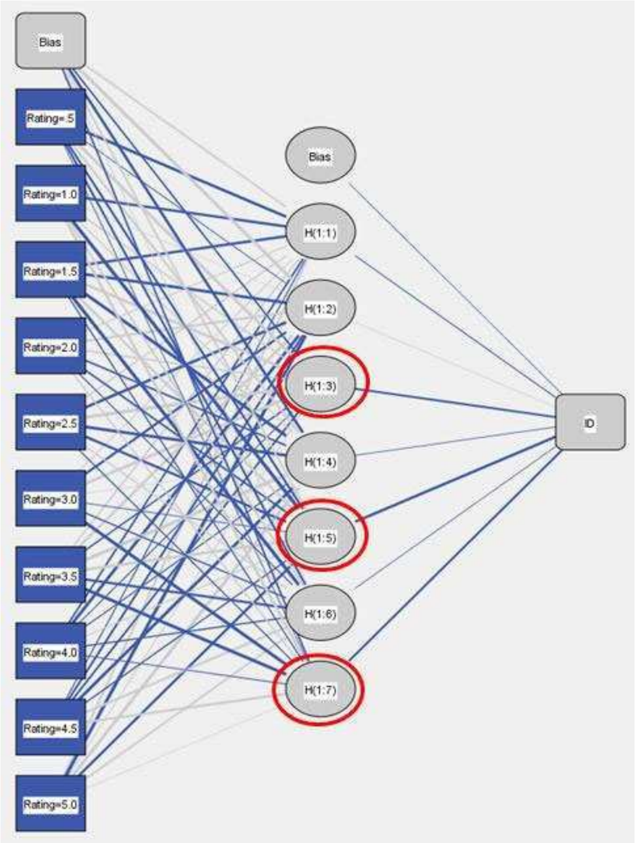


Fig. 8 Recommended Virtualization

6 Conclusion and Future Works

The main purpose of this paper is to propose intelligent multi-agent model for resource virtualization (IMAV)that automatically allocates service resources suitable for mobile devices in cloud computing environment supporting social media services. Our system is very promising since it provides intelligent virtualization rule based multi-agent model for resource virtualization in cloud computing. In fact, the proposed model recommends very suitable virtualization by analyzing user context and the state of system. Furthermore, our model analyzes social media service resource in real time, learning user context for virtualization. Therefore, cloud systems that apply our model can prevent resource bottlenecks and enhance resource availability. In addition, users are able to use reliable services because multi-agent model provides appropriate services for users depending on user situation.

Another purpose of this paper is to verify if our IMAV can be implemented efficiently in mobile cloud environments. In the section of performance evaluation, we demonstrated good performance of our model in terms of UC (User Context), SC (System Context), and DR (Direct relation). IMAV learning user context and user behavior recommend service resources for virtualization that is highly related to users. Therefore, our model is able to carry out resource virtualization suitable for requirements and needs of users in cloud systems.

In the future work, we will build our model into SaaS and PaaS of cloud computing. Furthermore, we are going to focus on developing upgraded model for recommendation service that is able to search large amount of data in real-time in mobile environment.

Acknowledgement. This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency (NIPA-2011 – (C1090-1101-0008)).

References

1. Wikipedia, http://en.wikipedia.org/wiki/Social_media
2. Kaplan, A.M., Haenlein, M.: Users of the world, unite! The challenges and opportunities of SocialMedia. Journal of business horizons 53(1), 59–68 (2010)
3. Fernando, A.: Social media change the rules. Communication world 24(1), 9–10 (2007)
4. Hypponen, M.: Malware Goes Mobile. Scientific American 295(5), 70–77 (2006)
5. Mui, L.: Computational Models of Trust andReputation: Agents, Evolutionary Games, and SocialNetworks. PhD Thesis, Massachusetts Institute ofTechnology (2002)
6. Bharadwaj, K.K., Al-Shamri, M.Y.H.: Fuzzy computational models for trust andreputation systems. Electronic Commerce Researchand Applications 8(1), 37–47 (2009)
7. Yoon, H., Lee, H.: An Intelligence Virtualization Rule based on multi-layer to support social mediacloud service. In: CNSI 2011(2011)
8. Goldberg, R.P.: Survey of Virtual Machine Research. IEEE Computer Magazine, 34–45 (1974)

9. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding High-Quality Content in Social Media. In: Web Search and Data Mining, pp. 183–194 (2008)
10. Yoon, H., Lee, M., Gatton, T.M.: A multi-agent based user context Bayesian neural network analysis system. Artificial Intelligence Review 34(3), 261–270 (2010)
11. Yoon, H., Kim, E., Lee, M., Lee, J., Gatton, T.M.: A Model of Sharing Based Multi-Agent to Support Adaptive Service in Ubiquitous Environment. In: Proceedings of the 2008 International Conference on Information Security and Assurance (ISA 2008), pp. 332–337 (2008)
12. `http://www.fipa.org/index.html`

Compiler-Assisted Maximum Stack Usage Measurement Technique for Efficient Multi-threading in Memory-Limited Embedded Systems*

Sung Ho Park, Dong Kyu Lee, and Soon Ju Kang

Summary. One of the reasons why it is hard to use multi-threading in memory-limited embedded systems is the difficulty of stack optimization. Some solutions for this problem have been proposed in prior research, but the proposed solutions were not totally effective. This paper proposes the compiler-assisted maximum stack usage measurement technique as a new solution for this problem. This technique measures the maximum stack usage of each thread with special code that is automatically inserted at the beginning of each function by the compiler. With the help of the operating system, the special code records the maximum stack usage of each thread in run-time. Also, the special code predicts and prevents stack overflow in run-time. Therefore, with this technique, the maximum stack usage of each thread can be precisely determined during testing, and thus allowing the stack of each thread to be accurately optimized. Unlike the solutions proposed in previous research, this technique does not have problems such as limited availability, the possibility of undetectable stack usage, and memory fragmentation. Although this technique adds some overhead, the overhead is added only during the stack optimization process in the development phase. Also, despite the necessity for modification of the compiler and operating system, this technique is easy to implement. It can be implemented by slight modification of the existing compiler and operating system. To verify this technique, it was implemented and evaluated on the ARM platform by modifying the GNU ARM C compiler and the Ubinos, which is an operating system for memory-limited embedded systems.

Sung Ho Park · Dong Kyu Lee · Soon Ju Kang
Kyungpook National University - EECS
1370, Sangyuk-dong, Buk-gu, Daegu, Korea
e-mail: slblue@ee.knu.ac.kr

* This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2010-0003387).

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 113–129.
springerlink.com

1 Introduction

Recently, many embedded systems have a memory management unit and adequate memory. However, a considerable number of embedded systems still do not have a memory management unit and have less than tens of kilo-bytes of memory because of energy-efficiency, cost, and size. Also, in the near future, the number of such memory-limited embedded systems is expected to increase as wireless sensor networks and ubiquitous computing that need massive energy-efficient, low-cost, and small embedded systems emerge.

If multi-threading is used in such a memory-limited embedded system, the stack of each thread has to be allocated when the thread is created, and the stack size cannot be changed during run-time. Therefore, if the stack size of each thread assigned at thread creation time is less than the maximum stack usage of the thread, the system malfunctions. Also, if the stack size of each thread is greater than the maximum stack usage of the thread, it is a waste of memory which is a limited and important resource. In other words, for efficient multi-threading in a memory-limited embedded system, the stack size of each thread has to be optimized to the proper size to prevent system malfunction without wasting memory.

However, until now, there has been no clear solution for stack optimization. Stack analysis techniques have been proposed, but the techniques are unavailable if there is a recursive or an indirect function call, and often overestimate the maximum stack usage. Stack pollution check techniques and dynamic stack resizing techniques that do not have such problems have been proposed, but they also had problems such as the possibility of undetectable stack usage and memory fragmentation. Because of these problems, many developers still optimize stack by a trial and error method that is unclear and time-consuming, or are forced to give up the use of multi-threading.

This paper proposes the compiler-assisted maximum stack usage measurement technique as a novel solution for stack optimization. This technique measures the maximum stack usage of each thread with special code that is automatically inserted at the beginning of each function by the compiler. With the help of the operating system, the special code records the maximum stack usage of each thread in run-time. Also, the special code predicts and prevents stack overflow in run-time. Therefore, with this technique, the maximum stack usage of each thread can be precisely assessed during testing, and thus allowing the stack of each thread to be accurately optimized. Unlike the solutions proposed in prior research, this technique does not have problems such as limited availability, the possibility of undetectable stack usage, and memory fragmentation. Also, despite the necessity of modifying the compiler and operating system, this technique is easy to implement. It can be implemented by slight modification of the existing compiler and operating system. To verify this technique, it was implemented and evaluated on the ARM platform by modifying the GNU ARM C compiler and the Ubinos, which is an operating system for memory-limited embedded systems.

The rest of this paper is organized as follows: Section 2 introduces related research. Section 3 explains the compiler-assisted maximum stack usage measurement technique with the implementation example on the ARM platform. Section 4 shows the evaluation of this technique. Section 5 explains problems with this technique and proposes some solutions for the problems. Finally, section 6 concludes the paper.

2 Related Research

2.1 Stack Analysis Techniques

Some researchers have proposed maximum stack usage measurement techniques by static analysis of control flow graph as a solution for stack optimization [8, 9, 18, 24, 25]. In other papers, these techniques are often dubbed the control flow analysis or data flow analysis or stack size analysis. All control flows of each thread can be found by analyzing executable binary code. The stack usage of each control flow can also be found by analyzing executable binary code. The maximum stack usage of each thread is the maximum value among stack usages of control flows of the thread. Thus, it is possible to find the maximum stack usage of each thread by analyzing executable binary code. The stack analysis techniques work on this principle. However, these techniques are unavailable if the executable binary code contains a recursive or an indirect function call and often overestimate the stack usage [7, 18, 24, 25]. An indirect function call means calling a function by address instead of symbol.

2.2 Stack Pollution Check Techniques

There have also been researchers who propose maximum stack usage measurement techniques by stack pollution check in run-time [14, 22, 23]. A system using these techniques sets an initial value to the stack area, like Fig. 1-(1). Then, the initial

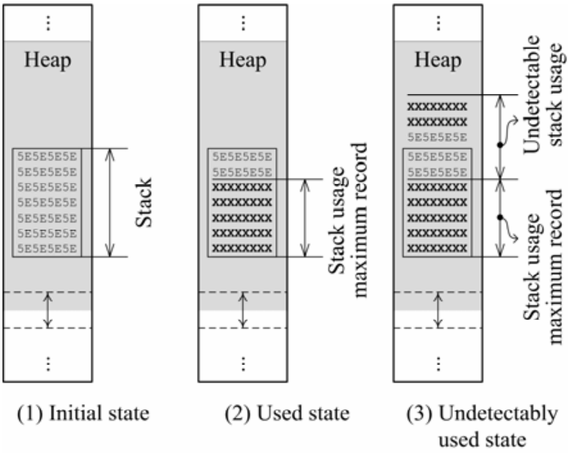


Fig. 1 Use case of the stack pollution check techniques

value of the stack area is polluted as much as the stack is used, like Fig. 1-(2). Therefore, it is possible to find out the maximum stack usage by checking the amount of polluted area after testing.

However, these techniques are unavailable if stack is used as shown in Fig. 1-(3). Also, when the stack usage exceeds the assigned stack size, the system could malfunction before detecting and handling the stack overflow, because these techniques check stack usage after the stack is used.

2.3 Dynamic Stack Resizing Techniques

Some other researchers have proposed dynamic stack resizing techniques as a solution for stack optimization [5, 7]. These techniques make it possible to use a set of discontinues memory fragments as a stack. Thus, with these techniques, it is possible to dynamically resize a stack in run-time just when the stack becomes insufficient. Developers do not need to optimize stack by themselves, if they use these techniques. However, these techniques decrease performance and cause memory fragmentation problems. In addition, these techniques are unavailable in a system that needs real-time characteristics, because they make the executing times of functions unpredictable.

2.4 Event-Driven Architecture

As stated above, it is hard to use multi-threading in memory-limited embedded systems because of the difficulty of stack optimization. Also, there has been no clear solution for stack optimization until now. So some researchers have proposed event-driven architecture as an alternative to the multi-thread architecture [11, 15, 16] to run multiple tasks at the same time in memory-limited embedded systems. If the event-driven architecture is used instead of the multi-thread architecture, the stack optimization becomes easy, because a system made with the event-driven architecture uses only one stack. However, with the event-driven architecture, it is hard to develop systems that use algorithms of which the execution time is long, such as compression and encryption. [6, 11, 12, 17, 18, 19].

3 Compiler-Assisted Maximum Stack Usage Measurement Technique

This paper proposes compiler-assisted maximum stack usage measurement technique as a novel solution for stack optimization. This technique uses two algorithms. The first is a run-time stack usage maximum record measurement algorithm. It is an algorithm for measuring the maximum record of stack usage in run-time, as its name indicates. The second is run-time stack overflow prediction algorithm, which is for predicting and preventing stack overflow in run-time. In this section, we will explain these two algorithms in detail with the implementation example on the ARM platform.

3.1 Run-Time Stack Usage Maximum Record Measurement Algorithm

The principle of this algorithm is as follows: The compiler inserts code in Fig. 2 at the beginning of each function during the compile time. Whenever context switching, the operating system switches the stack usage maximum record (saves the record of current thread and restores the record of next thread), like other contexts. Then, the stack usage maximum record of each thread can be found from the context backup area (thread control block) of the thread. This algorithm is possible because all compilers know the stack usage of each function. It is not an assumption but a fact because the compiler itself decides how to use the stack.

```
predicted stack usage =
    current stack usage + stack usage of function

if predicted stack usage > stack usage maximum record, then
    update stack usage maximum record
```

Fig. 2 Pseudo code for run-time stack usage maximum record measurement

The GNU ARM C compiler [13] and the Ubinos [20] were modified to support this algorithm. The Ubinos is an operating system that we developed for memory-limited embedded systems. It supports multi-threading, rich inter-thread communication functions, and yet it consumes few resources. The modified GNU ARM C compiler [21] and the Ubinos implement this algorithm as follows: The executable binary code generated by the GNU ARM C compiler uses a stack as shown in Fig. 3. As previously stated, the GNU ARM C compiler knows the stack usage of each function during the compile time. Thus, the GNU ARM C compiler could be easily modified to insert the code in Fig. 2 at the beginning of the executable binary code of each function.

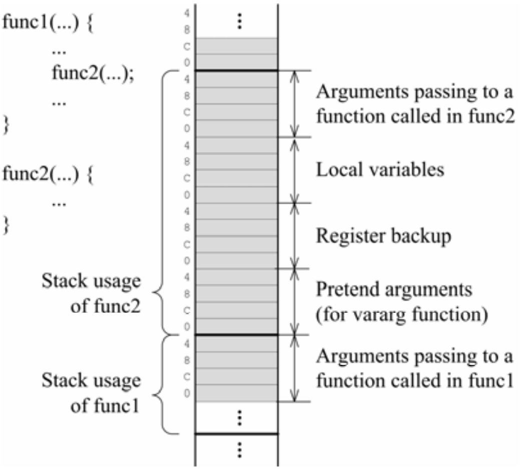
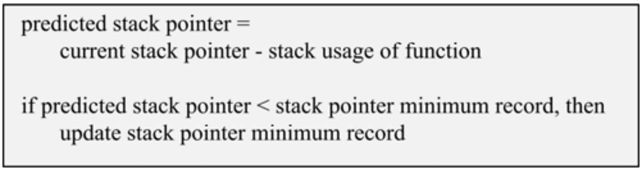

func1(...) {
...
func2(...);
...
}
func2(...) {
...
}
Stack usage of func2
Stack usage of func1
4 8 C 0 4 8 C 0 4 8 C 0 4 8 C 0 4 8 C 0 4 8 C 0 4 8 C 0
Arguments passing to a function called in func2
Local variables
Register backup
Pretend arguments (for vararg function)
Arguments passing to a function called in func1

Fig. 3 Stack use case of a code generated by the GNU ARM C compiler

Fig. 4 shows the optimized version of the pseudo code in Fig. 2 for easy implementation on the ARM platform. The stack pointer value decreases as much as stack usage increases, because the GNU ARM C compiler uses stack by a full descending method. Thus, the stack pointer after the execution of a function can be predicted by subtracting the stack usage of the function from the current stack pointer, and the stack pointer value becomes the lowest when the stack usage is at a record high. Also, the Ubinos keeps the stack start address and the stack size of each thread. The stack usage maximum record of each thread that is desired can be calculated using the stack pointer minimum record, the stack start address, and the stack size of each thread.

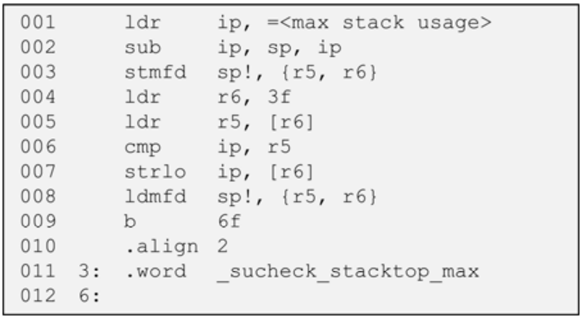


```
predicted stack pointer =
    current stack pointer - stack usage of function

if predicted stack pointer < stack pointer minimum record, then
    update stack pointer minimum record
```

Fig. 4 Pseudo code for run-time stack usage maximum record measurement (Optimized version for the ARM platform)

Fig. 5 shows the run-time stack usage maximum record measurement code that is implemented with the ARM instructions. Fig. 6 shows the code that is implemented with the THUMB instructions.

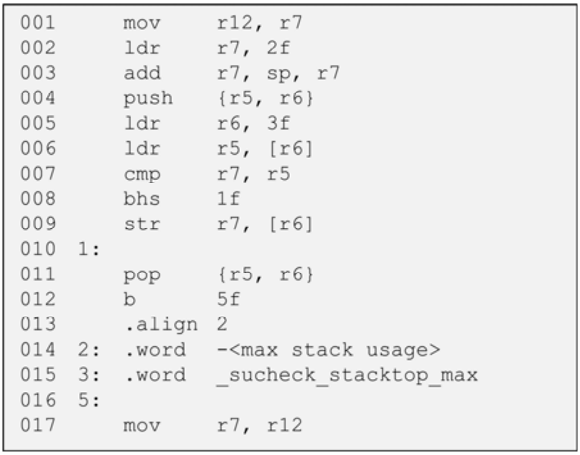


```
        ldr     ip, =<max stack usage>
        sub     ip, sp, ip
        stmfd   sp!, {r5, r6}
        ldr     r6, 3f
        ldr     r5, [r6]
        cmp     ip, r5
        strlo   ip, [r6]
        ldmfd   sp!, {r5, r6}
        b       6f
        .align  2
    3:  .word   _sucheck_stacktop_max
    6:
```

Fig. 5 The run-time stack usage maximum record measurement code implemented with the ARM instructions

The symbol, "_sucheck_stacktop_max" in Fig. 5 and 6 is the global variable that keeps stack pointer minimum record of current thread. The symbol, "<max stack usage>" is replaced with constant integer in real code. It is the stack usage of a function to which this code is inserted. The register, "ip(r12)" is used without being saved and restored because it is the inter-procedure-call scratch register. This register does not need to be preserved [3]. If the new compile option, "-msucheck" is used, the modified GNU ARM C compiler inserts the code in

Fig. 5 at the beginning of each function that consists of the ARM instructions, and inserts the code in Fig. 6 at the beginning of each function that consists of the THUMB instructions. Whenever context switching occurs, the Ubinos saves the value of the global variable, "_sucheck_stacktop_max" into the thread control block of the current thread, and changes the value of the variable to the value in the thread control block of the next thread, like Fig. 7.



```
     mov     r12, r7
     ldr     r7, 2f
     add     r7, sp, r7
     push    {r5, r6}
     ldr     r6, 3f
     ldr     r5, [r6]
     cmp     r7, r5
     bhs     1f
     str     r7, [r6]
1:
     pop     {r5, r6}
     b       5f
     .align  2
2:   .word   -<max stack usage>
3:   .word   _sucheck_stacktop_max
5:
     mov     r7, r12
```

Fig. 6 The run-time stack usage maximum record measurement code implemented with the THUMB instructions

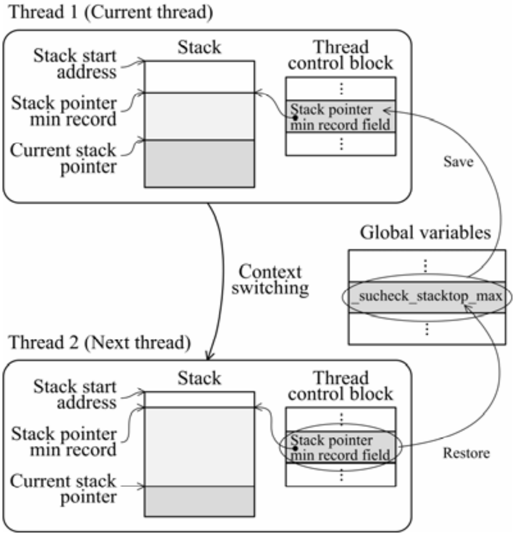
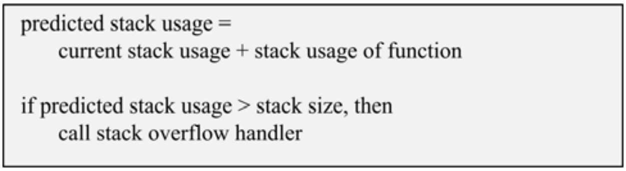


Fig. 7 Switching stack pointer minimum record

With this algorithm, the maximum stack usage of each thread can be found. However, this algorithm cannot handle stack overflow exceptions. So we also propose the run-time stack overflow prediction algorithm, which will be explained in the following section.

3.2 Run-Time Stack Overflow Prediction Algorithm

The principle of this algorithm is as follows: The compiler inserts the code in Fig. 8 at the beginning of each function. The operating system switches the stack size information, whenever context switching occurs. A system made this way can predict and prevent stack overflow exceptions.

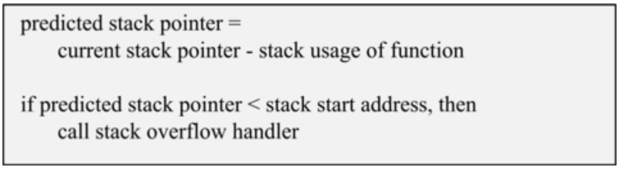


```
predicted stack usage =
    current stack usage + stack usage of function

if predicted stack usage > stack size, then
    call stack overflow handler
```

Fig. 8 Pseudo code for run-time stack overflow prediction

The modified GNU ARM C compiler and the Ubinos implement this algorithm as follows: Fig. 9 shows the optimized version of the pseudo code in Fig. 8 for easy implementation on the ARM platform. Because the GNU ARM C compiler uses a stack by a full descending method, if stack overflow occurs, then the stack pointer value becomes less than the stack start address. Thus, stack overflow can be predicted by comparing the predicted stack pointer and the stack start address, like the pseudo code in Fig. 9.



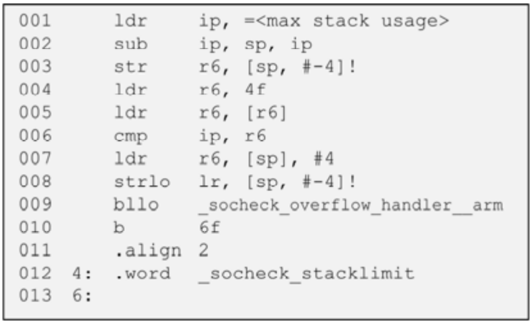
```
predicted stack pointer =
    current stack pointer - stack usage of function

if predicted stack pointer < stack start address, then
    call stack overflow handler
```

Fig. 9 Pseudo code for run-time stack overflow prediction (Optimized version for the ARM platform)

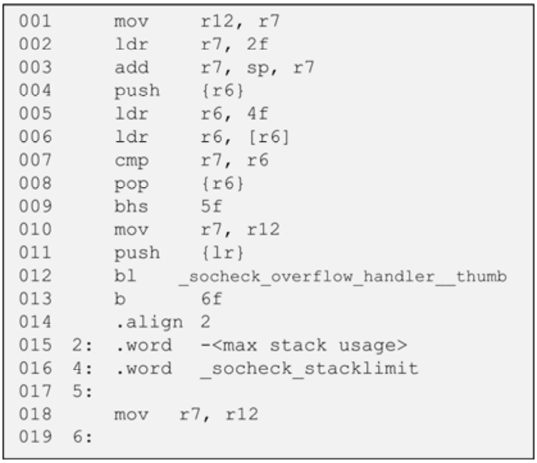
Fig. 10 shows the run-time stack overflow prediction code that is implemented with the ARM instructions. Fig. 11 shows the code that is implemented with the THUMB instructions.

As stated above, the symbol, "<max stack usage>" in Fig. 10 and 11 is replaced with constant integer in real code, and its value is the stack usage of the function to which this code is inserted. The register, "ip(r12)" is used without being saved and restored because it is inter-procedure-call scratch register. The symbol, "_socheck_overflow_handler__arm" and "_socheck_overflow_handler__thumb" are the stack overflow handler functions. The symbol, "_socheck_stacklimit" is the global variable that keeps the stack start address of current thread. If the new compile



```
    ldr     ip, =<max stack usage>
    sub     ip, sp, ip
    str     r6, [sp, #-4]!
    ldr     r6, 4f
    ldr     r6, [r6]
    cmp     ip, r6
    ldr     r6, [sp], #4
    strlo   lr, [sp, #-4]!
    bllo    _socheck_overflow_handler__arm
    b       6f
    .align  2
4:  .word   _socheck_stacklimit
6:
```

Fig. 10 The run-time stack overflow prediction code implemented with the ARM instructions



```
    mov     r12, r7
    ldr     r7, 2f
    add     r7, sp, r7
    push    {r6}
    ldr     r6, 4f
    ldr     r6, [r6]
    cmp     r7, r6
    pop     {r6}
    bhs     5f
    mov     r7, r12
    push    {lr}
    bl    _socheck_overflow_handler__thumb
    b       6f
    .align  2
2:  .word   -<max stack usage>
4:  .word   _socheck_stacklimit
5:
    mov   r7, r12
6:
```

Fig. 11 The run-time stack overflow prediction code implemented with the THUMB instructions

option,“-msocheck” is used, the modified GNU ARM C compiler inserts code in Fig. 10 at the beginning of each function that consists of the ARM instructions, and inserts code in Fig. 11 at the beginning of each function that consists of the THUMB instructions. Whenever context switching occurs, the Ubinos changes the value of the global variable, “_socheck_stacklimit” to the stack start address of the next thread, like Fig. 12.

The idea of this algorithm is not new. Some researchers proposed similar algorithms, although they lack considering for multi-threading [7]. The old ARM procedure call standard [1] also included a similar algorithm, although the current standard [3] does not. The old version of C compiler made by ARM Limited [2] supported the option ‘-apcs /swst’ for the similar algorithm, but the current version does not support this option [4]. Presumably, it is because the current major products based on the ARM architecture tend to have memory management unit and adequate memory. In addition, the GNU ARM C compiler does not support

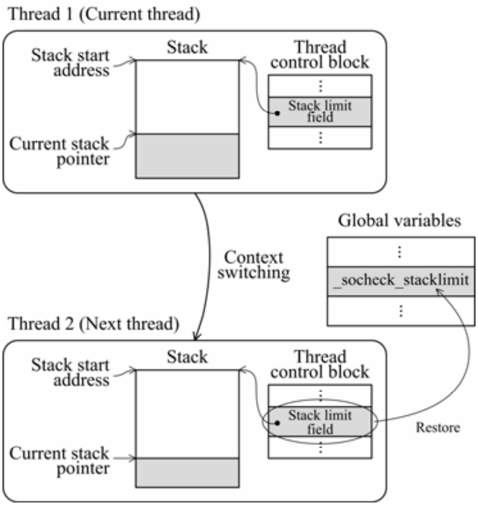


Fig. 12 Switching stack limit information

this algorithm. The options, "-mapcs-stack-check", "-fstack-check", "-fstack-limit-register=reg", and "-fstack-limit-symbol=sym" for similar purposes are defined, but these options cannot be used in memory-limited embedded systems which are the target of this paper, because these options are not implemented or are implemented using a memory management unit. In summary, we modified the GNU ARM C compiler to support this algorithm, as stated above.

4 Evaluation

Let's assume the following situation: An embedded system including TCP echo service function is being developed. The system has to run the following threads at the same time. The first is "ip_task" that is a thread for Ethernet and TCP/IP service. The second is "tcpecho_task" that is a thread for echo service through the TCP. The third is "shell_task" that is a thread for simple shell service. The maximum stack usage of each thread is 424(0x1A8), 296(0x128), and 404(0x194) bytes in that order. The available memory that can be used as stack is only 1152(0x480) bytes, and the default stack size of a thread is 384(0x180) bytes. And the system has a bug that makes itself crash. If a traditional way that does not use the technique proposed in this paper is used, the process to optimize stack is presumably as following:

1. The developer sets the stack sizes of all threads to the default value. Then, he/she builds and tests the executable binary code.
2. The system crashes during testing.
3. The developer guesses and sets the proper stack size of each thread. Then, he/she builds and tests the executable binary code again.

4. Although the developer repeats step 3 many times, the problem is not solved.
5. The developer thinks that the reason of the system crash may not be a stack overflow, and reviews the source code.
6. The developer finds and fixes a suspicious part of the source code. He/she then builds and tests the executable binary code again.
7. But the system still crashes during testing.
8. Although the developer keeps reviewing the source code many times, he/she cannot find a suspicious part of the source code anywhere.
9. The developer thinks again that the system crashes may be due to the stack overflow, and repeats step 3 again.
10. After repeating step 3 many times, accidentally, the developer finds out that the system does not crash if each stack size of "ip_task", "tcpecho_task", and "shell_task" is 432(0x1B0), 304(0x130), and 416(0x1A0) bytes in that order.
11. Feeling insecure, the developer releases the executable binary code.

The step 4 and 9 of this scenario need good intuition and good luck. Due to this, it is hard to predict the time and effort needed to optimize stacks in this scenario. It normally takes a lot of time and effort. In other words, this scenario is uncertain and inefficient. On the other hands, if the technique proposed in this paper is used, the stack optimization process is presumably as following:

1. The developer sets the stack sizes of all threads to the default value. Then, he/she builds and tests the executable binary code with the technique proposed in this paper.
2. During testing, a message notifying that stack overflow will occur in the "ip_task" is displayed. Immediately after, the current stack usage maximum record of each thread is displayed. Each stack usage maximum record of "ip_task", "tcpecho_task", and "shell_task" is 396(0x18C), 224(0xE0), and 224(0xE0) bytes in that order.
3. The developer changes each stack size of "ip_task", "tcpecho_task", and "shell_task" to 512(0x200), 320(0x140), and 320(0x140) bytes respectively. Then, he/she builds and tests the executable binary code again.
4. During testing, a message notifying that stack overflow will occur in the "shell_task" is displayed. Each stack usage maximum record of "ip_task", "tcpecho_task", and "shell_task" is 400(0x190), 224(0xE0), and 348(0x15C) bytes in that order.
5. The developer changes each stack size of "ip_task", "tcpecho_task", and "shell_task" to 464(0x1D0), 224(0x0E0), and 464(0x1D0) bytes respectively. Then, he/she builds and tests the executable binary code again.
6. During testing, a message notifying that stack overflow will occur in the "tcpecho_task" is displayed. Each stack usage maximum record of "ip_task", "tcpecho_task", and "shell_task" is 400(0x190), 244(0x0F4), and 404(0x194) bytes in that order.
7. The developer changes each stack size of "ip_task", "tcpecho_task", and "shell_task" to 432(0x1B0), 304(0x130), and 416(0x1A0) bytes respectively. Then, he/she builds and tests the executable binary code again.
8. Although the message notifying stack overflow is not displayed any more, the system still crashes during testing.
9. The developer reviews the source code.
10. The developer finds out and fixes a suspicious part of the source code. He/she then builds and tests the executable binary code again.
11. The system does not crash any more during testing.
12. After all tests, the developer checks the stack usage maximum record of each thread. Each stack usage maximum record of "ip_task", "tcpecho_task", and "shell_task" is 424(0x1A8), 296(0x128), and 404(0x194) bytes in that order.

13. The developer confirms that all stack sizes are proper. Then, he/she builds and tests the executable binary code without the technique proposed in this paper.
14. The developer confirms that there is no error during testing. Then, he/she releases the executable binary code.

There are no inefficient or uncertain steps in this second scenario. All steps are clear and easy. In fact, the first scenario is contrived. But, in all probability, almost all developers who have used multi-threading in memory-limited embedded systems have similar experiences. The second scenario (except the assumption of a bug) describes a real stack optimization process of the TCP echo server example developed with the modified GNU ARM C compiler and the Ubinos.

Table 1 The optimal stack size of the TCP echo server example

ARM / THUMB	Optimization option	Optimal stack size (byte)			
		ip task	tcpecho task	shell task (idle task)	Interrupt (svc)
ARM	-O0	528	344	572	376
	-Os	448	288	404	344
THUMB	-O0	576	336	620	380
	-Os	424	296	404	312

Table 1 shows the optimal stack size (maximum stack usage) of each thread of the TCP echo server example that is measured with the technique proposed in this paper. Table 2 shows the memory (RAM) usage of the example.

Table 2 Memory (RAM) usage of the TCP echo server example

ARM / THUMB	Optimization option	With proposed technique	Memory (RAM) usage			
			Static (byte)	Dynamic (byte)	Total (byte)	Increase (%)
ARM	-O0	no	2152	2276	4428	-
		yes	2192	2276	4468	0.90
	-Os	no	2120	1972	4092	-
		yes	2160	1972	4132	0.98
THUMB	-O0	no	2156	2364	4520	-
		yes	2196	2364	4560	0.88
	-Os	no	2088	1956	4044	-
		yes	2128	1956	4084	0.99

The stacks of the example were optimized through the process described in the second scenario. As a result, although the example runs three threads and uses the TCP/IP stack [10], it uses a very small amount of memory (only about 4K bytes). Table 3 shows the ROM (Flash) usage and performance of the example. Table 4 shows the ROM (Flash) usage and result of the Dhrystone performance test example.

Table 3 ROM (Flash) usage and performance of the TCP echo server example

ARM / THUMB	Optimization option	With proposed technique	ROM (Flash) usage		Performance	
			(byte)	Increase (%)	Round trip time (us)	Decrease (%)
ARM	-O0	no	90332	-	772	-
		yes	112040	24.03	884	12.67
	-Os	no	50720	-	442	-
		yes	64764	27.69	498	11.24
THUMB	-O0	no	58356	-	672	-
		yes	77188	32.27	761	11.70
	-Os	no	36740	-	391	-
		yes	50580	37.67	448	12.72

Table 4 ROM (Flash) usage and result of the Dhrystone performance test example

ARM / THUMB	Optimization option	Use proposed technique	ROM (Flash) usage		Performance	
			(byte)	Increase (%)	Dhrystones per second	Decrease (%)
ARM	-O0	no	115836	-	8375	-
		yes	136208	17.59	7350	12.24
	-Os	no	71928	-	23650	-
		yes	87144	21.15	21325	9.83
THUMB	-O0	no	79464	-	10425	-
		yes	97520	22.72	9175	11.99
	-Os	no	55276	-	30075	-
		yes	70520	27.58	25675	14.63

Fig. 13 and Fig.14 show overheads of the proposed technique. As shown in these figures, this technique has some overheads. However, the overhead of the memory that is the most important resource of the systems which are the target of this paper is less than 1%, and the performance overhead is not significant. The ROM overhead is relatively high, but it is at a manageable level and current research group anticipates that the ROM overhead will decrease to half in the next version of the implementation that is currently being developed. Furthermore, the overheads are not added to the final release of the executable binary code, because the proposed technique is necessary only during the stack optimization process in the development phase.

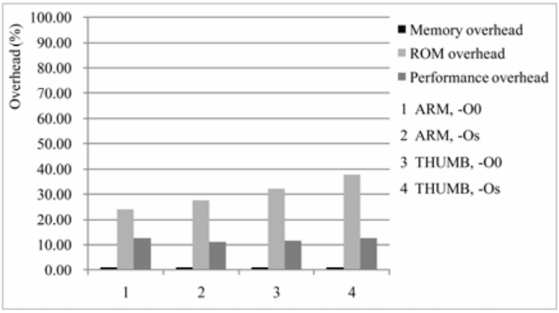


Fig. 13 Overheads of the proposed technique in the TCP echo server example

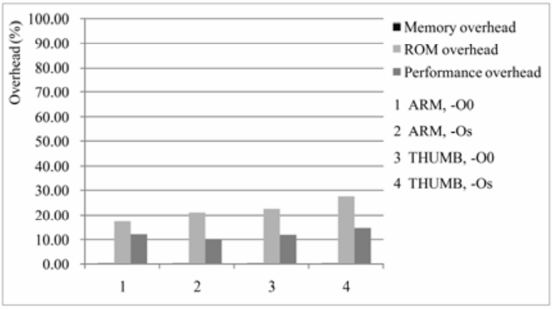


Fig. 14 Overheads of the proposed technique in the Dhrystone test example

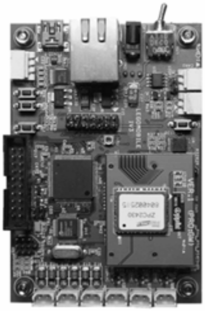
Table 5 shows problems of techniques for stack optimization. The techniques proposed in prior research have problems that make them unpractical, but the compiler-assisted maximum stack usage measurement technique proposed in this paper does not. As stated above, the technique has some performance and ROM overhead, but these overheads are added only during the stack optimization process in the development phase and not significant. Thus, it is not a serious problem. Also, the technique needs testing to measure maximum stack usage, but it is also not a serious problem because the testing is always necessary to make reliable systems even though other stack optimization techniques are used. The possibility of stack overflows by untested control flow can also be removed. This problem and the solution for it will be explained in detail in the next section. The solution for the problem that the technique cannot address the stack usage of software segments developed with assembly language will also be explained in the next section.

Table 5 Problems of techniques for stack optimization

-	Stack analysis technique	Stack pollution check technique	Dynamic stack resizing technique	Compiler-assisted maximum stack usage measurement technique
Possibility of overestimation of stack usage	●	X	X	X
Cannot use with recursive function call	●	X	X	X
Cannot use with indirect function call	●	X	X	X
Possibility of undetectable stack usage	X	●	X	X
Possibility of failure to detect stack overflows	X	●	X	X
Need testing	X	●	X	●
Possibility of stack overflows by untested control flow	X	●	X	●
Cannot address the stack usage of software segments developed with assembly language	X	X	X	●
Memory overhead	X	X	X	X
Performance overhead	X	▲	●	▲
ROM overhead	X	X	●	▲
Memory fragmentation problem	X	X	●	X
Unpredictable response time	X	X	●	X

▲ Exist only during the stack optimization process

This evaluation was accomplished on the ESPS mobile board shown in Fig. 15. It is a multi-purpose board that we developed for embedded system prototyping.



- Atmel AT91SAM7X256
  ARM7TDMI (48MHz)
  64KB SRAM
  256KB Flash
- 3 sensor input port
  (10-bit ADC)
- 3 DC motor driver port
- 1 Ethernet port
- 1 CAN port
- USB / ZigBee /RS232 /
  ...

Fig. 15 ESPS mobile board (Multi-purpose board for embedded system prototyping)

5 Problems and Solutions

First, the technique proposed in this paper has the possibility of stack overflow by untested control flow. To find the exact maximum stack usage with this technique, all possible control flows have to be reproduced during testing. However, sometimes it is difficult because there can be control flows that occur very rarely. The untested control flow can make the measurement incorrect and cause stack overflows. This problem can be solved by using dynamic stack resizing technique in conjunction with the proposed technique. If the dynamic stack resizing technique is used with the run-time stack overflow prediction algorithm proposed in this paper, the stack overflow caused by the untested control flows can be prevented. Also, the performance overhead problem of the dynamic stack resizing technique is solved naturally if it is used with the technique proposed in this paper, because stack resizing action will be performed very rarely in systems in which stack is optimized with the technique proposed in this paper. At present, the current researchers are developing algorithms that will make it possible to resize stack in a short fixed time without memory fragmentation. It is believed by the current authors that these algorithms can clearly solve this problem.

Second, the technique proposed in this paper cannot address the stack usage of the software developed with assembly language. Sometimes, a few segments of software are developed with assembly language to improve performance or to precisely control hardware. The technique proposed in this paper cannot manage the stack usage of such software segments, because they do not go through the compile process. However, it is not difficult to manually insert the special code for supporting this technique at the beginning of these software segments, because only a very few segments of uncomplicated software code are developed with assembly language these days, as stated above. It can also be a practical solution for this problem to manually calculate and add the maximum value among the stack usages of these software segments to the maximum stack usage of each thread in advance, because almost all stack usages of such segments are small and not accumulated.

6 Conclusion

This paper proposes the compiler-assisted maximum stack usage measurement technique as a new solution for stack optimization. With this technique, we can clearly find out the maximum stack usage of each thread during testing, and thus clearly allowing stack optimization. Also, unlike the solutions proposed in prior research, this technique does not have problems such as limited availability, the possibility of undetectable stack usage, and memory fragmentation. This technique adds some overhead, but the overhead is added only during the stack optimization process in the development phase. Also, despite the necessity for modification of the compiler and operating system, this technique is easy to implement. It can be implemented by slight modification of the existing compiler and operating system. Furthermore, this technique was verified by implementing and evaluating it on the ARM platform. Although this technique still has the possibility for stack overflow by untested control flow, the current research group believes that the short and fixed time-consuming stack resizing technique that is currently being developed can solve this problem.

Many embedded systems currently have a memory management unit and adequate memory, but a considerable number of embedded systems still do not have a memory management unit and have less than tens of kilo-bytes of memory because of energy-efficiency, cost, and size. Also, in the near future, the number of such memory-limited embedded systems is expected to increase as wireless sensor networks and ubiquitous computing that need massive energy-efficient, low-cost, and small embedded systems come into use. However, the people who develop such memory-limited embedded systems are still having a hard time using multi-threading because of the difficulty of stack optimization. The technique proposed in this paper can be a clear solution for this problem. Although the technique has not been perfected yet, it is still a practical solution, and the current research group anticipates perfecting the technique in their next work.

References

1. Arm ltd. The ARM-THUMB Procedure Call Standard - A-05. ARM Limited (1998)
2. Arm ltd, ARM Ltd. Homepage (2010), http://www.arm.com
3. Arm ltd, Procedure Call Standard for the ARM Architecture. ARM Limited (2008)
4. Arm ltd, RealView Compilation Tools Version 4.0 Essentials Guide. ARM Limited (2008)
5. Behren, R.V., Condit, J., Zhou, F., et al.: Capriccio: Scalable Threads for Internet Services. ACM SIGOPS Operating Systems Review 37, 268–281 (2003)
6. Bhatti, S., Carlson, J., Dai, H., et al.: MANTIS OS: An Embedded Multithreaded Operating System for Wireless Micro Sensor Platforms. Mobile Networks and Applications 10, 563–579 (2005)

7. Biswas, S., Carley, T., Simpson, M., et al.: Memory Overflow Protection for Embedded Systems using Run-time Checks, Reuse, and Compression. ACM Transactions in Embedded Computing Systems 5, 719–752 (2006)
8. Brylow, D., Damgaard, N., Palsberg, J.: Static Checking of Interrupt-driven Software. In: Proceedings of the 23rd International Conference on Software Engineering, pp. 47–56. IEEE Computer Society Press, Toronto (2001)
9. Chatterjee, K., Ma, D., Majumdar, R., et al.: Stack size analysis for interrupt-driven programs. Information and Computation 194, 144–174 (2004)
10. Dunkels, A.: The uIP Embedded TCP/IP Stack. (2006), http://www.sics.se/~adam/uip
11. Dunkels, A., Gronvall, B., Voigt, T.: Contiki - A Lightweight and Flexible Operating System for Tiny Networked Sensors. In: Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks, pp. 455–462. IEEE Computer Society Press, Tampa (2004)
12. Dunkels, A., Schmidt, O., Voigt, T., et al.: Protothreads: Simplifying Event-driven Programming of Memory-constrained Embedded Systems. In: Proceedings of the 4th international conference on Embedded networked sensor systems, ACM, Boulder (2006)
13. Free software foundation, GNU ARM toolchain 4.3.2 (2010), `http://www.gnuarm.org/home.html`
14. Guillemin, P.: ST Application Note, Stack overflow detection using the ST9 watchdog timer. SGS-THOMSON Microelectronics Group of Companies (1994)
15. Han, C.-C., Kumar, R., Shea, R., et al.: A Dynamic Operating System for Sensor Nodes. In: Proceedings of the 3rd international conference on Mobile systems, applications, and services, pp. 163–176. ACM, Seattle (2005)
16. Hill, J., Szewczyk, R., Woo, A., et al.: System architecture directions for networked sensors. ACM SIGPLAN Notices 35, 93–104 (2000)
17. Kasten, O., Romer, K.: Beyond Event Handlers: Programming Wireless Sensors with Attributed State Machines. In: Proceedings of the 4th international symposium on Information processing in sensor networks, pp. 45–52. IEEE Press, Los Alamitos (2005)
18. Kim, H., Cha, H.: Multithreading Optimization Techniques for Sensor Network Operating Systems. Lecture Notes In Computer Science 4373, 293–308 (2007)
19. Ousterhout, J.: Why Threads Are A Bad Idea (for most purposes). In: Usenix Annual Technical Conference, San Diego, California, USA (1996)
20. Park, S.H.: Ubinos - A Multi-threaded Operating System for Resource-limited Embedded Systems (2010), `http://www.ubinos.org`
21. Park, S.H.: Ubitools - The Develpment Tools for the Ubinos (2010), `http://www.ubinos.org/mediawiki/index.php/Ubitools`
22. Ralf, S.E.: Portable multithreading: the signal stack trick for user-space thread creation. In: Proceedings of the annual conference on USENIX Annual Technical Conference. USENIX Association, San Diego, California, USA,
23. Real time engineers ltd, FreeRTOS - A FREE open source RTOS for small embedded real time systems (2010), `http://www.freertos.org`
24. Regehr, J.: Say no to stack overflow. Embedded Systems Programming 17, 10–20 (2004)
25. Regehr, J., Reid, A., Webb, K.: Eliminating Stack Overflow by Abstract Interpretation. ACM Transactions on Embedded Computing Systems 4, 751–778 (2005)

Fault-Tolerant Clock Synchronization for Time-Triggered Wireless Sensor Network

Dong-Gil Kim and Dongik Lee

Summary. Wireless sensor networks have been employed in a wide range of industry thanks to the benefits of mobility, flexibility and low cost. However, the unreliability of wireless networks is a great challenge to be used in feedback control. In order to offer reliable and deterministic wireless communications, time-triggered mechanisms are commonly used. This paper presents a fault-tolerant clock synchronization method that can be used for a time-triggered wireless network. The effectiveness of the proposed method is demonstrated through a set of experiments.

1 Introduction

Wireless technology has many advantages over wired network [1]. First of all, wireless links between computing components reduce the install costs and the complexity of wiring harness. Enhanced mobility and flexibility of the system are also beneficial for the easy modification leading to the reduced time-to-market. Thanks to these advantages, several researchers have paid attention to apply wireless networks into feedback control [2-4]. However, wireless networks, such as Zigbee [5,6], have drawbacks with communication reliability to be used in feedback control. One of the serious problems with Zigbee is the non-deterministic temporal behavior caused by CSMA/CA. Although CSMA/CA provides an efficient solution to resolving message collisions at low bus traffic, randomly varying message latency results in the degradation of control performance [7]. Another drawback with CSMA/CA is the difficulty in detecting the loss of message because the expected arrival time is unknown to the receiving nodes.

Several researchers have employed a time-triggered mechanism to achieve reliable and deterministic wireless networks. For example, Kim *et al.* [8] implemented a time-triggered mechanism on top of IEEE802.15.4. The key idea is to assign timeslots in which designated nodes have exclusive rights to broadcast on the bus. As a result, even at high bus traffic, latency of each message with a low priority can be bounded.

Dong-Gil Kim · Dongik Lee
Kyungpook National University
1370 Sankyug-Dong, Daegu, 702-701, Republic of Korea
e-mail: dilee@ee.knu.ac.kr

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 131–142.
springerlink.com

Since a time-triggered protocol depends on the notion of common time reference, synchronization of all clocks in the network must be achieved. Otherwise, message collision could occur due to overlapped timeslots, leading to excessive message delays. Since every node in the network operates in a physically separated place, local clocks can drift away from each other. However synchronization of local clocks is not an easy task as clock properties are varying with time and environmental variables such as temperature and humidity. For the last three decades, the clock synchronization problem has been mainly studied in computer science [9,10]. However, these methods have drawbacks in applying to a wireless sensor network because of the computing overhead, needs for external hardware, or the risk of faulty master clocks.

This paper addresses a software-based fault-tolerant clock synchronization technique that can be used for a time-triggered wireless sensor network based on IEEE802.15.4. The proposed method requires very low computing load while offering microsecond-precision and the ability of tolerating faulty master clocks. In order to achieve the high precision and fault-tolerance, it exploits MAC-layer timestamping and a mater-slave structure with multiple masters. The effectiveness of the proposed method is demonstrated through a set of experiments.

2 Time-Triggered Wireless Sensor Network

2.1 Needs for Time-Triggered Wireless Network

Control systems nowadays employ digital communication networks to satisfy various demands on performance, maintenance, and reliability [12]. However, for implementing and/or modifying a networked control system, wired networks cause significant limitations in terms of mobility, flexibility and extensibility. The use of wireless communication networks can be considered as an effective solution to overcoming these problems [2,3,13].

The ZigBee protocol based on IEEE802.15.4 is one of the most commonly used wireless sensor networks [5,6,14]. However, a ZigBee network has two drawbacks to be used in a feedback control system. One is the non-deterministic temporal behavior under high level of bus traffic, and the other is undetected message collisions between nodes. The non-deterministic behavior is mainly caused by the arbitration mechanism based on CSMA/CA. Although CSMA/CA offers an efficient solution to resolving message conflicts at a low level of bus traffic, lower priority messages may prevent higher priority messages from accessing to the bus channel. Also the randomly varying message latency results in the degradation of control performance, and even an unstable system [7]. Another difficulty is the undetected message collisions caused by the hidden terminal problem as shown in Fig.1. Since messages transmitted by either node D1 or node D2 are not recognized by the other node, both nodes may try to send messages simultaneously, resulting in a conflict at node C.

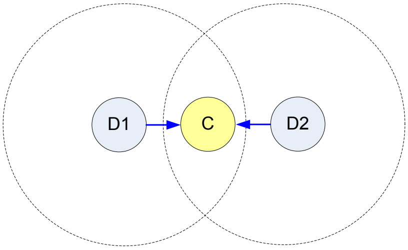


Fig. 1 An example to show hidden terminal problem

2.2 Time-Triggered Network Based on IEEE802.15.4

In order to overcome these problems, the authors have previously developed a deterministic wireless network for feedback control [8]. As illustrated in Fig.2, the protocol is based on a time triggered mechanism implemented on top of the MAC layer of IEEE 802.15.4. Time-triggered approaches are mainly based on the time division multiple access (TDMA) mechanism, in which exclusive timeslots are designated to each node of the system. Therefore, not only can message latency be constant virtually independent of the level of bus traffic, but also the hidden terminal problem can be avoided.

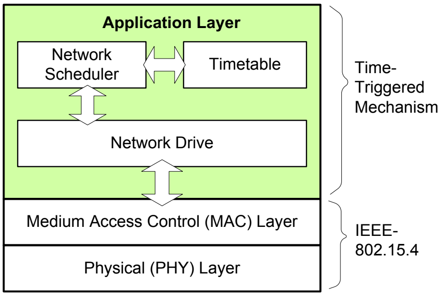
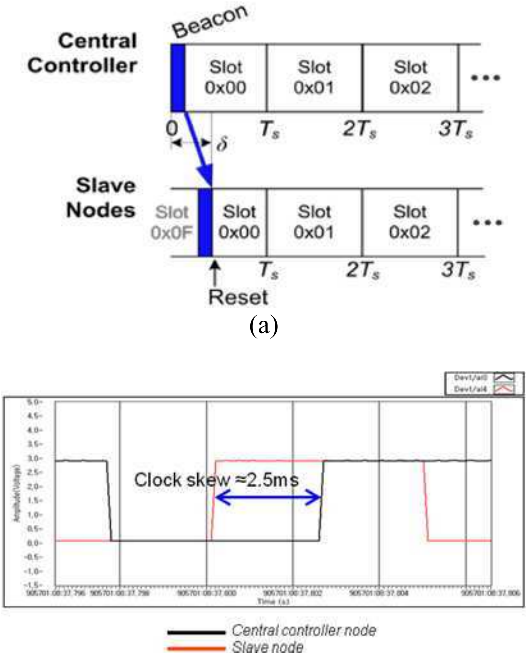


Fig. 2 Time-triggered wireless network architecture [8]

2.3 Clock Synchronization

The availability of reliable and accurate time references across the whole network nodes is very critical for the success of time-triggered communications. A simple synchronization strategy based on master-slave architecture has been employed in the previous work. The central controller periodically broadcasts beacon messages to which entire nodes of the system synchronize their local time. As depicted in Fig.3, the central controller broadcasts beacon messages at the predetermined period of time. On the reception of a beacon message, each slave node resets its local time. However, this approach has obvious limitations: the lack of fault-tolerance in the event of faulty master, and the excessive synchronization errors of 2.5ms due to the transmission delay with beacon signals [14].



(a)

(b)

Fig. 3 Clock synchronization for the previous work [14]; (a) beacon message broadcast by master, and (b) synchronization error between master and slave nodes

3 Fault-Tolerant Clock Synchronization for Time-Triggered Wireless Sensor Network

3.1 Overview

It is obvious that a clock synchronization mechanism for feedback control systems must be able to provide high degree of precision and the ability of tolerating faulty clocks while requiring low computing load. However, the existing synchronization mechanism depicted in Fig.3 suffers the problems of excessive synchronization errors due to the transmission delay as well as the lack of fault-tolerance.

In order to improve the above mentioned two problems, this paper presents a novel multi-master based fault-tolerant clock synchronization method that can be applied to time-triggered wireless sensor networks. The proposed method adopts a master-slave structure with multiple masters and MAC-layer timestamping that can greatly improve the precision. Although the use of multiple masters is commonly found in a master-slave structure, the uniqueness of the proposed method is the simplicity in selecting a correct master candidate. The main feature of the proposed method is summarized as follow:

Tolerating faulty master: The proposed method has a master-slave structure in order to make the algorithm as simple as possible. The master-slave structure can drastically reduce the number of messages for synchronization if a broadcast network is used. The ability of tolerating faulty master is achieved through a novel method to determine a correct master clock by comparing the message arrival times.

Improving synchronization accuracy: All clock nodes in the system synchronize to the time of a selected master clock. In order to eliminate all uncertainty including the effect of transmission delay, the MAC-layer timestamping technique [11] is used.

3.2 Broadcasting Synchronization Messages by Multi-master Nodes

Fig.4 shows the proposed synchronization method in a multi-hop network environment. Each hop has a small number of master nodes that are preselected to broadcast synchronization messages in their exclusive transmission timeslots. Note that, even in the same hop, the master clock that broadcasts the synchronization message keeps changing according to the progress of timeslots. All other nodes excepting the current master node play as slave nodes. Therefore, any single faulty master does not cause any synchronization failure.

Synchronization across a multi-hop network is achieved by assigning the *relay master clock nodes* that transfer the reference time of one hop to the next hop. The number of relay nodes between two hops is determined based on the demand for fault-tolerance.

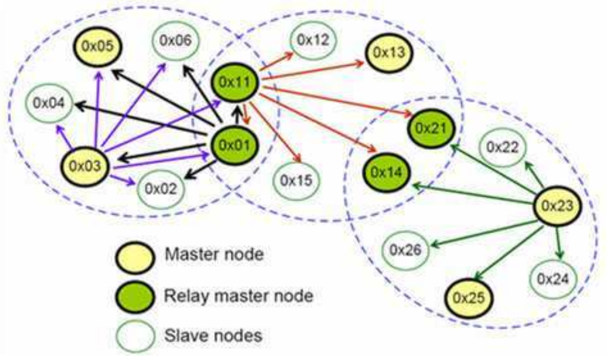


Fig. 4 Broadcasting of synchronization messages by multi-master clocks in a multi hop network

3.3 Fault-Tolerance with Multi-master

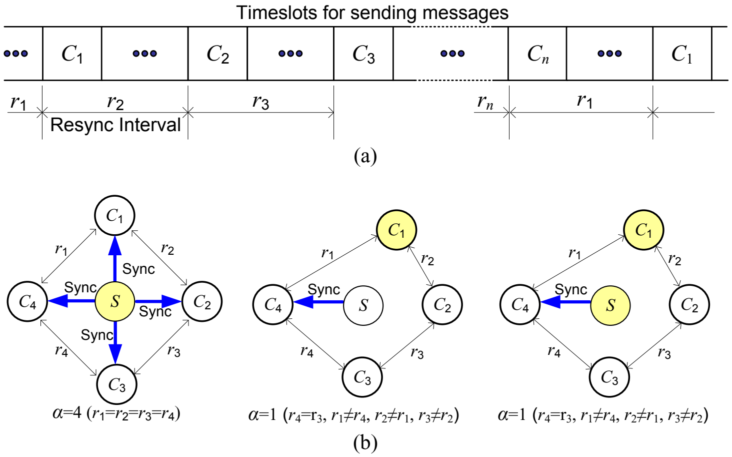
A multi-master approach is employed in the proposed algorithm in order to tolerate faulty master clocks. As shown in Fig.5(a), a set of master candidate clocks (C_i) are predetermined and assigned to timeslots at which they transmit synchronization messages (i.e., timestamps). The time interval (r_i) between any two successive master candidate clocks, C_{i-1} and C_i, is set to equal. In a multi-master approach, the minimum size (N) of master candidate group to tolerate upto f faulty clocks is given by $N=2f+1$. Therefore, each clock has to run a voting mechanism to select a correct master clock among the candidate master clocks.

The proposed method for identifying a correct clock is based on the time interval between two successive synchronization messages. Firstly, each clock calculates the time interval r_i as below:

$$r_i = C_i - C_{i-1} \tag{1}$$

where, C_i and C_{i-1} represent the arrival times of synchronization messages being sent by master candidate clocks. Since all master candidate clocks are supposed to send their synchronization messages in a uniform time interval, a correct master candidate clock C_i must satisfy the following requirement:

$$r_i = r_{i-1} \tag{2}$$



(a)

α=4 (r_1=r_2=r_3=r_4)

α=1 (r_4=r$_3$, r_1≠r_4, r_2≠r_1, r_3≠r_2)

α=1 (r_4=r$_3$, r_1≠r_4, r_2≠r_1, r_3≠r_2)

(b)

Fig. 5 Proposed method for selecting a correct master to tolerant faulty clocks; (1)schedule for sending synchronization messages; and (b) example of identifying correct master clocks for f=2 (Shaded clocks are faulty. Bold arrows indicate possible synchronization by clock S.)

In contrast, if the clock C_i is faulty, the above condition cannot be satisfied. Therefore correct master candidate clocks can be identified by examining eqn (2). Since every clock has a nonzero drift rate, eqn (2) can be rewritten as follow:

$$\left| r_i - r_{i-1} \right| \leq D \tag{3}$$

where, D denotes an acceptable clock skew between any two correct clocks.

Fig.5(b) shows an example of determining faulty clocks using the proposed method. It is assumed that the total number of faulty clocks is f=2, and two successive master candidate clocks cannot be faulty. In this example, α denotes the number of pairs having equal time intervals, and satisfies α>0 by the assumption. The minimum number of clocks in the network should be 5 including non-master clocks. The clock S denotes an arbitrary clock receiving synchronization messages. It is important to notice that the receiving clock S indicates not only slave clocks but also all master candidate clocks excepting the one sending a synchronization message. It is straightforward that the clock C_i satisfying eqn.(3) is a correct master clock to which the clock S can synchronize.

3.4 Synchronization Precision

The synchronization precision (δ) can be defined as the worst skew between any two correct clocks. The synchronization precision that can be achieved with a master-slave structure is given by:

$$\delta = 2\rho R + \xi, \quad r \le R \le r(2f+1) \tag{4}$$

where, ρ and R denote the clock drift rate and the interval between two effective synchronizations, respectively. ξ is the clock reading error caused by clock resolutions and delays for computing and message transmissions. For example, assume that the required synchronization precision is δ=10μs with f=1 and ξ=5μs. Then the maximum resynchronization interval r can be given by 83ms.

To improve the synchronization precision, it is necessary not only to use a short resynchronization interval, but also to minimize the clock reading error. Since the primary cause for the clock reading error is delays for message arbitration, the MAC-layer timestamping technique [11] is used in this work.

4 Experimental Results

The experiments conducted in this work are two fold: one is to verify the performance of the proposed synchronization method in terms of fault tolerance and synchronization precision; and the other is to demonstrate the performance of DC motor control with the time-triggered wireless sensor network using the proposed synchronization method. Two different demonstrator setups for each experiment are used as shown in Fig.6 and Fig.8. The key features of the nodes on which the time-triggered mechanism is implemented are summarized in Table 1.

Table 1 Key specifications of IEEE802.15.4 nodes used for experiments

Item	Description
Microcontroller	Atmega 128L
Clock [MHz]	7.37
ROM [KB]	128
RAM [KB]	2
Storage [KB]	4
Radio [GHz]	CC2420 2.4
Bit rate [Kbps]	250
Max range [m]	125
Power source	2 AA batteries
Operating system	TinyOS ver. 1.1.7

4.1 Results for Clock Synchronization

The proposed method is examined using the experimental setup illustrated in Fig.6. The number of faulty clocks is assumed to *f*=1, and thus three clocks including two master candidates (C_1 and C_2) are used. The resynchronization interval *r* is chosen by 100ms. For the purpose of collecting experimental data, each node is equipped with a CAN interface. Two master candidate clocks are supposed to periodically broadcast their synchronization messages, while all of the three clocks transmit their local time to a PC that runs CANalyzer®.

The experiments carried out in this work are two fold: one is to demonstrate the ability of fault-tolerance against a faulty master clock C_1; and the other is to determine the synchronization precision achieved.

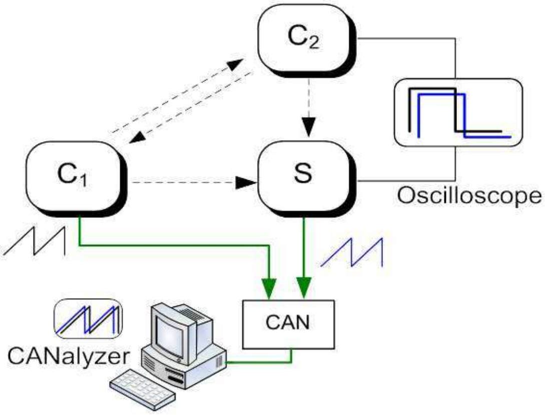
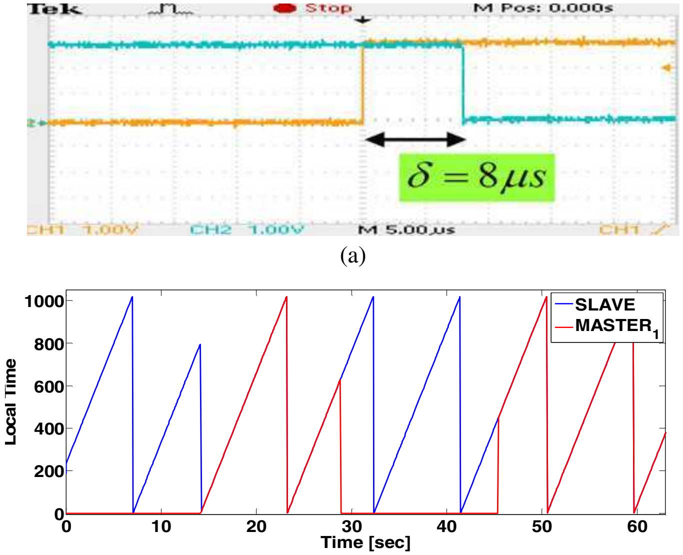


Fig. 6 Experimental setup for clock synchronization

Clock synchronization precision: Fig.7(a) shows the synchronization precision between the clocks C_2 and *S*. Each clock generates a periodic pulse according to its local time. The synchronization precision is determined by comparing two pulse signals with an oscilloscope. The precision achieved is around 8μs thanks to MAC-layer timestamping.

Fault-tolerance with faulty master clock: The ability of tolerating a faulty master clock (C_1) is verified by examining the synchronization between two clocks, C_1 and *S*, as shown in Fig.7(b). The clock C_1 is enabled at t=15*s* and then the slave clock *S* immediately synchronizes to the time of C_1. While the clock C_1 is removed at t=28*s*, the slave clock *S* still maintains its local time by synchronizing to the other master clock C_2. The disabled clock C_1 is reactivated at t=45*s*. Note that the recovered master clock C_1 is now reset to the time of C_2 and the slave clock *S* still retains its local time which is synchronized to C_2, instead of resynchronizing to C_1. This result implies that the slave clock *S* is not interrupted by the faulty master clock C_1 thanks to the second master clock C_2, proving the ability of tolerating faulty master clocks with the proposed method.



(a)

(b)

Fig. 7 Results of clock synchronization; (a) synchronization precision measured with oscilloscope, and (b) synchronization between the slave clock S and fault master C_1

4.2 Results for Feedback Control of DC Motor

Fig.8 depicts the experimental setup for these experiments. The central controller (node SYNCH_1) is communicating with the PC through CAN to provide set-points which are desired motor speed. The setpoint information is then transmitted to the motor controllers (nodes PID_1 and PID_2) through the time-triggered wireless network which is synchronized with the proposed method. MOT_1 and MOT_2 nodes act as the interface to the DSP which generates commanded PWM signals as well as collects the encoder outputs. The encoder outputs are then fed back to node PID_1 and PID_2 through wireless communications, forming the closed-loop control. The experiments are carried out under approximately 80% of channel load.

Fig.9 illustrates the response of angular velocity control with the time-triggered wireless communication network which is synchronized using the proposed method. When the time-triggered mechanism with the proposed method is employed it can be seen that the response is relatively stable and time-invariant. In contrast, using the standard Zigbee network without the time-triggered mechanism, the response represents a larger settling time and time-varying behavior due to the excessive jitter in message latency as shown in Fig.10. The excessive overshoots in both results are introduced by proportional control with a large gain.

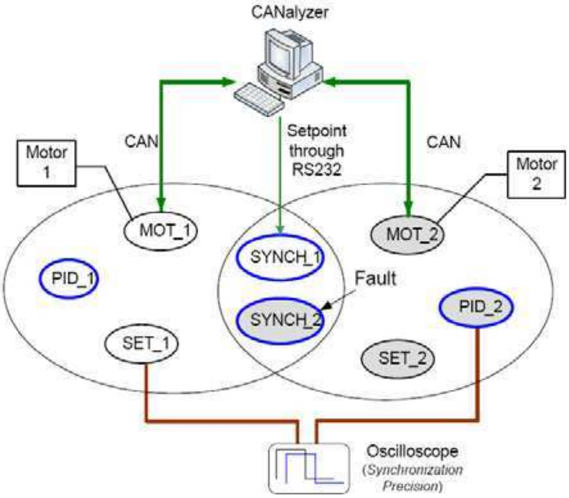
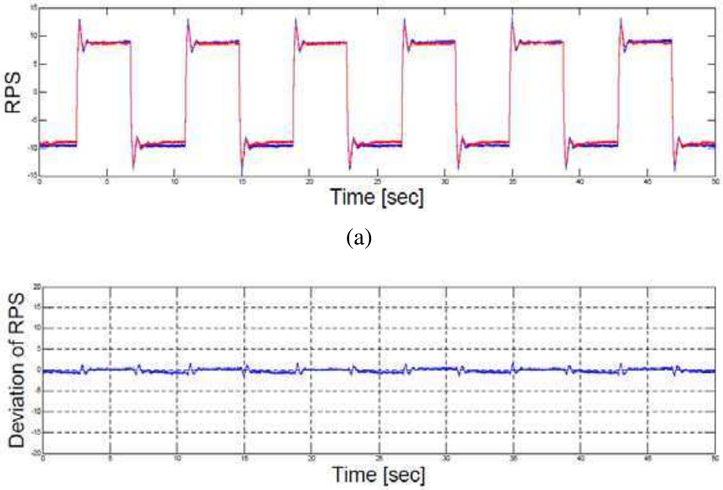


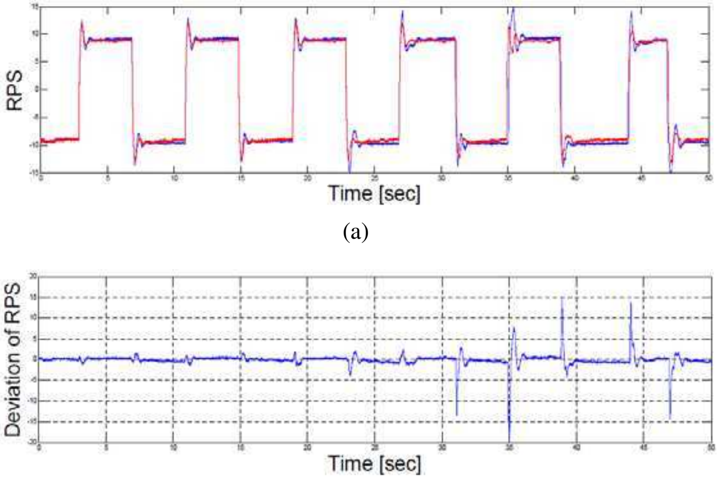
Fig. 8 Experimental setup for feedback control of DC motor



(a)

(b)

Fig. 9 Result of motor control with time-triggered communications; (a) actual trajectory of angular velocity, and (b) errors between desired and actual trajectories



(a)

(b)

Fig. 10 Result of motor control with standard zigbee network; (a) actual trajectory of angular velocity, and (b) errors betweens desired and actual trajectories

5 Concluding Remarks

In order to offer reliable and deterministic communications on a wireless sensor network, time-triggered approaches are commonly used in a wide range of industry. In this paper a fault tolerant clock synchronization method that can be used for time-triggered wireless sensor networks has been proposed. The proposed method is based on a multi-master structure with an efficient mechanism to determine correct master clocks to tolerate any faulty master clocks. The microsecond precision has been also achieved by adopting MAC-layer timestamping. The proposed method is so simple that can be implemented on a low-cost embedded processor requiring very low level of computing load. The effectiveness of the proposed method has been demonstrated through a set of experiments.

Acknowledgments. Following are results of a study on the "Human Resource Development Center for Economic Region Leading Industry" Project, supported by the Ministry of Education, Science & Technology (MEST) and the National Research Foundation of Korea (NRF).

References

1. Irwin, D.W., Colandairaj, J., Scanlon, W.G.: An overview of wireless networks in control and monitoring. In: Huang, D.-S., Li, K., Irwin, G.W. (eds.) ICIC 2006. LNCS (LNAI), vol. 4114, pp. 1061–1072. Springer, Heidelberg (2006)
2. Liu, X., Goldsmith, A.: Wireless network design for distributed control. In: Proc. IEEE Conf. Decision and Control (CDC 2004), pp. 2823–2829 (2004)

3. Thompson, H.: Wireless and internet communication technologies for monitoring and control. Control Engineering Practice 12(6), 781–791 (2004)
4. Drew, M., Liu, X., Goldsmith, A., Hedrick, K.: Networked control system design over a wireless LAN. In: Proc. IEEE Conf. Decision and Control (CDC 2005), pp. 6704–6709 (2005)
5. IEEE, IEEE Std. 802.15.4-2003 wireless medium access control (MAC) and physical layer (PHY) specification for low rate wireless personal area networks (LR-WPANs), IEEE Press, New York (2003)
6. ZigBee Alliance, ZigBee Specification (December 2006), http://www.zigbee.org
7. Tipsuwan, Y., Chow, M.: Control methodologies in networked control systems. Control Engineering Practice 11(10), 1099–1111 (2003)
8. Kim, D., Park, S., Kang, K., Lee, D.: Time-triggered wireless sensor network for feedback control. In: IEICE ELEX, vol. 4(21), pp. 640–644 (2007)
9. Rhee, I., Lee, J., Kim, J., Serpedin, E., Wu, Y.: Clock synchronization in wireless sensor networks: an overview. Sensors 9, 56–85 (2009)
10. Kopetz, H., Ochsenreiter, W.: Clock synchronization in distributed real-time system. IEEE Trans. Computers 36, 933–940 (1987)
11. Ganeriwal, S., Kumar, R., Srivastava, M.B.: Timing sync protocol for sensor network. In: Proc. 1st Conf. Embedded Networked Sensor System, pp. 138–149 (2003)
12. Sink, P.: Device networking 101: your first installation. Control Engineering 110 (May 2001)
13. Baronti, P., Pillai, P., Chook, V., Chessa, S., Gotta, A., Hu, Y.: Wireless sensor networks: a survey on the state of the art and the 802.15.4 and ZigBee standards. Computer Communications, 1655–1695 (2007)
14. Kim, D., Park, S., Kang, K., Lee, D.: A deterministic wireless network for feedback control based on IEEE802.15.4. In: Proc. 7th IFAC Conf. Fieldbuses & Networks in Industrial & Embedded Systems (FeT2007), Toulouse, France,

A Preliminary Empirical Analysis of Mobile Agent-Based P2P File Retrieval

Naoki Fukuta

Abstract. In this paper, I present a preliminary empirical analysis of P2P-based semantic file sharing and retrieval mechanism. The mechanism enables us to utilize private ontologies for flexible concept-oriented semantic searches without loss of privacy in processing semantic matching among private metadata of files and the requested semantic queries. The private ontologies are formed on a certain reference ontology with differential ontologies for personalization. In my approach, users can manage and annotate their files with their own private ontologies. Reference ontologies are used to find out semantically relevant files for the given queries that include semantic relations among existing files and the requested files. Mobile agent approach is applied for both implementing a system with less use of network bandwidth and coding it into a set of simple and small programs. I show the effectiveness of the use of private ontologies in metadata-based file retrieval. Also I show that the mobile agent approach has rather less overheads in execution time when the network latency is relatively high, while it is small enough even when the network is ideally fast.

1 Introduction

P2P(Peer to Peer) approach is a way to share data among peers with little control of central servers. In P2P approach, data are mainly managed by each peer and sufficient retrieval mechanisms are prepared to find out necessary data from peers[6]. For such purpose, DHT(Distributed hash table) and other mechanisms have been investigated for efficient and secure data retrieval[6]. However, there are further frontiers to improve such mechanisms when we need non-exact, flexible matching in the retrieval[14]. Furthermore, protecting unwanted reveals of secure data that includes protection of privacy for users is an important issue to be investigated[14][6].

Naoki Fukuta
Faculty of Informatics, Shizuoka University,
3 5 1 Johoku Hamamatsu Shizuoka 432-8011 Japan
e-mail: fukuta@cs.inf.shizuoka.ac.jp

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 143–154.
springerlink.com

Mobile agent is an approach that makes each software agent capable to move from an execution environment to another when such environments are not on the same computer but connected via a certain network that might be sometimes disconnected[15]. This approach is especially important to build P2P-based communication software. There are many approaches to implement better mobile agents and their platforms to realize faster execution and migration of agents, ease of programming, smaller footprint of agents and platforms, better security, etc[9].

P2P-based approaches are quite different from a traditional information sharing approach. For example, one traditional information sharing approach could be the use of a shared file server which stores common files for team members. Such file servers are often capable to prepare separate accounts for each member and to give proper access permissions for each file and folder. While file sharing policies should be shared for the project team members, it is also very difficult to enforce all members to follow the given policies when each member joined in many projects and each of which has different policies. Also in such case, many members might not have enough time to learn and review all policies. Furthermore, when such policies have formed in ad-hoc manner and they cannot be shared in all members, it makes difficult to find out and access appropriate information, especially in the final phases of the project. This also makes it difficult to find out which data are deprecated, inconsistent, and being removed. Solving these issues may cause communication overheads that cannot be neglected[6].

Mobile agent-based approach is especially effective to be applied to build a software system that supports teaming tasks that may cover many organizations including companies, universities, communities, etc[27][10][28]. On such teaming tasks, better information sharing is crucial to reach better results in the tasks[16]. For example, there are many types of information to be shared in there. Their granularity is varied, from fine-grained low data files or messages to tacit knowledge that cannot easily be formed into a document that has explicit descriptions of them. There are already many researches about information sharing on organization[16][17][7][12]. However, early discussions are for a static organization whose members may have different interests. Therefore, there are increasing discussions to share information among people who are in different organizations with different information sharing policies and the structures are dynamically changing[24][27].

In [8], I discussed about the basic idea and its implementability about P2P-based file retrieval mechanism when each user has private, personalized ontologies for their storage and annotation of the shared files. In this paper, I present a preliminary empirical analysis of P2P-based semantic file sharing and retrieval mechanism that enables us to use private ontologies for flexible concept-oriented semantic searches without loss of privacy in processing semantic matching among private metadata of files and the requested semantic queries. Mobile agent approach is applied for the purpose to implement a system with less use of network bandwidth and also to code it into simple programs. In this paper, I show the effectiveness of the use of private ontologies in metadata-based file retrieval. Also I show that the mobile agent approach has rather less overheads in execution time when the network latency is relatively high, while it is small enough even when the network is ideally fast.

2 Ontology-Based Data Retrieval

2.1 Search Approaches with Semantic Metadata

In many text retrieval approaches, traditional indexing approaches have been used that allows us to find out all documents that include a certain set of terms. Such approaches often have been used for Web search engines and other local data storage software. On the other hand, in P2P-based file retrieval, sometimes it is difficult to make proper index for the files before the search query has issued[20]. In this paper, since the discussions are about issues on P2P-based file retrieval, we consider a metadata-based semantic retrieval approach as an approach that does not need pre-indexing for the content of documents.

On retrieval approaches that are using semantic metadata, search targets are not the files which include a certain keyword in their filenames or texts. Rather, they try to retrieve the files by the relations among the target files with other related files which are associated with certain conceptual backgrounds[22]. For example, this enables us to find out a file that contain the original data about a certain figure that are used in a specific file. Such retrieval can be realized by preparing sufficient semantic metadata for the files to be retrieved[22][23]. Furthermore, such technologies can be extended for the use of service retrieval and automated compositions of them[4].

On P2P-based file retrievals, there are many granularities of targets, e.g., a peer, a set of files, and a part of file that might be encrypted[23]. While it is very difficult to cover all of them in a discussion within a single paper, I start the discussion in a case that seeks a set of files that can normally be managed by ordinary operating systems.

2.2 Ontology

Ontology based approaches are often discussed for a realization of flexible retrieval of data. In Gruber's definition, ontology is defined as an explicit process of conceptualization[11]. In recent works in the area, a simple taxonomy-like structure that captures conceptual hierarchies and relations is also considered as a (light-weight) ontology. Available ontologies can be found on the web by using Swoogle[2] and other ontology search services. Furthermore, conceptual networks in the Wordnet[18] can also be treated as an ontology. Also some automatically generated ontologies have been published that are mainly generated from Wikipedia or other large-scale media that contain numerous instances (called *individual*s in the ontology research field)[3][25]. Additionally, upper ontologies have also been proposed(e.g., SUMO[21], YATO[19], etc). These ontologies have been used for defining certain business concepts and also for combining their business processes[5][13]. Ontology description languages and frameworks have also been discussed and having growth in the use with OWL[1] and other description languages.

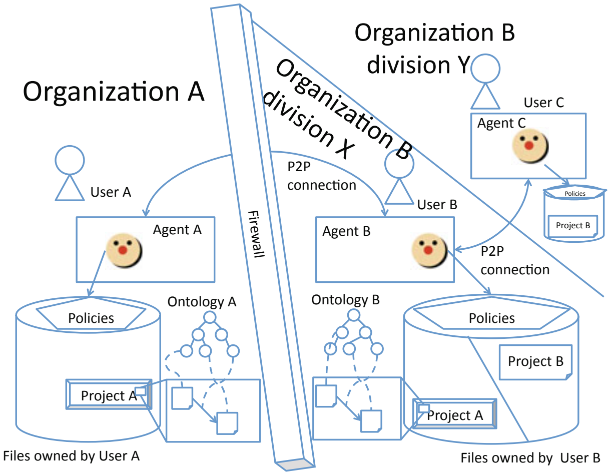


Fig. 1 An Overview of Prototype System

In this paper, ontologies that will have concepts, individuals, properties that denote relations among them are used in the discussion. This construction is mostly equivalent to the OWL-DL framework, I omit some discussions about strict inference mechanisms and restrictions but rather focus on building and using mappings among ontologies.

2.3 Personalized Ontologies: Their Use and Issues

By associating semantic relations from ontologies to files, it is demanded to realize flexible semantic retrievals on searching files[22]. However, when such files are owned and maintained by their owners, the given annotations for such files and even the used ontologies themselves deeply depend on their owners' thoughts and policies.

Furthermore, there are some annotations and associated part of ontologies that should be kept secret for others even for members of the same division who can share them. For example, a file F is associated to a project X, and an annotation for the file F has associated that relates to another secret project Y in the company, the file F might be shared to a member of X but the relations to Y and details about the project Y should be kept secret when the member of X is not the employee of the company. There are demands to keep such metadata or the files secret without complicated management of policies and permissions to such metadata and files while they seek effective use of them for better retrievals. There are also demands about good implementations to realize such mechanisms efficiently and securely[8]. In the next section, I describe a rough idea about them.

3 A Prototype System

3.1 Overview

I have implemented a prototype system that has some functions that can be useful to discuss about the issues presented in the earlier sessions. Figure1 show an overview of the system.

The system realizes a P2P-based file sharing mechanism among same project members. A peer is assigned to each project member and the sharing policies are managed and maintained by each software agent that are associated with each peer. Thus, the peers can store and manage registered files to be shared with other members.

This system also allows users to make annotations for files and relations among files. For example, users can add a relation to a file to another file that denotes the file is an older version of the another file to prevent unexpected or unwanted sharing of older files, and denote two files as the one be the original figure and the other be the document that uses the figure to find out the original figures to modify them.

3.2 Privatized Ontology

To make relations among concepts in different ontologies easily, the system prepared a reference ontology to relate such ontologies as a starting point of the discussion. Of course the reference ontology might not be shared to all users and multiple reference ontologies can be existed. However, in this case, at least one reference ontology is assumed to be associated the personalized ontologies and all reference ontologies are assumed to be accessed in open environment. This is a possible way to associate privatized ontologies by using one or several reference ontologies as a kind of scaffolding.

To this way, each peer has at least one reference ontology and associated private ontologies as differential ontologies to the reference ontology. Therefore, in this approach, each private ontology can be reproducible by the differential ontology and the associated reference ontology. This approach has another advantage that each associated software agent does not have to keep the whole reference ontology in them but just refer it when it needs. There are some big ontologies such as Wordnet. On the system, such ontologies are not fully loaded in each agent but just referred when necessary, and only some necessary parts of it is referred. In this model, the important part of the ontology is the ontology that denotes differences from its reference ontology.

3.3 File Retrieval and Transmission

Retrieval of files have been operated by the following way. First, the agent of peer that wants to retrieve files issues a query as a mobile agent which has differential ontology and moves to the peers which might have the target files. Then, the

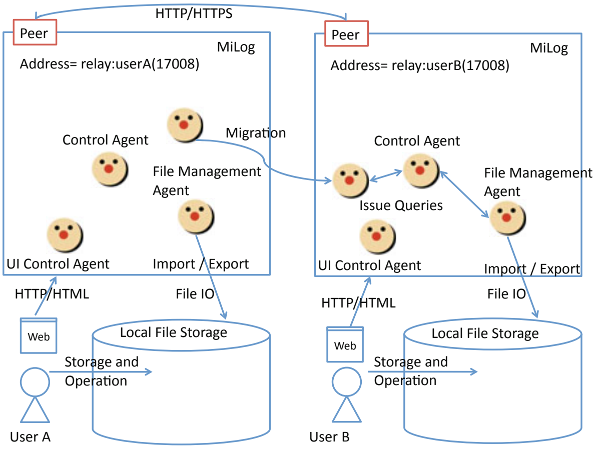


Fig. 2 The Structure of the Implemented System

destination peer allows the agents to access reference ontologies and the agent can reproduce the complete ontologies from them. And then the agent will try to find out the files that are relevant to the query. The files found and permitted by the peer can be transmitted to the original peer and the obtained files are automatically associated to metadata that are inherit from the original one when it should be so (i.e., a metadata that denotes the file cannot be shared to other project members, it should be inherited to the transmitted files).

3.4 Implementation

The system has been implemented by using logic-based mobile agent platform *MiLog*[9]. The abstract structure of the implemented system is shown in Figure2.

Each peer has realized as an instance of *MiLog* execution platform. Each peer has Web-based user interface that has been implemented by *MiPage* Web API built on *MiLog*. The *MiLog* execution platform runs on each user's computer and it can be manipulated by the Web-based UI even when the computer has no connection to the network such as *google gear*.

Each peer has individual address and communicates to other peers based on the address. Here, the addresses for peers should be given when they have possibilities to share files each other. Accesses to the peers have implemented by HTTP or HTTPS access and tunneling capability has also implemented when peers are in the firewalls that prevent direct access to the peers via a relay server on the Internet[26].

Each peer has three types of agents. One is to control interactions to the users, another one is to control the whole behavior of the system, and the other is the agent that moves to another peer and retrieve files. When the system receives a request

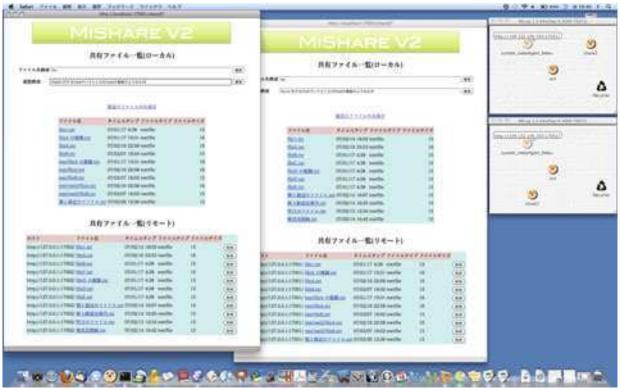

MISHARE V2
共有ファイル一覧(ローカル)
共有ファイル一覧(リモート)
MISHARE V2
共有ファイル一覧(ローカル)
共有ファイル一覧(リモート)

Fig. 3 An Implemented Web-based User Interface

for retrieval, a clone agent is produced for the file retrieval and then moved to, and perhaps walked through some peers, and then back to the peer with obtained files and their metadata.

Migration among peers has been implemented by mobility function implemented in *MiLog* platform. *MiLog* has implemented a very higher class of mobility, *strong mobility*, that allows all agents to move to other execution platforms while preserving the ones' whole internal execution states, including stacks and local variables in the code. On the *MiLog* implementation, communications for agent migration can be put on HTTP or HTTPS so it can easily be passed through proxy servers and tunneling through firewalls when there is a relay server that is sufficiently configured and has a global IP address[1].

Figure 3 shows the web-based user interface implemented on the prototype system[2]. Each browser window is associated to each peer and the window at right side is the monitoring window implemented in *MiLog* that indicates how many agents are in the platform and what they are doing. Since the user interface has built as Web-based one, it is possible to use with any software or extensions that can capture and use a part of Web pages, including the capability to convert a part of Web page into a small widget. Further implementation details are shown in [8].

4 Evaluation

This section describes a preliminary empirical evaluation to the proposed system. First of all, I evaluated the performance of search precision. Especially, I evaluated how the private ontology raise up the precision of the search. Since it is difficult

[1] Further extensions about this capability have been discussed in [26].

[2] Here, to easily capture the screenshot with multiple peers, these runtimes were running on the same computer. But they also work even when they are deployed in different computers connected by a certain network.

to scale the real ontologies for the evaluation, artificially generated ontologies are used in the evaluation. Here, I considered parameters that will approximate the real ontologies that are widely used in various purposes. To keep generality and simplicity of experiments, an ontology is constructed only by concept hierarchy that contains n_c of children in each nodeand has depth n_d. In the experiment, I prepared artificially generated ontologies by combinations of parameter $n_c \in \{2,5,10\}$ and $n_d \in \{3,4,5,6,10,13,16\}$. Note the number of concepts in SUMO is nearly equal to the case of depth and siblings $< n_c, n_d >$ as $< 13, 2 >$, $< 4 : 10 >$, and $< 6 : 5 >$. Also in Wordnet case, it is equal to the case that uses the depth and siblings as $< 16 : 2 >$, $< 5 : 10 >$, and $< 7 : 5 >$. Also, private ontologies are generated from the base ontology by applying the specified number of edit operators $n_e \in \{3,5,10\}$. In the experiment, I only used the edit operator that replaces its superclass. Then, a specified number n_e of virtual documents are prepared which have some conceptual tags that are automatically assigned to them based on reference ontology. In the experiment, the number of assigned tags for each document is at most n_t concepts. In each search process, a virtual document is selected for the search target and the search query is generated based on the tags assigned to the target document. To emulate natural annotations, the tags used for the query are randomly shifted to closer concepts in the user's private ontology. Also, the tags assigned to documents are also randomly shifted to closer concepts in the document owner's private ontology. Therefore, both each query and assigned tags are not the original ones. Then the system retrieves the target documents at the document owner's peer. The experiment compares the rank of target documents in the search results whether it is using private ontologies in search algorithm or not. To cover both sparse and dense document case, I prepared experiments with $n_f \in \{10, 100 1000\}$ depending on the purpose of the experiments. The parameters for generating ontologies and virtual documents are summarized in Table 1.

Table 1 Parameters to generate ontologies

n_d	$\{3,4,5,6,10,13,16\}$
n_c	$\{3,5,10\}$
n_e	$\{1,2,3,4\}$
n_t	$\{10,100,1000\}$

The similarity between a document and the conceptual tags in a query is calculated in the following formula.

$$\frac{\sum^{qc_i \in Q} \max_{vc_j \in V} sim(qc_i, vc_j)}{|Q|}$$

Here, $qc_i \in Q$ means the conceptual tags in the query Q, $vc_j \in V$ are conceptual tags assigned to the documents V, and $|Q|$ is the number of tags in the query, respectively. Although there are several appraoches to compute conceptual similarity $sim(c_1, c_2)$, to keep simplicity of the discussion, I use the depth difference i to the deepest shared upper concept. For instance, when c_j is a depth i of upper concept of c_{j+i}, $sim(c_j, c_{j+1}) = \gamma_i$, where γ_i was set to $\gamma_0 = 1, \gamma_1 = 0.75, \gamma_2 = 0.10, \gamma_j = 0 \;\; s.t. \;\; i > 2$.

Table2 shows the cases which have differences of ranks of retrieved documents when private ontology is considered. In this experiment, $n_t = 10$ is used so that the conditions could be harder for the approach. However, in all cases the ranks are better when private ontology is considered in the retrieval. When the size of ontology is large, due to the limited number of operation to generate private ontology, the difference to private ontology is relatively small. Therefore, there is no difference in the experiment in such case. The experiment shows that the approach works well even when the prepared ontology is small and documents are sparse in the peer.

Table 2 Search result differences with and without mobile agents

-n_d-n_c-n_e-n_t	with mobile agents	without mobile agents
-10-2-5-2	1st	9th
-3-10-5-1	1st	6th
-3-10-5-2	1st	5th

When the system uses mobile agent to retrieve documents, it may take some computational and transfer overheads compared to use simple remote querying. Below I compared overheads of using mobile agents in some situations. Note that the experiments have been done with a preliminary implementation that is not deeply optimized for better speed or lower data transfer.

The experiment conditions are equal to the previous experiments. Here, I used $n_f \in \{100, 1000\}$ and $n_c = 5$, respectively. The computer used for the experiment is a laptop computer running on MacOSX 10.6.4, with 8GB memory and 3.06GHz dual-core processor. For each peer, 256MB of memory is assigned for the Java VM. For experiments throughout firewalls, a relay host is used that is geographically located in approximately 800km from the experiment place.

Table 5 shows the comparison in time when two peers are locally communicated. Here, the shown values are the averages in 100 times of the experiments. Here, the results in “remote query” do not use private ontology so it shows the case when private ontologies are shared before querying and there are no computational overheads to retrieve with private ontologies. So the shown “remote-query” case is the ideal case when it does not use mobile agent-based retrieval. In the case of $n_f = 100$, mobile agent-based approach took 50msec of overheads compared to the remote query case. However, in $n_f = 1000$ the overhead is still there but only 160msec in 1976msec of total execution time. It can be said that when the number of documents are not too small, the overhead is relatively small to the overall execution time.

Table 4 shows the result on the situation that requires communications through firewalls to a distanced host. Here, the number of documents is $n_f = 1000$ and the results are average of 100 times of experiments. Since the mobile agent approach is robust for network latency, the overhead is far less than remote query approach and the value is almost equal to the case of local communication.

Table 3 Overheads of file retrieval within single computer

	$n_f = 100$	$n_f = 1000$
using mobile agents[msec]	264.15	1976.57
using remote query[msec]	211.42	1816.28

Table 4 Overheads of file retrieval through firewall

($n_f = 1000$)	through firewall[msec]	local[msec]
using mobile agents	1972.68	1976.57
using remote query	4428.72	1816.28

Table 5 shows communication overheads in the previous experimental condition. Here, although the transferred data are still larger in mobile agent approach, it is small enough to be used.

Table 5 Data transfers of file retrieval between different hosts

($n_f = 1000$)	sent[kbytes]	received[kbytes]
using mobile agents	21.88	20.05
using remote query	20.88	6.02

5 Conclusion

In this paper, I presented a preliminary empirical analysis of ontology-based file retrieval approach that uses mobile agents. I demonstrated that the use of private ontology was effective in metadata-based file retrieval. Furthermore, the overhead of mobile agent approach is rather small in response time when the communication has done through a firewall with relatively long distance / high latency networks and it is also small enough even in a local communication environment.

Acknowledgment. This work is partly supported by Grants-in-Aid of The Telecommunications Advancement Foundation(TAF), and Grants-in-Aid of Artificial Intelligence Research Promotion Foundation in Japan.

References

1. Owl web ontology language reference. W3C Recommendation (February 10, 2004), `http://www.w3.org/TR/owl-ref/`
2. Swoogle, `http://swoogle.umbc.edu/`
3. Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: Proc. 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference(ISWC2007 + ASWC2007), pp. 722–735 (2007)
4. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American, 35–43 (2001)
5. de Bruijn, J., Fensel, D., Keller, U., Lara, R.: Using the web service modeling ontology to enable semantic e-business. Commun. ACM 48(12), 43–47 (2005)

6. Cudré-Mauroux, P., Budura, A., Hauswirth, M., Aberer, K.: Picshark: mitigating metadata scarcity through large-scale p2p collaboration. The VLDB Journal 17(6), 1371–1384 (2008)
7. Fox, T.L., Spence, J.W.: The effect of decision style on the use of a project management tool: an empirical laboratory study. SIGMIS Database 36(2), 28–42 (2005)
8. Fukuta, N.: A mobile agent approach for flexible peer-to-peer file retrieval. In: Proc. of IEEE/ACIS International Conference on Computer and Information Science (ICIS 2010), pp. 599–604 (2010) doi:10.1109/ICIS.2010.83
9. Fukuta, N., Ito, T., Shintani, T.: A logic-based framework for mobile intelligent information agents. In: Poster Proc. of the Tenth International World Wide Web Conference (WWW 2010), pp. 58–59 (2001)
10. Gamma, E.: Agile, open source, distributed, and on-time: inside the eclipse development process. In: Proc. the 27th international conference on Software engineering (ICSE 2005), pp. 4–4 (2005)
11. Gruber, T.R.: A translation approach to portable ontologies. Knowledge Acquisition 5(2), 199–220 (1993)
12. Grudin, J.: Why cscw applications fail: problems in the design and evaluation of organizational interfaces. In: Proc. the 1988 ACM conference on Computer-supported cooperative work (CSCW 1988), pp. 85–93 (1988)
13. Haller, A., Oren, E., Kotinurmi, P.: An ontology for internal and external business processes. In: Proc. the 15th international conference on World Wide Web (WWW 2006), pp. 1055–1056 (2006)
14. Karnstedt, M., Sattler, K.-U., Hauswirth, M., Schmidt, R.: A dht-based infrastructure for ad-hoc integration and querying of semantic data. In: Proc. the 2008 international symposium on Database engineering & applications (IDEAS 2008), pp. 19–28 (2008)
15. Lange, D.B., Oshima, M.: Seven good reasons for mobile agents. Communications of the ACM 42(3), 88–89 (1999)
16. Malone, T.W., Grant, K.R., Turbak, F.A., Brobst, S.A., Cohen, M.D.: Intelligent information-sharing systems. Commun. ACM 30(5), 390–402 (1987)
17. Maqsood, M.e., Javed, T.: Practicum in software project management: an endeavor to effective and pragmatic software project management education. In: Proc. the the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering (ESEC-FSE 2007), pp. 471–480 (2007)
18. Miller, G.A.: WordNet: A lexical database for English. Communications of the ACM 38(11), 39–41 (1995)
19. Mizoguchi, R.: Yet another top-level ontology: Yato. In: Proc. of the Second Interdisciplinary Ontology Meeting, pp. 91–101 (2000)
20. Moulin, C., Lai, C.: Issues in semantic file sharing. In: Wegrzyn-Wolska, K.M., Szczepaniak, P.S. (eds.) Advances in Intelligent Web Mastering, vol. 43, pp. 242–247. Springer, Heidelberg (2007)
21. Niles, I., Pease, A.: Towards a standard upper ontology. In: Proc. the 2nd International Conference on Formal Ontology in Information Systems, FOIS-2001 (2001)
22. Siebert, M., Smits, P., Sauermann, L., Dengel, A.R.: Increasing search quality with the semantic desktop in proposal development. In: Reimer, U., Karagiannis, D. (eds.) PAKM 2006. LNCS (LNAI), vol. 4333, pp. 279–290. Springer, Heidelberg (2006)
23. Tran, N., Beydoun, G., Low, G.: Design of a peer-to-peer information sharing mas using mobmas (ontology-centric agent oriented methodology). In: Magyar, G., Knapp, G., Wojtkowski, W., Wojtkowski, W.G., Zupancic, J. (eds.) Advances in Information Systems Development, vol. 2, pp. 63–76. Springer, Heidelberg (2007)

24. Wang, X., Zhang, L., Xie, T., Anvik, J., Sun, J.: An approach to detecting duplicate bug reports using natural language and execution information. In: Proc. the 30th international conference on Software engineering (ICSE 2008), pp. 461–470 (2008)
25. Yamaguchi, T., Morita, T.: Building up a large ontology from wikipedia japan with infobox and category tree. In: Proc. the 3rd Interdisciplinary Ontology Meeting (InterOntology 2010), pp. 121–134. Keio University, Japan (2010)
26. Yamaya, T., Shintani, T., Ozono, T., Hiraoka, Y., Hattori, H., Ito, T., Fukuta, N., Umemura, K.: MiNet: Building ad-hoc peer-to-peer networks for information sharing based on mobile agents. In: Karagiannis, D., Reimer, U. (eds.) PAKM 2004. LNCS (LNAI), vol. 3336, pp. 59–70. Springer, Heidelberg (2004)
27. Yu, L., Ramaswamy, S.: Mining cvs repositories to understand open-source project developer roles. In: Proc. of the Fourth International Workshop on Mining Software Repositories(MSR 2007), p. 8 (2007)
28. Zhou, Y., Davis, J.: Open source software reliability model: an empirical approach. In: Proc. the fifth workshop on Open source software engineering(5-WOSSE), pp. 1–6 (2005)

An Approach to Sharing Business Process Models in Agile-Style Global Software Engineering

Takayuki Ito, Naoki Fukuta, and Mark Klein

Abstract. The globalization of information technology and the improvement of telecommunication facilities have facilitated software development business processes worldwide. Despite this increasingly popular trend, the initial expectations of the cost reductions of offshore outsourcing have not been realized. Many software development companies are facing difficulties caused by many hidden costs, including translation efforts in language gap, transition risks, learning needs, communication overheads, setup times, ramping up durations, scope creeps, etc. In this paper, we propose an approach to improving knowledge sharing in global software development. In addition, the trend of software development methodology has been changing to iterative and agile style from the classic waterfall model. In iterative and agile software developments, requirement specification, coding, testing, are mainly interested in a relatively short term. The rationale of this change came from the higher software quality and higher customer satisfaction. Iterative process can involve customers and developers into its software development mode carefully and deeply. One of the challenges is to apply an iterative or agile development model into global software developments. The main problem is such iterative and agile

Takayuki Ito
Dept. of Computer Science, School of Techno-Business Administration,
Nagoya Institute of Technology, Gokiso, Showa-ku, Nagoya 466-8555
e-mail: ito.takayuki@nitech.ac.jp

Mark Klein
Center for Collective Intelligence, MIT Sloan School of Management,
5 Cambridge Center, NE25-749A, Cambridge 02139, USA
e-mail: m_klein@mit.edu

Naoki Fukuta
Faculty of Computer Science, Shizuoka University,
Hamamatsu Shizuoka 4328011
e-mail: fukuta@cs.inf.shizuoka.ac.jp

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 155–170.
springerlink.com

process might need the tremendous number of communications and documents for collaboration because of iteration. Thus, software collaboration tools will be valuable for such situations in global software developments.

1 Introduction

Global offshore software development[1] has been receiving attention as a new trend in the global software engineering field. The globalization of information technology (IT) and the improvement of telecommunication facilities have made it possible to facilitate software development business processes worldwide. Countries such as India and China have been recognized as significant players in global offshore software development business. Despite this increasingly popular trend, the initial expectations of cost reduction of offshore outsourcing are not realized.

The main advantage of offshore software development has been cost reduction. The other benefit of offshore has been discussed in the literature[1]. As much as possible, software companies move development tasks offshore to countries with a highly educated and cheap workforce. In particular, the advancement of IT and telecommunication facilities is pushing software development offshore. So naturally in the global world, tasks and decision making are distributed around the world. Also, in Japan, offshoring is one of the most important software developing methodologies for surviving the competition of software development in the global world. However, based on several studies of general offshoring (including data entry, customer service, etc.), half of the organizations failed to generate the expected benefits[1].

The followings are the popular reasons for failures[1]. First they do not spend time evaluating which processes they should offshore and which they should not. Second most organizations do not consider all the risks that accompany offshoring. Third most companies do not realize that offshoring is no longer an all-or-nothing choice. Rather they have continuum of the other options. This analysis[1] also pointed out that some smart companies have gained strategic advantage by offshoring processes. Since R&D and product design have high operational risk compared with other simple tasks[1], offshoring software development should also have high operational risk. This is because much communication is necessary, and some implicit knowledge, concept, and information should be shared between both companies. Sharing precise information and knowledge about the requirement specification on what to make and how to make it is the key for better offshoring.

When considering offshore software development processes, all players must think the software development model, which is adopted by the other countries. In the Paper[3], Japanese and United States tend to adopt semi-waterfall model. But, India tends to adopt waterfall and formalized development model compared with the other countries. A typical offshore software development process can be described as follows[4]: For the entire project period, a small developer team ”invades” the client’s place and handles system integration, installation, and testing through direct communication with the client. Initial requirements are usually determined at the client site, and more detailed requirement specifications are conducted offshore. Next, after the project leader and senior designers assemble the core team, development begins. Once the software is ready, it is shipped to the on-site members, who integrate the components of the system and carry out acceptance testing. The interface between the client and the offshore team is managed using a variety of mechanisms such as information requests, resolution of open issues, changing specifications, and status review video / tele-conferences. Even though a small team has ensconced itself at the client site, communication between the client and break developer sites becomes difficult due to the distance and time differences.

In order to share more precise information and knowledge on the requirement specification for software development, we have to resolve the main two issues: (1) enough communication, and (2) flexible development process. In (1), players should have enough communication to reach clear understanding with each other. In (2) as the current research result says waterfall model often fail to meet with the customer requirements even if they are in same country. For offshoring, flexible development process, like agile or iterative development processes, should be adopted for sharing and meeting the precise specification. In iterative and agile software developments, requirement specification, coding, testing, are mainly interested in a relatively short term. The rationale of this change came from the higher software quality and higher customer satisfaction. Iterative process can involve customers and developers into its software development mode carefully and deeply. One of the challenges is to apply an iterative or agile development model into global software developments[2]. The main problem is such iterative and agile process might need the tremendous number of communications and documents for collaboration because of iteration. Thus, software collaboration tools will be valuable for such situations in global software developments.

Global software engineering can be seen as a global collaborative design. Collaborative design has been widely studied for designing buildings, airplanes, cars, etc. These collaborative design concepts and methodologies can be applied to the global software design fields[8].

We first need to represent business processes so that developers and customers can share clearly what they will make. In this paper, we propose a new business process representation, which can represent task transition, resource dependency (fit and share), rationale, and modality, for software specification. Transition is required for realizing workflows in business processes. The most well known work on representing, sharing, and organizing business processes is MIT Process Handbook[10]. Recently most ontology description languages and XML related markup languages have possibility to represent business processes. But, the advantage of the MIT Process Handbook is its simplicity and its scale. It just employs the simple concepts like specifications and dependencies including flow, fit, and share. These concepts are widely well acceptable for real world. We have been inspired by its idea of resource dependency (fit and share), which can represent dependencies to resources.

2 A New Methodology for Sharing Business Processes for Agile Software Development

2.1 The Outline of the Methodology

In global software development process, it is really hard to clearly know the specification of the software product. In particular, if adopting agile process for development, such the shared specification could be changed, modified, or improved many times. If there is a large language barrier, it becomes the endless story to share clear understanding on the correct specification between developers and the customer. In this sense also, waterfall model tend to be failed because in general the customer's demand might change after fixing the specification.

In this paper, we propose a new methodology to sharing business process specification for global agile software development process. The challenge is to make a methodology for externalizing and sharing the business process in a simple way. In order to externalize and share the business process, we utilize the concepts of transition, sharing and fitting, rationale, and the modality, and represent and share them in a XML based language.

Our methodology focuses on specification of a business process, which will be implemented as a software program. In iterative or agile software development process, a specification will be changed, modified and improved many times. Such modifications are quite valuable to increase the quality of software products since they will include the customer's detailed demands. In order to realize such modifications on a specification, our methodology provides a state-transition model as a intuitive representation of a business process.

We assume our methodology will be employed in global software developments, which utilize Agile style program development. For these reason, a web-based system, which supports our methodology should be prepared. In Agile software development process, developers and customers first try to make a overall plan. Then, they will enter into a cycle of requirements specification, design, coding, and testing. In the process of requirements specification, in our methodology, developers and customers will share the business process, which will be implemented as a product.

Overall, the characteristics of our methodology are described as follows:

- It explicitly represents a business process as a state transition.
- It provides an easy method for changing, modifying, and improving the business process.
- OWL representation provides semantics, which resolve a burden to share meaning of words. This might resolve language barrier.

2.2 Business Process Representation

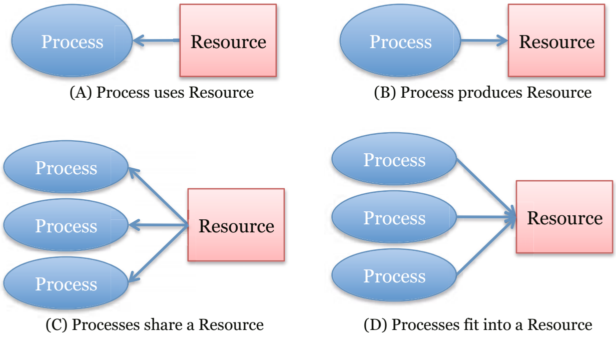
We propose an expressive model, which can represent a business process, which is the target to be produced as a software. The model includes some important elements : transition, fit& share dependency, modality and rationale. These elements are needed to represent business processes, which will be designed and implemented as software.

Transition (Flow)

This element shows transitions or flows among processes.

Process and Resource Dependency

Tasks and Resources hare dependency, which have been inspired from the concept of fit and share dependencies in the MIT Process handbook, can be captured in our model1. In Figure1, we present 4 types of relations: (A) A process uses a resource, (B) A process produce a resource, (C) processes share a resource, and (D) processes produces a resource. For example, when several workers share one scissor, this is the relation (C). When managers schedule a meeting, this relation is (D).



(A) Process uses Resource

(B) Process produces Resource

(C) Processes share a Resource

(D) Processes fit into a Resource

Fig. 1 Process and Resource dependency

Modality

It must be very helpful to add ”modalities” to a business process representation, if we are using it as a requirements specification language. A modality is a piece of type information (”must”, ”can”, or ”must not”) that is added to relationships in the process model. For example, in Figure 2, the above process must use papers as resources. The middle process must-not output toxic waste. The bottom process can output water.

Rationale

When we develop requirements collaboratively with a distributed team, it can be very helpful to capture the rationale for the requirements in the same formalism that we capture the requirements themselves[8]. In our system, the rationale is represented in argumentation-graph based language, which has been inspired from DRCS

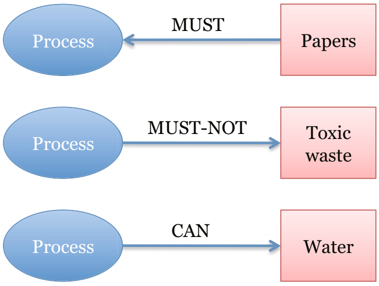


Fig. 2 An Example of Modality

(the Design Rational Capturing System)[8]. DRCS captures the requirements, decision problems and options using a structured language with explicit semantics. We provide a simple tool for capturing rationales for a part of process design.

Figure 3 shows an example of rationales for building a business process. The customers and the developers cooperatively add the rationales, which support the reason why the processes and the resources are required. These rationales are valuable for sharing and reusing the business process. In Figure 3, when the customer noticed ”Estimation” is required before ”order”, then the customer put the ”Estimation” process with the rationale of it. Then, the customer also noticed that ”Inquiry” is also needed before ”Estimation”. The customer put ”Inquiry” process with the rationale of it.

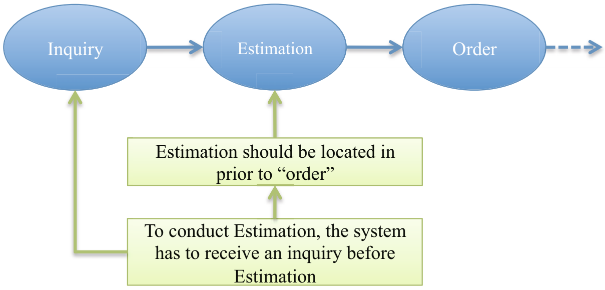

Inquiry
Estimation
Order
Estimation should be located in prior to “order”
To conduct Estimation, the system has to receive an inquiry before Estimation

Fig. 3 An Example of Rationale

2.3 *Graph-Based Mark-Up Language for Representing Business Process*

In order to share the business process among remotely distributed people, we provide a mark-up language for make business processes be computationally readable. Based on the business process represented in our mark-up language, users can have their own user interface for showing them. Also, the mark-up language enables users to compare the business process and search similar business processes automatically.

Our mark-up language is based on a graph representation since any business process in our model can be represented in a graph. Thus the mark-up language basically provides nodes and edges which has special labels. Actually, the current mark-up language has been inspired from RuleML and a lot of XML-based graph markup languages(for example, [5]). RuleML, the Rule Markup Language, is a markup language for publishing and sharing rule bases on the World Wide Web[7]. The current language example is intuitive.

Transition

We can represent a simple transition shown in (A) in Figure 1 in our graph-based mark-up language.

```
<node>
 <id>1</id>
 <name>Process</name>
 <type>process</type>
</node>
<node>
 <id>2</id>
 <name>Resource</name>
  <type>resource</type>
</node>
<link>
 <id>3</id>
 <fromNode>2</fromNode>
 <toNode>1</toNode>
 <type>transition</type>
</link>
```

Fit and Share Dependency

We can represent a simple share relation shown in (C) in Figure 1 in our graph-based mark-up language. We allow arguments-based expression of each tags.

```
<node id="1" name="Process" type="process"/>
<node id="2" name="Process" type="process"/>
<node id="3" name="Process" type="process"/>
<node id="4" name="Resource" type="resource"/>
<link id="5" fromNode="4" toNode="1" type="share"/>
<link id="6" fromNode="4" toNode="2" type="share"/>
<link id="7" fromNode="4" toNode="3" type="share"/>
```

Modality

Modalities are represented as a link with the label about modality. We can represent a simple MUST relation shown in the top of Figure 2 in our graph-based mark-up language.

```
<node id="1" name="Process" type="process"/>
<node id="2" name="Papers" type="resource"/>
<link id="3" fromNode="2" toNode="1" type="modality" name="MUST"/>
```

Rationale

Rationales are all represented in a graph style in our model. Thus, We can represent a simple rationales shown in Figure 3 in our graph-based mark-up language.

```
<node id="1" name="Inquiry" type="process"/>
<node id="2" name="Estimation" type="process"/>
<node id="3" name="Order" type="process"/>
<link id="4" fromNode="1" toNode="2" type="transition"/>
<link id="5" fromNode="2" toNode="3" type="transition"/>
<link id="6" fromNode="3" toNode="4" type="transition"/>
...
<node id="7" label="Estimation should be located in
  prior to Order" type="rationale"/>
<node id="8" label="To conduct Estimation, the system
  has to receive an inquiry before Estimation" type="rationale"/>
<link id="9" fromNode="7" toNode="2" type="rationale-link"/>
<link id="10" fromNode="8" toNode="7" type="rationale-link"/>
<link id="11" fromNode="8" toNode="1" type="rationale-link"/>
```

3 Example

This is an example to create a model of activities of a small company, which provides piecework at home, which make the software design be explicit. The following are basic parts in their business process.

1. Inquiry from the customer: a customer makes an inquiry to the company on a specific task (like creating 10,000 small calendars).
2. Estimation by the company: The company estimates the cost on the specific task. (like one calendar is 20 cents).
3. Order from the customer: Based on the estimation, the customer places the order.
4. Materials from the customer: The customer gives materials for producing the products (like papers, wood stands, etc.)
5. Assign tasks to workers (e.g., each worker will complete 100-300 calendars in a week).
6. Check completed tasks (checking the completed products and the failed products).
7. Check overall completed tasks (checking the number of all products).
8. Deliver to the customer
9. Payment by the customer

These processes are also divided into sub-processes (and also be decomposed into sub-sub-processes). Also, they might be integrated to more abstract level processes.

Modeling Transition

First we show ho these process can be represented in a transition diagram (Figure4). In the real cases, it is not necessary to create a transition model first. For simplicity of explanation in this paper, we show modeling a transition first.

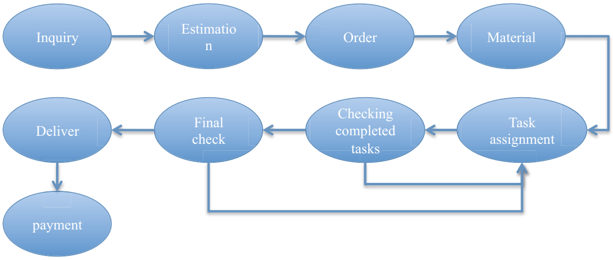


Fig. 4 An Example of A Transition Model : The entire model

As an example, when starting making a requirement specification, participants could have the simplest shared model (Figure 5). This is one of the top abstraction level models of production business. Many production businesses can be represented in this simple model and decomposed them into more specific processes. This concept is very suitable for sharing business processes in a repository.

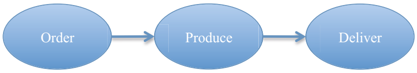


Fig. 5 An Example of A Transition Model : The first abstract model

Actually, Order includes Inquiry, Estimation, and Order. Produce includes Material, Task assignment, Checking completed tasks, and Final check. Deliver includes Deliver and Payment. As the first step, to view widely and share the entire process, they should have the simplest entire process.

Then they will try to make more specific specifications based on the simplest model. They need to communicate with different languages. For managing such communications with different languages, using ontology like OWL[13] is one approach. In the ontology, the semantics of Order, Produce, and Deliver were defined in each language by the experts on both of the language. This reduces work load and real cost for translating each communication.

Order could be decomposed into Inquiry, Estimation, and Order (Figure 6). This decomposition might be done with discussion among developers and customers. The representations in RuleML or Prolog will be automatically re-written by the software program based on what they changed in the web interface. While decomposing into sub processes, the ontology also should be re-written by an expert.

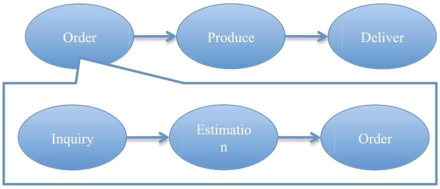


Fig. 6 An Example of A Transition Model : Decomposing Order to subprocesses

Also, Produce could be decomposed into Material, Task Assignment, and Checking completed tasks (Figure 7). Because this is a transition model, there is a backward transition between Task assignment and Checking completed tasks. Also, this decomposition does not match the entire process we showed at first in this example. But, in this story, after some agile processes happen, the customer will make a change on their demand.

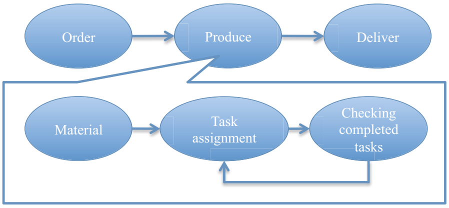


Fig. 7 An Example of A Transition Model : Decomposing Produce to subprocesses

Further Deliver could be also decomposed into Deliver and Payment (Figure 8).

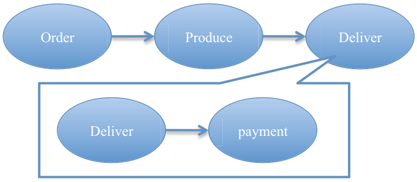


Fig. 8 An Example of A Transition Model : Decomposing Deliver to subprocesses

Now, the entire process can be described in Figure 9.

This is a very basic process. Based on this basic process, developer could start their design phase, or of course, it is also possible to have deeper discussion on specialization.

Here, as an example, after moving to the design phase, the customer changes Produce process like ”there need to be Final Checking process.” Then, they can change the specification as like as Figure 10.

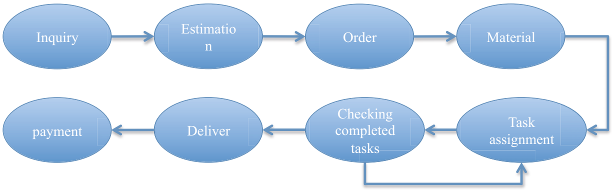


Fig. 9 An Example of A Transition Model : A Temporally Completed Business Process

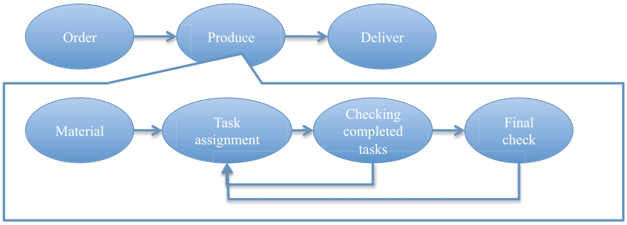


Fig. 10 An Example of A Transition Model : Updating the Business Process

The change is very easy on the graph representation. However, the important point is there might be a lot of sub effects when some small parts are changed in the software process. When considering waterfall model, the specification is fixed when the design phase started. Thus it is difficult to change the parts in the specification even if it is a small change because it is also difficult to judge the change is really small or not after completed the entire specifications. On the other hand, our proposed method allows participants to return back to the previous process. In order to realize such iterative and agile style in the global development, we believe we need a methodology and software, which support modifying, changing, and improving procedures on the business processes as we mentioned here.

Modeling Resource Relation

Figure 11 shows an example of modeling resource relations. For checking processes, which include Checking completed tasks, and Final check, a courting tool, which is for counting the number of completed products, is required. Thus these processes need to share this resource. Also, for Delivery process, it is required to have cars or motorbikes. These are required because people need to carry the products to the customer.

Modeling Modality

Figure 12 shows an example of modeling modality. When modeling resource relation above, we assumed cars and motor bikes are required for delivery. By using modality we can represent the customer's ambiguous demands, like ”if a management module for motor bikes can be included in the system, then add it. But cars

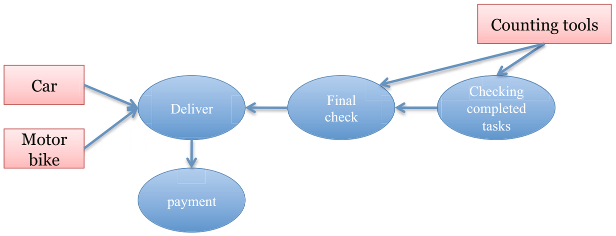


Fig. 11 An Example of Modeling Resource Relation

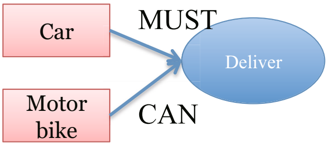


Fig. 12 An Example of Modeling Modality

must be managed in the system". In this case, we can represent this demand in Figure 12. Cars are linked as MUST. Motor bikes are linked as CAN.

Describing the Rationale

The rationales are added to the business process graph. In this example, as we explained in the previous section, the rationale for having Estimation and Inquiry has been added. Also, the rationale for having the Final Check process is described as "Final check was put by the customer's demand", and also "The entire final check

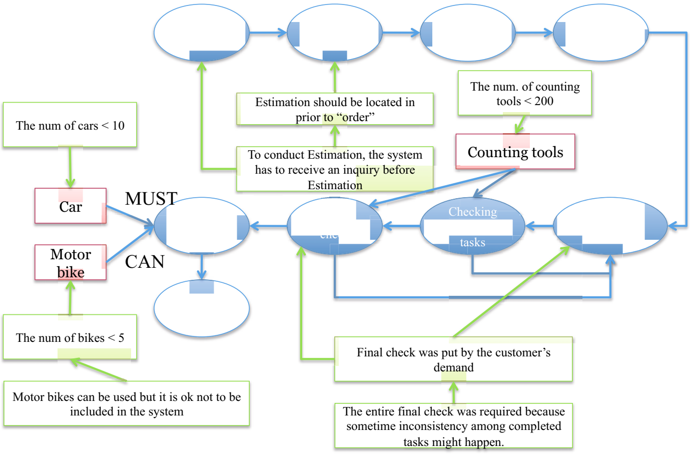

Estimation should be located in prior to "order"
The num. of counting tools < 200
The num of cars < 10
Counting tools
To conduct Estimation, the system has to receive an inquiry before Estimation
Car
MUST
Checking
che
tasks
Motor bike
CAN
The num of bikes < 5
Final check was put by the customer's demand
Motor bikes can be used but it is ok not to be included in the system
The entire final check was required because sometime inconsistency among completed tasks might happen.

Fig. 13 An Example of the Entire Business Process

was required because sometime inconsistency among completed tasks might happen." These rationale information are important for developing software interface and business logics, reusing this business process and update or modify it.

Figure 13 shows an example of the entire business process completed. Some more rationales, which explain why cars are MUST and why motorbikes are CAN, have been added. Rationales can be used like as the condition for design. One rationale shows "the num. of cars $<$ 10," means the number of cars is less than 10.

4 Example of Business Process Retrieval

Let us show an example of a business process retrieval in search and recommendation. When there is a business process with rationales in the repository shown in Figure 14 (This is the same as Figure 3). The followings are the XML-based representation.

```
<node id="1" name="Inquiry" type="process"/>
<node id="2" name="Estimation" type="process"/>
<node id="3" name="Order" type="process"/>
<link id="4" fromNode="1" toNode="2" type="transition"/>
<link id="5" fromNode="2" toNode="3" type="transition"/>
<link id="6" fromNode="3" toNode="4" type="transition"/>
...
<node id="7" label="Estimation should be located in
  prior to Order" type="rationale"/>
<node id="8" label="To conduct Estimation, the system
  has to receive an inquiry before Estimation" type="rationale"/>
<link id="9" fromNode="7" toNode="2" type="rationale-link"/>
<link id="10" fromNode="8" toNode="7" type="rationale-link"/>
<link id="11" fromNode="8" toNode="1" type="rationale-link"/>
```

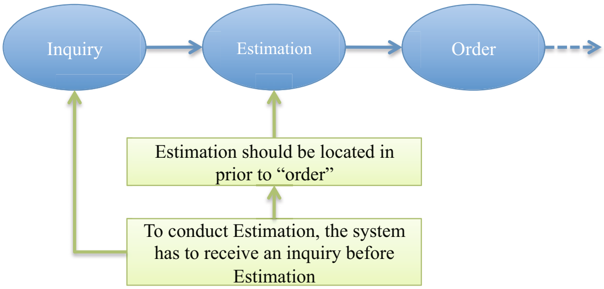

Inquiry
Estimation
Order
Estimation should be located in prior to "order"
To conduct Estimation, the system has to receive an inquiry before Estimation

Fig. 14 An Example of Retrieval

Let assume that a user has a query shown in Figure 15. This query actually can be written down as follows:

```
<node id="1" name="Estimation" type="process"/>
<node id="2" name="Order" type="process"/>
<link id="3" fromNode="1" toNode="2" type="transition"/>
<link id="4" fromNode="2" toNode="3" type="transition"/>
...
<node id="5" label="Estimation is needed before Order" type="rationale"/>
<link id="6" fromNode="5" toNode="1" type="rationale-link"/>
```

Our system will measure the similarity between the business process in Figure 14 and the query in Figure 15 by counting how many nodes, links, and types are matched. In this case, two process nodes, one rationale node, two transition links, and one rational-link are matched. It can be said that all elements in the query can be matched as a graph structure. Thus, in this case, the system can return the business process in Figure 14 as a response of the query. This is a very simple example, and we are able to introduce the threshold values, which define the degree of similarity among the business processes. The details of the similarity measure is out of the scope of this paper, and will be included in our future work.

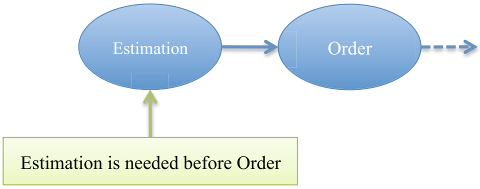


Fig. 15 An Example of Retrieval : A Query

5 Discussion and Related Work

Goal oriented software design tools have been studied very successively. The paper[12] has proposed the model so that people can manage frequent requirement changes by using goal-oriented tree like model. They present two topics related to change management of goal graphs; 1) version control of goal graphs and 2) impact analysis on a goal graph when its modifications occur.

There have been several languages and frameworks for representing business processes (for example, BPML (Business Process Modeling Language) and its modeling language BPMN(Business Process Modeling Notation)[6]). These languages and frameworks have very generic purpose, and they might be utilized in software developments, also. Our focus is to make a simple framework and methodology for modeling business processes in iterative and agile-style software developments.

The goal oriented analysis methods[9] [11] have been employed for eliciting requirements. In goal-oriented analysis, the users' abstract goals to be achieved are decomposed into more concrete sub-goals, and requirements analysts produce a requirements specification documents based on this analysis. In particular, the paper[11] proposed to define enactive rules, which form the basis of a software environment to guide the requirements elicitation process through interleaved goal modeling and scenario authoring. The focus of the paper was on the discovery of goals from scenarios.

6 Conclusion and Future Work

The globalization of information technology and the improvement of telecommunication facilities have facilitated software development business processes worldwide. Despite this increasingly popular trend, the initial expectations of the cost reductions of offshore outsourcing have not been realized. Many software development companies are facing difficulties caused by many hidden costs, including translation efforts in language gap, transition risks, learning needs, communication overheads, setup times, ramping up durations, scope creeps, etc.

In this paper, we proposed an approach to improving knowledge sharing in global software development. In particular, we discussed on a new methodology for requirement specification in agile software development processes.Our methodology has been inspired by MIT Process Handbook, and utilizes its concept on well-structured classification of business processes, which adopts the object-oriented programming like model. In particular, the hierarchical concept of abstraction and specialization of processes are adopted in our methodology.

References

1. Aron, R., Sigh, J.V.: Getting offshoring right. Harvard Business Review, 135–142 (2005)
2. Cusumano, M.A.: Managing software development in globally distributed teams. Communications of The ACM 51(2), 15–17 (2008)
3. Cusumano, M.A., MacCormack, A., Kemerer, C.F., Crandall, W.: Software development worldwide: The state of the practice. IEEE Software, 2–8 (2003)
4. Gopal, A., Mukhopadhyay, T., Krishnan, M.S.: The role of software process and communication in offshore software development. Communications of The ACM 45(4), 193–200 (2002)
5. Herman, I., Marshall, M.: Graphxml - an xml based graph interchange format (2000), `http://ftp.cwi.nl/CWIreports/INS/INS-R0009.pdf`
6. Business Process Management Initiative, Bpmi.org (2008), `http://www.bpmi.org/`
7. The RuleML Initiative, Ruleml tutorial (2005), `http://www.ruleml.org/papers/tutorial-ruleml-20050513.html`
8. Klein, M.: Capturing design rationale in concurrent engineering teams. IEEE Computer. Special Issue on Computer Support for Concurrent Engineering 26(1), 39–47 (1993)
9. van Lamsweerde, A.: Goal-oriented requirements engineering: A guided tour. In: RE, pp. 249–262 (2001)
10. Malone, T.W.: The Future of Work: How the New Order of Business Will Shape Your Organization, Your Management Stype, and Your Life. Harvard Business School Press (2004)
11. Rolland, C., Roll, C., Souveyet, C., Souveyet, C., Ben Achour, C., Ben Achour, C.: Guiding goal modeling using scenarios. IEEE Transactions on Software Engineering 24, 1055–1071 (1998)

12. Tanabe, D., Uno, K., Akemine, K., Yoshikawa, T., Kaiya, H., Saeki, M.: Supporting requirements change management in goal oriented analysis. In: Proceedings of the 2008 16th IEEE International Requirements Engineering Conference RE 2008, pp. 3–12. IEEE Computer Society Press, Los Alamitos (2008)
13. W3C: Owl web ontology language guide (2004), http://www.w3.org/TR/owl-guide/

Modeling for Gesture Set Design toward Realizing Effective Human-Vehicle Interface

Cheoljong Yang, Jongsung Yoon, Jounghoon Beh, and Hanseok Ko

Abstract. Intuitive driver-to-vehicle interface is highly desirable as we experience rapid increase of vehicle device complexity in modern day automobile. This paper addresses the gesture mode of interface and proposes an effective gesture language set capable of providing automotive control via hand gesture as natural but safe human-vehicle interface. Gesture language set is designed based on practical motions of single hand gesture. Proposed language set is optimized for in-vehicle imaging environment. Feature mapping for recognition is achieved using hidden Markov model which effectively captures the hand motion descriptors. Representative experimental results indicate that the recognition performance of proposed language set is over 99%, which makes it promising for real vehicle application.

Keywords: language set, gesture recognition, driver-vehicle interface, HMM.

1 Introduction

In the last decade, research activity on vision-based driver assistance systems (VDAS) has been proliferating. Most VDAS efforts, however, are focused on vehicle safety features such as automatic lane detection, pedestrian protection, etc [1-3]. These new attributes have gradually moved to include the features that provide driver's convenience such as automatic parking, head-up display capable of overlaying the augment reality objects, etc [4-6]. A change of VDAS features is currently due as we experience further rise of in-vehicle device complexity promulgated by new devices such as navigation system connected with internet, and integrated vehicle control system, etc. Since the display of in-vehicle device is not all located with the driver's line of sight, it is difficult to conveniently control the complex in-vehicle devices while driving. In this case, a gesture based control mechanism may provide as convenient interface to driver.

Cheoljong Yang · Jongsung Yoon · Hanseok Ko
Vision Information Processing, Korea University, Seoul, Korea
e-mail: {cjyang,jsyoon,hsko} @ispl.korea.ac.kr

Jounghoon Beh
Institute for Advanced Computer Studies, University of Maryland,
College park, MD 20742, USA
e-mail: jhbeh@umics.umd.edu

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 171–180.
springerlink.com

Many studies have been conducted on the subject of developing gesture recognition systems [7-12]. For example, sign language recognition is known as one of the most rewarding and successful applications of gesture recognition [8, 13]. Since driving requires driver's constant attention to driving, it is desirable for the driver-to-vehicle interface to be simple and easy to learn, while not imposing driver distraction. We also note that gestures in vehicle should be limited to just one hand motions because the other hand should be on the steering wheel. For this reason, the sign language that includes both hand motion and pose is not considered and they are simply too excessive with respect to complexity for in-vehicle device control. On the other hand, hand static pose is not enough to express a sufficient number of interface commands. In this paper, we propose a gesture language set for in-vehicle device. The complexity of hand gesture in the envisioned language set stands in the middle between sign language and static hand pose. Thus we choose hand gesture set that consists of 14 motions for meaningful motion.

Many gesture recognition methods have been proposed in the past. But most of all, Hidden Markov Model (HMM) is known advantageous to represent spatio-temporal property [7-11]. As hand signal consists of continuous motion in sequential time, hand gesture motions can be effectively mapped by HMM. In general, the issue of HMM-based recognizer is discernability, e.g. how to distinguish well between real gestures. To accurately distinguish gestures in continuous flow, there is a need to minimize meaningless motions between meaningful gestures. In essence, the starting point and ending point of an intended gesture should be predefined. We have devised a convenient way of achieving this by fixing a common point so that every hand gesture returns to its original position at the end of a gesture trajectory.

This paper is organized as follows. In Section 2, an overall scenario of a gesture control system in-vehicle device is described. In Section 3, an efficient language set of gesture command suitable for human-vehicle interface is proposed. Next, dynamic model and observation model of gesture are described in Section 5. The experimental results are discussed in section 6. In 7, we provide our concluding remark of this paper.

2 Gesture Control as Human-Vehicle Interface

In order to develop a language set suitable for as an in-vehicle device, we constructed a camera arrangement in car as shown in Fig 1 (a). A camera is located next to room mirror where it can capture driver's right hand well as depicted in Fig 1 (b). If the camera acquires image, the hand region is separated from background and the system is expected to transfer the trajectories of hand coordinates as motion descriptors. Then the recognition module performs classification of inputs in accordance to the gesture set constructed through HMM-based training. In the following section, language sets suitable for in-vehicle device are described.

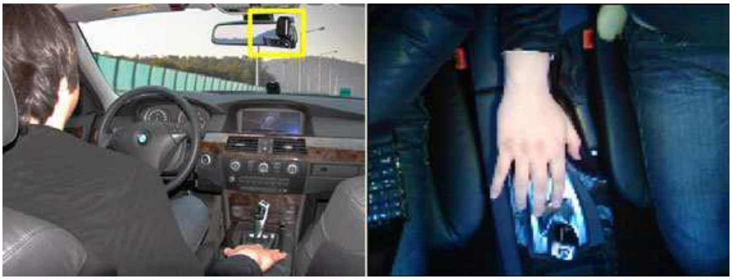


Fig. 1 (a) Camera arrangement of hand gesture control system and (b) hand image from the camera

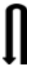
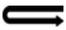
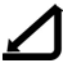
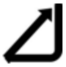
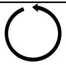
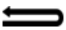
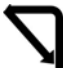
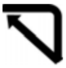
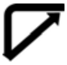
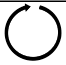
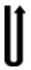
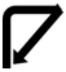
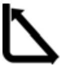
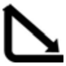
3 Medium Level Complexity Gesture Set for Driver Safety

If the gesture set is too complex for drivers to learn and perform during driving, drivers may have hard time learning the gestures and simultaneously become distracted from safe driving. To avoid too complex gesture sets for suitable human-to-vehicle interface device, we propose a gesture set consisting of simple motions and yet enough to express all of the desirable commands in vehicle situations. We analyzed American Sign Language Data Base provided from Boston University as well as those hand signal set of official referee and crane operator [13-15]. It can be seen through the database that most gesture motions used by practical human actions are made of the cardinal direction strokes (e.g. 0°, 90°, 180°, 270°). Based on this information on cardinal direction strokes, we design a unit gesture (UG) set for in-vehicle device as shown in Table 1. Note that diagonal direction and circular strokes are added in the Table for capturing the diversity of strokes. In addition, to minimize the trans-motions between valid hand gestures' starting and ending point are fixed to identical position.

To apply the gesture set to in-vehicle environment, each UGs described in Table 1 is mapped to a specific command as shown in Table 2. Each command is categorized as either "action" or "object".

A full gesture command can be constructed by a sequential arrangement of UG. The order of arrangement, for example, is determined depending on the category of

Table 1 Unit Gesture (UG) set for in-vehicle device

Idx	Gesture	Idx	Gesture	Idx	Gesture	Idx	Gesture	Idx	Gesture
UG 01		UG 04		UG 07		UG 10		UG 13	
UG 02		UG 05		UG 08		UG 11		UG 14	
UG 03		UG 06		UG 09		UG 12			

UG. We propose a language model based on a two step word network; {Object – Action}. Then the number of possible commands is 48(6X8) for in vehicle interface. Fig 2 shows an example of “Volume + Up” that is consisted of ‘UG07’ and ‘UG01’.

Table 2 Command list of each UG

Index	Meaning	Category	Index	Meaning	Category
UG01	Up/Enlarge	Action	UG08	Temperature	Object
UG02	Right/Next	Action	UG09	Right-side mirror	Object
UG03	Down/ Reduce	Action	UG10	Left-side mirror	Object
UG04	Left/ Previous	Action	UG11	Speech Recognition	Object
UG05	Navigation	Object	UG12	Radio player	Object
UG06	Music Player	Object	UG13	On	Action
UG07	Volume	Object	UG14	Off	Action

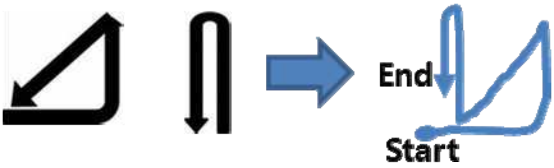


Fig. 2 Two step gesture of “Volume up” command

4 Dynamics of Hand Gesture Motion

Hand gesture is considered as a physical system and thus can be modeled by physical dynamics [16]. According to the Newtonian physics, the equation of motion can be expressed in the following form,

$$\ddot{x} = M^{-1} \cdot F \tag{1}$$

where ***M*** is the mass matrix, ***x*** is position of the object and **F** is the vector of forces. Considering moving object is affected by two types of force; applied force and constrained force, the expression can be modified by the form as follows:

$$\ddot{x} = M^{-1} \cdot (F_a + f_c(x(t), t)) \tag{2}$$

where F_a is the vector of applied force and $f_c(x(t), t)$ is the vector of constrained force dependent on position ***x*** and time *t*. Note that the constrained force does not add energy. So it can be shown that constrained forces should lie in the null space complement of the constraint Jacobian:

$$f_c(x(t),t) = \eta \frac{\partial \dot{f}_c}{\partial x} \tag{3}$$

Substituting Eqn(3) into Eqn(2), we have a system of linear equations with only the vector of unknown, η.

$$-\left[\frac{\partial f_c}{\partial x}^T M^{-1} \frac{\partial f_c}{\partial x}\right]\eta = \frac{\partial f_c}{\partial x}^T M^{-1} F_a + \frac{\partial \dot{f}_c}{\partial x}\dot{x} + \frac{\partial^2 f_c}{\partial t^2} \tag{4}$$

This is a system of linear equations with the constraint force vector with only unknown, η. In the hand gesture model, the freedom of hand motion is related with the position of elbow joint. Also, the constrained force $f_c(x(t),t)$ is limited within the movable hand range dependent on the elbow joint position.

5 Gesture Observation Model

In this section, we describe our method of extracting the features from hand signal motions using vision sensing mechanism. To detect hand motions, the region-of-interest (ROI) is bounded to hand region. We also present the relevant implementation of the HMM-based hand gesture.

5.1 ROI Detection

YCbCr color space is used for skin color segmentation. As input image is flexible for variation of illumination, the YCbCr color space has an advantage because illumination component of color space is concentrated in a Y channel [17]. RGB color space image captured by camera translates to YCbCr color space.

The following equations are segmentation of each color regions. Means and standard deviations are obtained by sampling manually of hand region pixels in image recording conditions.

$$\bar{Y} - n*\sigma_Y < Y_{Range} < \bar{Y} + n*\sigma_Y \tag{5}$$

$$\overline{Cb} - n*\sigma_Y < Cb_{Range} < \overline{Cb} + n*\sigma_Y \tag{6}$$

$$\overline{Cr} - n*\sigma_Y < Cr_{Range} < \overline{Cr} + n*\sigma_Y \tag{7}$$

We apply ***n*** flexibly based on the experimental conditions. The sequence of center of gravity is regarded as hand trajectory.

5.2 Feature Extraction

We choose a 2-dimensional feature vector (θ_t, V_t) as input for the recognition system. This Feature vector is based on a hand signal trajectory: angle and length. In

Fig 3, (X_t, Y_t)indicates the center of gravity of the hand at time t, l_t and θ_t are length and angle respectively obtained from consecutive images.

$$\theta_t = \tan^{-1}(\frac{Y_t - Y_{t-1}}{X_t - X_{t-1}}) \tag{8}$$

$$l_t = \sqrt{(X_t - X_{t-1})^2 + (Y_t - Y_{t-1})^2} \tag{9}$$

Our aim is not to translate the hand pose but to examine the hand trajectory of hand signal consisting of line and curve strokes. So angle and length between consecutive the center of gravity can fully describe hand trajectory in 2D image.

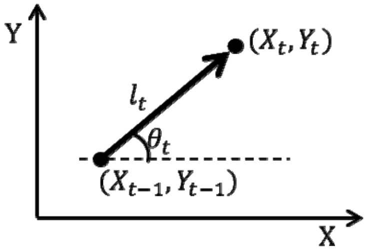


Fig. 3 Feature extraction of consecutive coordinates

5.3 Recognition

Through the feature extraction of sequential images, recognition part is the transmitted observation vector **O**.

$$\mathrm{O} = \{o_1, o_2, .. \, o_t\} \tag{10}$$

Using the observation vector, the recognition of hand gesture is made via left-to-right without skip topology HMM training and decoding, that is a stochastic process that includes three parameters $\lambda = \{A, B, \pi\}$ given N state vector [18, 21].

$$S = \{s_1, s_2, .. \, s_N\} \tag{11}$$

Note that the observation vector, $\mathrm{A} = \{a_{ij}\}$ is a set of state transition probabilities from i to j. $\mathrm{B} = \{b_j(o_t)\}$ is a probability distribution in each of the states and $\pi = \{\pi_i\}$ represents the initial state distribution. These parameters are estimated by Baum-Welch Algorithm [19]. When observation vector **O** entered, evaluation can be solved by using forward algorithms. State sequence ***S*** associated with the **O** is decoded by Viterbi algorithm [20].

By Bayes' theorem, the recognition process of hand signal is concisely presented by:

$$i^{*} = \mathrm{argmax}\{P(\mathrm{O} \mid \lambda_i)\} \tag{12}$$

where i denotes the index of the hand signal model. Calculating $P(\mathrm{O} \mid \lambda_i)$ based on HMM is carried out with the transition probability $\mathrm{A} = \{a_{ij}\}$ and the observation probability $\mathrm{B} = \{b_j(o_t)\}$.

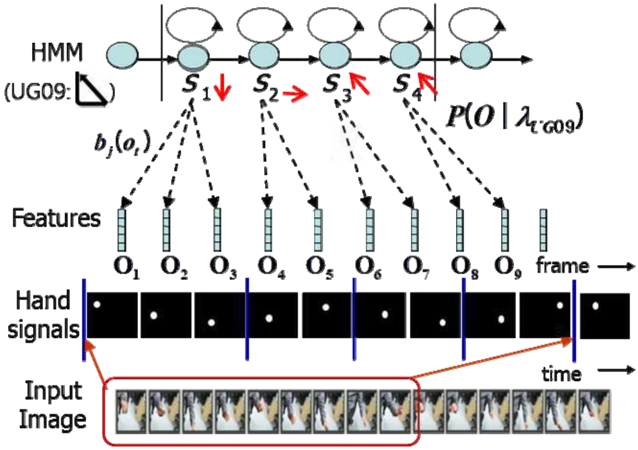


Fig. 4 HMM-based hand –signal recognition process

$$P(\mathrm{O} \mid \lambda_i) = \sum_{all\ S} P(\mathrm{O}, \mathrm{s} \mid \lambda) \tag{13}$$

A sample flow of HMM recognition process of UG09 is shown in Fig 4.

6 Evaluations

The goal of our experiments is to validate the effectiveness of the proposed language set. To achieve this goal, we conducted relevant off line experiments and obtained recognition rate. We first obtained the recognition result of UG and then expanded it to the language set of command lists.

We collected the hand gesture database from 24 people wherein each person tried 3 times, accumulating a total of 1008 UG gesture samples. First half of the database is used in training for the HMM parameter and the second half database is used for test database.

The overall recognition rate is obtained by averaging the recognition rate for each gesture set (individual UG and connected UG).

6.1 UG Recognition

First, we evaluated 14 UG's from Table 1 on recognition performance. To assess the influence of HMM state size variation, we obtained the UG recognition rate at different state number with the Gaussian mixture fixed at 2.

The best recognition rate achieved was 99.21% and at state5 and 7 respectively as shown in Fig 5. To assess for the sources of error among the UG set, the recognition result of each UG at state 7 was obtained and delineated in Table 3. Note that among the UG set tested, 10 UG's represent 100% recognition rate while the remaining four UG's mis-recognized at 2.8% error rate.

From these two results, the proposed UG set is verified that it can provide reasonably good discrimination. Based on this observation, we expanded the experiment to include the command (connected UG) recognition performance.

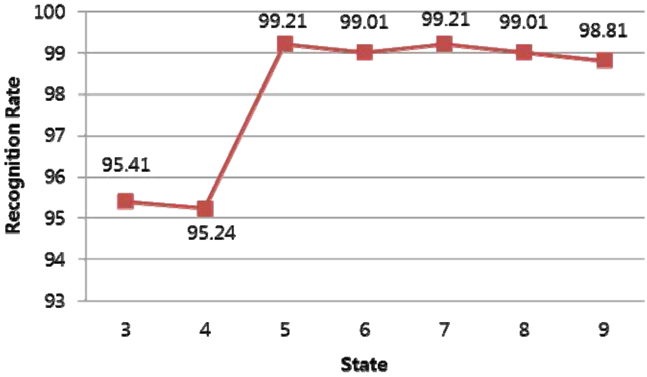


Fig. 5 Gesture recognition rate of UG

Table 3 Confusion matrix of UG recognition

	UG 01	UG 02	UG 03	UG 04	UG 05	UG 06	UG 07	UG 08	UG 09	UG 10	UG 11	UG 12	UG 13	UG 14
UG01	97.2	0	0	0	0	2.8	0	0	0	0	0	0	0	0
UG02	0	100	0	0	0	0	0	0	0	0	0	0	0	0
UG03	0	0	97.2	0	0	0	0	0	0	0	0	0	2.8	0
UG04	0	0	0	100	0	0	0	0	0	0	0	0	0	0
UG05	0	0	0	0	97.2	2.8	0	0	0	0	0	0	0	0
UG06	0	0	0	0	0	100	0	0	0	0	0	0	0	0
UG07	0	0	0	0	0	0	100	0	0	0	0	0	0	0
UG08	0	0	0	0	0	0	0	100	0	0	0	0	0	0
UG09	0	0	0	0	0	0	0	0	100	0	0	0	0	0
UG10	0	0	0	0	0	0	0	0	0	100	0	0	0	0
UG11	0	0	0	0	0	0	0	0	0	0	100	0	0	0
UG12	0	0	0	0	0	0	0	0	0	0	0	97.2	2.8	0
UG13	0	0	0	0	0	0	0	0	0	0	0	0	100	0
UG14	0	0	0	0	0	0	0	0	0	0	0	0	0	100
Average Recognition Rate = 99.21%														

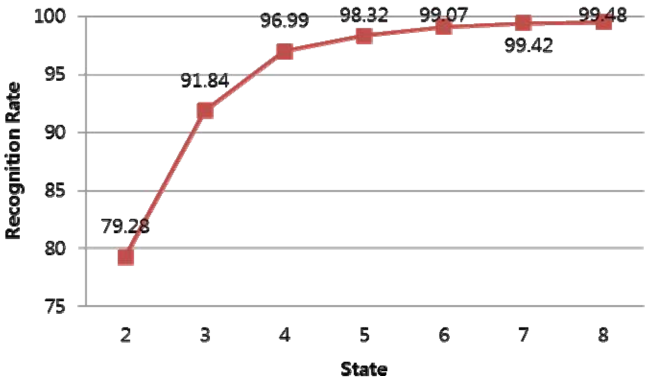


Fig. 6 Command (connected UG) recognition rate

6.2 *Command Recognition*

Similar to the UG experiment, we tested 48 commands from the list in Table 2 for recognition performance. Fig 6 shows the result of command recognition. As shown by the figure, the recognition rate achieved consistently stands above 99%. Through these experiments, not only the single UG but also language set (combination of UG) shows excellent performance which confirms for its suitability as an in-vehicle human interface.

7 Conclusions

We proposed and evaluated a gesture language set in the form of gesture units and connected gesture units for suitability as in-vehicle human interface. Representative experiments have demonstrated its effectiveness in terms of recognition rate and consistency. The proposed gesture based human-vehicle interface was shown promising for real world vehicle application.

For future work, the environmental robustness of hand detection using infra-red camera needs be addressed. By combining gesture to speech recognition, more intuitive human-vehicle interface is anticipated.

Acknowledgements. This work was supported by the ONR Program N000141010068.

References

1. McCall, J., Trivedi, M.: Video-Based Lane Estimation and Tracking for Driver Assistance: Survey. System, and Evaluation. IEEE trans. intelligent transportation systems 7(1), 20–37 (2006)
2. Gavrila, D.: Sensor-based pedestrian protection. IEEE intelligent systems 16, 77–81 (2001)

3. Handmann, U.: An image processing system for driver assistance. Image and Vision Computing 18, 367–376 (2000)
4. Zu, J., et al.: Vison-guided automatic parking for smart car. IEEE intelligent Vehicles symposium, 725–730 (2000)
5. Alpern, M., Minardo, K.: Developing a car gesture interface for use as a secondary task. In: CHI 2003 extended abstracts on Human factors in computing systems, pp. 932–993 (2003)
6. Tonnis, M., et al.: Experimental Evaluation of an Augmented Reality Vsualization for Direction a Car Driver's Attention. In: International Symposium on Mixed and Augmented Reality (2005)
7. Lee, H., Kim, J.: An HMM-Based Threshold Model Approach for Gesture Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 21(10), 961–973 (1999)
8. Liang, R., Ouhyoung, M.: A Real-Time Continuous Gesture Recognition System for Sign Language. In: Proceedings of the Third Int. Conference on Automatic Face and Gesture Recognition, Nara (Japan), pp. 558–565 (1998)
9. Elmezain, M., et al.: A Hidden Markov Model-based continuous gesture recognition system for hand motion trajectory. In: International Conference on Pattern Recognition, pp. 1–4 (2008)
10. Vogler, C., Metaxas, D.: Parallel hidden Markov models for American Sign Language recognition. In: Proc. Seventh International Conference on Computer Vision, vol. 1, pp. 116–122 (1999)
11. Shon, S., Beh, J., Wang, H.: Robot User Control System using Hand Gesture Recognizer. Journal of Institute of Control, Robotics and Systems 17(4) (in press 2011)
12. Zhenyao, M., Neumann, U.: Real-time hand pose recognition using low-resolution depth images. In: IEEE Computer Vision and Pattern Recognition, pp. 1499–1505 (2006)
13. Dreuw, P., et al.: Speech Recognition Techniques for a Sign Language Recognition System. Interspeech, 705–708 (2007)
14. Neitzel, L., et al.: A reviews of crane safety in the construction industry. Applied Occupational and Environmental Hygiene 16(12), 1106–1117 (2001)
15. Shon, S., Beh, J., Wang, H., Yang, C., Ko, H.: Hand Motion Design for Performance Enhancement of Vision Based Hand Signal Recognizer. Journal of IEEK, SP 48 (in press 2011)
16. Pentland, A., Liu, A.: Modeling and prediction of human behavior. Neural Computation 11, 229–242 (1999)
17. Chai, D., Bouzerdoum, A.: A Bayesian approach to skin color classification in YCbCr color space. In: Proceedings of TENCON, vol. 2, pp. 421–424 (2000)
18. Liu, N., et al.: Model Structure Selection and Training Algorithms for an HMM Gesture Recognition System. In: International Workshop on Frontiers in Handwriting Recognition, pp. 100–105 (2004)
19. Baum, L.: An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3(9), 1–8 (1972)
20. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–285 (1989)
21. Young, S., et al.: The HTK Book (for HTK version 3.4) (2009), `http://htk.eng.cam.ac.uk`

A Power Efficiency Based Delay Constraint Mechanism for Mobile WiMAX Systems

Nga T. Dinh

Abstract. In Mobile WiMAX systems, a power saving mechanism, which operates using sleep- and wake- modes, extends the battery lifetime of a Mobile Station (MS) but at the expense of the frame response delay. These two performance metrics, power consumption and delay, are reciprocally affected by sleep mode parameters such as initial sleep window Tmin, and final sleep window Tmax. This paper focuses on minimizing energy consumption under a guaranteed delay performance using a Power Efficiency based Delay Constraint (PEDC) mechanism which finds jointly-optimal parameters Tmin and Tmax. Simulation results validate that our proposed PEDC mechanism indeed minimizes consumed power for a specified frame response delay. In addition, in some cases, it can save up to 60% of consumed power compared to the original power saving mechanism in Mobile WiMAX.

Keywords: Mobile WiMAX, sleep mode, power consumption, frame response delay.

1 Introduction

The explosive growth of the internet over recent decades has led to increasing demands for high speed and ubiquitous internet access. To address these demands, broadband wireless access (BWA) is a potential technology supplies broad bandwidth at a low cost for small business and residential users. World Interoperability for Microwave Access (WiMAX) or IEEE 802.16 [1] is a standard which enables fixed and mobile convergence through BWA technology and flexible network architecture. The original IEEE 802.16 standard defines a common medium access control (MAC) to support only fixed BWA where the locations of subscriber station are stationary. As an enhancement of this standard, IEEE 802.16e or Mobile WiMAX [2] targets service provisioning to Mobile Stations (MSs). It adds mobility components and defines physical and MAC layers for combining fixed and mobile operations in licensed bands. Due to mobility support, a power saving mechanism (PSM) is extremely important because an MS is generally powered by a limited battery.

Nga T. Dinh
Bell Labs Seoul, Seoul, Korea
e-mail: Dinh.Nga@alcatel-lucent.com

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 181–191.
springerlink.com

In a PSM, an MS repeatedly goes between sleep modes and wake modes in order to conserve power. Sleep mode is a state in which an MS conducts pre-negotiated periods of absence from its serving Base Station (BS) air interface. The sleep mode is intended to minimize MS power usage and decrease usage of the BS air interface resource. Specifically, in sleep mode operation, an MS alternatively enters between a sleep state, where it does not communicate with its serving BS, and a listening state, where it checks whether or not there is a frame addressed to it during sleep time. During sleep states, the MS powers down and it only powers up during listening states [2]. Since power consumed in the sleep state is much smaller than that in the listening state, how to effectively reduce the duration of the listening state is the key in power saving. Previously, durations of sleep and listening states were determined by sleep mode parameters such as initial sleep state window T_{min} and final sleep state window T_{max} [3]- [6], and thus determining suitable values for T_{min} and T_{max} is essential.

Much of research on Mobile WiMAX has mainly focused on improving power efficiency in the MAC layer. In reference 4, sleep mode operation is analyzed for uplink and downlink traffic. Reference 5 investigates the queuing behavior of sleep mode in terms of dropping probability and mean waiting time of packets in the BS queue. Also, PSM is numerically analyzed with a Markov chain pertaining to the power consumption and average frame response delay in consideration of sleep mode parameters [6]. In reference 7, the initial sleep window is changed during low frame arrival rates to reduce energy consumption. Unfortunately, this approach results in a worse delay than the original PSM when traffic is high. Reference 3 evaluates the performance of the standard PSM in terms of energy consumption in sleep mode and frame response delay. All these studies conclude that the network performance metrics are mainly affected by sleep mode parameters; hence, it is reasonable to consider them as key to performance enhancements in Mobile WiMAX systems.

Although a PSM extends battery lifetimes of MSs by reducing their power consumption, it simultaneously induces a frame response delay due to arrival of frames during the MSs sleep mode. Furthermore, due to the trade-off relationship between consumed power and frame response delay [3], they cannot be improved simultaneously. Interestingly, these metrics are mainly affected by sleep windows T_{min} and T_{max}. Long sleep windows lower the power consumption but increase the frame response delay [3]. In our previous study [10], the optimal T_{max} is determined with the assumption that T_{min} is already known. In this paper, we propose a Power Efficiency based Delay Constraint (PEDC) mechanism which minimizes MS's power in its sleep mode while guaranteeing a specified delay by finding jointly suitable T_{min} and T_{max}. Based on the observation that consumed power is inversely proportional to T_{min} and T_{max} while delay is proportional and that T_{min} is more important than T_{max} in affecting consumed power and delay, the PEDC mechanism first finds the maximum T_{min} that satisfies delay constraint. After that, the mechanism determines the optimal T_{max}.

2 Power Saving Mechanism in Mobile WiMAX

The IEEE 802.16e standard specifies a PSM in the MAC protocol by defining three Power Saving Classes (PSC). PSC of Type I is recommended for Best Effort (BE) and Non Real Time–Variable Rate (NRT-VR) while PSC of Type II is recommended for Unsolicited Grant Service (UGS) and Real Time-Variable Rate (RT-VR). Type III PSC is used for multicast connections as well as management operations [2]. In this paper, we consider operations of PSM in Type I PSC, which is also examined in [3]-[8].

To save power, an MS repeatedly goes into wake modes and sleep modes by communicating with its serving BS. In wake modes, the MS can receive all downlink (DL) transmission and send data to its BS. On the other hand, in sleep modes, the MS cannot send or receive any incoming frame or frame fragment to its BS during pre-negotiated intervals. Before switching from a wake mode to a sleep mode, the MS sends a sleep request message (MOB-SLP-REQ) which includes sleep mode parameters such as T_{min}, T_{max}, and listening window T_L, to its serving BS. With a sleep response message (MOB_SLP-RES) from the BS, the MS enters sleep mode. In sleep modes, the MS first sleeps for the first sleep window T_1, which equals T_{min}. Then, it temporarily wakes up for duration of T_L to listen indication message (MOB_TRF-IND) from its BS. If this message is negative meaning that there is no appearance of DL traffic during MS sleep states, the MS will sleep again. The second sleep window T_2, doubles T_1 and after T_2, the MS moves to listening state. Generally, each sleep window is twice the size of the previous one, but not greater than T_{max} [2]. Once the sleep window reaches the T_{max}, the sleep window is kept constant. This procedure is repeated until the MOB_TRF-IND is positive indicating the presence of buffered traffic destined to the MS. Then, the MS moves to normal operation. Furthermore, the MS can transition to wake mode by sending a bandwidth request message to its BS whenever it has some packet data units (PDUs) for the BS, whether the MS is in sleep state or in listening state.

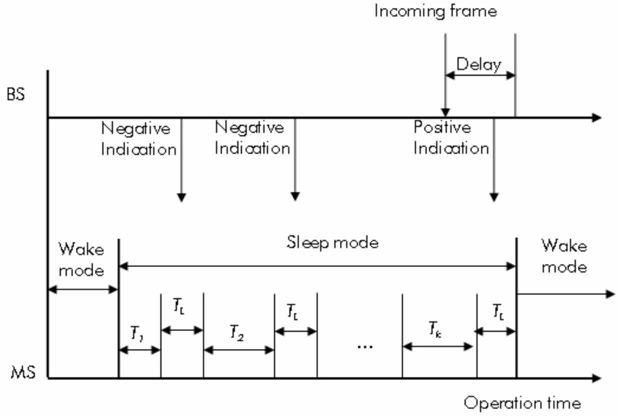


Fig. 1 Sleep mode operation, showing sleep- and listening- states, and frame response delay

In the case where the BS terminates the MS sleep mode by sending a positive MOB_TRF-IND message, a frame response delay will occur since the BS has to wait until the listening state of MS, as shown in Fig.1. Since consumed power in sleep states is much smaller than that in listening states, longer sleep windows produce better power management performance but cause an increased frame response delay. Therefore, in order to achieve a targeted performance, we need to carefully consider the effects of sleep windows, especially T_{min} and T_{max}.

3 Numerical Analysis

In the sleep mode of Type I PSC, sleep-state windows are measured in number of frames and increase binary exponentially [2]. In the IEEE 802.16e standard, the kth sleep-state window is

$$T_k = \min\left\{2^{k-1}T_{min};T_{max}\right\} \tag{1}$$

The initial sleep-state window T_{min} and final sleep-state window T_{max} have the following relationship

$$T_{max} = 2^{M-1}T_{min} \tag{2}$$

where M is an integer.

In PSC of Type I, the NRT-VR is suitable for Internet access with minimum guaranteed rate such as FTP. The BE is suitable for World Wide Web (WWW) traffic [11]. These two classes are proven to be self-similar where the connection arrival follows Poisson process [11]. We assume that frames arrive to an MS with a rate λ (the average number of frames per time unit) and consider sleep cycle k, which includes a variable sleep-state window T_k and a fixed listening-state window T_L. Then, the duration of sleep cycle k is $(T_k+ T_L)$. Let $Pr_k(i)$ be the probability that there are i frames arriving in sleep cycle k. The probability that there is at least one frame arriving during $(T_k+ T_L)$ is

$$\sum_{i=1}^{\infty} Pr_k(i) = 1 - Pr_k(0) \tag{3}$$

where

$$Pr_k(0) = \frac{\left(\lambda\left(T_k+T_L\right)\right)^0}{0!}e^{-\lambda(T_k+T_L)} = e^{-\lambda(T_k+T_L)} \tag{4a}$$

Based on the operations of PSC of Type I, an MS transitions from sleep mode to wake mode after k sleep cycles if and only if there is at least one frame addressed to the MS during sleep cycle k and no frames addressed to the MS in previous cycles. Let Pr_k be the probability of this event, which is determined as follows

$$Pr_k = \begin{cases} 1-e^{-\lambda(T_1+T_L)} & with\ k=1 \\ e^{-\lambda\sum_{i=1}^{k-1}(T_i+T_L)}\left(1-e^{-\lambda(T_i+T_L)}\right) & otherwise \end{cases} \tag{4b}$$

We next derive a numerical model for the frame response delay. We assume that an MS wakes up after k sleep cycles. This means that incoming frames arrive only in sleep cycle k. If we divide the kth sleep cycle into N small subintervals where N is sufficient large, the duration that a new frame arriving at the nth subinterval has to wait for an MS to transition to wake mode is calculated as follows

$$\tau_n = (T_k + T_L)\left(1-\frac{n}{N}\right) \tag{5}$$

Since a Poisson process is memoryless, the probabilities of frames arriving in each subinterval are the same. Trivially then, the average time that each frame has to wait is

$$\frac{1}{N}\sum_{n=1}^{N}(T_k + T_L)\left(1-\frac{n}{N}\right) = \frac{T_k + T_L}{2} \tag{6}$$

Let T_D be the delay that a frame has to wait due to sleep mode operation. Because an MS transitions from sleep mode to wake mode after k sleep cycles with probability Pr_k, the expectation of T_D is calculated as follows

$$E[T_D] = \sum_{k=1}^{\infty} Pr_k \frac{T_k + T_L}{2} \tag{7}$$

We let P_S and P_L denote the consumed power of an MS in sleep-state window and listening-state window, respectively. Then, the consumed power, which is defined as energy consumed per time unit, of an MS in its sleep mode is,

$$P = \frac{\sum_{i=1}^{\infty} Pr_i \sum_{k=1}^{i}(T_k P_S + T_L P_L)}{\sum_{i=1}^{\infty} Pr_i \sum_{k=1}^{i}(T_k + T_L)} \tag{8}$$

where the numerator is the energy that an MS consumes in its sleep mode and the denominator is the time that an MS spends in its sleep mode.

From (8), $E[T_D]$ linearly increases with T_{min} and T_{max} and from (8) one can infer that the power consumption decreases with T_{min} and T_{max} [3-8].

Looking at Pr_k calculated by (4b), we see that for a fixed T_{min}, Pr_k diminishes exponentially as k increases. Therefore, the terms with smaller k are likely to be more dominant in determining $E[T_D]$. First, $E[T_D]$ is proportional to T_{max}. However, when T_{max} reaches a certain value, $E[T_D]$ reaches constant. In other words, this T_{max} makes $E[T_D]$. stay constant, called $E[T_D]_{max}$ [12]. The reason is that with sufficiently large T_{max}, Pr_k is almost zero or in physical meaning, there is no frame arriving in sleep cycles whose sleep-state window is T_{max} while no frame arrives in previous sleep cycles.

4 Proposed Power Efficiency Based Delay Constraint Mechanism

This section presents a simple but effective mechanism, called Power Efficiency based Delay Constraint (PEDC) mechanism, to minimize the consumed power of an MS in its sleep mode while satisfying the constraint that $E[T_D]$ is no greater than a specified value, T_{Dmax}. by finding jointly suitable T_{min} and T_{max}. Moreover, the frame arrival rate λ depends on customer behavior and we assume for simplicity that λ is already known. Therefore, we have the set of parameters (T_{Dmax}, λ, T_L) from which we are able to find T_{min} and T_{max} that achieve the lowest power consumption.

We infer from (7) that

$$E[T_D] \geq \frac{T_{min} + T_L}{2} \qquad (9)$$

and

$$E[T_D]_{min} = \frac{T_{min} + T_L}{2} \qquad (10)$$

In the analytical and simulation results, times are normalized to T_L, which is of fixed duration, typically one frame. T_{min} measured in the number of frames [2] is an integer [3-10], and satisfies the following inequality

$$T_{min} \leq \lfloor 2T_{Dmax} - T_L \rfloor \qquad (11)$$

where $\lfloor A \rfloor$ is the largest integer not exceeding a real number A. Therefore, there will be no solution for T_{min} and consequently for T_{max} if $2T_{Dmax} - T_L < 1$. Otherwise, T_{min} must belong to the set $X = \{1, 2, ..., \lfloor 2T_{Dmax} - T_L \rfloor\}$. In this paper we ignore the trivial case where $2T_{Dmax} - T_L < 1$.

With any given Tmin, there exists a corresponding $E[T_D]_{min}$ calculated by (11) and a $E[T_D]_{max}$ that saturates with T_{max}[12]. The solution for T_{max} that satisfies T_{Dmax} condition is as follows: (1) if $T_{Dmax} < E[T_D]_{min}$, no T_{max} exists; (2) if $E[T_D]_{min} < T_{Dmax} \leq E[T_D]_{max}$, there is an unique T_{max}, and (3) if $E[T_D]_{max} \leq T_{Dmax}$, any T_{max} will satisfy T_{Dmax} [12]. Moreover, as the expected frame response delay $E[T_D]$ is proportional to T_{min} while the consumed power P is inversely proportional, the best way to minimize P under a given T_{Dmax} is to find the maximum T_{min} that satisfies T_{Dmax}. Next, as $E[T_D]$ and P are non-decreasing and non-increasing functions of T_{max}, respectively, we will find the largest T_{max} that satisfies T_{Dmax} or the T_{max} at which $E[T_D]$ saturates.

Based on the above observations, the idea for our proposed PEDC mechanism is as follows. The PEDC mechanism starts with the smallest T_{min} in set X. Since T_{min} is more important than T_{max} in determining $E[T_D]$, this step finishes after several iterations. This process is described in lines 3~10 of the pseudo code below. After line 10, T_{min} is known. When T_{min} is determined, we will find the corresponding $E[T_D]_{\max}$ by starting T_{max} equals to $2T_{min}$ and then increasing T_{max} until $E[T_D]_{max}$ does not change. In fact, this step will stop if the difference between current $E[T_D]$ and the previous $E[T_D]_{max}$ is lower than a very small value ε. After that, we compare $E[T_D]_{max}$ with T_{Dmax}. If $T_{Dmax} > E[T_D]_{max}$, the optimal T_{max} will be the value that saturates $E[T_D]$. Otherwise, starting with T_{max} equal to $2T_{min}$ and then increasing T_{max} gradually, we continue to calculate the corresponding $E[T_D]$ by (7) as long as $E[T_D]$ is lower than T_{Dmax}. Then, we divide T_{max} by 2 because this T_{max} violates the frame delay constraint. From this, we get the T_{min} and T_{max} which achieve the lowest consumed power of an MS in its sleep mode. The following is the pseudo code of the PEDC mechanism.

Mechanism 1 The operation of proposed PEDC mechanism

```
Require: $T_{Dmax}$, $\lambda$, $T_L$
1:  $\varepsilon = 10\text{-}6$
2:  Assign $T_{min} = 1$
3:  **while** $T_{min} \in X$ **do**
4:      Calculate $E[T_D]_{min}$ by Eqn (10)
5:      **while** $\left(E[T_D]_{min} < T_{Dmax}\right)$
6:          $T_{min} = T_{min} + 1$;
7:          Go back to 3:
8:      **end while**
9:  **end while**
10: $T_{min} = T_{min} - 1$
11: $T_{max} = 2\ T_{min}$
12: Find $E[T_D]_{max}$ by increasing $T_{max}$ in Eqn (7) until the difference between current $E[T_D]$ with its previous value is lower than $\varepsilon$. Keep the $T_{max}$ that makes this condition happen as $T_{max}$ *
13: **if** $E[T_D]_{max} \le T_{Dmax}$ **then**
14:     $T_{max} = T_{max}$ *
15: **else**
16:     Calculate $E[T_D]$ by Eqn (7)
17:     **while** ($E[T_D] < T_{Dmax}$)
18:         $T_{max} = 2T_{max}$;
19:         Go back to 17:
20:     **end while**
21:     $T_{max} = ½\ T_{max}$;
22: **end if**
23: Calculate $P$ by Eqn (8)
```

The PEDC mechanism will converge after several iterations in each loop for determining T_{min} and T_{max}, thus making PEDC mechanism is simple. As explained in the following section, PEDC indeed minimizes the consumed power P. In addition, while PEDC guarantees the expected delay, there will still be some instances where the delay may exceed T_{Dmax}. After T_{min} and T_{max} are obtained by our proposed PEDC mechanism, we let i be the index for the smallest sleep-state window where $\frac{T_i + T_L}{2} > T_{Dmax}$. Furthemore, let Pr_{excess} be the probability that the instant frame response delay is equal to or greater than T_{Dmax} , then Pr_{excess} is calculated as follows

$$Pr_{excess} = \sum_{k=i}^{\infty} Pr_k \frac{T_k + T_L}{2} \tag{12}$$

where Pr_k is determined by (4b). Pr_{excess} converges as Pr_k diminishes rapidly with k. Furthermore, Pr_{excess} is small because Pr_k decreases exponentially with k.

5 Performance Evaluations

In this section we show the effects of the sleep-state windows on consumed power and frame response delay of an MS in its sleep mode. In addition, we validate our proposed mechanism through computer simulations based on the following parameters: $T_L = 1$ (frame), $P_S = 1$; and $P_L = 30$ (unit of power) as in Reference [13]. The incoming frame arrival rate, λ, is varied for different scenarios. The simulation runs for 1,000,000 time units (frames) and the results are the mean values from 100 different runs with 100 different seed values.

Let T be the mean inter-arrival time of incoming frames. Then, we will have $\lambda = 1/T$. Fig.2 shows the effects of T_{max} on consumed power when λ is 0.0625, meaning that the average inter-arrival time is 16 (frames), with different T_{min}. The effects can be divided into two regimes. In the first regime, P is inversely proportional with T_{max} while in the second regime, P saturates at a minimum value for large T_{max}. The minimum power results from the near-zero probability of frames not arriving in previous sleep windows for large T_{max}. In addition, we can infer the effects of T_{min} on consumed power from this figure. Similar to the effects of T_{max} on P in the first regime, P is a strictly decreasing function of T_{min}.

The effects of T_{max} on the expected frame response delay is illustrated in Fig. 3 when λ is 0.0625 with different T_{min}. Similar to how Tmax affects consumed power, the relationship between T_{max} and P is divided into two regimes. First, $E[T_D]$ is proportional to T_{max}. This is because when T_{max} increases, the duration for a sleep cycle, which includes a sleep-state window and a listening-state window, will increase. This makes frames arriving during the sleep state wait longer thereby increasing $E[T_D]$. In the second regime, $E[T_D]$ stops increasing for large T_{max} for the same reason that P saturates with T_{max} in figure 2. We can also see that $E[T_D]$ is a strictly increasing function of T_{min}. Moreover, small changes of T_{min} result in large changes of $E[T_D]$, as shown in figure 3. In other words, T_{min} is more important than T_{max} on affecting the frame response delay. Furthermore, as illustrated in these figures, the simulation results are very close to the analytical results.

Figure 4 presents the performance comparison between PEDC mechanism with all cases of original PSM in Mobile WiMAX with respect to T_{Dmax} when the mean inter-arrival time T is 20 (frames) and $\lambda = 0.05$. The figure shows that our proposed PEDC mechanism achieves a better power performance compared to all cases of IEEE 802.16e standard. Under a given T_{Dmax}, there may exist several (T_{min} ; T_{max}) pairs which result in an $E[T_D]$ that is equal to or less than T_{Dmax}. The PEDC mechanism always finds the maximum values of T_{min} and suitable values of T_{max}. Since P is a strictly decreasing function of T_{min} and non-increasing function of T_{max}, the (T_{min}, T_{max}) pair obtained by our proposed mechanism guarantees that P is the lowest. In addition, this also ensures that the PEDC converges. Furthermore, the difference between the consumed power obtained by our proposed algorithms and the one obtained by the IEEE 802.16e is large, especially when T_{Dmax} is large.

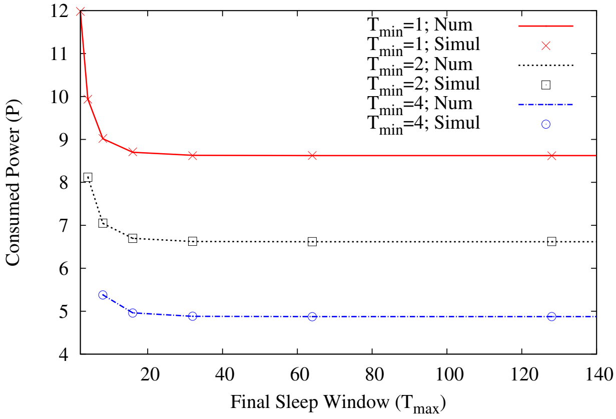


Fig. 2 Effects of sleep windows on consumed power, showing how T_{min} and T_{max} affect consumed power

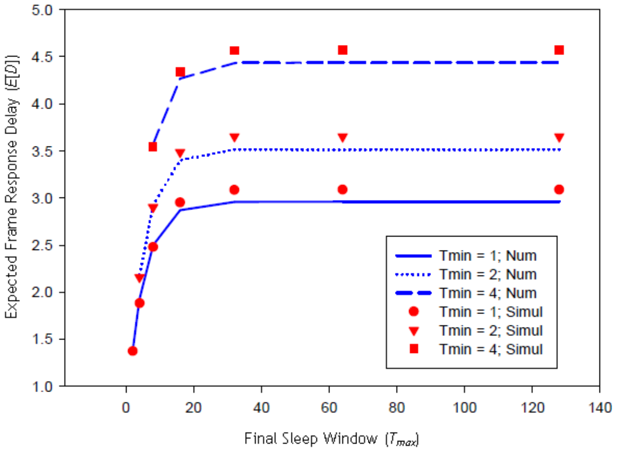


Fig. 3 Effect of sleep windows on frame response delay- showing how T_{min} and T_{max} affect expected frame response delay

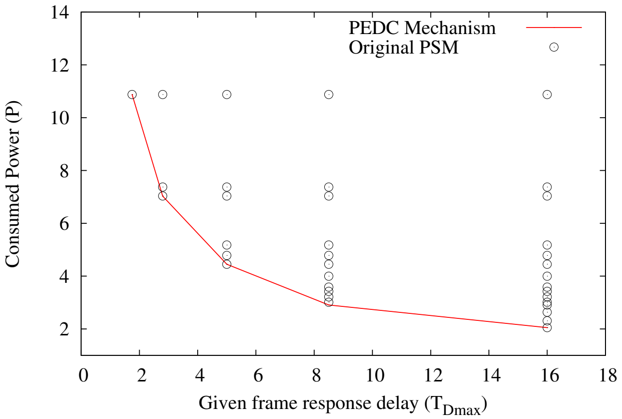


Fig. 4 Consumed power comparisons- showing how the proposed PEDC mechanism improves power saving compared to the original PSM in IEEE 802.16e standard

In some cases, for example when T_{Dmax} is greater than 4.5 (frames), the proposed PEDC can save up to 60% of MS's consumed power in its sleep mode compared to the original PSM in IEEE 802.16e. Although the IEEE 802.16e standard can achieve the same energy consumption as our proposed mechanism, this probability is very low.

6 Conclusion

In Mobile WiMAX, power saving mechanisms can prolong battery lifetime in the mobile station by reducing the consumed power but may adversely affect the frame response delay. A Power Efficiency based Delay Constraint (PEDC) mechanism is proposed to minimize consumed power of an MS by calculating jointly the initial sleep-state window T_{min} and the final sleep-state window T_{max}, constrained by the maximum allowed frame response delay. The PEDC algorithm first finds the maximum T_{min} and then determines the suitable T_{max}. Simulations results match well with analytical results and the proposed PEDC mechanism indeed minimizes consumed power. In some cases, the PEDC mechanism can save up to 60% of consumed power of MS in its sleep mode compared to the original PSM of the IEEE standard.

Acknowledgment. This work is supported by the Seoul Metropolitan Government R&BD Program WR080951. In addition, the author would like to thank Dr. Randy Giles for his review of the manuscript.

References

[1] IEEE Std. 802.16/Cor1/D3-2005, IEEE Standard for Local and Metropolitan Area Network – Part 16: Air Interface for Fixed Broadband Wireless Access Systems, (May 2005)

[2] IEEE Std 802.16e-2005, Part 16: Air Interface for Fixed Broadband Wireless Access Systems (February 2006)

[3] Xiao, Y.: Energy Saving Mechanism in the IEEE 802.16 Wireless MAN. IEEE Commun. Lett., 595–597 (July 2005)

[4] Zhang, Y., Fujise, M.: Energy Management in the IEEE 802.16e MAC. IEEE Commun. Lett., 311–313 (April 2006)

[5] Jeong, D.G., Jeon, W.S.: Performance of adaptive sleep period control for wireless communications systems. IEEE Trans. Wireless Commun. 5(11), 3012–3016 (2006)

[6] Choi, H.-H., Lee, J.-R., Cho, D.-H.: Hybrid power saving for VoIP services with silence suppression in IEEE 802.16e systems. IEEE Commun. Lett. 11(5), 455–457 (2007)

[7] Xiao, J., Zou, S., Ren, B., Cheng, S.: An Enhanced Energy Saving Mechanism in IEEE 802.16e. In: Proc. IEEE Global Communication (2006)

[8] Jung, E.S., Vaidya, N.H.: An energy efficient MAC protocol for wireless LAN. In: Proceeding of INFOCOM 2002, vol. 3, pp. 1756–1764 (2002)

[9] Seo, J.-B., Lee, S.-Q., Park, N.-H.: Performance Analysis of Sleep Mode Operation in IEEE 802.16e. In: Proc. IEEE Vehicular Technology Conference, pp. 1169–1173 (2004)

[10] Han, K., et al.: Performance Analysis of Sleep Mode Operation in IEEE 802.16e Mobile Broadband Wireless Access Systems. In: Vehicular Technology Conference, pp. 1141–1145 (2006)

[11] Wang, H., He, B., Agrawal, D.P.: Admission Control and Bandwidth Allocations above Packet Level for IEEE 802.16 Wireless MAN. In: Proceeding of the International Conference on Parallel and Distributed Systems (2006)

[12] Nga, D.T.T., Kim, M.-G., Kang, M.: A Novel Energy Saving Algorithm with Frame Response Delay Constraint in IEEE 802.16e. IEICE Transaction on Communications E91-B(4) (April 2008)

[13] Kim, M.-G., Choi, J., Kang, M.: Performance Evaluations of the sleep Mode Operations in the IEEE 802.16e MAC. In: International Conference on Advanced Communication Technology, pp. 602–605 (2007)

An Adaptive Scheduling Algorithm for Scalable Peer-to-Peer Streaming

Longzhe Han and Hoh Peter In*

Summary. Because of high scalability and low cost, Peer-to-Peer (P2P) video streaming has been the promising and interesting approach for delivering multimedia content over the Internet. However very different capabilities of end users' systems, unpredictable network bandwidth and the rigorous nature of the video streaming impose design challenges on P2P streaming systems. In this paper, an adaptive scheduling algorithm has been proposed to overcome these limitations. This algorithm is based on H.264 Scalable Video Coding (H.264/SVC) and P2P paradigm to provide scalable video streaming services for heterogeneous users and self-adapt the dynamic network environment.

1 Introduction

As continuous improvement of computing power, memory capacity and network bandwidth, multimedia applications, such as video streaming, VoIP, and video conferencing, over the Internet has been hot research topic and industrial practice for last few years. Because of high scalability, self-management and low cost, Peer-to-Peer (P2P) video streaming becomes a promising and interesting approach for delivering multimedia content over the Internet [1, 2]. The key idea of P2P streaming is to utilize end-users' resources, for example out-going bandwidth, memory and CPU, to construct an overlay on the application layer and forward data to connected users. Figure 1 demonstrates a P2P streaming system.

In contrast to the client-server architecture, peers (end-users) self-organize and maintain an overlay topology on the application layer, and each peer has two roles: receiving data for playback and forwarding data to the required peers. Because of this unique characteristic, P2P streaming paradigm has superior scalability than the client-server architecture. Currently wide-used P2P streaming systems [3, 4, 5] deliver the same bitrate video stream to all users. As the advance in communication technology and embedded system, different network

Longzhe Han · Hoh Peter In
Korea University
Seoul, 136-713, South Korea
e-mail: {lzhan,hoh_in}@korea.ac.kr

* Corresponding author.

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 193–200.
springerlink.com

connections (HSDPA, ADSL, LAN, WiBro and WiMax) and diverse end-users' systems (Desk-top, Lap-top PC, Mobile Phone and PDA) have been used to access the Internet. Delivering the pre-fixed bitrate video stream in this heterogonous network makes the high and low capability users receive the same quality of service, which is usually adapted to low capability users. Also best-effort service of the Internet makes the available bandwidth of end-users fluctuation unexpectedly. To overcome these limitations, in this paper, we propose an Adaptive Scheduling Algorithm for scalable peer-to-peer Streaming (ASA). The ASA is based on H.264 Scalable Video Coding (H.264/SVC) [6] and P2P paradigm to provide adaptive video streaming service in the heterogeneous network environment.

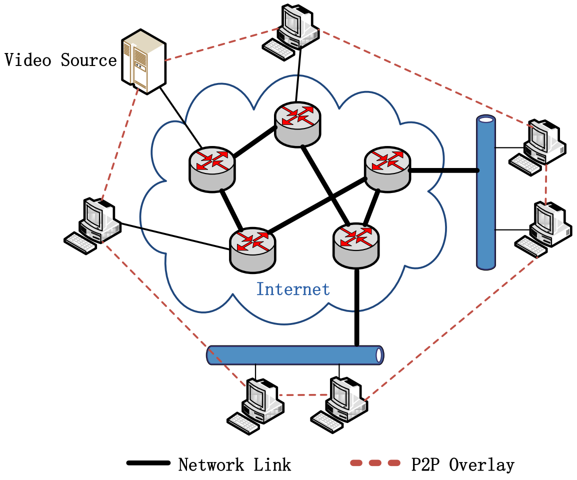


Fig. 1 The overlay structure for P2P streaming

2 Background

The proposed P2P streaming system adopts BitTorrent-like (BT) P2P paradigm and H.264/SVC. There are several considerations for choosing BT-like P2P overlay architecture. BT [7] is currently most popular and successful file sharing system in the Internet. Its tit-for-tat incentive strategy can effectively reduce free-riding. The copyright issue is the serious problem of the P2P Streaming applications, by using the centralized component, such as the tracker in BT; P2P Streaming systems can deploy user management, authentication and access monitoring to avoid copyright issues. Figure 2 depicts the architecture of BT. It consists of four main components: tracker, seed, leecher, and torrent file server. The tracker traces all peers (including seeds and leechers), and provides a list of random selected peers to the new peer. The seed is a peer that holds the entire content for sharing, and a leecher is a peer that needs to download the content. The new peer first needs to download the metadata (torrent file) of the content which is stored on the torrent file server.

Although BT works well for the file swapping purpose, it is not efficient for the video streaming in the dynamic and heterogeneous network environment. First, BT is not designed for video streaming applications; the piece selection algorithm cannot meet temporal requirements for the play back. Second, the different capability and available bandwidth fluctuation of each peer is not considered in the algorithm [8].

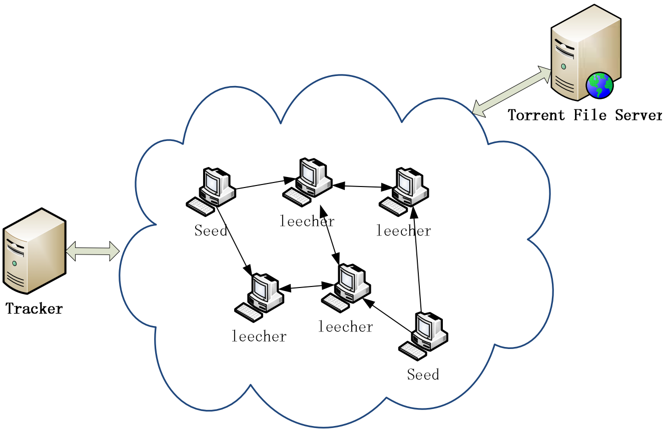


Fig. 2 The architecture of BitTorrent

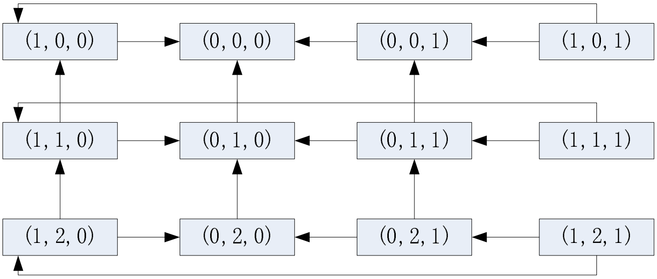


Fig. 3 Dependent relationship of NALUs

In our system, we preserve the spirit of BT and extensively redesign the architecture and algorithms to providing adaptive H.264/SVC P2P streaming service. H.264/SVC is the latest video encoding standard proposed for heterogeneous network environment. It uses layered coding scheme to generate multi-layer bit stream: One base layer (BL) and several enhancement layers (EL) are coded in a SVC bit stream.

Base layer provides a basic quality of video while the enhancement layers are mainly used to refine the video quality. In order to facilitate the network streaming applications, the video data is further encapsulated into Network Abstraction Layer Units (NALUs). The scalability of the NALU can be identified from the NALU header by three fields (DID, TID, QID): DID is for the spatial scalability, TID represents the temporal scalability, and QID stands for the quality scalability. Figure 3 shows the dependent relationship between the SVC NALUs. For decoding one NALU, the dependent units of that NALU have to be received successfully.

3 Adaptive Scheduling Algorithm

3.1 System Architecture

As shown in Figure 4, the system consists of three main components: peer swarm, video source and system management server. Each SVC layer data is separately extracted and divided into a series of pieces. Each layer data will be transported by the different group. In Figure 4, the SVC video is encoded as one base layer and two enhancement layers.

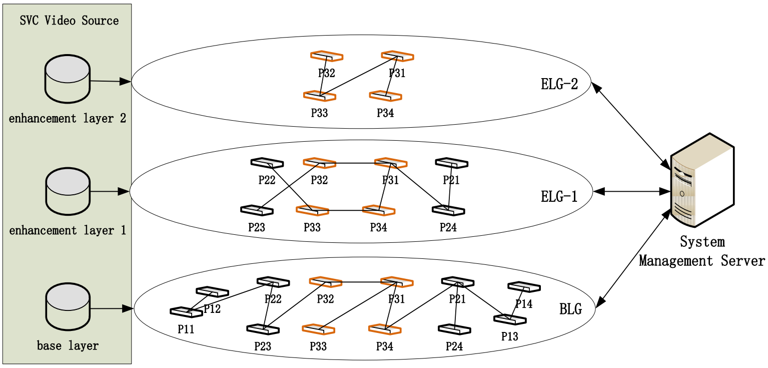


Fig. 4 System Architecture

According to the number of layer in SVC stream, the same number of group will be constructed: Base Layer Group (BLG), Enhancement Layer 1 Group (ELG-1), …, Enhancement Layer N Group (ELG-N). One group only swaps one layer data based on SVC coding scheme. For example, In Base Layer Group (BLG), peers only transfer base layer data. For better video quality, peers can join multiple groups. However, if one peer takes part in ELG-N, this peer must also appear in the under-layer groups: ELG-(N-1), ELG-(N-2), …, BLG. By this hierarchical grouping mechanism, each peer can optionally select the different quality of video streaming. The mesh-based connection overlay structure is constructed inside of each group. The System Management Server (SMS) trace the statues of all connected peers in the swarm: download/upload statics and group ID. When a peer first time joins the swarm, it will retrieve a peer list from the SMS. The list is made by randomly selected peers from request groups. Finally each peer will maintain a fix number of concurrent connections.

3.2 Scheduling Algorithm

The streaming buffer is divided into two parts: zone of urgency and zone of demand, as presented in figure 5. The SVC video is divided into chunks, currently one chunk is one NALU. The peer constantly measures its data download bandwidth. According to the bandwidth, the peer estimates the number of layers it can afford. The chunks in the urgent zone are requested by using strict layered selection algorithm. The selection order is based on the inter-dependent priority. All based layer chunks in the urgency zone are fetched first, then the chunks of enhancement layer 1 are fetched, the higher layer chunks will be requested if all lower layer chunks are downloaded. This process can ensure the continuous video playback, if the peer has not enough bandwidth to download the chunks of all SVC layers. The chunks in the demand zone are requested by using random layered selection algorithm. Each layer assigns a weight value for calculating the probability of the layer chunks to be selected. The lower layer chunks will get more chance be selected than higher layers. The reason is the dependent relationship of SVC layers. The higher layer is non-decodable if the lower layer is not existed. The detail algorithm is presented in figure 6.

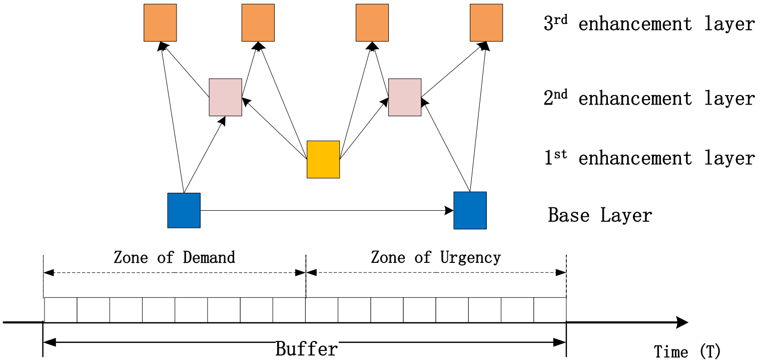


Fig. 5 Streaming Buffer

```
Initialization;
//calculate bandwidth
receive_datasize = receive_piece * piece_size;
current_bandwidth = receive_datasize / time_interval;
bandwidth = (1 – bweight) * current_bandwidth + bweight * bandwidth;

//According to the bandwidth, decide how many layers can //afford.
If (the urgency zone has missing chunks) {
    For (i = 0; i < number_of_layers; i++) {
        If (layer i has missing chunks)
            Download(layer i chunks);
    }
}
// each layer has a weight value
// base Layer Weight (LW1) is 4, enhancement Layer 1
// Weight (LW2) is 7, LW3 is 9, LW4 is 11, …, LWn is 15
// LW0 is 0
If (the demand zone has missing chunks) {
    no = random(0, 15);
    if (no >= $LW_i$ && no <= $LW_{i+1}$) {
        Download(layer i chunks);
    }
}
```

Fig. 6 Pseudo code of scheduling algorithm

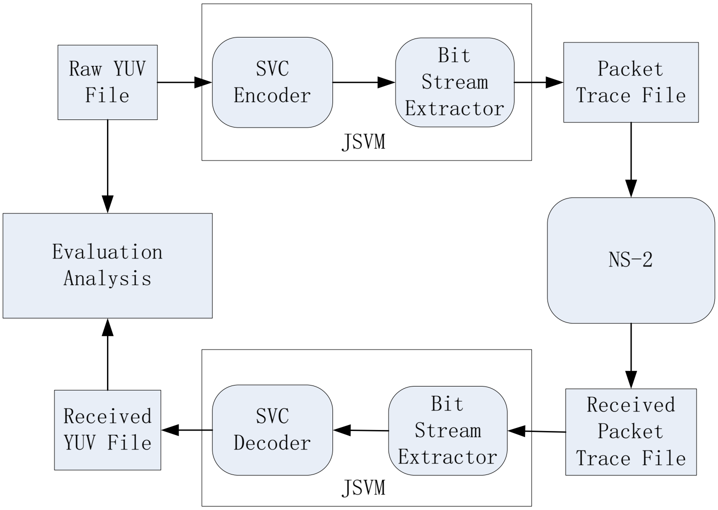


Fig. 7 Simulation process

4 Simulation Results

We use the SVC reference software (JSVM 9) [9] and the NS-2 simulator [10] to verify the proposed scheduling algorithm. The simulation process is shown in figure 7. The raw video sequence [11], SOCCER, in CIF format, is encoded by using JSVM. The frame rate of the video is 30 frames per second with two spatial enhancement layers and one MGS enhancement layer. There are 1060 NALUs in the encoded stream. The UDP packet size is 1000 bytes. Total number of source video packets is 1317.

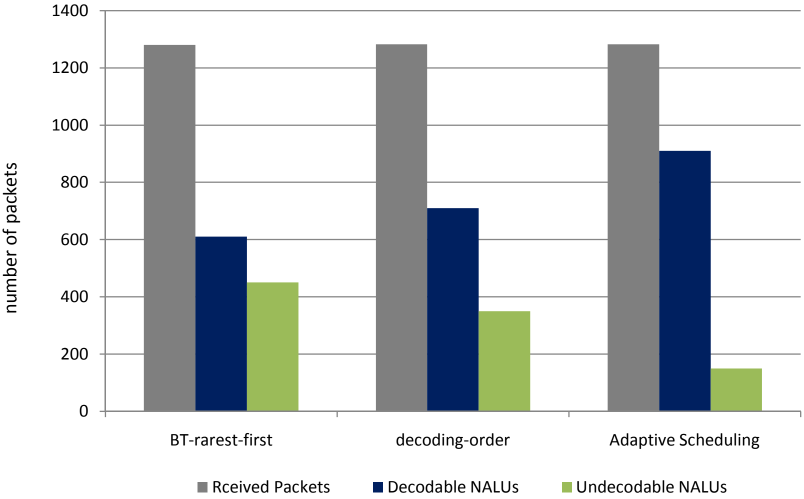


Fig. 8 Simulation result

As shown in Figure 8, the proposed algorithm achieves better performance than BT-rarest-first and decoding-order based scheduling algorithms. BT-rarest-first and decoding-order scheduling algorithms treat different priority NALUs as the same, more packets are not decodable than the proposed algorithm even though they are correctly arrived. This is caused by the dependent relationship between different SVC layers.

5 Conclusion

This paper proposed an adaptive scheduling algorithm for H.264/SVC based P2P streaming in the heterogeneous network environment. In proposed algorithm, the streaming buffer is divided into two parts: zone of urgency and zone of demand. According to the SVC layered characteristics, the adaptive scheduling algorithm dynamically selects SVC video data to meet various network conditions. The experiment results demonstrate that the proposed algorithm achieves better performance than the conventional methods. As a future work, we plan to conduct further experiments for SVC video quality evaluation, and develop a prototype system to validate the proposed algorithm in the real network environment.

Acknowledgments. This work was made possible by a gift from Microsoft Research Asia. And this work was jointly funded by Seoul R&BD Program(WR080951), and by Mid-career Researcher Program through NRF grant funded by the MEST (No. 2010-0000142).

References

1. Liu, Y., Guo, Y., Liang, C.: A survey on peer-to-peer video streaming systems. In: Peer-to-Peer Networking and Applications, vol. 1(1), pp. 18–28 (2008)
2. Liu, J., Rao, S., Li, B., Zhang, H.: Opportunities and challenges of peer-to-peer Internet video broadcast. Proceedings of the IEEE 96(1), 11–24
3. TVAnts, `http://www.tvants.com`
4. SopCast, `http://www.sopcast.com`
5. PPLive, `http://pplive.com`
6. Schwarz, H., Marpe, D., Wiegand, T.: Overview of the Scalable Video Coding Extension of the H. 264/AVC Standard. IEEE Trans. Circuirts and Systems for Video Technology 17(9) (Septemper 2007)
7. BitTorrent, `http://www.bittorrent.com`
8. Erman, D.: Extending bittorrent for streaming applications. In: Proceedings of the 4th Euro-FGI workshop on New Trends in Modelling, Quantitative Methods and Measurements (2007)
9. Joint Scalable Video Model - reference software
10. The Network Simulator 2, `http://www.isi.edu/nsnam/ns/`
11. H.264/SVC Test Sequences. In: Institute of Signal Processing, Hannover University, Germany, `ftp://ftp.tnt.unihannover.de/pub/svc/`

Enterprise Data Loss Prevention System Having a Function of Coping with Civil Suits*

Youngsoo Kim, Namje Park**, and Dowon Hong

Summary. More and more enterprises and organizations are adopting Data Loss Prevention (DLP) systems to detect and prevent the unauthorized use and transmission of their confidential information. Usually, DLP systems identify, monitor, and protect data in use, data in motion, and data at rest through deep content inspection, contextual security analysis of transaction and with a centralized management framework. Electronic documents, e-mails or network logs that are produced within enterprise, can be included in enterprise's confidential information. On the other hand, as ESI is included in extent of evidence that become discovery's target in FRCP taken effect on December 1, 2006, enterprises been always vexing in several litigations are hurrying to adopt systematic ESI administration and confrontation system to prevent a lawsuit from losing owing to failure in duty of presenting related evidences and to maintain their confidences. This paper is about enterprise DLP system having a function of coping with civil suits. We loaded various functions needed at discovery process of civil suit to conventional enterprise DLP system. The proposed system can reduce enterprise's damage by coping spontaneously about enterprise's litigation dispute.

1 Introduction

More and more enterprises and organizations are adopting Data Loss Prevention (DLP) systems to detect and prevent the unauthorized use and transmission of their confidential information. Usually, DLP systems identify, monitor, and protect data in use (e.g., endpoint actions), data in motion (e.g., network actions),

Youngsoo Kim · Dowon Hong
Cryptography Research Team, Electronics & Telecommunications Research Institute
138 GaJeong-no, YuSeong-Gu, Daejeon, 305-700, Korea
e-mail: {blitzkrieg,dwhong}@etri.re.kr

Namje Park
Department of Computer Education, Teachers College, Jeju National University
61 Iljudong-ro, Jeju-si, Jeju-do, 690-781, Korea
e-mail: namjepark@jejunu.ac.kr

* This work was supported by the IT R&D program of MKE/KEIT[10035157, Development of Digital Forensic Technologies for Real-Time Analysis].

** Corresponding author.

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 201–208.
springerlink.com

and data at rest (e.g., data storage) through deep content inspection, contextual security analysis of transaction and with a centralized management framework[1]. Electronic documents, e-mails or network logs that are produced within enterprise, can be included in enterprise's confidential information.

On the other hand, as ESI (Electronically Stored Information) is included in extent of evidence that become discovery's target in FRCP taken effect on December 1, 2006, enterprises been always vexing in several litigations are hurrying to adopt systematic ESI administration and confrontation system to prevent a lawsuit from losing owing to failure in duty of presenting related evidences and to maintain their confidences[2].

FRCP (Federal Rules of Civil Procedure) has a procedure of asking an opposing part to open related evidences and information through discovery[3]. A litigant opens and collects information and evidences to clarify a point at issue of litigation, by legal method out of court in order to prepare trial. By asking each other to open an opposing party's evidences, documents, and witnesses, it can help litigants proceed this lawsuit under the same condition.

Litigants should open all evidences they have by themselves prior to trial and can request the other party or the third party to make public theirs at the same time[4]. The purpose of this requesting procedure for opening evidences is to make clear a point at issue of suit and secure all evidences which might be hidden purposely on trial, and there are a lot of cases that compromise is achieved prior to trial because each party knows about the other party's evidences in detail. Discovery is made in writing such as a written request, a written answer, or a written protest and all documents need a lawyer's signature. This process is fulfilled between litigants without a court's participation. However, if a dispute occurs which a litigant rejects requests of the other litigant, a court participates in it. If litigants make excessive or expensive discovery requests on purpose, a court can revoke them, conversely, they do not their duty of discovery in good faith, a court can imposes mandatory sanctions.

As ESI is included in extent of evidence that become discovery's target in FRCP taken effect on December 1, 2006, terminology named e-Discovery was appeared[5]. Enterprises been always vexing in several litigations are hurrying to adopt systematic ESI administration and confrontation system to prevent a lawsuit from losing owing to failure in duty of presenting related evidences and to maintain their confidences[6].

To satisfy this enterprise's necessity, some tools and solutions that can shorten litigation costs and time radically having various automated functions that is necessary in e-Discovery process are released. However, it can be very inefficient that enterprises which already have their own e-mail and documents management system or information protection solution additionally adopt these tools and solutions, the price increases greatly according to enterprise's scale.

In this paper, we propose an efficient system having a function of integrated security and a function of coping with civil procedure simultaneously that loads new functions needed at discovery process of civil suit to conventional enterprise DLP system.

It includes functions of information management, identification, preservation, collection, processing, analysis, review, production, and presentation of EDRM

(Electronic Discovery Reference Model) and integrated security function of existing enterprise DLP system, as well. For efficiency, we select some modules applicable to preparation of civil suit from conventional enterprise DLP system and use it as common modules of our system[7].

Enterprises which have existing enterprise DLP system can prepare civil lawsuits without adopting other automating tools for e-Discovery additionally[8]. On the other hand, enterprises which have a plan to purchase automating tools or solutions for e-Discovery can save costs using this proposed system having a function of preparing civil lawsuit and a function of integrated security as well.

A general structure of conventional enterprise DLP system is described in chapter 2, and EDRM defining detailed procedures to provide a function of coping with civil suit is introduced in chapter 3. We propose enterprise DLP system having a function of preparing e-Discovery in chapter 4. It is described in separate ways, common modules and unique modules. In chapter 5, we conclude it and show some future plans.

2 Conventional Enterprise DLP System

Fig.1 depicts a structure of general enterprise DLP system. Enterprise's notebooks or PCs at endpoints, storage servers, database servers, or file servers can be a target of this system. To identify, monitor, and protect enterprises' confidential data from these devices, enterprise DLP modules are loaded. Electronic documents, e-mails or network logs that are produced within enterprise, can be included in enterprises; confidential information. Since these devices are linked with networks, and data leakage can happen on networks, DLP functions support networks, too.

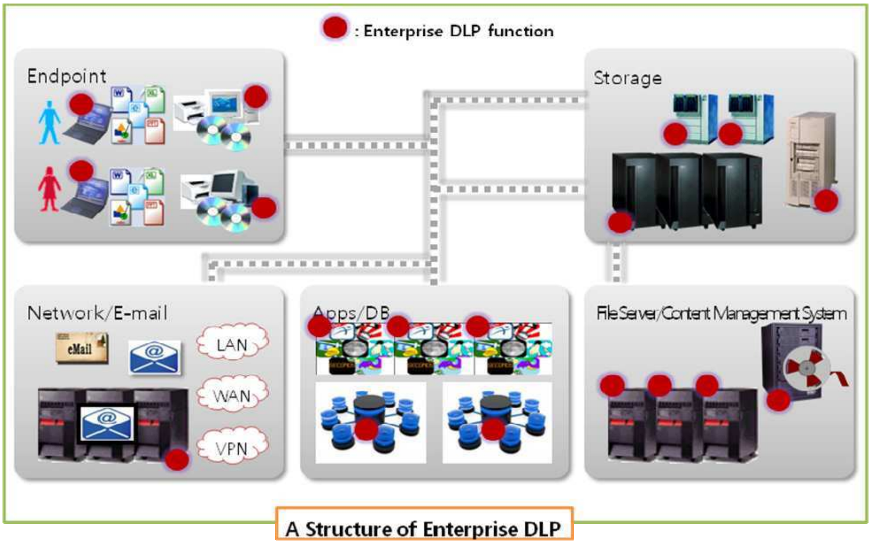

● : Enterprise DLP function
Endpoint
Storage
Network/E-mail
eMail
LAN
WAN
VPN
Apps/DB
FileServer/Content Management System
A Structure of Enterprise DLP

Fig. 1 A structure of enterprise DLP

3 EDRM Defining Detailed Procedures Providing a Function of Coping with Civil Suits

E-Discovery related tools or solutions are designed and made referring EDRM of fig.2. This reference model standardizes proceedings and defines each step's functional specification to effectively follow guidelines and recommendations described in FRCP[9].

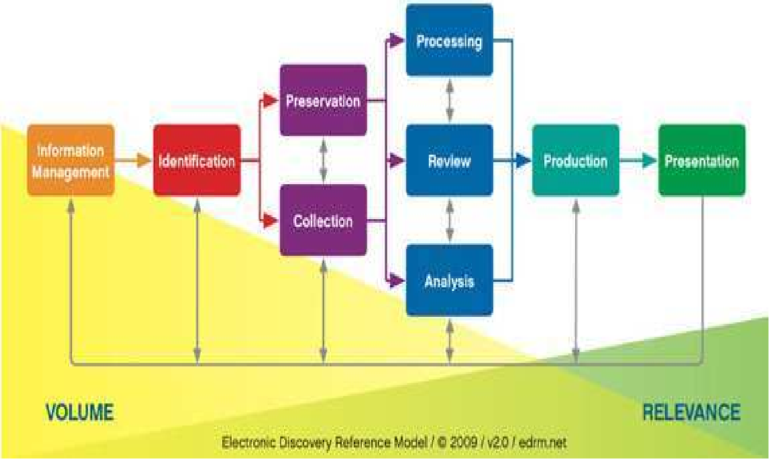


Fig. 2 Electronic Discovery Reference Model

EDRM offers general, scalable, and flexible frameworks to develop e-Discovery related products and services and evaluate them. This is used by general standard about e-Discovery that is authorized, because it was developed by co-works of various related organizations. Tab.1 shows simple descriptions of each function.

Table 1 Functional Specification of EDRM

Function	Description
Information Management	It manages documents of enterprises or organizations with documents control and preservation policy
Identification	As step of deciding a scope of discovery-related documents, it prepares documents can be used in discovery potentially and decides documents should be collected and preserved.
Preservation	It secures that documents do not change or destroy.
Collection	It collects ESI from various media such as tapes, drives, portable storage, networks, etc.
Processing	It filters duplicated or unrelated documents and changes format of ESI to be able to review them more effectively.
Analysis	As step of making related summaries (Related subjects, persons, or documents) by analyzing ESI, it should be done to enhance productivity prior to detailed review step.
Review	As step of establishing strategy on court, it evaluates collected ESI via relations and privileges and selects sensitive documents.
Production	It stores ESI to various media and submits it to a court and opposing litigants.
Presentation	It considers methods that can be seen effectively in trial.

4 Enterprise DLP System Having a Function of Coping with Civil Suits

Fig 3 shows functional modules of our enterprise DLP system. It can be divided into enterprise DLP functional module and e-Discovery supporting functional module.

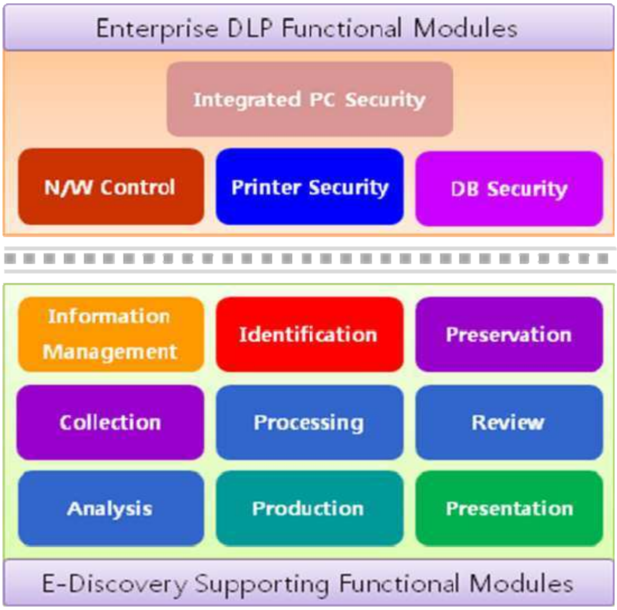


Fig. 3 A Structure of Functional Module

First, enterprise DLP function consists of integrated PC security module, printer security module, network control module, and database security module. Integrated PC security module plays a role of protecting data included at employees' personal computers or notebooks and comprises various functions such as device control, file search, detection of private firewall and hacking tools, file leakage prevention, storing original file, automatic file encryption, notebook/PC taking out management, secure USB management, PC diagnostics, service usage logging, and e-mail attachments control, etc. Printer security module which prevents information leakage from printers, representative output equipment, includes authority control function for printer usage, watermarking function and original copy storing function. Network control module which stops the outflow of information using networks comprises logging function storing contents of e-mails or messengers, usage control function based on services, and keyword-based monitoring and blocking function. Database security module which offers protecting function for enterprise information that is stored at databases includes functions of database access control, certification, audit, and database encryption/decryption, etc.

On the other hand, e-Discovery supporting function consists of information management module, identification module, preservation module, collection module, processing module, review module, analysis module, production module, and representation module. Information management module which makes relationship of each member's roles and business in order to manage data efficiently and produce ESI rapidly and exactly, includes related policy establishment and management function, policy observance monitoring function, automatic ESI classification function, datamap creation and management function, ESI life cycle management function, etc. Identification module which decides a scope of targeting documents for e-Discovery plays a role of preparing potentially available documents and selecting some documents from them to collect and preserve. It consists of various search functions, ECA (Early Case Assessment) function for locations, scopes, and timelines, and reasonable accessibility discernment function, etc.

Preservation module which protects, maintains, and secures documents that obtained from identification step not to be changed or destroyed until production step, includes litigation hold control and management function, preservation policy establishment and management function, and preservation log management function. Collection module directly extracts litigation-related documents identified and preserved from data backup system of enterprise or organization with magnetic tape, hard drive, movable type storage, to examine and analyze. Imaging can be the main function of this module. Processing module shortens size of stored information and changes formats of them properly for analysis and examination, includes ESI indexing function, word frequency list creation function, filtering function by condition, document de-duplication function, document near de-duplication function, metadata extraction function, e-mail thread extraction function, and review format export function, etc.

Review module which examines relativity and privilege for ESI, includes search function and automatic Tagging function, and sampling function, etc. Analysis module which evaluates ESI and makes related summaries such as related subjects, persons, or documents is certainly necessary module for productivity elevation prior to detailed review. This module includes ESI relation analysis function and e-mails/attachments relation analysis function. Production module, that aims to reduce costs, risk, and error and observe specification and timeline, prepares to produce related ESI in available format. This module includes production history log creation and management function, access log creation and management function, specific file format production function, edit function, chain of custody log creation and management function, and load file creation function, etc. Representation module which aims to submit final version of processed ESI as evidence, includes report and diagram creation function.

Fig.4 shows some detailed functions belonging to enterprise DLP system can be used for e-Discovery supporting function and we call them as common detailed functions. E-mail/messenger logging function of network control module logs contents of e-mail or messenger and can be used by information management module of e-Discovery supporting function. File search function on PC which searches files on PCs in various ways from fundamental keyword-based search to contents-based search, can be applied to diverse e-Discovery modules such as identification module, review module, or analysis module, etc. E-mail attachments control function which examine contents of file that is attached in e-mail and classify and save them, can be used to e-mail thread extraction function of e-Discovery's process module or emails/ attachments relation analysis function of e-Discovery's analysis module.

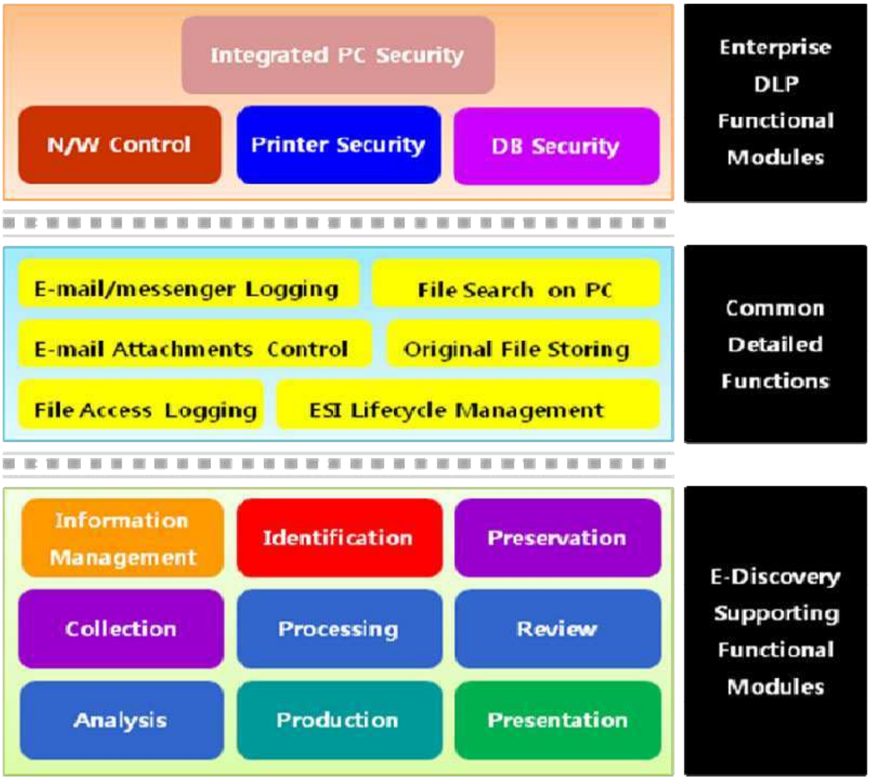


Fig. 4 Common Detailed Functions

Original file storing function which preserves and stores files that exist on various devices in enterprises safely, so that the original text may not be changed, can be used to information management module or preservation module of e-Discovery supporting function. File access logging function which stores logs that authorized persons accessed and used specific files, can be used to information management module and production module of e-Discovery supporting function. ESI lifecycle management function which manages ESI's lifecycle from creating to changing, storing and destroying, can be applicable to e-Discovery's information management module.

5 Conclusion and Further Work

In this paper, we propose an efficient system having a function of integrated security and a function of coping with civil procedure simultaneously that loads new functions needed at discovery process of civil suit to conventional enterprise DLP system. Enterprises which have existing enterprise DLP system can prepare civil lawsuits without adopting other automating tools for e-Discovery additionally. On the other hand, enterprises which have a plan to purchase automating tools or solutions for e-Discovery can save costs using this proposed system having a function of preparing civil lawsuit and a function of integrated security as well.

To provide e-Discovery supporting function, we adopted all procedures of EDRM. We have a plan to reconstruct required modules in order to promote efficiency. We are analyzing various e-Discovery solutions/tools and e-Discovery related project like the Sedona Conference[10]. We expect our new analyzing result can be proposed to more highly sophisticated system.

References

1. DLP. Wikipedia, http://en.wikipedia.org/wiki/Data_loss_prevention_software
2. FRCP. Federal Rules of Civil Procedure, http://www.law.cornell.edu/rules/frcp
3. FRCP. Federal Rules of Civil Procedure. Wikipedia, http://en.wikipedia.org/wiki/FRCP
4. Volonino, L., Redpath, I.: e-Discovery for Dummies. Wiley, Chichester (2010)
5. ESI. Wikipedia, http://en.wikipedia.org/wiki/ESI
6. Cohen, A.I., Edward Kalbaugh, G.: ESI Handbook: Sources. In: Technology and Process, 2010 Edition. Aspen Publishers (2010)
7. Kim, Y., Hong, D., Shin, S.: FRCP and e-Discovery. Weekly Technology Trends. No. 1467 (2010)
8. Kim, Y., Shin, S., Hong, D.: Management, Identification and Preservation of ESI in e-Discovery. In: Digital Forensic Technology Workshop, p. 85 (2010)
9. Electronic Discovery Reference Model, http://edrm.net
10. The Sedona Conference, http://www.thesedonaconference.org
11. Park, N., Kwak, J., Kim, S., Won, D.H., Kim, H.W.: WIPI mobile platform with secure service for mobile RFID network environment. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 741–748. Springer, Heidelberg (2006)
12. Park, N., Kim, S., Won, D.H., Kim, H.W.: Security analysis and implementation leveraging globally networked rFIDs. In: Cuenca, P., Orozco-Barbosa, L. (eds.) PWC 2006. LNCS, vol. 4217, pp. 494–505. Springer, Heidelberg (2006)
13. Park, N., Lee, H., Kim, H., Won, D.: A Security and Privacy Enhanced Protection Scheme for Secure 900MHz UHF RFID Reader on Mobile Phone. In: Consumer Electronics, ISCE 2006, pp. 1–5 (2006)

Optimal Sizing of Hybrid Wind-PV-Tide System

Kun Hyun Park, Chul Ung Kang, and Jong Hwan Lim

Summary. Hybrid generation system is basically merging systems of two or more different types of generation systems. Hybrid Generation System is more effective than utilization of single renewable energy resource. This study presents a methodology to perform the optimal sizing of a new and renewable hybrid generation system. The methodology aims at finding the configuration, among sets of system components, that meets the desired system requirements, with the lowest value of the energy cost. The availability of the methodology is demonstrated with the field data acquired from sets of experiments.

1 Introduction

Hybrid system consists of two or more different types of new and renewable energy generation systems. In recent years, the hybrid generation system has become significant because of the complementary characteristics among the new and renewable energy resources. To use the energy resources of hybrid system more efficiently, the optimal sizing of the hybrid system including battery is very important. However, the sizing of the hybrid system is performed on the basis of experience and intuition, which is not attained the optimum efficiency.

Since solar, wind and tide energy resources have stochastic behavior, the major aspects in the design of the hybrid system are the reliable power supply of the consumer under varying atmospheric conditions and the cost of energy. In order to use wind, solar, tide energy resources more efficiently and economically, the optimal sizing of hybrid system with battery plays an important role in this respect.

Various optimization techniques of hybrid system sizing have been reported in the literature. Kellogg [1] and Chedid [2] reported the linear programming technique. On the other hand, Karaki [3] and Bagul [4] developed the probabilistic approach, and iterative technique was developed by Kellogg [1]. Musgrove [5] presented dynamic programming, and Yokoyama et al. [6] developed multi-objective method. Yang et al. [7] and Beyer et al. [8] have obtained the set of

Kun Hyun Park · Chul Ung Kang · Jong Hwan Lim
Department of Mechatronics
Jeju national University, Jejusi, Korea
e-mail: {3313park,cukang,jhlim}@jejunu.ac.kr

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 209–218.
springerlink.com

different configurations which meet the load using the autonomy level of the system. Protogeropoulos et al. [9] presented general methodology by considering design factor such as autonomy for sizing and optimization.

Recently, Diaf at al. [10] suggest very accurate mathematical approach for characterizing PV module, wind generator and battery.

This paper presents a methodology for the optimal sizing of hybrid Wind-PV-Tide system with storage batteries. The methodology adopted LPSP concept which was presented by Diaf [10]. By modifying the LPSP, we present a methodology to perform the optimal sizing of a new and renewable hybrid generation system. The methodology aims at finding the configuration, among sets of system components, that meets the desired system requirements, with the lowest value of the energy cost. The availability of the methodology is demonstrated with the field data acquired from sets of experiments.

2 Mathematical Modeling of Hybrid System Components

2.1 Modeling of Wind Generator System

There are many types of wind generators that have different power output performance curves, so that the model used to describe the performance of wind generators is expected to be different. Some authors assume that the turbine power curve has a linear, quadratic or cubic form. Other authors approximate the power curves with a piecewise linear function with a few nodes.

In this study, we use the original mathematical model of output power for wind generation system. This may be somewhat different from the actual power curves. The model, however, can be applied any types of wind generation system.

The mathematical model of wind turbine output can be defined as:

$$\mathrm{P_w} = \frac{1}{2}(\rho A \upsilon) \cdot \upsilon^2 = \frac{1}{2}\rho A \upsilon^3 \tag{1}$$

where ρ is air density, A is diameter of rotor, ν is wind speed.

If the height of the wind turbine is different from that of the wind speed measurement, the adjustment of the wind profile for height can be taken into account by using a height adjustment equation. The following power law is applied for the adjustment of the wind profile.

$$\mathrm{V} = \mathrm{V_0}\left(\frac{\mathrm{H}}{\mathrm{H_0}}\right)^{\alpha} \tag{2}$$

where V is the wind speed at hub height H, V_0 is the wind speed measured at the height H_0 , and α is the power law exponent which varies with the climate and environmental conditions. The typical value of 1/7 corresponding to low roughness surfaces and well exposed sites, is used in this study.

2.2 Modeling of PV System

If the solar radiation on the tilted surface, the ambient temperature and the manufacturers data for the PV modules are available, the power output of the PV generator, P_{PV}, can be calculated according to the following equations.

$$P_{PV} = \eta_g N A_m G_t \tag{3}$$

where η_g is the instantaneous PV generator efficiency, N is number of modules, A_m is the area of a single module used in a system and G_t is the global irradiance incident on the titled plane. The instantaneous PV generator efficiency, η_g, is represented by the following equation.

$$\eta_g = \eta_r \eta_{pt}[1 - \beta_t(T_c - T_r)] \tag{4}$$

where, η_r is the PV generator reference efficiency, η_{pt} is the efficiency of power tracking equipment which is equal to 1 if a perfect maximum power point tracker is used, T_c is the temperature of PV cell (°C), T_r is the PV cell reference temperature and β_t is the temperature coefficient of efficiency, raging from 0.004 to 0.006 per °C for silicon cells. However, to simplify the model, we use the general PV generator efficiency that is used many practical approach.

2.3 Modeling of Tide Generator System

There are many types of tidal generation system. Among them the most popular one is a horizontal axis blade type of tidal generation system which is basically no different from the horizontal axis wind turbine system. In this study, the horizontal axis blade type of tidal generation system is assumed. It, therefore, has the same mathematical model as that of wind generation system stated in Eq.(1). That is,

$$P_t = \frac{1}{2} \rho_{sea} A \upsilon^3{}_{cur} \tag{5}$$

where ρ_{sea} is sea water density, A is diameter of rotor, ν_{cur} is the speed of current. Since ρ_{sea} is about $1052.2 \mathrm{kg/m^3}$ the tidal energy is much greater than wind energy when the current speed is equal to wind speed.

2.4 Modeling of Battery System

Since the state of battery is related to the previous state of charge and to the energy production and consumption situation of the system during the time from $t - 1$ to t, it should be modeled differently according to the generation and load conditions.

When the total power from the hybrid generation system is greater than the load required, the battery is in charging state and modeled as follows;

$$\mathrm{C(t) = C(t-1)(1-\sigma) + \left(E_G(t) - \frac{E_L}{\eta_{inv}}\right)\eta_{bat}} \tag{6}$$

Where, $\mathrm{C(t)}$ is battery bank capacity, $\mathrm{E_G(t)}$ is total power of the hybrid system, $\mathrm{E_G(t)}$ is the power needed by the load at time t, σ is self discharge rate of the battery, η_{bat} is the battery efficiency, and η_{inv} is the inverter efficiency. During discharging process, the battery discharging efficiency was set equal to 1, and during charging, the efficiency is 0.65 to 0.85 depending on the charging current.

On the other hand, when total power is less than the load demand, the battery is in discharging state and modeled as follow;

$$\mathrm{C(t) = C(t-1)(1-\sigma) - \frac{E_L(t)}{\eta_{inv}} E_G(t)} \tag{7}$$

3 Design Model of Optimal Sizing

3.1 The RLP Model

Several approaches are used to achieve the optimal configurations of hybrid system in term of technical analysis. In this study, the RLP method is used that is modified from the method of LPSP, and can be summarized in the following steps.

The total power, $\mathrm{P_{tot}}$, generated by the wind turbine, PV generator and tide generator at time t is calculated as follow:

$$\mathrm{P_{tot}(t) = P_{pv}(t) + P_w + P_t} \tag{8}$$

Then, the inverter input power, $\mathrm{P_{inv}(t)}$, is calculated using the corresponding load power requirements.

$$\mathrm{P_{inv}(t) = \frac{P_{load}(t)}{\eta_{inv}}} \tag{9}$$

where $\mathrm{P_{load}(t)}$ is the power required by the load at time t, η_{inv} is the inverter efficiency.

The following two different situations may appear during the operation of the hybrid wind-PV-tide system.

① The total power generated by the hybrid generators is greater than the power needed by the load $\mathrm{P_{inv}}$. In this case, the energy surplus is stored in the batteries and the new battery capacity is calculated using Eq.(5) until the full capacity is obtained, the remainder of the available power is not used.

② The total power generated by the hybrid generators is less than the power needed by the load, $\mathrm{P_{inv}}$, the batteries supply the energy deficit, and a new battery capacity is calculated using Eq.(7).

In case when the total power generated by the hybrid generators is equal to the power needed by the load $\mathrm{P_{inv}}$, the batteries remains unchanged, and this case can be considered as special case of ①.

If the power generated by the hybrid system is less than the load demand, the batteries should supply the energy deficit. However, if the battery capacity reaches to the minimum capacity stat, C_{min}, in which the batteries cannot discharge anymore, the hybrid system can no more supply energy deficit. In this case the power deficit must be supplied from the external energy system. The power deficit in this case is called as ' Lack of power', P_{LP}, and can be defined as:

$$P_{LP}(t) = P_{load}(t)\Delta t - (P_G(t)\Delta t + C(t-1) - C_{min})\eta_{inv} \tag{10}$$

Where, $P_G(t)$and$P_{load}(t)$ are total power and load power requirement. $P_{load}(t)\Delta t$ represents total load demand power, and the last term represents the power consumed by the load. In Eq. (10), it is assumed that power generated by the hybrid system during Δt is unchanged. The amount of power discharged until the battery capacity reaches C_{min}, C_{out}, is written as,

$$C_{out} = P_{load}(t)\Delta t - P_G(t)\Delta t\eta_{inv} \tag{11}$$

The ratio of lack of power (RLP), P_{LP}, for a period T, can be defined as the ratio of total lack of power over the total load required during that period.

$$P_{LP} = \frac{\sum_{t=1}^{T} P_{LP}(t)}{\sum_{t=1}^{T} P_{load}(t)} \tag{12}$$

Using Eq. (12), the optimal sizing of the hybrid system components is performed.

3.2 Economical Model

For hybrid generation systems the most important concern is to achieve the lowest energy cost, and the economical approach can be the best benchmark of cost analysis. Several methods are used to get different options for energy system; the levelised cost of energy is often the preferred indicator [11]. However, the method is not easy to apply in a practical application because it is very complicated.

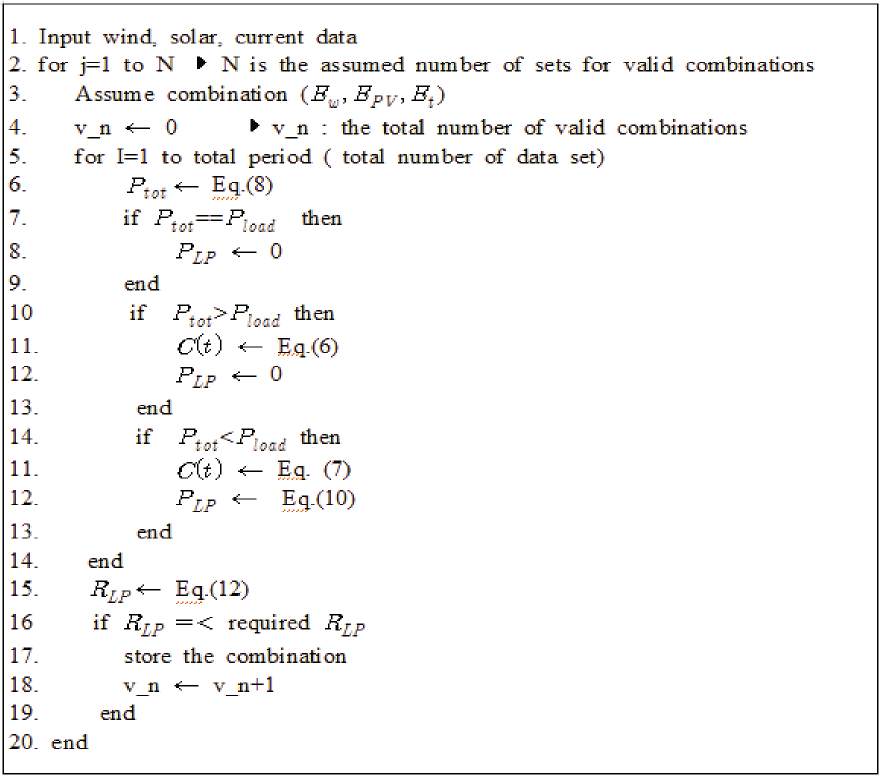
In this study, a simple economical model is developed. Let $T_w, T_{pv}, T_t, T_{bat}$ be costs of wind generation, PV generation, tide generation, and battery per Wh respectively. The total cost per Wh, T_{tot} , can be expressed as follows.

$$T_{tot} = T_w E_w + T_t E_t + T_{pv} E_{pv} + T_{bat} C \tag{13}$$

Where, E_w, E_t and E_{pv} is Wind, tide and PV generation capacity respectively, and C is battery capacity.

4 Simulation Results

Fig.1 shows the optimal design algorithm. First, it assumes combination of the size of each component of the hybrid system. Using the given data, it calculates the total power generated by the hybrid system. The power is then compared with the power required by the load. During the process P_{LP} is calculated and summed for total period T. Finally, R_{LP} is calculated and if the resulting R_{LP} satisfies the required R_{LP}, the assumed combination of the size is a candidate for optimal sizing. Among the many candidates, it finds optimal combination of size by applying the economical model.



```
Input wind, solar, current data
for j=1 to N   ▶ N is the assumed number of sets for valid combinations
    Assume combination ($E_w$, $E_{PV}$, $E_t$)
    v_n ← 0      ▶ v_n : the total number of valid combinations
    for I=1 to total period ( total number of data set)
        $P_{tot}$ ← Eq.(8)
        if $P_{tot}$==$P_{load}$  then
            $P_{LP}$ ← 0
        end
        if  $P_{tot}$>$P_{load}$ then
            $C(t)$ ← Eq.(6)
            $P_{LP}$ ← 0
        end
        if  $P_{tot}$<$P_{load}$ then
            $C(t)$ ← Eq. (7)
            $P_{LP}$ ←  Eq.(10)
        end
    end
    $R_{LP}$← Eq.(12)
    if $R_{LP}$ =< required $R_{LP}$
        store the combination
        v_n ← v_n+1
    end
end
```

Fig. 1 Optimal design algorithm

Fig.2 through Fig. 4 show wind data, solar irradiance data, and tidal current data respectively. The data were acquired for 3 days at somewhere in Jeju island. Design condition is shown in table 1. The load is assumed as 20W and operates for 24 hours a day. The size of wind turbine is set to 3W because it is the smallest one among the commercialized wind system. The required R_{LP} is 0, which meet the stand-alone system that need no external supply of energy.

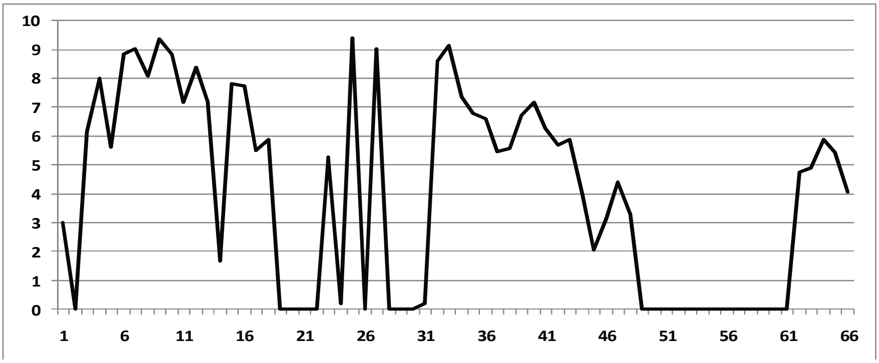


Fig. 2 Wind speed data

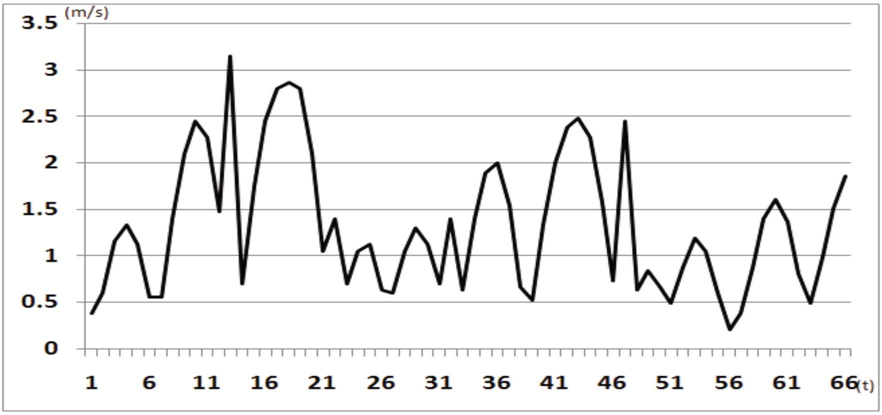


Fig. 3 Current data

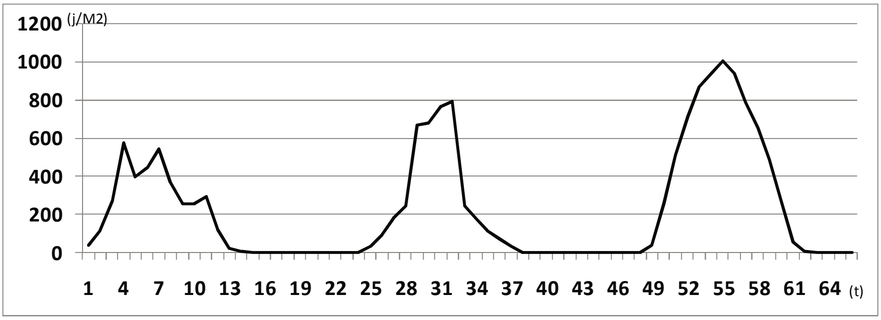


Fig. 4 Irradiance data

Table 1 Design Conditions

type	Unit Size	Total Size	cost(USD/Wh)
wind	3W	3W	0.09
PV	30W	to be designed	0.58
tide	5W	//	0.19
battery	5Wh	//	0.29
load	-	20Wh	-

Fig.5 and Fig. 6 show the design results. In the figures all the dots represent the combinations of sizes that satisfy the required R_{LP}. Among the combinations the optimal one that can satisfies the economical model is shown in table 2. The algorithm yields only one combination for the optimum solution, where the cost of Wh energy is a minimum. These results can be different when the cost of Wh for each component is changed. However, the algorithm can still find the optimal solution with consistency.

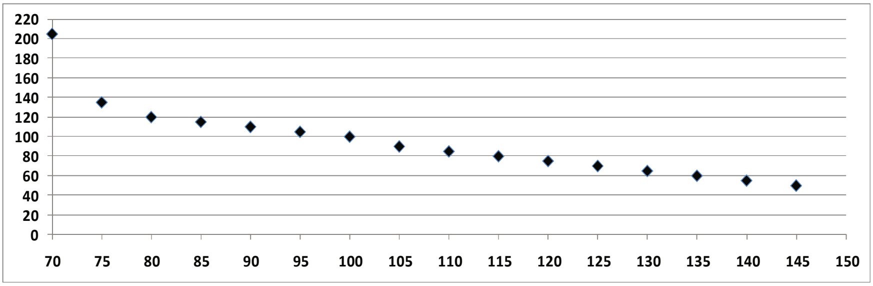


Fig. 5 Design results (PV 30W)

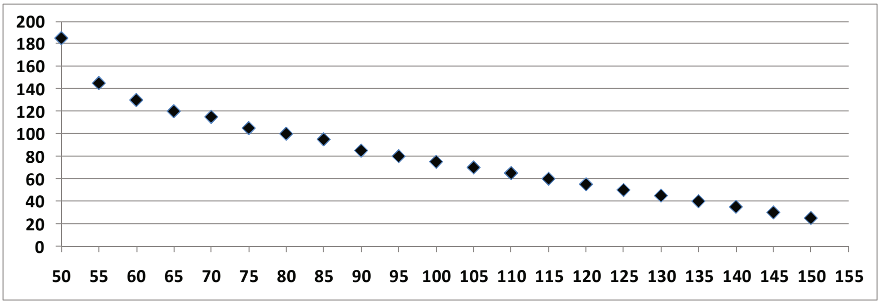


Fig. 6 Design results (PV 60W)

Table 2 Result of optimal sizing

Type	Size	remark
wind	3W	1 set
PV	30W	1 panel
tide	180W	blade radius : 0.107m
battery	80Wh	12V-6.7Ah

5 Conclusions

In this paper, method for optimal sizing of hybrid renewable generation system has been studied. A simple mathematical modeling for each component of hybrid wind-PV-tide generation system was developed. Using the models, the method of optimal sizing of the hybrid system components was developed. The method is based on RLP (Ratio of Lack of Power) and economical model.

This aims at finding the configuration, among a set of system components, that meets the desired system requirements, with the lowest value of the energy cost. The method was applied to hybrid wind-PV-tide generation system. The availability of the methodology was successfully demonstrated with the field data acquired from sets of experiments.

Acknowledgement. This work was supported by the research grant of the Jeju National University in 2006.

References

1. Kellogg, W., Nehrir, M.H., Venkataramanan, G., Gerez, V.: Optimal unit sizing for a hybrid PV/wind generating system. Electric power system research 39, 35–38 (1996)
2. Chedid, R., Saliba, Y.: Optimization and control of autonomous renewable energy systems. Int. J. Energy Res. 20, 609–624 (1996)
3. Karaki, S., Chedid, R., Ramadan, R.: Probabilistic performance assessment of autonomous solar-wind energy conversion systems. IEEE Trans Energy Conv. 14(3), 766–772 (1999)
4. Bagul, A.D., Salameh, Z.M., Borowy, B.: Sizing of stand-alone hybrid PV/wind system using a three-event probabilistic density approximation. Solar Energy 56(4), 323–335 (1996)
5. Musgrove, A.R.D.: The optimization of hybrid energy conversion system using the dynamic programming model - RAPSODY. Int. J Energy Res. 12, 447–457 (1988)
6. Yokoyama, R., Ito, K., Yuasa, Y.: Multi-objective optimal unit sizing of hybrid power generation systems utilizing PV and wind energy. J. Solar Energy Eng. 116, 167–173 (1994)
7. Yang, H.X., Burnett, J., Lu, L.: Weather data and probability analysis of hybrid photovoltaic wind power generation systems in Hong Kong. Renewable Energy 28, 1813–1824 (2003)

8. Beyer, H.G., Langer, C.: A method for the identification of configurations of PV/wind hybrid systems for the reliable supply of small loads. Solar Energy 57(5), 381–391 (1996)
9. Protogeropoulos, C., Brinkworth, B.J., Marshall, R.H.: Sizing and techno-economical optimization for hybrid solar PV/wind power systems with battery storage. Int. J. Energy Res. 21, 465–479 (1997)
10. Diaf, S., Diaf, D., Belhamel, M., Haddadi, M., Louche, A.: A methodlogy for optimal sizing of autonomous hybrid PV/wind system. Energy Policy 35(11), 5708–5718 (2007)
11. Nelson, D.B., Nehrir, M.H., wang, C.: Unit Sizing of Stand Alone Hybrid Wind /PV/Fuel Cell Power Generation Systems. IEEE Power Engineering Society General Meeting 3, 2116–2122 (2005)

A Smart Web-Sensor Based on IEEE 1451 and Web-Service Using a Gas Sensor

Jeong-Do Kim, Jung-Ju Kim, Sung-Dae Park, Chul-Ho Hong, Hyung-Gi Byun, and Sang-Goog Lee

Summary. The purpose of a web-sensor is to transmit recorded data and related information to a remote user. Since the user is at a remote location, the sensor information must be reliable and secure, and the diagnosis for sensors should be easy to handle. To ensure these outcomes, the IEEE 1451 has been used in the past for smart sensor. This paper proposes a new smart web sensor model. Since the proposed smart web sensor is based on IEEE 1451.0, most of the existing sensor interfaces may be used, and the smart web sensor can be achieved using TEDS information. In addition, as XML is used, the web service is user friendly and a remote user can easily handle all kinds of information related to the sensor. This research presents a reference model for a smart web sensor and, to prove how valuable it is, a web-service using a gas sensor is utilized.

1 Introduction

The usage of web-sensors that can obtain sensor information at remote locations through web technology has allowed geographical boundaries to be overcome in automation technology. Existing web-sensors show the sensor information through a static HTML web-page by using TCP/IP and HTTP. Here a web-page is provided to the remote user by a web-server which is directly connected to the web-sensor.

Jeong-Do Kim · Jung-Ju Kim · Sung-Dae Park
Dept. of Electronic Engineering, Hoseo University, Korea
e-mail: jdkim@hoseo.edu, ichromosome@nate.com,
sdpark@control.hoseo.ac.kr

Chul-Ho Hong
Dept. of System and Control Engineering, Hoseo University, Korea
e-mail: chhong@hoseo.edu

Hyung-Gi Byun
Dept. of Information & Communication Eng., Kangwon National Univ., Korea
e-mail: byun@kangwon.ac.kr

Sang-Goog Lee
Dept. of Multimedia System Eng., The catholic University of Korea, Korea
e-mail: sg.lee@catholic.ac.kr

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 219–235.
springerlink.com

In 2003, A. Flammini constructed a web-sensor using an inexpensive microcontroller. Although this web-sensor utilized a cheap microcontroller, it perfectly supported the internet protocol of TCP/IP and HTTP, and provided the sensor information to the user through a HTML page[1]. In 2004, Castaldo wrote an article on a distributed measurement system based on a smart web-sensor, but he focused on aspects of intelligence with regard to the concept of the distributed network rather than the smart sensor itself[2]. Also, in 2005 G. Bucci proposed a web-service method and application component based on the user's purposeful use of XML, and solved compatibility problems between different types of platforms [3].

In 2004, Janecek presented research on how to simplify communications by using SOAP with web-service technology [4]. As mentioned above, several methods have been proposed on how web-sensors should be used and how they should be actualized. But most research has focused on networking methods and how to implement the web- service.

As well as efficiency, one of the most important aspects regarding the purpose of a web-sensor is measurement. Being a sensor, accuracy and specifications are the most important factors, not the web-service. Many sensors have nonlinear elements and limitation factors and show characteristic changes due to factors like temperature. Therefore, the precise specification or information for calibrations should be provided. Recently, sensor standards with such the intelligent information have been established; namely IEEE 1451, a smart sensor standard. In many papers the term 'smart web-sensor' was used, yet no one has exactly defined the term. Many researchers simply used the term based on their research objective.

First, we would like to newly define the term 'smart web-sensor.' The definition of a smart web-sensor is: a smart web sensor that must provide a remote user with sensor status and intelligent information that is guaranteed to be reliable, together with measured data. This information is provided in a web-page format with actual data. And, when received, the remote user can process it based on the user's purposes. According to this definition, a web-sensor reference model based on IEEE 1451 will be presented first. Aside from this paper, few web-sensor researchers have presented a IEEE 1451 based web-sensor, which is constructed in the same manner as the IEEE 1451.2, IEEE 1451.3, and IEEE 1451.4. In this research, the smart web-sensor is structured based on IEEE 1451.0 and can expropriate IEEE 1451.X; and therefore there is no restriction on the interface with sensors. We also used a method to transmit TEDS information in a web-page format for web-service purposes, and all the information and data are transmitted by XML.

In 2003, A. Flammini constructed a web-sensor using an inexpensive microcontroller. Although this web-sensor utilized a cheap microcontroller, it perfectly supported the internet protocol of TCP/IP and HTTP, and provided the sensor information to the user through a HTML page[1]. To prove the efficiency of the proposed method, we implemented a smart web sensor system with a MOS type gas sensor, and designed a web-service with it.

2 Existing Web-Sensors and Web-Service

2.1 Structure of Existing Web-Service

A web-sensor transmits additional information, including notes and the user interface, as well as the measured data to a remote user through the web. This improves convenience of use for the remote user and a web-page can be received by just launching a web-browser. Fig. 1 shows the structure of a web-sensor and Fig. 2 shows an example of a web-page transmitted to a remote place by a web-sensor.

It seems convenient to send a web-page to a remote user, but sometimes the remote user has to process sensor data for his or her purposes. Therefore, XML is commonly used rather than HTML to increase user convenience. However, a web-service with text-based XML doesn't install components every time.

A web service also has to work independently on hardware platforms and program language [5]. XML web-service application software uses protocols like HTTP, XML, XSD (XML Schema Definition), SOAP(Simple Object Access Protocol), and WSDL(Web Services Description Language) to connect to a network [6]. The advantage of XML web-service is that the client does not need to learn the language the web-service is embodied in. The client just needs to know the XML web-service location and user method to obtain the service.

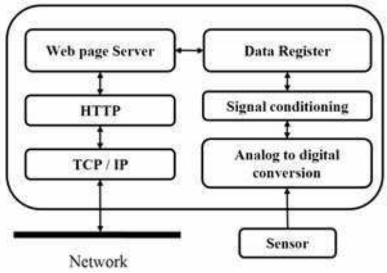
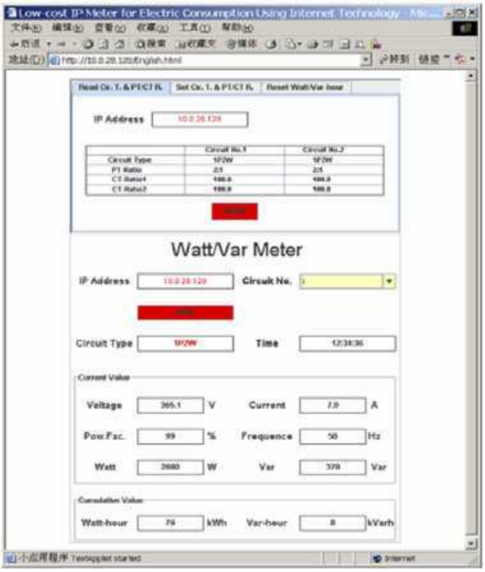


Fig. 1 A block diagram of existing web-sensor

SOAP is a network protocol for data transmission whatever the operating system and programming environment is. SOAP uses XML to distribute data and provides a light protocol for transmission of structural type information between computers. In other words, SOAP encodes data by using HTTP and XML and sends data. SOAP type data is covered by HTTP, so web-servers can understand it, and because of this network application is made easier.

WSD is a XML format which displays the network service. To obtain this service the server should provide WSDL and then the web-service client can find out how to use the web-service based on this document. To use a specific type of service, we must know which web-service is provided and from where it is provided. UDDI(Universal Description, Discovery, and Integration) could be called a search engine for web-services because the produced web-service can be registered on UDDI and the registered web-service can be searched.


Low-cost IP Meter for Electric Consumption Using Internet Technology - Mic...
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹 媒体
地址(D) http://10.0.28.120/English.html
转到 链接
Read Cir. T. & PT/CT R.
Set Cir. T. & PT/CT R.
Reset Watt/Var hour
小应用程序 TestApplet started
Internet

IP Address 10.0.28.120

	Circuit No.1	Circuit No.2
Circuit Type	1P2W	1P2W
PT Ratio	2/1	2/1
CT Ratio1	100.0	100.0
CT Ratio2	100.0	100.0

Watt/Var Meter

IP Address 10.0.28.120 Circuit No. 1

Circuit Type 1P2W Time 12:34:36

Current Value

Voltage 205.1 V Current 7.9 A

Pow.Fac. 99 % Frequence 50 Hz

Watt 2680 W Var 378 Var

Cumulative Value

Watt-hour 76 kWh Var-hour 8 kVarh

Fig. 2 An example of web-page

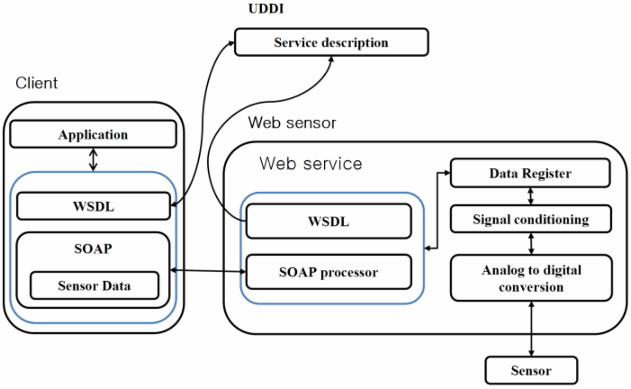


Fig. 3 The structure of web-service using XML

Fig. 3 is the structure of an existing web-service based on XML. XML web-server registers XML web-service to UDDI. When a client goes to UDDI and searches a XML web-service, UDDI returns URL for the XML web-service, then with the URL the client connects to the server, and through the SOAP processor asks for sensor module data as a XML document. The server returns the XML document with sensor data to the client. The client uses the information on the XML document to make an application.

2.2 Problems with Currently Existing Web-Sensors

Aside from efficiency, the most important aspect of a web-sensor is measurement. Being a sensor, accuracy and specification are the most important factors, not the web-service. Many sensors have nonlinear factors and limitation factors and show characteristic changes due to factors like temperature.

Therefore precise specification or information for diagnosis and calibrations should be provided. Recently a sensor standard with such intelligent information capacity has been created: IEEE 1451, which is a smart sensor standard. A detailed explanation concerning IEEE 1451 is given in chapter 3.

3 IEEE 1451 Standard for Smart Sensor

3.1 Introduction about IEEE 1451

IEEE 1451 standards were first developed in 1993 by IEEE standards Association (IEEE-SA) Standards Board's Technical Committee on Sensor Technology (TC-9) and IEEE Instrumentation and Measurement Society. The standard title is “A Smart Transducer Interface for Sensors and Actuators." This provides a standard interface between a transducer interface module (TIM) and network capable application processor (NCAP); it also provides intelligence and interoperability using common TEDS formats

Providing a standard interface brings the following advantages. Sensor and actuator manufacturers only have to provide a standard interface no matter what the network type or connection structure is. And a network can obtain information through a common interface no matter what the transducer type is.

TEDS information brings the following advantages. In the past, when a measurement system was installed and organized, major sensor parameters like measuring range, sensitivity, and magnifying factor had to be inserted so that software could analyze and convert sensor data. But using a sensor with built in TEDS data minimizes possible errors, and increases accuracy and credibility by using built in calibration data to calibrate sensor.

3.2 Basic Structure and Composition of IEEE 1451

IEEE 1451.1 to .4 were standardized by 2006, and in 2007 IEEE 1451.0 and IEEE 1451.5 were standardized. Currently IEEE P1451.6 and .7 are in the works. IEEE 1451.0 was approved as a standard after IEEE 1451.IEEE 1451.1, IEEE 1451.2, IEEE 1451.3, IEEE 1451.4 and IEEE 1451.5 had been approved. Therefore, the other standards have been improved upon. The basic purpose of IEEE 1451.0 is to provide a service based on API, which is the most important standard for a user.

Fig. 4 shows a reference model of IEEE 1451. Through the figure we can learn that IEEE 1451.0 can accommodate the rest of the IEEE 1451.x series. However, IEEE 1451.4 is a standard for an individual sensor only. As it defines a separate interface and uses its own TEDS, direct linkage is difficult. If a linkage is wanted, the user should convert IEEE 1451.4 TEDS to that of IEEE 1451.0 standard. Table 1 gives a brief description about the IEEE 1451 family.

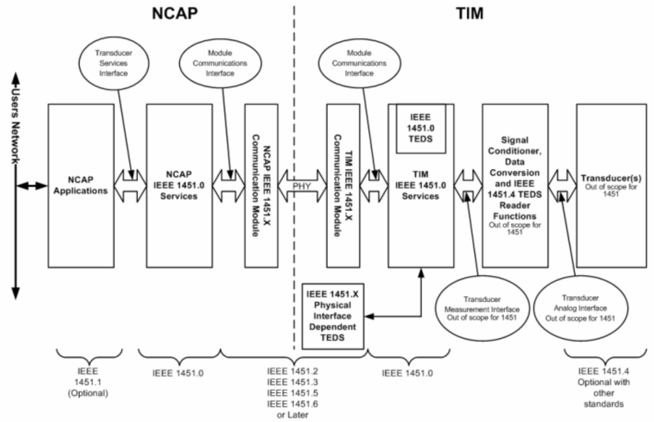

NCAP
TIM
Transducer Services Interface
Module Communications Interface
Module Communications Interface
Users Network
NCAP Applications
NCAP IEEE 1451.0 Services
NCAP IEEE 1451.X Communication Module
PHY
TIM IEEE 1451.X Communication Module
IEEE 1451.0 TEDS
TIM IEEE 1451.0 Services
Signal Conditioner, Data Conversion and IEEE 1451.4 TEDS Reader Functions
Out of scope for 1451
Transducer(s)
Out of scope for 1451
IEEE 1451.X Physical Interface Dependent TEDS
Transducer Measurement Interface Out of scope for 1451
Transducer Analog Interface Out of scope for 1451
IEEE 1451.1 (Optional)
IEEE 1451.0
IEEE 1451.2
IEEE 1451.3
IEEE 1451.5
IEEE 1451.6
or Later
IEEE 1451.0
IEEE 1451.4 Optional with other standards

Fig. 4 Reference model of IEEE 1451

Table 1 Characteristics of IEEE 1451 family and standardization status

standard	Characteristic	Status
IEEE 1451.0	Provides standard API for NCAP	Created in 2007
IEEE 1451.1	Defines NCAP model for smart sensor	Created in 1999
IEEE 1451.2	Connects by TII (transducer independent interface)	Created in 1997
	Converter information presentation by TEDS	
IEEE 1451.3	Connects by multidrop bus	Created in 2003
	Supports TEDS format for distributed multi system	
IEEE 1451.4	Connects by MMI (mixed mode interface)	Created in 2004
	Supports protected TEDS format	
IEEE 1451.5	Sensor interface and protocol by wireless	Created in 2007
IEEE P1451.6	Sensor interface and protocol by controller area network : CAN	in the works <pending>
IEEE P1451.7	Sensor interface and protocol by universal serial bus : USB	in the works <pending>

4 IEEE 1451.0 Based Smart Web-Sensor Design

4.1 Proposing a Smart Web-Sensor Reference Model

The goal of a web-sensor is to provide sensor data and related information in a web-page form to increase remote user convenience. A web-page is usually transmitted in XML format so that a remote user can process the sensor data based on the user's purpose. However, as explained in clause 2.2, the diagnostic technique of sensors is very important in an actual sensor. Moreover, in the case of a remote place, a user cannot actually check the sensor and therefore it is very important to know the sensor condition and to be able to trust the sensor data.

To solve these problems, we would like to newly define the term 'smart web-sensor. A web-sensor is a device that grasps the sensor condition and provides accurate and intelligent information for a remote user, and this information is provided in a web-page format with actual data. When received, the remote user can process it based on the user's purpose.

Based on this new definition, we present the IEEE 1451 based web-sensor reference model, which is shown in Fig. 5. This reference model is designed based on IEEE 1451 in order to provide related information about sensor conditions and calibration information in a standard method using TEDS to the remote user. Few web-sensor researchers have presented the IEEE 1451 based web-sensor, but most are designed in the same method as IEEE 1451.2, IEEE 1451.3, and IEEE 1451.4.

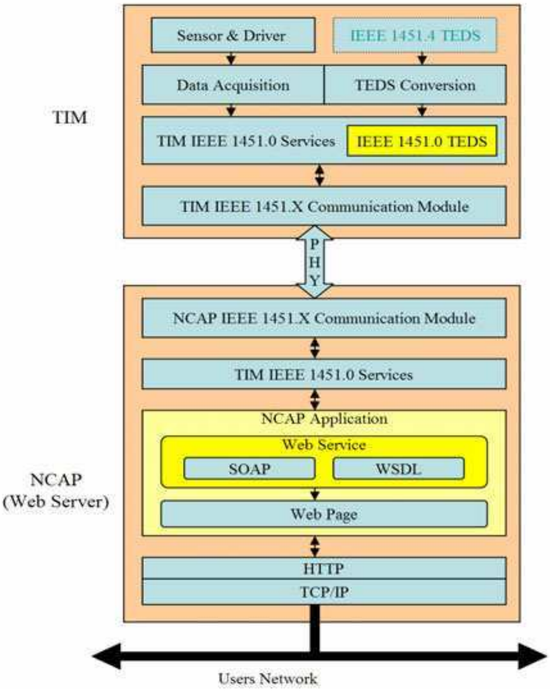


Fig. 5 Proposed smart web-sensor reference model

In this article, a web-sensor is designed based on IEEE 1451.0 to support most communication interfaces, and therefore any form of interface is able to be supported between sensor and server, which means that most of the IEEE 1451.X family can be supported. But IEEE 1451.0 does not support IEEE 1451.4. If IEEE 1451.4 TEDS is desired, a user needs to convert it separately or use a manufacturer-defined TEDS. Since TEDS in relation to an individual sensor is well defined by IEEE 1451.4, it is better to convert IEEE 1451.4 TEDS by a manufacturer-defined TEDS of IEEE 1451.0 than using a transducer Channel TEDS of IEEE 1451.0.

4.2 Presenting a Web-Service Method by Using a Smart Web-Sensor Reference Model

When a remote user receives sensor information, usually the user interface and data are processed based depending on the user's purpose. Then, why should a sensor related program be sent in a complex web-page format when it is going to be re-framed? Such concerns are unnecessary when TEDS information is transmitted in a web-page format as displayed in Fig. 6. TEDS information rarely changes and as the user only refers to it indirectly, a web-page format transmission is desirable, and the information is transmitted by XML.

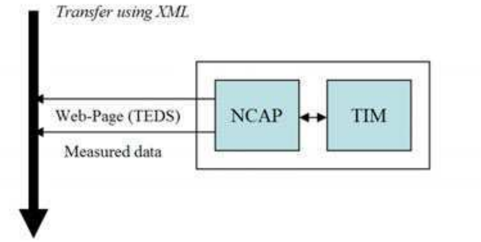
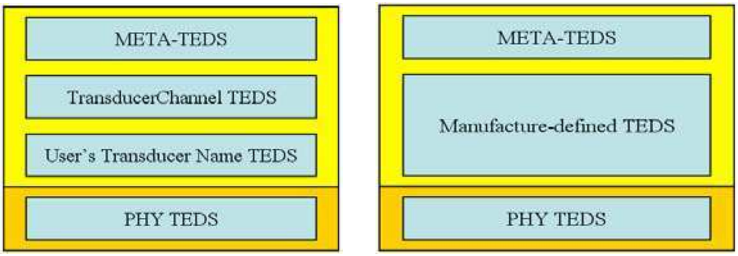


Fig. 6 Web-service sending TEDS information is in a web-page form

Transmitting data by using XML is explained in chapter 2 and will not be repeated here. TEDS information that is implemented by web-page is based on IEEE 1451.0, and therefore TIM should include IEEE 1451.0 TEDS information in its memory. In this case, NCAP has to organize the information into a web-page, which is difficult.

In the other method, TIM doesn't contain TEDS information but NCAP has virtual TEDS in the form of a web-page. In this way it minimizes loads in the actual manufacturing because only this TEDS web-page and measured data is sent in XML form.

The composition of IEEE 1451.0 TEDS information is shown in Fig. 7. In the figure, the User's Transducer Name - TEDS and PHY TEDS - are not the contents to be sent through the web-page and have no relation with the actual sensor, and therefore they will not be explained here.



(a) Not using IEEE 1451.4 TEDS

(b) When IEEE 1451.4 TEDS is composed of Manufacturer-defined TEDS

Fig. 7 Structure of IEEE 1451.0 TEDS

4.3 Design of META TEDS and TransducerChannel TEDS for Web-Service

Table 2 shows contents of META TEDS that will be implemented as web-page.

Table 2 The contents of META TEDS that will be designed as web-page

Field Type	Field Name	Description	
		Length	
3	TEDSID	TEDS Identification Header	
4	UUID	Globally Unique Identifier	
Timing-Related information			
10	OHoldOff	Operational Time-Out	
12	TestTime	Self-Test Time	
Number of implemented TransducerChannels			
13	Maxchan	Number of implemented TransducerChannels	
17	Proxies	TransducerChannel Proxy Definition sub-block	
Types 22,23 and 21 define one TransducerChannel Proxy			
22	ChanNum	TransducerChannel number of the TransducerChannel Proxy	
23	Organiz	TransducerChannel Proxy data-set organization	
21	MemList	TransducerChannel Proxy member list	
25-127	-	open to manufacturers	

Table 3 and 4 shows the contents of TEDS Identification Header and Globally Unique Identifier which is in META TEDS in Table 2. In particular, in the case of a web-sensor, it is necessary to locate the sensor because it is at a remote place.

Table 3 Contents of TEDS Identification Header

Field	Function
Family	IEEE 1451.0
Class	TEDS Access Code
Version	TEDS Version
Tuple Length	Number of octets

Table 4 Contents of Globally Unique Identifier

Field	Description	Description	Comment
1	Location Field	MSB : North(1) or south latitude(0)	
		The next 20 MSBs : the magnitude of the “latitude” as an integer number of arc seconds note : 1 arc second is about 30m	
		The next MSB : East(1) or West longitude(0)	
		The next 20 MSBs : the magnitude of the “longitude” as an integer number of arc seconds	
2	Manufacturer’s Field	Reserved (sensor ID may be included)	
3	Year Field	The year 0 to 4095 AD	manufactured date
4	Time Field	in unit seconds	

Table 5 shows TransducerChannel TEDS which will be included in the webpage. This TEDS is the actual information about the sensor and can be used by users in many ways. For example, by using a lower range limit and upper range limit, an alarm can be sounded when the limit is in excess.

Table 5 TransducerChannel TEDS contents will be presented in web-page form

Field	Field Name	Description	
		TEDS Length	
3	TEDSID	TEDS Identification	
TransducerChannel related information			
10	Calkey	Calibration key	
11	ChanType	TransducerChannel type key	
12	PhyUnits	Physical Units	
50	UnitType	Physical Units interpretation enumeration	

Table 5 (*continued*)

13	LowLimit	Design operational lower range limit	
14	HiLimit	Design operational upper range limit	
15	OError	Worst-case uncertainty	
16	SelfTest	Self-test key	
Data Converter related information			
18	Sample		
40	DatModel	Data model	
41	ModelLenth	Data model length	
42	SignBits	Model significant bits	
Timing-Related Information			
20	UpdateT	TransducerChannel update time (tu)	
21	WSetupT	TransducerChannel write setup time (tws)	
22	RSetupT	TransducerChannel read setup time (trs)	
23	Speriod	TransducerChannel sampling period (tsp)	
24	WarmUpT	TransducerChannel warm-up time	
25	RDelayT	TransducerChannel read delay time (tch)	
26	TestTime	TransducerChannel self-test time requirement	
Attributes			
31	Sampling	Sampling attribute	
48	SampMode	Sampling mode capability	
49	SDefault	Default sampling mode	
Sensitivity (Optional)			
37	Direction	Sensitivity direction	
38	Dangles	Direction Angles	

TEDS from IEEE 1451.0 at each field has its own number and fixed data type. Details of the contents of TEDS must be referred from IEEE 1451.0 standard document. But when this is organized as a web-page, it is necessary for it to be shown with integers, real numbers, and text so users can understand it. Also, when presented on a web-page, the whole defined necessary field doesn't need to be displayed. Depending on the situation, it can be omitted when the user designs the web-page.

4.4 Organization of a Manufacturer-Defined TEDS for Web-Service

It may be very difficult to make a smart sensor by using TransducerChannel TEDS in individual sensors. When data sheet information for each sensor is complex, transducerChannel TEDS of IEEE 1451.0 cannot contain them all. For example, just in the case of MOS gas sensor, there are several specifications which IEEE 1451.0 transducerChannel TEDS couldn't express. To provide a sensor's various electrical and physical characteristics to a user, using an IEEE 1451.4 standard template TEDS is recommended. IEEE 1451.4 TEDS provides standard TEDS from ID=30 to 39 and applies to various sensors.

The user can choose a TEDS that corresponds to the sensor. But when IEEE 1451.0 is used, IEEE 1451.4 TEDS cannot be used directly. IEEE 1451.0 takes most of the other IEEE 1451.x but IEEE 1451.4 is out of its scope.

A standard in order to apply IEEE 1451.4 TEDS to IEEE 1451.0 TEDS has not been presented yet, but it can easily be achieved by using an IEEE 1451.0 manufacturer-defined TEDS. Table 6 shows a standard template TEDS which is standardized in IEEE 1451.4. Table 7 shows an example of template ID at ID=39. A user can choose the desired sensor ID, and can use it by converting it into a manufacturer-defined TEDS. IEEE 1451.4's standard template ID uses fixed data type and bits to express each of their functions, but it doesn't have to be maintained - the manufacturer can decide which data is appropriate and use it accordingly. Only the function and the description must be included.

Table 6 IEEE 1451.4 standard template TEDS

Type	Template ID	Name of Template
Transducer Type Template	25	Accelerometer & Force
	26	Charge Amplifier (w/ attached accelerometer)
	27	Charge Amplifier (w/ attached force transducer)
	28	Microphone with built-in preamplifier
	29	Microphones (capacitive)
	30	High-Level Voltage Output Sensors
	31	Current Loop Output Sensors
	32	Resistance Sensors
	33	Bridge Sensors
	34	AC Linear/Rotary Variable Differential Transformer Sensors (LVDT/RVDT)
	35	Strain Gage
	36	Thermocouple
	37	Resistance Temperature Detectors (RTDs)
	38	Thermistor
	39	Potentiometric Voltage Divider

Table 7 IEEE 1451.4 standard template of ID=39

Function	Select	Description
ID	-	Template ID
Measurement	Select Case—-Physical Measurand	
	Case 0-45	Minimum physical value
		Maximum physical value
Electrical signal output		Transducer Electrical Signal Type
	Select Case—-Electrical Value Precision	
	Case 0	Minimum electrical output
		Maximum electrical output
	Case 1	Minimum electrical output
		Maximum electrical output
	-	Mapping Method
	-	Sensor input impedance
	-	Sensor Response Time
Excitation voltage supply	-	Excitation level, nominal
	-	Excitation level, min
	-	Excitation level, max
	-	Power-supply type
Calibration information	-	Calibration Date
	-	Calibration initials
	-	Calibration period
Misc	-	Measurement location ID

5 Smart Web-Service for Monitoring of Gas Sensor

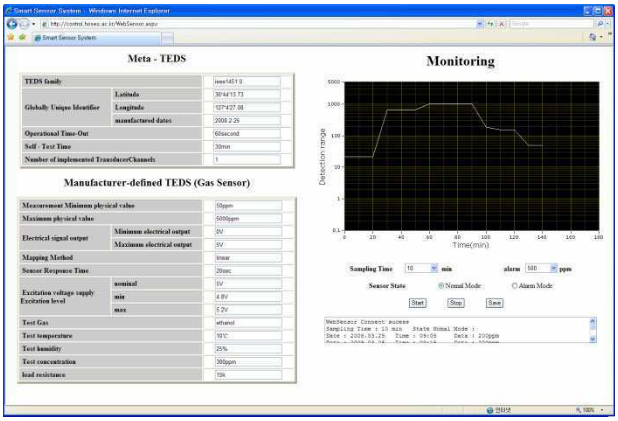
A smart web sensor must provide a remote user with sensor status and intelligent information to ensure reliability, together with measured data. If we deliver TEDS information using IEEE 1451.0 to a remote user, a remote user can diagnose the sensors using this TEDS information, and calibration works are also possible.

In this paper, we designed a smart a web-service using the Figaro gas sensor based on IEEE 1451.0. A designed smart web-service for a gas sensor uses RS485 communications for interface with the web-server. A web-server provides remote clients with the web-page.

In our experiment, a web-server is connected with 1 gas sensor. However, in actual application, a web-server is able to be connected with multiple-sensors and to provide remote clients with multiple web-pages. Fig. 8 shows TEDS information and monitoring display transmitted by a web-server. This web-page can set a base line for an alarm and sampling time based on a transmitted TEDS information. TEDS information and measured data is transmitted by a web-sensor using XML. Setting parts such as sampling time and the alarm is programmed by a remote client, based on transmitted TEDS information.

The client's web-page shown in fig.8 is drawn up based on WSDL documentation written in XML, which is connected with a web-sensor server. The web-server provides TEDS information and real data to the client's web-page through a WSDL document. The web-page of client must indicate TEDS information.

Using TEDS information and real data, various kinds of user-friendly information which is convenient to the user could be shown in the web-page. Since delivered WSDL documents are based on XML, various types of web-page composition are possible according to the demands of the user. The web service and web browser were realized using asp.net in this paper.



Meta - TEDS

TEDS family		ieee1451.0
Globally Unique Identifier	Latitude	36°44'13.73
	Longitude	127°4'27.08
	manufactured dates	2008.2.25
Operational Time-Out		60second
Self - Test Time		30min
Number of implemented TransducerChannels		1

Manufacturer-defined TEDS (Gas Sensor)

Measurement Minimum physical value		50ppm
Maximum physical value		5000ppm
Electrical signal output	Minimum electrical output	0V
	Maximum electrical output	5V
Mapping Method		linear
Sensor Response Time		20sec
Excitation voltage supply Excitation level	nominal	5V
	min	4.8V
	max	5.2V
Test Gas		ethanol
Test temperature		10℃
Test humidity		25%
Test concentration		300ppm
load resistance		15k

Sampling Time 10 min alarm 500 ppm

Sensor State ● Nomal Mode ○ Alarm Mode

```
WebSensor Connect sucess
Sampling Time : 10 min   State Nomal Mode :
Date : 2008.03.25   Time : 09:05     Data : 200ppm
```

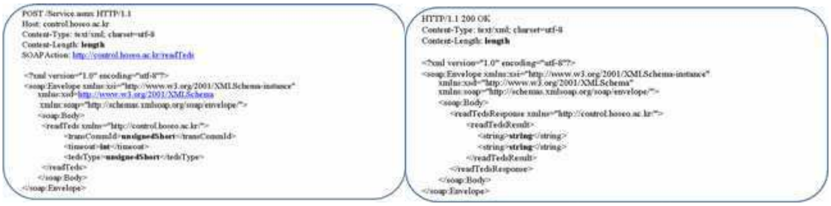
Fig. 8 Real web-page using web-service

In the web-page of this paper, we provide a standard number, the location of manufacture, the manufactured date, META-TEDS information which includes information of sensor channel, and manufacturer-defined TEDS which provides a datasheet of the actual sensors.

We were not able to provide a datasheet of an individual sensor which is based on the IEEE 1451.4, because this research is based on the IEEE 1451.0. In order to solve this problem, a datasheet of a gas sensor based on the IEEE 1451.4 is provided by using the manufacturer-defined TEDS.

In the web-page of this paper, monitoring information is also provided in addition to TEDS information. It is possible to plan the monitoring of information according to the goal of the user by referring to TEDS information, which is already delivered. In this paper, the sampling period of monitoring could be controlled, and it is possible to know the moment when the gas concentration is above the limited value. The graph in fig. 8 shows the monitoring data that was sampled every 10 minutes on Normal Mode. When the web-page is created, information is requested from the server through a SOAP message in order for TEDS information and measured data from a gas sensor to be received, and then the server answers through a SOAP message.

Fig.9 is shows the request and the answer of the SOAP regarding the read Teds method which provides TEDS information to the server. The host control is hoseo.ac.kr, and the transmission protocol is binding with HTTP.



```
POST /Service.asmx HTTP/1.1
Host: control.hoseo.ac.kr
Content-Type: text/xml; charset=utf-8
Content-Length: length
SOAPAction: http://control.hoseo.ac.kr/readTeds

<?xml version="1.0" encoding="utf-8"?>
<soap:Envelope xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xsd=http://www.w3.org/2001/XMLSchema
    xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/">
    <soap:Body>
     <readTeds xmlns="http://control.hoseo.ac.kr/">
            <transCommId>unsignedShort</transCommId>
            <timeout>int</timeout>
            <tedsType>unsignedShort</tedsType>
     </readTeds>
    </soap:Body>
</soap:Envelope>
```

(a) SOAP message for request from client

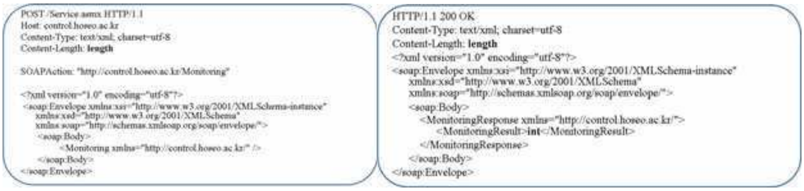
```
HTTP/1.1 200 OK
Content-Type: text/xml; charset=utf-8
Content-Length: length

<?xml version="1.0" encoding="utf-8"?>
<soap:Envelope xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema"
    xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/">
    <soap:Body>
      <readTedsResponse xmlns="http://control.hoseo.ac.kr/">
          <readTedsResult>
             <string>string</string>
             <string>string</string>
          </readTedsResult>
      </readTedsResponse>
    </soap:Body>
</soap:Envelope>
```

(b) SOAP message for response of web-server

Fig. 9 SOAP message for request and response of readTeds method

Fig.10 shows the request and the response of SOAP regarding the monitoring method which provides the monitoring data information of the server. Through the monitoring method, measured data is provided in a basic manner and various kinds of information can also be provided according to the needs of the user.



```
POST /Service.asmx HTTP/1.1
Host: control.hoseo.ac.kr
Content-Type: text/xml; charset=utf-8
Content-Length: length

SOAPAction: "http://control.hoseo.ac.kr/Monitoring"

<?xml version="1.0" encoding="utf-8"?>
 <soap:Envelope xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema"
    xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/">
     <soap:Body>
          <Monitoring xmlns="http://control.hoseo.ac.kr/" />
     </soap:Body>
</soap:Envelope>
```

(a) SOAP message for request from client

```
HTTP/1.1 200 OK
Content-Type: text/xml; charset=utf-8
Content-Length: length
<?xml version="1.0" encoding="utf-8"?>
<soap:Envelope xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema"
    xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/">
    <soap:Body>
      <MonitoringResponse xmlns="http://control.hoseo.ac.kr/">
         <MonitoringResult>int</MonitoringResult>
      </MonitoringResponse>
    </soap:Body>
</soap:Envelope>
```

(b) SOAP message for response of web-server

Fig. 10 SOAP message for request and response of monitoring method

6 Conclusion

In this research we defined in a new way what a smart web-sensor is. A smart web-sensor must provide a remote user with the sensor status and the intelligent information to ensure reliability, together with measured data. This information is provided in a web-page format with actual data. And, when received, the remote user can process it based on the user's needs. According to this definition, a web-sensor reference model was proposed based on IEEE 1451.0. It is possible to support most sensor interfaces based on IEEE 1451.X, because the proposed reference model is planned based on IEEE 1451.0.

In order to achieve the intelligence aspect of the sensor, information regarding the measured data and also information regarding TEDS are provided to a user in a remote place. The information transmission is realized using WSDL and SOAP based on XML. The use of XML provides flexibility in choosing a platform for the client and also achieves benefits regarding security matters. The benefit of flexibility may be maximized with proper planning, especially in the case of the use of gas sensors, wind direction sensors, BOD(biological oxygen demand) and COD(chemical oxygen demand) sensors which are located in remote places.

For chemical sensors (e.g. gas sensor), TEDS information is necessary in order to provide the reliability and the proper diagnosis of sensors. In order to show the usability of the smart web-sensor proposed in this paper, we designed a gas smart web- sensor and it was applied successfully. In order to show the usability of the smart web-sensor proposed in this paper, we designed a gas smart web- sensor and it was applied successfully.

The measurement of the sensor is more accurate and consequently it will be able to increase the reliability of the system, because we can transmit the difference between properties and manufacturer sensors by using TEDS. The most important thing is to augment the reliability of the sensors which are located in a remote place.

References

1. Flammini, A., Ferrari, P., Sisinni, E., Marioli, D., Taroni, A.: Sensor integration in industrial environment: from field bus to web sensors. Computer Standards & Interfaces 25, 183–194 (2003)
2. Castaldo, D., Gallo, D., Landi, C.: Collaborative Multisensor Network Architecture Based On Smart Web Sensors For Power Quality Applications. In: IMTC 2004, pp. 1361–1366 (2004)
3. Bucci, G., Ciancetta, F., Fiorucci, E., Gallo, D., Landi, C.: A Low Cost Embedded Web Services for Measurements on Power System. VECIMS 2005, 7–12 (2005)
4. Janecek, J.: Efficient SOAP processing in embedded systems. In: IEEE International Conference and Workshop on the Engineering of Computer-Based Systems, ECBS 2004 (2004)
5. Institute of Electrical, Electronics Engineers, IEEE Standard for a Smart Transducer Interface for Sensors and Actuators-Common Functions, Communication Protocols, and Transducer Electronic Data Sheet (TEDS) Formats, IEEE Std 1451.0-2007
6. Ding, H., Zhang, B., Ding, Y., Taob, B.: On a novel low-cost web-based power sensor via the Internet. Sensors and Actuators A: Physical 136(1), 456–466 (2007)
7. Mark Birbeck, Professional XML, wrox
8. Microsoft, Developing XML Web Services and Server Components with Microsoft Visual Basic. NET and Microsoft C#. NET Microsoft Press,Washington (2003)
9. National Institute of Standards and Technology, IEEE 1451 Website, http://ieee1451.nist.gov/
10. FIGARO Gas sensor TGS 2620datasheet, http://www.figarosensor.com/
11. Institute of Electrical, Electronics Engineers Inc., IEEE Std 1451.1-1999, IEEE Standard for a Smart Transducer Interface for Sensors and Actuators Network Capable Application Processor (NCAP) Information Model Institute of Electrical, Electronics Engineers Inc., New York(1999)
12. Institute of Electrical, Electronics Engineers Inc., IEEE Std 1451.2-1997, IEEE Standard for a Smart Transducer Interface for Sensors and Actuators - Transducer to Microprocessor Communication Protocols and Transducer Electronic Data Sheet (TEDS) Formats Institute of Electrical, Electronics Engineers Inc., Piscataway (1997)
13. Institute of Electrical, Electronics Engineers Inc., IEEE Std 1451.3-2004, IEEE Standard for a Smart Transducer Interface for Sensors and Actuators—Digital Communication and Transducer Electronic Data Sheet (TEDS) Formats for Distributed Multidrop Systems Institute of Electrical, Electronics Engineers Inc., New York (2004)

14. Institute of Electrical, Electronics Engineers Inc., IEEE Std 1451.4-2004, IEEE Standard for A Smart Transducer Interface for Sensors and Actuators Mixed-Mode Communication Protocols and Transducer Electronic Data Sheet (TEDS) Formats Institute of Electrical, Electronics Engineers Inc., New York (2004)
15. World Wide Web consortium, `http://www.w3c.org`
16. UDDI Specification –OASIS Group, `http://www.uddi.org/`
17. WSDL 1.1 Specification (March 2001), `http://www.w3.org/TR/wsdl`
18. SOAP 1.2 Specification (June 2003), `http://www.w3.org/TR/soap/`
19. Kim, J.-D., Kim, D.-J., Byun, H.-G., Ham, Y.-K., Jung, W.-S., Han, D.-W., Park, J.-S., Lee, H.-L.: The definition of basic TEDS of IEEE 1451.4 for sensors for an electronic tongue and the proposal of new template TEDS for electrochemical devices. Talanta, Septemper 7 (2006)
20. Sadok, E.F., Liscano, R.: A Web-Services Framework for 1451 Sensor Networks. In: Instrumentation and Measurement Technology Conference, vol. 17-19, pp. 554–559 (May 2005)
21. Microsoft, Developing XML Web Services and Server Components with Microsoft Visual Basic. NET and Microsoft C#. NET, Microsoft Press, Washington (2003)

CC Based Analysis Scheme for Evaluation Scope Models

Sun-Myung Hwang and Young-Hwan Bang

Abstract. In these days, many organizations try to manage their information system in safe way due to more rapidly change in information security system. The CC (Common Criteria) is scheme to secure evaluation for information security product/system. And the CC was approved by ISO/IEC 15408 in June, 1999 as international standard for information security system evaluation. The UK established C-TAS (CESG Tailored Assurance Service) that evaluate to IT product and software, and operational system. The Japan developed ISO/IEC 19791 for information security operating system security evaluation. Thus, we are preparing operating system evaluation. This paper is to propose evaluation scope computation model related with operating system evaluation to be enforced in the future.

1 Introduction

In these days, CCRA (Common Criteria Recognition Arrangement) is system to evaluation secure and reliable for information security product. And, that established in many of countries, such as America, UK, and France, etc. CCRA does CC and CEM base. Recognize evaluation result for information security product mutually. The CC was approved by ISO/IEC 15408 June, 1999 as international standard for information security system evaluation [5]. CC developed current v3.1 and is evaluated from September, 2009 to CC 3.1.

The UK established C-TAS (CESG Tailored Assurance Service) that evaluate to IT product and software and operational system. The Japan developed ISO/IEC 19791 for information security operating system security evaluation.

Each advanced countries are evaluation scope guide about target of evaluation (TOE). However, evaluation scope for composed model (i.e., operational system model) is complicated and was dispersed. Thus, that is hard to decide evaluation scope for composed model (i.e., operational system model).

According to, this problem, we establish to 3-dimensional evaluation scope model by related in security function of operational system.

Sun-Myung Hwang · Young-Hwan Bang
Department of Computer Engineering Daejeon University
Korea Institute of Industrial Technology, Korea
e-mail: sunhwang@dju.ac.kr

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 237–244.
springerlink.com

Therefore, we need study on evaluation scope computation model for operational system. This paper analyzed evaluation scope of schemes (i.e., ISO/IEC 19791, ISO/IEC 15408 and C-TAS). And we proposed by methods of 3-dimension evaluation scope model for operational system.

In this paper, explain about analysis each current scheme for evaluation scope model in chapter 2. And we describe definition of evaluation scope for composed model (i.e., operational system model) in chapter 3. Finally, chapter 4 has conclusion.

2 Current Guides and Schemes for Evaluation Scope Model

2.1 ISO/IEC TR 19791

ISO/IEC 19791 approved by DTR in Vienna meeting in last 2005, and was confirmed finally by technology document on July, 2005 [1].

ISO/IEC 19791 is document that emphasizes in product evaluation criteria for security evaluation of operational system. And, that referenced by ISO/IEC 17799 and CC.

ISO/IEC 19791 is simple, but there is evaluation scope guideline for operational system. According to this guideline, composes to security domain more than 1 to 'Security Domain' concept and need PP and ST in each security domain and should evaluate separately. Each domain constructs through various common use or itself development component. Operational system specifies physical, logical scope and is evaluated including security function within operating system in equal domain located

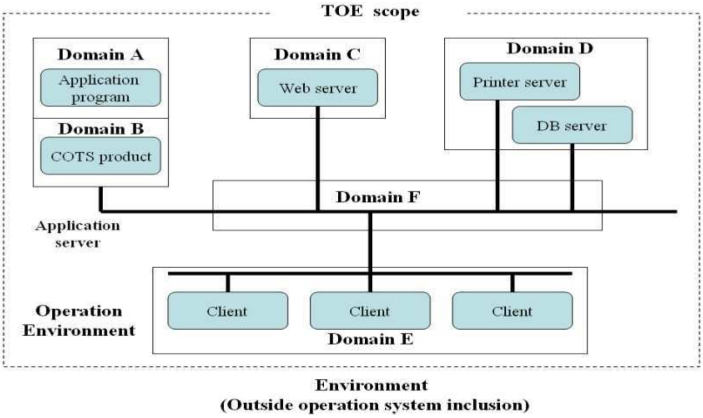


Fig. 1 Example of Domain

2.2 CC/CEM (ISO/IEC 15408)

CC is international standard for information security system evaluation. And, that was approved on June, 1999. Recently, that was revised by 3.1 in v2.3. The fundamental of CC is categorizing universal set of security function requirement doing requisitely in all information security systems hierarchically. Also, is categorizing universal set of assurance requirement hierarchically for accuracy of embodiment about security function. Addition to contents of composed component evaluation by CC v3.1 revision. So, what is composed component? That is combined to this use complete to evaluate product IT substance more than two. Figure 2 is example of composed TOE, that consists of basis component provide service and dependant component that offer service [6].

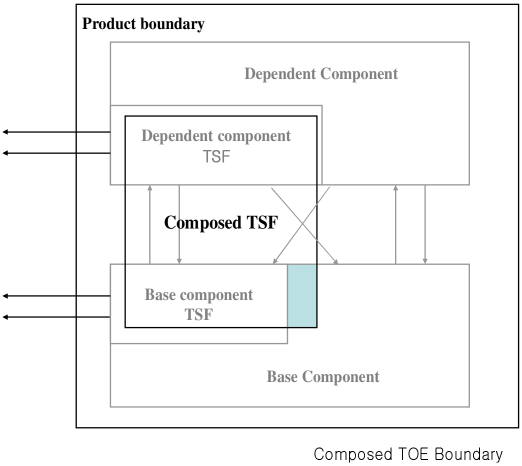


Fig. 2 Composed TOE abstractions

CEM is security evaluation and guidance about CC. And, include contents on minimum evaluation action item that when evaluate using standard and estimation proof that is defined in CC, evaluator should accompanies [2].

2.3 C-TAS

UK-IT security evaluation and certification system (UK-ITSEC) established C-TAS (CESG Tailored Assurance Service) to evaluate and warrants confidence for security property of Information Technology product and system. Following a fundamental reassessment of CESG Assurance Services influenced by current threats to HMG IT systems and valuable feedback from customers, the CESG Tailored Assurance Service was introduced in June 2007. This new flexible service takes the best from and replaces the existing Fast Track and System Evaluation (SYSn) services. It is designed to meet the needs of HMG Infosec Standard No.1 Residual Risk Assessment Method (IS1) [3, 4].

The service is intended for a wide range of IT products and systems ranging from simple software components to national infrastructure networks. Therefore, a toolbox of activities is provided that enables each evaluation to be tailored as appropriate. A summary of these components is provided in the table 1 below [3, 4].

Table 1 Assurance activities for C-TA

NO	Assurance Activities
1	Development Procedures Review
2	Product Functionality & Design Assessment
3	System Architecture and Design Review
4	Security Functional Testing
5	Installation & Operational Procedures
6	Vulnerability Analysis & Testing
7	Source Code Analysis
8	Assurance Maintenance Review

Under Figure 3 shows background of TOE security requirement. Outside of dot boxes are outside scope of TOE. Showing product include of interaction to TOE. And, each service systems have self-function and well-define interface. Between subsystems interaction is accomplished by interface.

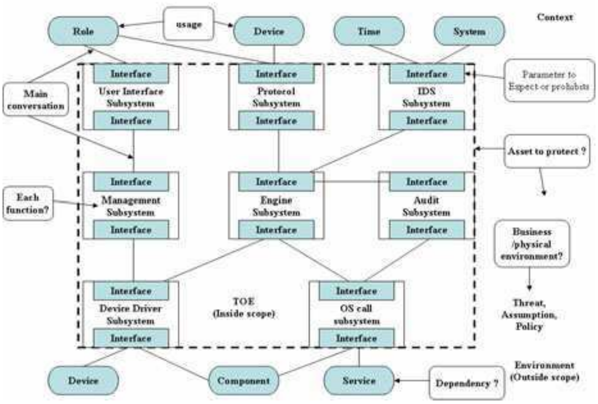


Fig. 3 Example of Background for security requirement by TOE

3 Definition of Evaluation Scope

Operation system have gotten put together complex hardware and software. Also, servers and various devices were scattered. And Operation system difficult evaluation scope selection because of many domains was contained hierarchically is not too easy. Therefore, should decide by TOE is involved with security function of operation system in principle. Suggests and defines three-dimensional evaluation scope model method with Figure 4. Table 2 gives three-dimensional evaluation scope model's example. When saw as temporal scope, TOE 1 is two current points of time and TOE 2 is model including operation duration to TOE 1.

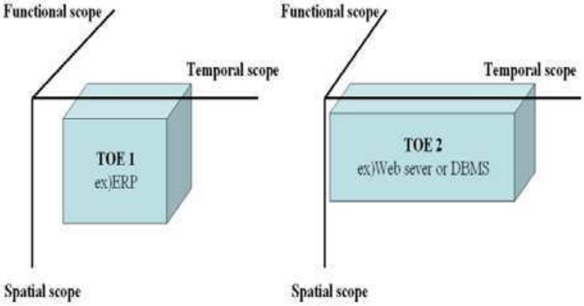


Fig. 4 Three-dimensional evaluation scope model

Table 2 Evaluation scope's example by three-dimensional evaluation scope model

Dimensional TOE	Temporal	Spatial	Functional
TOE 1	Current time point	HW that is scattered on the Internet	Security function in personage and FM
TOE 2	Inclusion operation duration	Web server in computer center	Security function in personage
TOE 3	Inclusion development duration	Inside smart card chip	Security function in card OS

3.1 Temporal Scope

Temporal scope life cycle (Analysis-Design-Implement-Test-Operation) duration of operation system or it means time point.

- Duration scope: Development duration, testing duration and operation duration.
- Evaluation time point: Evaluate Snapshot at evaluation time point of TOE. Evaluate development, test and information assurance level of applied current time point of information system.

3.2 Spatial Scope

Spatial scope is hardware that composed operation system. For example, spatial scope is the Internet because information system is operated to Internet base. Below, Figure 5 shows form of appropriate domain within physical scope. Figurer of spatial scope is same with Figure 1.

3.3 Functional Scope

Functional scope is deciding scope of TOE in functional side. TOE has various structures according to viewpoint in functional side. One TOE decides evaluation scope according to structure viewpoint, component viewpoint and function viewpoint.

3.3.1 Structure Viewpoint

Structure viewpoint include security function offer part in information system property. Security function operates because is scattered to hardware, system software, application software, and directory. Figure 5 shows structure viewpoint.

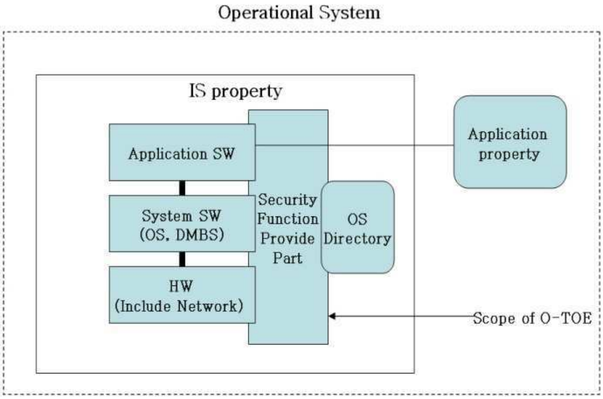


Fig. 5 Structure viewpoint

3.3.2 Component Viewpoint

Component viewpoint include security function and security-related function. And when decide TOE viewpoint apply to component viewpoint. Figure 6 shows component viewpoint.

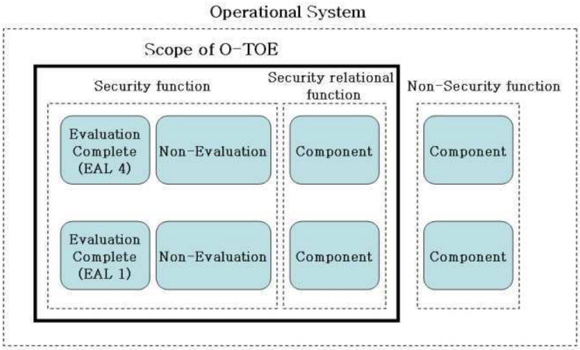


Fig. 6 Component viewpoint

3.3.3 Function Viewpoint

Function viewpoint includes all technology function and operation function. Specially, security function is function that is been common in all application functions. Figure 7 shows functional viewpoint.

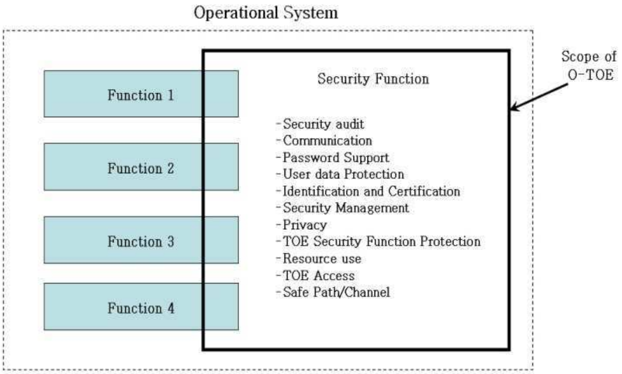


Fig. 7 Function viewpoint

3.3.4 Logical Domain Viewpoint

TOE is consisted of 'Logical domain' more than 1 and 'Connotation style domain' is included on domain inside. Figure 8 shows logical domain viewpoint

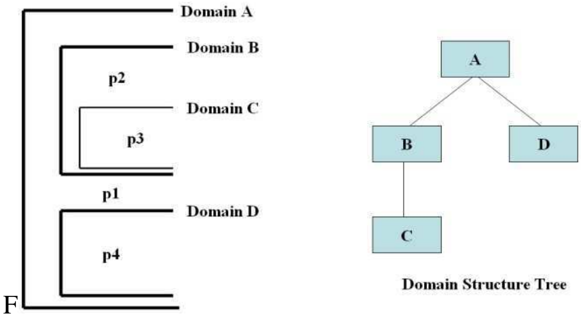


Fig. 8 Logical domain viewpoint

4 Conclusions

In this paper, we propose to computation model for evaluation scope of operation system. Thus, we selected ISO/IEC TR 19791, CC/CEM, and C-TAS by guideline for evaluation scope computation model proposal.

But, that is difficult to definition for scope TOE of operation system. Therefore, we must apply operation system evaluation scope computation model through over proposed evaluation scope model.

We suggest this study finding so that can be used in operation system security evaluation. And, that be carried out standardize operation system evaluation scope to base in the future. Also, this study should be proceeded so that can prove evaluation method relate with evaluation scope and level.

References

1. ISO/IEC TR 19791: 2006, Information technology – Security techniques – Security assessment of operational systems (May 15, 2006)
2. ISO/IEC 15408, Common Criteria for Information Technology Security Evaluation, Part 1, 2, 3, Version 3.1 (September 2006)
3. CESG, `http://www.cesg.gov.uk`
4. IACS (Information Assurance & Consultancy Services) , `http://www.cesg.gov.uk/products_services/iacs/ctas/index.shtml`
5. C.C., `http://www.commoncriteriaportal.org`
6. KISA (Korean Information Security Agency), Guide of Information Security System evaluation and certification (2006)

Integrating User-Generated Content and Spatial Data into Web GIS for Disaster History

Daniel Leonardo Niko, Hyunsuk Hwang, Yugyung Lee, and Changsoo Kim*

Abstract. Information on the damage area is critical in prompt disaster response. This information is supposed to be produced by organizations involved in disaster management. In reality, due to limited resources, comprehensive and updated data are not easily obtainable. Therefore, an alternative way to collect data is urgently needed. General public can play an important role in producing these types of data, but there are some drawbacks to resolve in this approach including validity and integration with existing spatial data. This paper presents a framework in performing geoprocessing to validate geographic references submitted by users and integrate them with the existing spatial database, i.e., parcel, hydrology, facility. The proposed framework will surely lead us to create a more complete spatial data of the damage that are to be utilized in a web based GIS application for effective disaster management.

Keywords: User generated content, Spatial Data, Disaster Management, Web GIS.

1 Introduction

More reliable and up-to-date spatial data are required for proper disaster responses by citizens and civil worker. Road networks, buildings, hospitals, fire stations, medical emergency stations, damaged areas and their associated attribute data are

Daniel Leonardo Niko · Hyunsuk Hwang
Interdisciplinary Program of Information Systems,
Pukyong National University, Korea
e-mail: daniel.gultom@gmail.com, hhs@pknu.ac.kr

Yugyung Lee
Dept. of Computer Science and Electrical Engineering,
University of Missouri at Kansas City, USA
e-mail: leeyu@umkc.edu

Changsoo Kim
Dept. of IT Convergence and Application Engineering,
Pukyong National University, Korea
e-mail: cskim@pknu.ac.kr

* Corresponding author.

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 245–255.
springerlink.com

some examples of required datasets for disaster management system [1]. Some of these datasets are static while others are dynamic and need continuous updates. Damaged areas need to be regularly observed and updated after the occurrence of a disaster. To achieve this, communities play an integral role in complementing the existing database from local government. Government can use contents submitted by publics and combine them with the existing spatial database to generate up-to-date information on past or ongoing disasters.

With the increasing use of smartphones equipped with GPS and wireless connectivity, we envision that disaster management systems would gain enormous benefits from this new trend. In addition, Geotagging, that is an advanced technique used for geographic references associated to various media types (texts, photos, sounds), has recently gained its popularity [2]. Geotagging disaster related content can be easily collected by providing users with mobile GIS applications on smartphones. Through this approach, a community can be empowered to act as one of the actors in sharing and reporting disasters happening in their areas. User reports should be integrated with existing spatial data. However, integration of spatial data is not a trivial task. In integrating spatial data with user submitted data, numerous important issues must be taken into account, such as different projection, different scales, or different topographic sources.

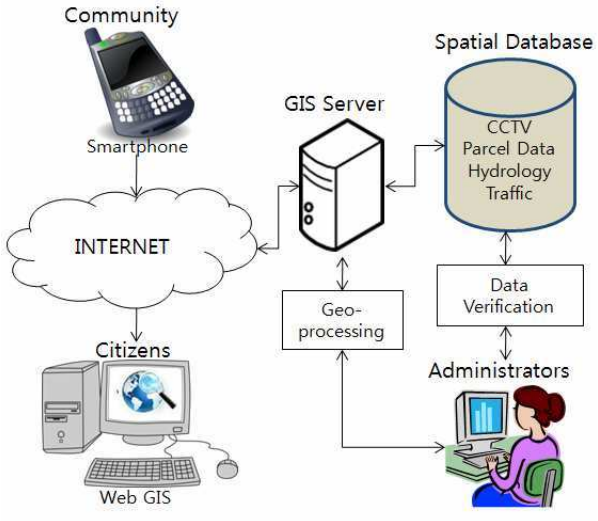


Fig. 1 Overview of the system

In this paper, we propose a system to combine disaster related contents, such as photos, locations, and descriptions that are produced by users with a spatial database provided by local government to accumulate better information on damaged areas, including the affected parcel, proximity with hydrology and public facilities, and traffic. As results, advanced searching can be provided for the analysis of potential risks associated with disasters such as dangerously flooded areas with regards to their proximity to rivers. Fig.1 depicts the system overview.

Community post the content using a web service hosted on the GIS server that is also a web server. Geo-processing will verify data to check validity as well as to integrate the content with the existing spatial database. The generated spatial data of damage is accessible to citizens through web services on GIS server.

The rest of the paper is organized as follows. Section 2 will describe the technical background of the proposed framework. In section 3, the details of system development will be described. This paper concludes with a conclusion and future research in Section 4.

2 Related Work

In this section, the related work will be presented. Furthermore, how these technologies support the proposed system will be described.

2.1 Disaster Management

Disaster management is defined as a cycle of activities including mitigation, preparedness, response and recovery [1]. All data are integral for whole phases in disaster management. These data can be defined in two categories and they are:

- Pre-disaster baseline data about the disaster history and risks of disaster
- Post-disaster real-time data about the impact of a disaster and the resources available to manage it

The ability to make a right decision on disaster management can be greatly enhanced by completeness of the information. However, information management and processing in disaster management are challenging due to the unique combination of characteristics of the data in this domain, which include [3]:

1. A large number of producers and consumers of information
2. Time sensitivity of the exchanged information
3. Various levels of trustworthiness of the information sources
4. Lack of common terminology
5. Combination of static (e.g., maps) and dynamic (e.g., damage history) datasets
6. Heterogeneous formats, ranging from free text, XML and multimedia data

Pre-disaster baseline data, such as damage history, are important in performing comprehensive spatial analysis for disaster management. It is noted that disaster or damage history needs a considerable amount of resources to produce and disseminate (1), to keep updated (2), to validate (3) and to integrate (4,5,6).

2.2 User-Generated Disaster History

The main idea of user-generated content is the paradigm that considers the user as not only the consumer but also the producer of the content. Social applications, such as Gowalla[1] or Foursquare[2], enable users to add geographic coordinates for identification of their current location. This approach (Geotagging) can support disaster management by allowing the user to report past or ongoing disasters using a smartphone application.

Considering the random nature of disasters, mobility is an important asset for producing and reporting disaster data. As the usage of smartphones becomes ubiquitous to society, disaster data collection can be feasible through the smartphones, which have the following capabilities:

- Position system: Smartphones should be able to determine the location of users in real time. This capability can be achieved by using a Global Positioning System (GPS) or by performing Geocoding on a submitted address.
- Camera: To provide more information on disasters, a disaster photo would be submitted. A smartphone is equipped with a camera to produce the photo.
- Wireless network: The network should be able to support transfer of information including multimedia data such as disaster photos.
- Internet connectivity: By using a wireless network, Internet connects a smartphone with the Web Service hosted in the server.

2.3 Spatial Data Infrastructure

Lemmens described five unique features that distinguish spatial data from other types of scientific data [4]. Those features include:

- Multiple versions - Versions of the same entities of the earth's surface may differ in terms of data models, scales that are mostly collected by different agencies
- Implicit linking - In general data, explicit references must be presented to combine information from multiple sources in a meaningful manner. Spatial data enable linking without explicit references, i.e., via a coordinate reference system.
- Massive datasets - Compared to general (administrative) information, spatial data would be massive. In case of satellite imagery, for instance, raster data volumes would be huge.
- Maps as implicit interfaces - Everyone is familiar with reading maps, so they are a natural manmade interface for representing spatial data.
- Spatial data is geometry based - It is possible to apply many mathematical tools in Geo-services (such as to compute the distance between two objects or compute the buffer around an object) whereas other data types use only limited operations such as string manipulation or statistical operation.

Before mass use of the Internet and its technologies, spatial data for a particular location had been stored in different physical locations and often used based on different standards or formats. This made it difficult for a potential user to access

[1] Gowalla.com (accessible on Feb.24, 2011).

[2] Foursquare.com (accessible on Feb. 24 2011).

and utilizes the data. Potential users of this disordered data might be an organization that could not afford to acquire data on their own, or access needed data from outside their organization.

Steigner and Hunter described Spatial Data Infrastructure (SDI) as a coordinate series of agreements on technology standards, institutional arrangements, and policies that enable discover and use of geospatial information by users [5]. An SDI can be used as an appropriate framework to facilitate disaster management and meet the need for collaboration in spatial data production and sharing for disaster management. It creates an environment in which agents can access, retrieve, and disseminate disaster data.

2.4 Web-Based GIS Application

To serve a broader audience for disaster history, a Web based application is preferable over its desktop counterpart. The data can be displayed in the form of a map so that users may have a better understanding of the data. Web GIS clients would be used to display and query spatial data stored at the remote locations that are accessible via the Internet or Intranet.

Web GIS can serve spatial data using two OGC (Open Geospatial Consortium) standards such as: WMS (Web Map Services) for the display of maps as images and WFS for featured data. In order for users to be able to access the data, GIS server is needed to process the request. Fig.2 shows the Web GIS Architecture that elaborates the collaborations of a GIS server, OGC standards, and Web browser in a full functioning GIS application. Each component in this architecture is designed as a Web Service.

Muehlen, et al. compared two types of Web Service protocol platforms, REST and SOAP [6]. REST is an architectural style described by Fielding [7]. In REST architecture, objects are identified by a URI and specific message protocols such as PUT, POST, GET, and DELETE are used to process the data objects. A request message sent to an object results in the processing for accessing or manipulating the objects, typically in the form of an XML document. This document provides the client with the ability to change the state of the data objects.

While SOAP based architecture is applicable in certain frameworks, it is inferior to REST based Web Service architecture in terms of both network bandwidth utilized when transmitting service requests over the Internet and the round trip latency incurred [8]. Therefore, for maximizing bandwidth for disaster reports, we develop our system based on the REST architecture.

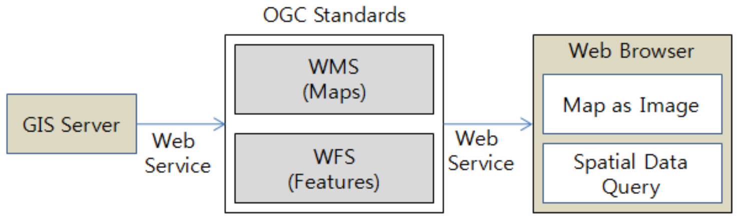


Fig. 2 Web GIS Architecture

3 System Development

3.1 Data Collection

User interface for our proposed smartphone application is depicted by Fig.3. It is capable of getting user coordinates from GPS as well as providing options to capture pictures of the disaster [9]. User can provide their own description and send the data to the server. In case of limited GPS capability, such as inside a building, the application provides a Google map based interface to select the location and will perform Geocoding to convert the address to the correct coordinate.

After the user fill in the form and send the information, application will call REST Web service as shown in Table 1.

The application will use a POST method and supply all the required parameters for the Web Service. We use a separate Web Service for disaster data and its photos because of the one-to-many relationship that disaster data had with its photos.

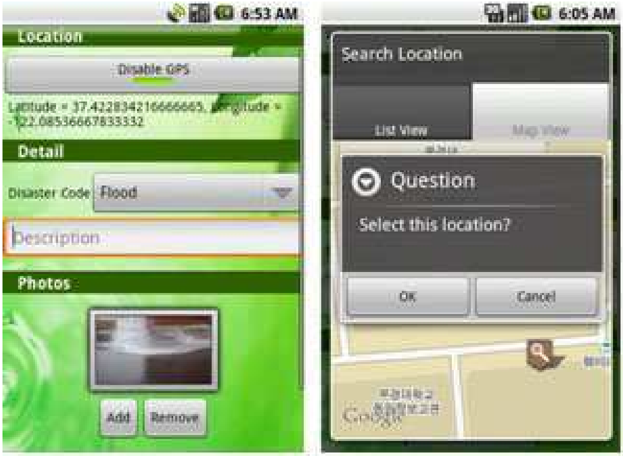
In case of more than one photos, service described in Table 2 will be used as many times as the number of photos. We utilized the method that can be used for transferring the information to a Web server for further processing in the database.

Table 1 Web service parameter for disaster information

Parameter	**Value**	**Description**
userId	string	Username of disaster reporter
discode	string	Disaster type (e.g. flood, tsunami, etc)
lat	double	Latitude
lng	double	Longitude
disdate	timestamp	Disaster occurrence time
description	string	Disaster description

Table 2 Web Service parameter for photo information

Parameter	**Value**	**Description**
phtId	string	Photo Id number
photo	input stream	Disaster photo in byte stream
phtDesc	double	Photo description
phtDate	double	Photo date
phtFname	timestamp	Photo file name
colId	string	Disaster Id number



Latitude = 37.422834216666665, Longitude = -122.08536667833332

Disaster Code: Flood

Description

Select this location?

Fig. 3 Smartphone application for disaster report

3.2 Spatial Data Integration

Integrating spatial data with normal data is complex because spatial data have specific properties. In the case of ordinary data, extracting relation from one data with the other can simply be performed with a string manipulation or statistical operation. However, for spatial data, to obtain a relation from one data to another, e.g., compute the distance between two objects or compute the buffer around an object, a geometry operation is needed. Often a geographic feature is represented using different geometric features (for instance, roads can be represented using polygons or lines, disaster data can represented by points), which adds to the complexity of the operation. Moreover, geo-datasets are retrieved from different sources in different formats (e.g., ESRI shaped files, Mapinfo files, PostGIS geoDB files). Therefore, to link disaster data submitted by users with the existing data, special approaches is needed.

In this paper, the spatial data that are used contain a topographic map of Busan Metropolitan City in South Korea as well as several other spatial data such as CCTV, rivers, parcel unit, and road networks will be saved in an ESRI ArcGIS domain. Hwang describe the preliminary analysis on framework for automatic generation of spatial area of damage which will be used in this paper[10].

In summary, to integrate disaster reports and existing spatial data, we will re-trieve location information from the disaster reports. Then, we will change data to the ESRI domain as geometry in form of point. Lastly, by performing geoprocessing on existing spatial data with the newly created point, we can extract the topographic relation of the reported disasters and existing data.

3.3 Process Flow

The flow of activities from community, system, administrators, and citizens are displayed in a swimlane diagram in Fig. 4. The process begins when the community which is smartphone users creates a new disaster record activity. The record, along with all the information, such as coordinates and photos, is saved in the database. Next, the coordinate is transformed according to existing projections. GPS coordinate uses WGS84 while the existing spatial data uses ITRF 2001. The point later will be created as geometry to enable spatial analysis with existing spatial data that are also in the form of geometry.

After the administrator chooses to generate a new area of damage based on the point, the system will analyze the area by intersecting the point data with the parcel data. If disaster points fall within a parcel, the parcel area becomes the new damage area and its attributes, such as disaster description and photos, become the damage area attributes.

Moreover, the system will also determine whether or not the area is in proximity with rivers or mountains that will provide more information on the back-grounds of any flood or landslide. The administrator then verifies the area of dam-age for validation. The administrator can edit the location or geometry and disaster attributes, such as photos and descriptions, to provide better information. This area can later be viewed by all citizens including community and administrators using Web GIS clients.

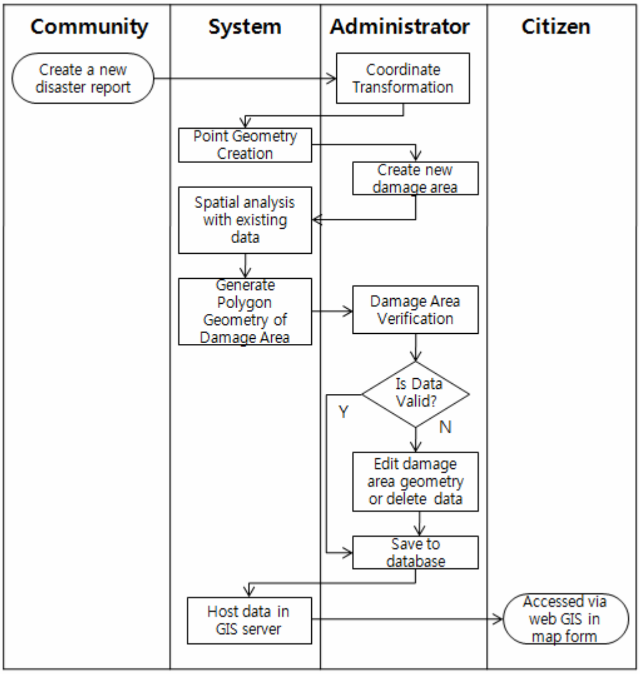

Community
System
Administrator
Citizen
Create a new disaster report
Coordinate Transformation
Point Geometry Creation
Create new damage area
Spatial analysis with existing data
Generate Polygon Geometry of Damage Area
Damage Area Verification
Is Data Valid?
Y
N
Edit damage area geometry or delete data
Save to database
Host data in GIS server
Accessed via web GIS in map form

Fig. 4 Swimlane Diagram of the process

3.4 System Architecture

The following are software used to develop the system:

- Desktop GIS: used for data creation, editing, analysis and map generation. We use ArcGIS suite from ESRI because the software suite arguably offers the most complete solution on GIS application development. ArcMap is used to create a map from several layers of spatial data features.
- Spatial Database Management Systems (Spatial DBMS): used for the storage of data. ArcSDE Spatial Database Server is used in this system.
- GIS Server: used to remotely process and analyze spatial data. For this capability, the ArcGIS server also provides WFS service for manipulating spatial data.
- Web clients: to display and query spatial data stored at remote locations that are only accessible via Internet or intranet. We develop the GIS clients using an ArcGIS server and a Microsoft .NET framework
- Web Server (IIS): to serve requests from the Web client. IIS is chosen because it is a platform in developing Web application on an ArcGIS server domain.
- Web Server (Tomcat): to host REST Web Service for smartphone applications to consume. We separate the Web server to make a distinction on the Web server for smartphone applications and the Web server for Web GIS application.
- Smartphone Application: This application is used by the community to report disasters in their area. In our paper, we decided to use an Android based application because of the open source nature of the development.

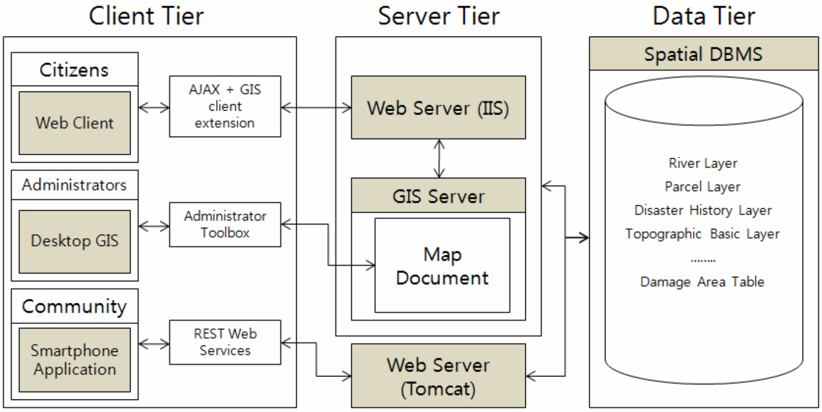


Fig. 5 System Architecture

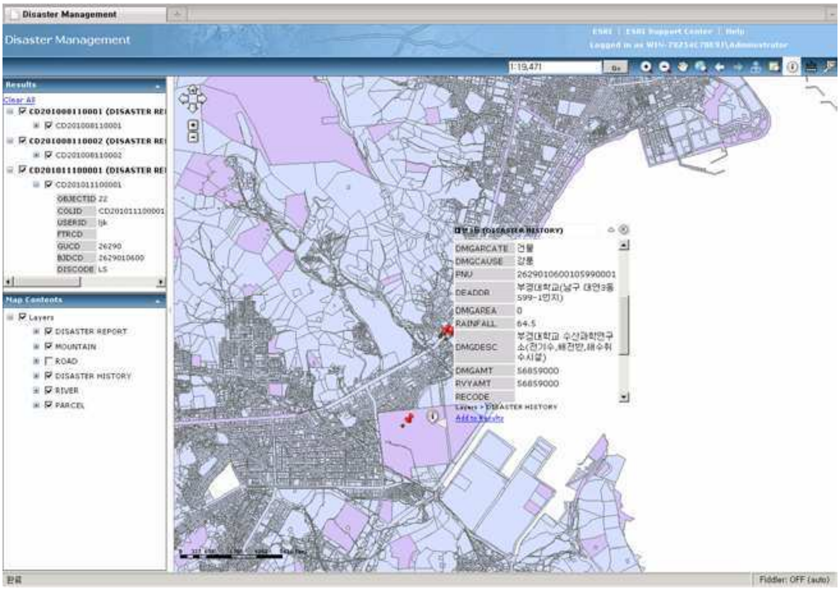
Fig.5 shows the interrelation of those GIS software in the proposed system architecture. Administrators can use desktop GIS equipped with administrator tool-boxes including Geoprocessing tools for performing: coordinate transformation; damage area creation; damage area verification and damage area edit. All of the data obtained from the process will be saved in Spatial DBMS including disaster reports from community.

Map document is the interface between the administrator and spatial DBMS. By using the administrator toolbox developed based on an ArcMap platform, the administrator can manipulate the data on the spatial DBMS from the desktop GIS application.

3.5 Web GIS Prototype

Fig 6 depicts the prototype of a Web based GIS application for the proposed system. There are several layers displayed in the application, namely disaster report, mountain, road, disaster history, river, and parcel.

The Disaster Report layer refers to a reported point of the disaster site submitted by users. Mountain, road, and river are self explanatory. These layers are part of the spatial data owned by governments and integrated with disaster information into the Web GIS application. After the reported data are processed and validated, additional information, such as description, disaster cause, losses, etc. will be sup-plied. The complete information on the disaster is displayed on the disaster history layer, which will be the final result of the system.



☑ CD201008110001 (DISASTER RE
☑ CD201008110001
☑ CD201008110002 (DISASTER RE
☑ CD201008110002
☑ CD201011100001 (DISASTER RE
☑ CD201011100001

OBJECTID	22
COLID	CD201011100001
USERID	ljk
FTRCD	
GUCD	26290
BJDCD	2629010600
DISCODE	LS

☑ Layers
☑ DISASTER REPORT
☑ MOUNTAIN
☐ ROAD
☑ DISASTER HISTORY
☑ RIVER
☑ PARCEL

(DISASTER HISTORY)

DMGARCATE	건물
DMGCAUSE	강풍
PNU	2629010600105990001
DEADDR	부경대학교(남구 대연3동 599-1번지)
DMGAREA	0
RAINFALL	64.5
DMGDESC	부경대학교 수산과학연구소(전기수,배전반,배수취수시설)
DMGAMT	56859000
RVYAMT	56859000
RECODE	

Layers > DISASTER HISTORY
Add to Results

Fig. 6 Web GIS prototype

4 Conclusion

This paper proposes a system for elucidation of disaster damage history by integrating user-generated content, spatial data and Web GIS. The system utilizes user submitted data and performs spatial analysis on the data to create more complete information on disaster damaged areas. To integrate the data with existing data, we addressed several issues, such as the nature of the spatial data. Through the proposed system, we were able to connect three stakeholders involved in the disaster management process: community, administrators, and citizens. Community is the one that provides past and ongoing disaster content. This information is edited and integrated with the existing spatial data by the administrator. The final generated damage area can be disseminated to citizens using the Web GIS.

Future direction of this research should include the improvement of existing Web GIS applications as well as provide information on smartphones by creating a mobile disaster management application.

Acknowledgement. This work (Grants No.00043527-1) was supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2010.

References

1. Mansourian, A., Rajabifard, M.J., Valadan Zoej, I., Williamson: Using SDI and web-based system to facilitate disaster management. Computers & Geosciences 32, 303–315 (2006)
2. Rinner, C., Kebler, C., Andrulis, S.: The use of Web 2.0 concepts to support deliberation in spatial decision-making. Computers, Environment and Urban System 32, 386–395 (2008)
3. Hristidis, V., et al.: Survey of data management and analysis in disaster situations. Journal of Systems and Software 83, 1701–1714 (2010)
4. Lemmens, R., et al.: Integrating Semantic and Syntactic Descriptions to Chain Geographic Services. IEEE Internet Computing 10, 42–52 (2006)
5. Steiniger, S., Hunter, A.J.S.: Free and open source GIS software for building a spatial data infrastructure. In: LNGC. Springer, Heidelberg (2009)
6. Muehlen, M., Nickerson, J., Swenson, K.: Developing web services choreography stan-dards-the case of REST vs SOAP. Decision Support System 40, 9–29 (2005)
7. Fielding, R.T.: Architectural styles and the design of network based software architectures. Doctoral Dissertation, University of California, Irvine (2000)
8. Mulligan, G., Gracanin, D.: A Comparison of SOAP and REST Implementations of a Service Based Interaction Independence Middleware Framework. In: Proceedings of the 2009 IEEE Winter Simulation Conference (2009)
9. Lee, J., Niko, D., Kim, C.: Design of User-generated Contents for Disaster Information System based on Smartphone. In: International Conference on Multimedia, Information Technology and Its Application, MITA (2010)
10. Hwang, H., Kim, C.: A design of automatic generation system for spatial data of damage. In: International Conference on Multimedia, Information Technology and Its Application, MITA (2010)

Frameworks for u-Health Bio Signal Controller

Haeng-Kon Kim and Roger Y. Lee

Summary. Control technologies based on biosignal manipulate devices such as computer and wheelchair. ElectroMyoGram(EMG), ElectroOculoGram(EOG), ElectroEncephaloGram(EEG) are typical important bio-signals for u-health care. In this paper, we approach the EMG signals from electrodes placed on the forearm and recognizes the four kinds of motion. We also develops the prototype of a u-health device controller that controls hardware devices using EMG signal. To analyze EMG with properties of non-stationary signal, time-frequency features are extracted by wavelet packet transform. For dimensionality reduction and nonlinear mapping of the features. We proposes a feature projection method composed of PCA and SVM. The dimensionality reduction simplifies the structure of the classifier, and reduces processing time for the pattern recognition. The nonlinear mapping using SVM transforms the PCA-reduced features to a new feature space with a highly class separatability. SVM is a pattern classifier to recognize various motions. We finally show the experimental results using the proposed method enhances the accuracy of pattern recognition. As a results, The proposed systems make the possible to control movements of u-healthcare signal device based on classified patterns.

Keywords: EMG signal, feature extraction, pattern classification, SVM, controller.

Haeng-Kon Kim
Department of Computer information & Communication Engineering,
Catholic Univ. of Daegu, Korea
e-mail: hangkon@cu.ac.kr

Roger Y. Lee
Software Engineering & Information Technology Institute,
Central Michigan University, USA
e-mail: lee1ry@cmich.edu

R. Lee (Ed.): Computers, Networks, Systems, & Industrial Eng. 2011, SCI 365, pp. 257–266.
springerlink.com

1 Introduction

The computer-based portable devices lead our societies to the world of u-healthcare, so that people may monitor their heart rate outside hospitals. The paper proposes an ultra-wearable smart sensor system combines electrocardiogram (ear-lead ECG), tri-axial accelerometer, and GPS sensors to measure normal or elderly personals daily activities. The other hand, the fields of virtual reality have studied the visual and hearing technologies in these days. It also is expected to be important areas of the study for the effective interface technology with computers by the medium of bio-signal and tactile sense. The study of developments interface using bio-signal has been actively progressing to carry out effective interactions between human and computer. Control technologies based on bio-signal manipulate devices such as computer and wheelchair. Electromyo-gram(EMG), electrooculogram (EOG), electroencephalo-gram(EEG) are typical bio-signals. It is true that the weak, the elderly and the handicapped persons have been excluded from the information society. But they can access and use the information if they use these technologies, and it will be helpful to alleviate the social gap. EMG signal is an electrical signal that is generated on the surface of the muscle according to physical movements, and its strength is below 10mV and its frequency range is less than 500Hz. EMG signal can be used as an input method for HCI(Human Computer interaction) because it can be simply measured and measured signal is a reflection of the users intention. Recently, HCI technologies that control devices such as wheelchair and information technology equipments are continuously being developed using EMG signals[1,2,3]. In this paper, we approach the EMG signals from electrodes placed on the forearm and recognizes the four kinds of motion. We also develops the prototype of a u-health device controller that controls hardware devices using EMG signal. To analyze EMG with properties of non-stationary signal, time-frequency features are extracted by wavelet packet transform. For dimensionality reduction and nonlinear mapping of the features. We proposes a feature projection method composed of PCA and SVM. The dimensionality reduction simplifies the structure of the classifier, and reduces processing time for the pattern recognition. The nonlinear mapping using SVM transforms the PCA-reduced features to a new feature space with a highly class separatability. SVM is a pattern classifier to recognize various motions. We finally show the experimental results using the proposed method enhances the accuracy of pattern recognition. As a results, The proposed systems make the possible to control movements of u-healthcare signal device based on classified patterns. In this paper, we extracts feature vectors from measured EMG signals, and classifies them into four kinds of motion(upward, downward, leftward, rightward) using Support vector machine in real time. This paper also develops the prototype of a device controller that can control application programs or hardware devices based on classified patterns. The structure of this paper is as follows: Section 2 presents the related works about signal processing and pattern recognition and SVM. In section 3 design issues of a device controller using EMG signal are introduced. Section 4 describes the implementation details of a device controller and experimental results. Finally Section 5 presents conclusion and future work.

2 Related Works

2.1 Signal Processing and Pattern Recognition

Attachment points and sticking methods of electrodes, experimental conditions, and movement of subjects tend to generate noises, when measuring biosignals. To build an efficient HCI system, we must go through preprocessing, feature extraction, and pattern classification stage because bio-signal with noises is inappropriate for HCI devices. Preprocessing stage removes noises and strengthens of the components. The representative methods are Auto correlation function, Independent component analysis, Band-pass filtering, Notch filtering, Ensemble average. Feature extraction stage discovers the features of input signal that have smaller dimensions than input data and can classify data. This stage reduces the computation for classification and improves the classification performance. The representative methods are Auto regressive, Power spectrum, Hjorth parameter, Principal component analysis, Linear discriminant analysis. After feature extraction stage, bio-signals go through pattern recognition stage. With the results of processing and analysis of bio-signals, pattern recognition stage discriminates predefined feature patterns. Pattern recognition falls into four categories such as character recognition, speech recognition, facial recognition, biometrics. The representative methods are Hidden Markov modes, Kalman filter model, multi-layer perception. Meanwhile, various pattern recognition methods have been suggested to grasp the users intention from EMG signal. Ajiboye et al. classified five kinds of patterns from four muscles of healthy people and assorted four kinds of patterns from three muscles of single arm amputees using a heuristic fuzzy system[4]. Englehart et al. analyzed patterns through wavelet analysis, and improved the stability of pattern recognition using a major vote algorithm[5]. Jingdong et al. classified patterns using auto regressive model and wavelet transform, and improved the learning speed using a neural network based on variable learning rates, and controlled a robot having prosthetic hands[6].

2.2 SVM

The SVM algorithm was invented by Valdimir Vapnik, and the current standard incarnation was proposed by Corinna Cortes and Vladimir Vapnik[7]. SVM is a set of related supervised learning methods that analyze data and recognize patterns. SVM constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression analysis. A good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class. SVM applies linear classification techniques to non-linear classification problems using kernel functions. SVM is one of the training models that provide excellent recognition performance.

3 Design Issues of a Device Controller

This section describes design issues of a device controller such as the overall system architecture and the functional designs of its components.

3.1 System Architecture

Figure 1 gives the overall system architecture of a device controller using EMG signal. It is composed of three modules such as sensor module, recognition module, and controller module. Sensor module measures EMG signal that is generated on the surface of muscles of arm according to motions of wrist. Filtering is essential because noise has a harmful effect on EMG signal. Amplifier is used to amplify and filter EMG signals. Amplified EMG signal is sent to AVR. AVR transforms analog signal into digital signal. The digital signals are sent to PC by a bluetooth transmitter. Recognition module processes signals, extracts feature vectors, and classifies four kinds of patterns using SVM. The patterns are transmitted to the controller module through the RS-232C port. Controller module controls the movement of LEGO Mindstorms NXT 2.0 using classified patterns.

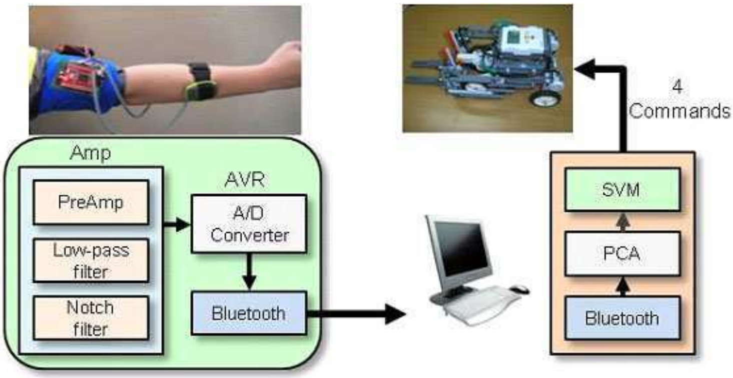


Fig. 1 The overall system architecture

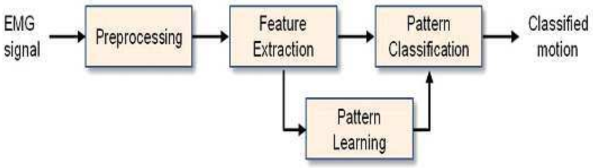


Fig. 2 The processing stages of EMG signal

$$k(x, x_i) = \exp\left(-\frac{1}{2\sigma^2} \left\| x - x_i \right\|^2 \right)$$

Fig. 3 ep.(1)

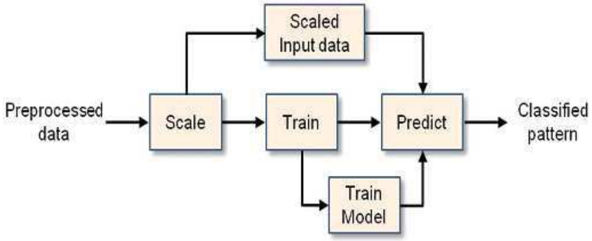


Fig. 4 The process of pattern classification

3.2 Functional Design of Recognition Module

The classification of the movements of the muscles needs EMG signals and a pattern classifier. Figure 2 shows the processing stages of EMG signal. The first stage eliminates noises and strengthens the components. The second stage extracts feature vectors of EMG signals. The third stage learns patterns and builds a train model, and then classifies input data into four kinds of patterns. paper uses Principal component analysis method to extract feature vectors from EMG signals. PCA reduces the dimension of time-frequency feature vectors extracted by wavelet packet transform. PCA outputs about 20 dimensions are necessary, although the dimension of reduced feature vectors does not have a serious effect on the results of pattern recognition[8]. The dimensionality reduction can simplify the structure of pattern classifier and learning process, shorten the computation time for pattern recognition. The output spaces of five dimensions are set for each channel.

This paper utilizes SVM that uses a kernel function of radial basis function type such as eq.(1) to learn and classify the patterns.

Figure 3 show the process of pattern classification. Scale stage normalizes preprocessed data. Train stage learns normalized data and builds a train model. To classify patterns, Predict stage compares normalized data sent from AVR with a train model. The classified patterns are sent to the controller module.

3.3 Functional Design of Controller Module

The motions of wrist control a LEGO Mindstorms NXT 2.0. That is, a movelessness of wrist comes to a standstill of Mindstorms. Four kinds of motions(Upward, downward, leftward, rightward) of wrist corresponds to four kinds of movements(forward, backward, leftward, rightward) of Mindstorms.

4 Implementation of a Device Controller

4.1 Implementation Environment

To implement the device controller, C and C++ are used as programming languages. Visual Studio 2008, CodeVision, LabVIEW 8.6 are used as development tools. CodeVision is used to program AVR ATmega128. LabVIEW is used to implement programs that extract feature vectors, recognize patterns, and control LEGO Mindstorms NXT 2.0.

4.2 Implementation Details

(1) Sensor Module

Sensor module is composed of electrodes, preprocessor, A/D converter, Bluetooth transmitter. To minimize hassles of sticking and removing electrodes, the dry type titanium sensors that need not gel or a tape are used. EMG signal should be amplified to make it easier to classify patterns since EMG signals are mostly weak below 10mV. EMG signals are preprocessed to remove noises of a power supply, and they become frequency domains with characteristics of EMG signal using a band-pass filter in the frequency range of 100 400 Hz as shown in Figure 4. The amplitude of preprocessed signals grows 7,000 times larger than measured signals through the Amplifier of Figure 5. After amplified EMG signals are transformed into digital signals,

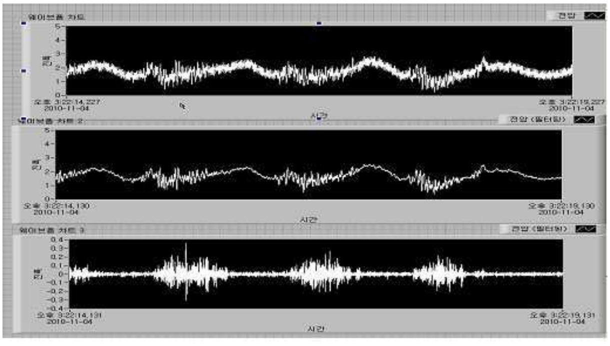


Fig. 5 The preprocessing process of EMG signal

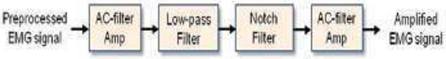


Fig. 6 The processing stage of Amplifier

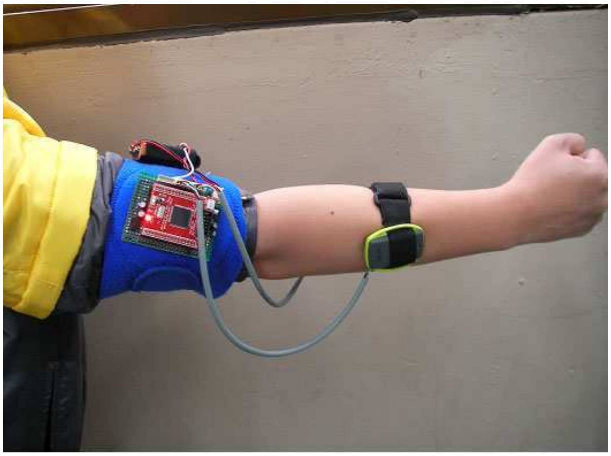


Fig. 7 The positions of electrodes

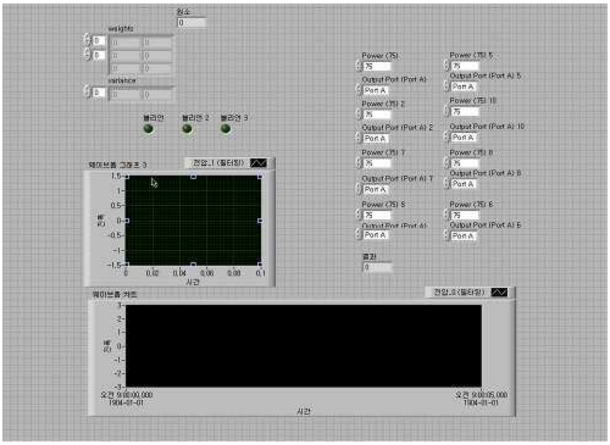
they are transmitted to a recognition module through a bluetooth transmitter. The muscles that take charge of motions of wrist are located in the forearm. To measure EMG signal, electrodes are put on Brachioradialis, Extensor carpi raialis longus, Palmaris longus, and Flexor carpi radialis as shown in Figure 6. Disposable electrodes cannot be moved once you have attached it. Titanium electrodes, however, can move its position to measure EMG signals. The attachment of band type sensor module reduces noise remarkably.

(2) Recognition and Contoller Modules

Recognition module is composed of a feature extractor and a pattern classifier. This paper implements recognition module and controller module with LabVIEW 8.6. These modules show feature vectors and EMG signals as shown in Figure 7. Feature vectors are sent to a pattern classifier that recognizes patterns using C-SVC type of SVM.

4.3 Experimental Results

To implement the device controller, C and C++ are used as programming languages. Visual Studio 2008, CodeVision, LabVIEW 8.6 are used as development tools. CodeVision is used to program AVR ATmega128. LabVIEW is used to implement programs that extract feature vectors, recognize patterns, and control LEGO Mindstorms NXT 2.0.



원소
0

weights
0 0 0
0 0 0
0 0

variance
0 0 0

Power (75)
75
Output Port (Port A)
Port A
Power (75) 2
75
Output Port (Port A) 2
Port A
Power (75) 7
75
Output Port (Port A) 7
Port A
Power (75) 9
75
Output Port (Port A)
Port A

Power (75) 5
75
Output Port (Port A) 5
Port A
Power (75) 10
75
Output Port (Port A) 10
Port A
Power (75) 8
75
Output Port (Port A) 8
Port A
Power (75) 6
75
Output Port (Port A) 6
Port A

결과
0

Fig. 8 Recognition and controller modules Controller module controls LEGO Mindstorms NXT 2.0 based on four kinds of patterns classified.

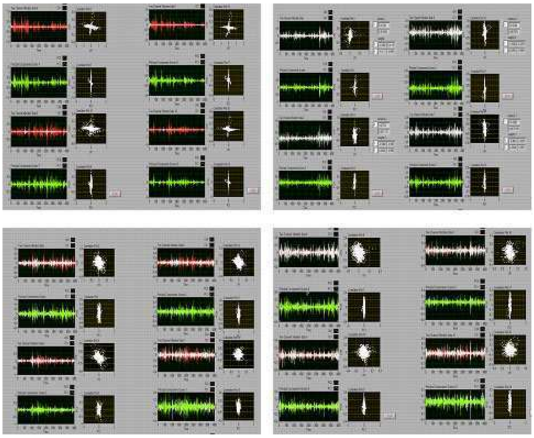


Fig. 9 Data distribution of feature vectors

(1) Feature extraction

Feature vectors are reassigned through the transformation of the reference axis of feature vectors based on non-correlative principal components. This process reduces variance of feature vectors and decreases the possibility of classification error. Figure 9 shows the result of the transformation of the reference axis based on principal components for feature vectors. In this figure, we can make certain the distributions of data are different in the motions of the wrist.

(2) Pattern classification

Feature vectors of four kinds of motion are extracted using PCA method and then a single point is divided into 13 dimensions based on feature values. A total of 40 data 10 data for each motion are prepared for machine learning. Non-linear SVM classifies 4 patterns through comparison between a train model and input data in real-time. Experimental results show that the accuracy of classification is 94.2.

5 Conclusions

In this paper, we develops the prototype of a device controller that controls a hardware devices using EMG signal. We also proposes a feature projection method composed of PCA and SVM. SVM is a pattern classifier to recognize various motions. Experimental results show that the proposed method improves the recognition accuracy, and makes it possible to control movements of a target device or application programs in real-time. In the future, We are going to use the controller as an input method of HCI or wearable computers and to improve its functionality and performance continuously.

References

1. Gianni, A., Caro, D., Ducatelle, F., Gambardella, L.M.: BISON: Biology-Inspired techniques for Self-Organization in dynamic Networks, `http://www.idsia.ch/~frederick/bison.pdf`
2. WirelessSensorNetworks Wiki
3. Heinzelman, W.R., Kulik, J., Balakrishnan, H.: Adaptive Protocols for Information Dissemination in Wireless Sensor Networks. In: Proc. ACM MobiCom 1999, Seattle, WA (1999)
4. Hedetniemi, S.M., Hedetniemi, S.H., Liestman, A.: A Survey of Gossiping and Broadcasting in Communication Networks. Networks 18 (1988)
5. Kulik, J., Heinzelman, W.R., Balakrishnan, H.: Negotiation base protocols for Disseminating Information in Wireless Sensor Networks. Wireless Networks 8, 169–185 (2002)
6. Intanagonwiwat, C., Govindan, R., Estrin, D., Heidemann, J., Silva, F.: Directed Diffusion for Wireless Sensor Networking. IEEE/ACM Transactions on Networking 11, 2–16 (2003)
7. Han, J.-S., Zenn Bien, Z., Kim, D.-J., Lee, H.-E., Jong-Sung: Kim Human-Machine interface for wheelchair control with EMG and its evaluation. In: Proc. of the Int'l Conf. of Engineering in Medicine and Biology Society, Jong-Sung, pp. 1602–1605 (2003)
8. Rosenberg, R.: The Biofeedback Pointer: EMG Control of a Two Dimensional Pointer. In: The 2nd Int'l Symposium on Werable Computers, pp. 162–163 (1998)
9. Tsuji, T., Fukuda, O., Murakami, M., Kaneko, M.: An EMG Controlled Pointing Device using a Neural Network. Trans. of the Society of Instrument and Control Engineers 37(5), 425–431 (2001)

10. Ajiboye, A.B., Weir, R.F.: A Heuristic fuzzy logic approach to EMG pattern recognition for multifunctional prosthesis control. IEEE Trans. on Neural Systems and Rehabilitation Engineering 13, 280–291 (2005)
11. Englehart, K., Hudgins, B., Parker, P.A.: A wavelet-based continuous classification scheme for multifunction myoelectric control. IEEE Trans. on Biomedical Engineering 48, 302–311 (2001)

Author Index